Randomized Confidence Bounds for Stochastic Partial Monitoring

Maxime Heuillet [1] **Ola Ahmad** [1 2] **Audrey Durand** [1 3]

Abstract

The partial monitoring (PM) framework provides a theoretical formulation of sequential learning problems with incomplete feedback. At each round, a learning agent plays an action while the environment simultaneously chooses an outcome. The agent then observes a feedback signal that is only partially informative about the (unobserved) outcome. The agent leverages the received feedback signals to select actions that minimize the (unobserved) cumulative loss. In contextual PM, the outcomes depend on some side information that is observable by the agent before selecting the action. In this paper, we consider the *contextual* and *non-contextual* PM settings with stochastic outcomes. We introduce a new class of PM strategies based on the randomization of deterministic confidence bounds. We also extend regret guarantees to settings where existing stochastic strategies are not applicable. Our experiments show that the proposed `RandCBP` and `RandCBPside`* strategies have competitive performance against state-of-the-art baselines in multiple PM games. To illustrate how the PM framework can benefit real world applications, we design a use case on the real-world problem of monitoring the error rate of any deployed classification system.

1. Introduction

Partial monitoring (Bartók et al., 2014) is a framework to formulate online learning problems where the feedback is partially informative. These online learning problems can be cast as partial monitoring (PM) games played between a learning agent and the environment over multiple rounds. At a given round, the agent selects an *action* and the environment simultaneously selects an *outcome*. The agent then incurs an instant *loss* and receives a *feedback* signal that is partially informative about the outcome. The challenge is that the agent does not observe the loss while the goal of the agent is to minimize the (unobserved) cumulative loss. To achieve this goal, the agent trades-off between actions associated to informative feedback signals (*exploration*) and small-loss actions (*exploitation*).

At a specific round, the agent's performance is measured by the *regret*. The regret corresponds to the difference between the loss of the selected action and the loss of the best action. The cumulative regret scales linearly with the horizon of rounds T if the agent fails to identify the best action. In this work, we consider the *stochastic PM setting* where outcomes are independent and identically distributed (i.i.d.) according to some (unknown) stationary outcome distribution. In the stochastic setting, Bartók et al. (2011) classified PM games into four categories based on achievable bounds on the cumulative regret: *trivial* games (no regret); *easy* games with poly-logarithmic upper bounds in $\tilde{\Theta}(\sqrt{T})$; *hard* games with upper bounds in $\Theta(T^{2/3})$; and intractable games with lower bounds in $\Omega(T)$. For example, the well-known multi-armed bandit problem (Auer et al., 2002) corresponds to an easy game. Additionally, hard games capture a valuable diversity of applications, such as learning from costly expert advice (Helmbold et al., 1997) and dynamic pricing (Kleinberg et al., 2003).

Deterministic PM strategies such as `CBP` (Confidence Bound Partial Monitoring) (Bartók et al., 2012b) and `PM-DMED` (Komiyama et al., 2015) have sub-linear regret guarantees on both easy and hard games. However, these are consistently outperformed empirically by stochastic strategies such as `BPM-Least` (Vanchinathan et al., 2014) and `TSPM` (Tsuchiya et al., 2020), for which regret guarantees are limited to easy games.

The *contextual* PM setting is an extension of the stochastic setting where the outcome distribution is a function of some *side information* (a *context*) observed by the agent before selecting the action at each round. Existing contextual PM strategies are fairly restrictive. The deterministic `CBPside` (Bartók et al., 2012a) (a contextual extension of `CBP`) is not applicable to hard games. On the other hand, the stochastic `IDS-FW` (Kirschner et al., 2023) has regret

[*]Equal contribution [1]Université Laval, Canada [2]Thales Research and Technology (cortAIx), Canada [3]Canada-CIFAR AI Chair, Mila, Canada. Correspondence to: Maxime Heuillet <maxime.heuillet.1@ulaval.ca>, Audrey Durand <audrey.durand@ift.ulaval.ca>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

guarantees on both easy and hard games but comes with several drawbacks: IDS-FW scales quadratically with the number of rounds; and IDS-FW requires the set of contexts to be finite and known in advance, a restriction that often does not hold in practice.

The primary aim of this study is to make progress for practical aspects of partial monitoring algorithms. We focus on CBP-based strategies. While CBP-based strategies have extensive theoretical regret guarantees (available in both easy and hard games in the non-contextual setting, and for easy games in the contextual setting), the empirical performance of CBP-based strategies is often dominated by stochastic strategies. Kveton et al. (2019) and Vaswani et al. (2020) show that the deterministic confidence bounds used in "optimistic in the face of uncertainty" (OFU) strategies can be randomized to improve empirical performance, while maintaining theoretical guarantees. We therefore formulate the following hypothesis: ***Can the randomization of confidence bounds also benefit strategies that are non OFU-based, such as CBP-based strategies?***

Contributions ① Algorithmic. We investigate the algorithmic mechanics restricting the applicability of CBPside to easy games. As a response, we propose CBPside* that is applicable to both easy and hard contextual PM games. Building upon CBP and CBPside*, we then show how to successfully randomize non-OFU based strategies by introducing the randomized variants RandCBP and RandCBPside*. ② Theoretical. In the non-contextual setting, we obtain a regret upper-bound for RandCBP that matches the bound of its deterministic counterpart CBP. Similarly, in the contextual setting, we analyze RandCBPside* and obtain an upper-bound for easy games that matches CBPside's. Our analysis of RandCBPside* introduces the first upper bound for hard contextual PM games (without assumptions on the set of contexts). ③ Empirical. Our experiments show that RandCBP and RandCBPside* have competitive empirical performance against state-of-the-art baselines in hard and easy PM games, both in the contextual and non-contextual settings. ④ Application. Currently, the PM field is predominantly theoretical and there is a notable scarcity of PM applications (Singla et al., 2014; Kirschner et al., 2023). To illustrate how the PM framework can benefit real world applications, we design a new use-case based on the real-world problem of monitoring the error rate of any deployed classification system. ⑤ Reproducibility. Our paper is the first to provide extensive reproducibility resources (open-source code for all strategies and environments, and game analyses in the Appendix) to facilitate future applied developments.

2. Preliminaries on Partial Monitoring

We consider finite PM games defined by N actions available to the agent and M outcomes available to the environment. A game is characterized by a loss matrix $\mathbf{L} \in [0,1]^{N\times M}$ and a feedback matrix $\mathbf{H} \in \Sigma^{N\times M}$. The feedback space Σ is finite, arbitrary, and not necessarily numeric. Similarly to Bartók et al. (2012a), we assume that the difference between greatest and lowest elements in the loss matrix is bounded by 1, i.e. $\max(\mathbf{L}) - \min(\mathbf{L}) \leq 1$. A table of notations is reported in Table 1 in the Appendix.

2.1. Finite stochastic partial monitoring games

A finite PM game is played over T rounds between a learning agent and the environment. The horizon T is unknown to both the agent and the environment. The matrices $\mathbf{L}$ and $\mathbf{H}$ are known. At each round $t = 1, 2, \ldots, T$, the environment samples an outcome J_t from a distribution $p^\star \in \Delta_M$, where Δ_M is the probability simplex of dimension M (column vector). We refer to $p^\star$ as the *outcome distribution* and the outcomes are independent and identically distributed (i.i.d.) according to $p^\star$. The agent then plays an action I_t. Following this action choice, the agent observes a feedback $\mathbf{H}[I_t, J_t]$ and incurs a deterministic loss $\mathbf{L}[I_t, J_t]$, where $[i, j]$ denotes the element at row i and column j. We emphasize that the loss and the outcome are never revealed to the agent.

Non-contextual setting The expected loss of action i is noted $\ell_i = L_i p^\star$, where the notation L_i corresponds to the i-th row of matrix $\mathbf{L}$. The optimal action is given by $i^\star = \operatorname{argmin}_{1\leq i\leq N} \ell_i$. The performance of the agent is evaluated using the cumulative regret (to minimize):

$$R(T) = \sum_{t=1}^{T} (L_{I_t} - L_{i^\star}) p^\star. \tag{1}$$

Contextual setting In the non-contextual setting, the optimal action does not depend on side information (context). In the contextual setting (Bartók et al., 2012a), also known as PM with side information, the optimal action depends on side information (context). Let $p^\star(x)$ denote the outcome distribution as a function of a context $x \in \mathcal{X}$, with $\mathcal{X}$ denoting the unknown and possibly continuous context space. The optimal action in context x minimizes the expected loss in that context: $i^\star(x) = \operatorname{argmin}_{1\leq i\leq N} L_i p^\star(x)$. The agent aims to minimize the cumulative contextual regret:

$$R(T) = \sum_{t=1}^{T} (L_{I_t} - L_{i^\star(x_t)}) p^\star(x_t), \tag{2}$$

where $L_{i^\star(x_t)}$ is the loss vector of the optimal action in x_t.

Other relevant settings We focus on stochastic settings where the outcome distribution is stationary over the rounds, which differs from the adversarial settings studied in Piccolboni et al. (2001); Cesa-Bianchi et al. (2006); Lattimore et al. (2020); Lattimore (2022); Tsuchiya et al. (2023) that assume the outcome distribution may change over the rounds.

We also assume finite action and feedback spaces, unlike Kirschner et al. (2020) who focus on settings with continuous action and feedback spaces. Finally, we consider a contextual setting where the context space $\mathcal{X}$ is unknown and can be continuous, whereas Kirschner et al. (2023) assume that $\mathcal{X}$ is finite and known in advance.

2.2. Structure of partial monitoring games

The optimality and informativeness of actions are respectively defined by loss matrix $\mathbf{L}$ and feedback matrix $\mathbf{H}$.

Definition 2.1 (Cell decomposition, Bartók et al. (2012b))**.** The *cell* $\mathcal{C}_i$ of action i is defined as the subspace in the probability simplex Δ_M where action i is optimal. Formally, $\mathcal{C}_i = \{p \in \Delta_M, j \in \{1, ..., N\}, (L_i - L_j)p\rangle \leq 0\}$.

Based on the cell, one can tell that an action i is: (i) *dominated* if $\mathcal{C}_i = \emptyset$ (i.e. there is no outcome distribution s.t. the action would be optimal); (ii) *degenerate* if the action is not dominated and there exist action i' such that $\mathcal{C}_i \subsetneq \mathcal{C}_{i'}$ (i.e. actions i and i' are duplicates, and therefore, both are jointly optimal under some outcome distributions); (iii) *Pareto-optimal* if the action is neither dominated nor degenerate. The set of Pareto-optimal actions is denoted $\mathcal{P}$.

Let σ_i denote the number of unique feedback symbols on row i of $\mathbf{H}$. Let $s_1, ..., s_{\sigma_i} \in \Sigma$ be an enumeration of the unique feedback symbols induced by action i (i.e. symbols in row H_i), sorted by order of appearance (columns) in H_i.

Definition 2.2 (Signal matrix, Bartók et al. (2012b))**.** Given action i, the elements of *signal matrix* $S_i \in \{0, 1\}^{\sigma_i \times M}$ are defined as $S_i[u, v] = \mathbb{1}_{\{\mathbf{H}[i,v]=s_u\}}$.

The signal matrix S_i is binary and it can be thought of as a one-hot encoding over the unique feedback symbols induced by action i. The signal matrices verify the important relation $\pi_i^\star = S_i\, p^\star \in \Delta_{\sigma_i}$, where $\pi_i^\star$ (respectively $\pi_i^\star(x)$ in the contextual setting) is the distribution over the feedback symbols induced by action i.

Difference between easy and hard games A PM game is *easy* if it suffices to play Pareto-optimal actions to minimize the regret. In hard games, minimizing the regret requires to play actions that can be dominated and degenerate. Formal definitions of easy and hard games are in Appendix B.

3. Towards a Randomized CBP

CBP (Bartók et al., 2012b) (Confidence Bound Partial Monitoring) currently stands out as the only strategy offering regret guarantees in both easy and hard games for non-contextual PM, and a practical extension in easy contextual games. In terms of empirical performance, CBP is outperformed by stochastic PM strategies. Similar limitations have been identified in the bandits setting for deterministic strategies (Chapelle et al., 2011). Randomizing OFU-based strategies has proven to be helpful (Vaswani et al., 2020; Kveton et al., 2019) for improving empirical performance while preserving the theoretical analysis. Here, we extend these ideas to CBP, a non-OFU based strategy instantiating successive elimination (Even-Dar et al., 2002). Algorithm 1 jointly displays the pseudo-codes of CBP and the proposed RandCBP. Differences are highlighted in purple. Implementation details are reported in Appendix A.

Algorithm 1 CBP (Bartók et al., 2012b) and RandCPB

input : $\mathcal{P}, \mathcal{N}, \alpha, \eta_a, f(\cdot), K, \sigma, \varepsilon$
\# Notation $e(\cdot)$ is a σ_{I_t} dimensional one-hot encoding.
for $t = 1, 2, \ldots, N$ **do**
  Play action $I_t = t$ (play each action once)
  Observe feedback $\mathbf{H}[I_t, J_t]$
  Init $n_{I_t}(N) = 1$, $\nu_{I_t}(N) = e(\mathbf{H}[I_t, J_t])$
for $t > N$ **do**
  for *each neighbor pair* $\{i, j\} \in \mathcal{N}$ **do**
    $\hat{\delta}_{ij}(t) \leftarrow \sum_{a \in V_{ij}} v_{ija}^\top \frac{\nu_a(t-1)}{n_a(t-1)}$
    $B \leftarrow \sqrt{\alpha log(t)}$
    Sample Z_{ijt} with Algorithm 2 (Appendix A.1)
    $c'_{ij}(t) \leftarrow \sum_{a \in V_{ij}} \|v_{ija}\|_\infty Z_{ijt} \sqrt{\frac{1}{n_a}}$
    if $|\hat{\delta}_{ij}(t)| > \cancel{c_{ij}(t)}\ c'_{ij}(t)$ **then**
      Add $\{i, j\}$ to $\mathcal{U}(t)$
  Compute $D(t)$ based on $\mathcal{U}(t)$
  Get $\mathcal{P}(t)$ and $\mathcal{N}(t)$ given $\mathcal{P}, \mathcal{N}$ and $D(t)$
  $\mathcal{N}^+(t) \leftarrow \bigcup_{ij \in \mathcal{N}(t)} N_{ij}^+$
  $\mathcal{V}(t) \leftarrow \bigcup_{ij \in \mathcal{N}(t)} V_{ij}$
  $\mathcal{R}(t) \leftarrow \{a \in N : n_a(t-1) \leq \eta_a f(t)\}$
  $\mathcal{S}(t) \leftarrow \mathcal{P}(t) \cup \mathcal{N}^+(t) \cup (\mathcal{V}(t) \cap \mathcal{R}(t))$
  Select action $I_t = \text{argmax}_{a \in \mathcal{S}(t)} \frac{W_a^2}{n_a(t-1)}$
  Observe feedback $\mathbf{H}[I_t, J_t]$
  $n_i(t) \leftarrow n_i(t-1) + \mathbb{I}[i = I_t], \forall i$
  $\nu_i(t) \leftarrow \nu_i(t-1) + \mathbb{I}[i = I_t] e(\boldsymbol{H}[I_t, J_t]), \forall i$

3.1. The CBP Strategy

Recall that the unknown parameter of the game is the outcome distribution $p^\star \in \Delta_M$. The expected loss difference between two actions i and j is defined as

$$\delta_{i,j} = (L_i - L_j)p^\star = \ell_i - \ell_j. \tag{3}$$

The sign of the expected loss indicates which action is better: action j is better than action i when $\delta_{i,j} > 0$.

Definition 3.1 (Neighbor pairs, Bartók et al. (2012b))**.** Two Pareto-optimal actions i and j are *neighbors* if $\mathcal{C}_i \cap \mathcal{C}_j$ is an $(M-2)$-dimensional polytope. The set of all neighbor pairs is denoted $\mathcal{N}$.

Two actions are neighbors when they cannot be jointly optimal under a given outcome distribution. CBP exploits that

it suffices to compute $\delta_{i,j}$ for the pairs in $\mathcal{N}$, instead of computing $\delta_{i,j}$ for all the action pairs in the game.

Successive elimination confidence bounds CBP computes expected loss difference estimates for all action pairs in $\mathcal{N}$. Pairs with low confidence estimates are then eliminated based on a successive elimination (Even-Dar et al., 2002) criterion. Based on Eq. 3, any estimate $\hat{\ell}_i(t)$ of the expected loss of action i at round t admits an upper confidence bound, denoted $\text{UCB}_i(t) = \hat{\ell}_i(t) + c_i(t)$ and a lower confidence bound, denoted $\text{LCB}_i(t) = \hat{\ell}_i(t) - c_i(t)$, where $c_i(t)$ is a confidence width that holds with some probability. One can tell with confidence that action j has a lower expected loss (i.e. is confidently better) than action i if $\text{UCB}_j(t)$ is strictly lower than $\text{LCB}_i(t)$:

$$\begin{aligned}\text{UCB}_j(t) < \text{LCB}_i(t) &\Leftrightarrow \hat{\ell}_j(t) + c_j(t) < \hat{\ell}_i(t) - c_i(t) \\ &\Leftrightarrow |\hat{\delta}_{ij}(t)| > c_{i,j}(t), \end{aligned} \tag{4}$$

where $\hat{\delta}_{i,j}(t) = \hat{\ell}_i(t) - \hat{\ell}_j(t)$ and $c_{i,j}(t) = c_i(t) + c_j(t)$.

At each round, action pairs $\{i, j\}$ that do not satisfy the elimination criterion of Eq. 4 correspond to low confidence estimates that are eliminated by CBP. Action pairs that satisfy the criterion are added to a set of high confidence pairs, denoted $\mathcal{U}(t)$. The set $\mathcal{U}(t)$ is then used to compute a subspace of the probability simplex, denoted $D(t)$, that summarizes the current knowledge of CBP about the outcome distribution $p^\star$. Formally, $D(t) = \{p \in \Delta_M, (i,j) \in \mathcal{U}(t), \texttt{sign}(\hat{\delta}_{i,j})(L_i - L_j)p > 0\}$. The sub-space $D(t)$ gathers constraints based on the signs of the confident estimates, which inform on the relative quality of actions (as explained below Eq. 3).

Unfortunately, one cannot empirically estimate the losses ℓ_i and ℓ_j to compute $\hat{\delta}_{i,j}(t)$ and $c_{i,j}(t)$. Indeed, that would require to estimate the outcome distribution $p^\star$, but the agent never observes the outcomes. To address this challenge, CBP exploits a connection between the outcome and feedback distributions.

Connecting outcome and feedback distributions Computing $\hat{\delta}_{i,j}$ and $c_{i,j}$ in practice requires two definitions.

Definition 3.2 (Observer set, Bartók et al. (2012b))**.** The set $V_{i,j}$ associated with action i and j contains the actions required to verify the relation $(L_i - L_j)^\top \in \oplus_{a \in V_{i,j}} \text{Im}(S_a^\top)$, where $\oplus$ corresponds to the direct sum.

Definition 3.3 (Observer vectors, Bartók et al. (2012b))**.** The observer vector of the action pair $\{i, j\}$ with respect to action a in observer set $V_{i,j}$, denoted $v_{ija} \in \mathbb{R}^{\sigma_a}$, is selected to satisfy the relation $(L_i - L_j)^\top = \sum_{a \in V_{i,j}} S_a^\top v_{ija}$.

The set $V_{i,j}$ identifies all the actions that induce informative feedback signals about a loss difference. Actions in $V_{i,j}$ allow $L_i - L_j$ to be expressed as a linear combination of their corresponding signal matrix images, with observer vectors as coefficients.

From Def. 3.2 and Def. 3.3, one can express the expected loss difference $\delta_{i,j}$ as a function of the feedback distributions $\pi_a^\star$ associated with every action a in $V_{i,j}$:

$$\delta_{i,j} = \langle L_i - L_j, p^\star \rangle = \sum_{a \in V_{i,j}} v_{ija}^\top S_a p^\star = \sum_{a \in V_{i,j}} v_{ija}^\top \pi_a^\star,$$

where we used Eq. 3 and $\pi_a^\star = S_a\, p^\star$ for action a. As a result, CBP computes $\hat{\delta}_{i,j}(t)$ using the estimates $\hat{\pi}_a(t) = \frac{\nu_a(t)}{n_a(t)}$, where the vector $\nu_a(t) \in \mathbb{N}^{\sigma_a}$ counts the number of times each unique feedback symbol observable with action a up to time t was observed, and $n_a(t)$ is the number of times that action a was played up to time t. The confidence bound over the estimate $\hat{\delta}_{i,j}(t)$ is defined as (Bartók et al., 2012b):

$$c_{i,j}(t) = \sum_{a \in V_{i,j}} \|v_{ija}\|_\infty \sqrt{\frac{\alpha \log(t)}{n_a(t)}}, \tag{5}$$

s.t. $\mathbb{P}[|\hat{\delta}_{ij}(t) - \delta_{ij}| \geq c_{ij}(t)] \leq 2|V_{ij}|t^{1-2\alpha}$ where $\alpha > 1$, and $|V_{ij}|$ is the size of the observer set V_{ij}. Greater values of α result in more exploration, as it causes less eliminations from the criterion in Eq. 4.

Exploration and exploitation in CBP At round t, CBP identifies plausible subsets of $\mathcal{P}$ and $\mathcal{N}$, denoted $\mathcal{P}(t)$ and $\mathcal{N}(t)$, based on the constrained probability space $D(t)$. The set $\mathcal{P}(t)$ contains all Pareto-optimal actions $i \in \mathcal{P}$ whose cell $\mathcal{C}_i$ intersects with $D(t)$. Similarly, the set $\mathcal{N}(t)$ contains all neighbor pairs $\{i, j\} \in \mathcal{N}$ whose cell intersection $\mathcal{C}_i \cap \mathcal{C}_j$ also intersects with $D(t)$. When $\mathcal{P}(t)$ contains only one action, the set $\mathcal{N}(t)$ is automatically empty and therefore CBP exploits. When $\mathcal{P}(t)$ contains more than one action, $\mathcal{N}(t)$ is not empty and CBP needs to explore. The following definitions characterize exploration:

Definition 3.4 (Underplayed actions, Bartók et al. (2012b))**.** The set $\mathcal{R}(t) = \{a = 1, \ldots, N : n_a(t) \leq \eta_a f(t)\}$ contains actions that are underplayed according to a play rate function $f(t)$ and a constant $\eta_a > 0$.

Definition 3.5 (Neighbor action set, Bartók et al. (2012b))**.** The neighbor action set of a neighbor pair $\{i, j\}$ is defined as $N_{i,j}^+ = \{k = 1, \ldots, N : \mathcal{C}_i \cap \mathcal{C}_j \subseteq \mathcal{C}_k\}$. Note that $N_{i,j}^+$ naturally contains i and j. If $N_{i,j}^+$ contains another action k, then $\mathcal{C}_k = \mathcal{C}_i$ or $\mathcal{C}_k = \mathcal{C}_j$ or $\mathcal{C}_k = \mathcal{C}_i \cap \mathcal{C}_j$.

Based on $\mathcal{N}(t)$ and Def. 3.5, CBP computes $\mathcal{N}^+(t) = \bigcup_{i,j \in \mathcal{N}(t)} N_{i,j}^+$ for the neighbor pairs. Similarly, CBP computes $\mathcal{V}(t) = \bigcup_{i,j \in \mathcal{N}(t)} V_{i,j}$ for the observer actions.

The final set of actions considered by CBP, denoted $\mathcal{S}(t)$, contains potentially optimal actions ($\mathcal{P}(t) \cup \mathcal{N}^+(t)$) and informative underplayed actions ($\mathcal{V}(t) \cup \mathcal{R}(t)$). CBP selects the action with the smallest action count,

i.e. $I_t = \text{argmax}_{a \in \mathcal{S}(t)} \frac{W_a^2}{n_a(t)}$, weighted by $W_a = \max_{\{i,j\} \in \mathcal{N}} \|v_{ija}\|_\infty$.

3.2. Instantiating `RandCBP`

We now introduce `RandCBP`, a randomized counterpart of the `CBP` strategy. The main idea behind `RandCBP` is to replace deterministic confidence bounds (Eq. 5) by randomized confidence bounds:

$$c'_{i,j}(t) = \sum_{a \in V_{i,j}} \|v_{ija}\|_\infty \frac{Z_{ijt}}{\sqrt{n_a(t)}},$$

where Z_{ijt} is sampled for each action pair $\{i, j\}$ from a discrete probability distribution supported over K bins in the interval $[A, B]$. Note that the `CBP` strategy corresponds to the specific case of $K = 1$ and $A = B = \sqrt{\alpha \log(t)}$.

Randomization procedure Let $\rho_1 = A, \ldots, \rho_K = B$ denote K equally spaced values, and p_k denote the probability of sampling the value ρ_k, with $k = 1, \ldots, K$. The probabilities assigned to the remaining $K - 1$ points are shaped according to the positive side of a discretized Gaussian distribution centered at 0. Formally, for $k \leq K - 1$, let $\bar{p}_k := \exp(-\rho_k^2 / 2\sigma^2)$. Then, p_k corresponds to the normalized probabilities, that is, $p_k := (1 - \varepsilon)\bar{p}_k / (\sum_k \bar{p}_k)$. The above distribution from which the Z_{ijt} are sampled is a truncated (between A and B) and discretized (into K points) Gaussian distribution with tunable hyper-parameters ε, $\sigma > 0$, and K. A pseudo-code of the randomization procedure is provided in Algorithm 2.

Algorithm 2 Randomization Procedure

Input : $A, B, K, \varepsilon, \sigma$
Output : Z

Initialize an array ρ of size K with equally spaced values in $[A, B]$
for $k \leftarrow \{1 \ldots K\}$ **do**
  Calculate $\bar{p}_k$ using $\bar{p}_k = \exp\left(-\frac{\rho_k^2}{2\sigma^2}\right)$
Initialize an array p of size K
for $k \leftarrow \{1 \ldots K - 1\}$ **do**
  Calculate p_k using $p_k = \frac{(1-\varepsilon)\bar{p}_k}{\sum_k \bar{p}_k}$
Set $p_K = \epsilon$
Sample Z in ρ with probabilities p

This randomization procedure was introduced by Vaswani et al. (2020) for randomizing Upper Confidence Bound (UCB) strategies in the bandit setting. We generalize these ideas to the broader PM setting, where confidence bounds articulate a successive elimination criterion. This requires to define the randomized confidence bounds on quantities estimated for each action pair $\{i, j\}$ in $\mathcal{N}$. We will now show how this mechanism can be considered seamlessly in the theoretical analysis of `CBP`, allowing to maintain the regret guarantees with `RandCBP`.

Regret analysis The analysis, follows the structure of `CBP`'s analysis (Bartók et al., 2011), and involves upper bounding the expected number of times the confidence bounds succeed and fail, as detailed in Appendix C.2 and C.3 respectively. For the failure case, we leverage that the probability of the randomized SE criterion failing becomes negligible over time following Kveton et al. (2019). For the success case, we adapt lemmas from Bartók et al. (2012b) by observing that the randomized bounds are always upper bounded by their deterministic counterparts. The detailed analysis is reported in Appendix C.

Theorem 3.1. Consider the interval $[A, B]$, with $B = \sqrt{\alpha \log(t)}$ and $A \leq 0$. Set the randomization over K bins with a probability ϵ on the tail and a standard deviation σ. Set $f(t) = \alpha^{1/3} t^{2/3} \log(t)^{1/3}$, $\eta_a = W_a^{2/3}$ and $\alpha > 1$. On easy games, `RandCBP` achieves

$$\mathbb{E}[R_T] \leq N \left[2(1 + \frac{1}{2\alpha - 2})|\mathcal{V}| + 1 \right] + \sum_{k=1}^{N} \delta_k + \sum_{k=1, \delta_k > 0}^{N} 4 W_k^2 \frac{g_k^2}{\delta_k} \alpha \log(T),$$

with $\mathcal{V} = \bigcup_{i,j \in \mathcal{N}} V_{ij}$ and g_k being game dependent constants. On hard games, assuming positive constants C_1 and C_2, `RandCBP` achieves

$$\mathbb{E}[R_T] \leq C_1 N + C_2 T^{2/3} \log^{1/3}(T).$$

Similarly to the guarantees of `CBP` (Bartók et al., 2012b), the bound on easy games is problem dependent while the bound on hard games is problem independent. In both cases, the expected regret of `RandCBP` grows at the same rate as `CBP` on the horizon T. On easy and hard games, our bound is equivalent to CBP's in N, up to a constant. The dependency on the other terms is equivalent. We will see in the experiments that `RandCBP` empirically outperforms `CBP`.

4. The Contextual Setting

`CBPside` (Bartók et al., 2012a) extends `CBP` to the linear and logistic contextual PM settings. `CBPside` was initially hard-coded for easy games. Then, (Lienert, 2013) showed that the exploration mechanism developed for `CBP` (in the non-contextual setting) can be leveraged to extend `CBPside` to hard games. However, (Lienert, 2013) proposed an exploration based on action counts, which is inadequate for the contextual setting. Indeed, action counts do not reflect the fact that actions are played in specific contexts (i.e. observations), in the contextual setting. As a response, we introduce `CBPside`*, a variation that weights action counts

based on the current context and the history of contexts that have been previously observed. This variation enables us to derive regret guarantees for easy and hard games, as well as an applicable strategy to hard games. We then propose `RandCBPside`*, a stochastic counterpart of `CBPside`*, that enjoys regret guarantees on both easy and hard games in the linear setting, while empirically outperforming its deterministic counterpart. Pseudo-codes of `CBPside`* and `RandCBPside`* are reported in Appendix A.

4.1. The linear `CBPside`* strategy

Recall that under the contextual PM setting, $p^\star(x)$ denotes the outcome distribution given d-dimensional contexts $x \in \mathcal{X} \subseteq \mathbb{R}^d$. In the linear setting, $p^\star(x) = \theta x$, where $\theta \in \mathbb{R}^{M\times d}$ is an unknown parameter matrix. Similarly to the non-contextual setting, it is not possible to estimate the outcome distribution directly (see Section 3.1). Consequently, `CBPside`* exploits the connection between outcome and feedback distributions. In the contextual setting, the feedback distribution is $\pi_i^\star(x) = S_i p^\star(x) \in \Delta_{\sigma_i}$ for all actions $i \in \{1,\dots,N\}$. If we denote $\theta_i = S_i\theta \in \mathbb{R}^{\sigma_i\times d}$ as the per-action unknown parameter of the regression, then the contextual feedback distribution is $\pi_i^\star(x) = \theta_i x$.

`CBPside`* estimates θ_i with a ridge estimator defined as $\hat{\theta}_i(t) = Y_{i,t}X_{i,t}^\top(\lambda I_d + X_{i,t}X_{i,t}^\top)^{-1}$, where $X_{it} \in \mathbb{R}^{d\times t}$ is the history of contexts, $Y_{it} \in \{0,1\}^{\sigma_i\times t}$ is the history of one-hot-encoded feedback symbols for action i, and I_d the d-dimensional identity matrix. The following confidence bound on $\hat{\delta}_{i,j}(x)$ holds with probability $1/t^2$:

$$c_{i,j}(x) = \sum_{a\in V_{ij}} \|v_{ija}\|_2 \times\sigma_a\left(\sqrt{d\log(t) + 2\log(1/t^2)} + \sigma_a\right)\|x\|_{G_{a,t}^{-1}}, \quad (6)$$

where $G_{a,t} = \lambda I_d + X_{at}X_{at}^\top$ is the Gram matrix and $\|x\|^2_{G_{a,t}^{-1}} = x^\top G_{a,t}^{-1}x$ is the weighted 2-norm.

Remark 4.1. The confidence bound of `CBPside`* (Eq. 6) corrects the bound $c_{i,j}(x) \propto d(\sqrt{d\log(t)\dots})$ used in Bartók et al. (2012a); Lienert (2013). We multiply by σ_a instead of d to correctly instantiate Theorem 3 in Bartók et al. (2012a) over matrix traces. The corrected confidence bound is less conservative as σ_a is usually smaller than d.

Exploration based on pseudo-counts As opposed to `CBPside` (Lienert, 2013), the exploration of `CBPside`* is based on a new definition of underplayed actions suitable for the contextual setting. The definition enables the applicability of `CBPside`* to hard games, and the derivation of regret upper-bounds in both easy and hard games. In the non-contextual setting, underplayed actions are based on the number of times that action a was played up to time t, i.e. $n_a(t)$. A natural extension would consist in counting the number of times action a was played in context x_t. Unfortunately, given that contexts are usually sampled from a continuous domain $\mathcal{X}$, each context is typically encountered only once over a game, making such counters irrelevant.

Definition 4.1 (Underplayed actions (contextual case))**.** At round t, the set $\mathcal{R}(x_t) = \{a = 1,\dots,N : 1/\|x_t\|^2_{G_{a,t}^{-1}} < \eta_a f(t)\}$ contains actions that are underplayed at context x_t given a play rate function $f(t)$ and a constant $\eta_a > 0$.

The quantity $1/\|x\|^2_{G_{a,t}^{-1}}$ is a *pseudo-count* of the number of selections of action a at a given context x. In the specific case of orthogonal contexts sampled from the finite set of d-dimensional one-hot vectors, $1/\|x\|^2_{G_{a,t}^{-1}}$ corresponds to the exact number of selections of action a in context x. When contexts are not orthogonal, the pseudo-count increases proportionally to the frequency of the action being played in similar contexts.

4.2. Instantiating linear `RandCBPside`*

The randomized counterpart of `CBPside`*, namely `RandCBPside`*, relies on randomized confidence bounds defined at time t for a pair $\{i,j\}$, defined as

$$c'_{i,j}(x) = \sum_{a\in V_{ij}} ||v_{ija}||_2\sigma_a(Z_{ijt} + \sigma_a)\|x\|_{G_{a,t}^{-1}},$$

where Z_{ijt} is a random variable bounded in $[A,B]$ and follows the randomization procedure presented in Section 3.2.

Regret analysis We leverage the non-contextual analysis of `CBP` (Bartók et al., 2012b). We adapt the analysis to the contextual case by introducing pseudo-counts and simplifying the obtained expressions with the Cauchy-Schwartz inequality. The contextual confidence bounds are simplified by considering the total number of feedback symbols in the game. To upper bound the expected number of times the confidence bound fails, we consider some worst-case probability over all actions of the game. The detailed analysis is reported in Appendix D.

Theorem 4.2. Consider the interval $[A,B]$, with $B = \sqrt{d\log(t) + 2\log(1/t^2)}$ and $A \le 0$. Set the randomization over K bins with a probability ϵ on the tail and a standard deviation σ. Let $f(t) = \alpha^{1/3}t^{2/3}\log(t)^{1/3}$, $\eta_a = W_a^{2/3}$ and $\alpha > 1$. Assume $\|x_t\|_2 \le E$ and positive constants E,C_1, C_2, C_3, and C_4. On easy games, `RandCBPside`* achieves:

$$\mathbb{E}[R_T] \le C_1N + C_2Nd\sqrt{T}\log(T)$$

and, on hard games, `RandCBPside`* achieves:

$$\mathbb{E}[R_T] \le C_3N + C_4\sqrt{d}\log(T)^{1/3}T^{2/3}.$$

Similarly to `CBPside`, the guarantee of `RandCBPside`* is problem independent and grows at the same rate on easy games. However, `RandCBPside`* presents a new problem dependent guarantee on hard games. On easy games,

(Bartók et al., 2012a) uses a different equation for the confidence bound, which influences the dependencies in N and d. Accounting for Remark 4.1, we found an improved dependency in N, rather than a dependency in $N^{3/2}$, and an improved dependency in d rather than d^2, as reported in (Bartók et al., 2012a). Additionally, since `CBPside`* is a special case of `RandCBPside`* with $A = B$ and $K = 1$, our bounds for `RandCBPside`* also hold for `CBPside`*. We will see in the experiments that `RandCBPside`* empirically outperforms `CBPside`* on easy and hard games.

Applicability to other contextual settings. The focus of this manuscript is on the linear partial monitoring setting. Beyond this linear contextual setting, randomization offers an interesting perspective to online learning strategies that rely on statistical deviation bounds that are overly conservative in practice, which can be of particular interest for bounds in the logistic (Bartók et al., 2012a), and neural settings (Zhou et al., 2020; Xu et al., 2022).

5. Numerical Experiments

We conduct experiments to validate the empirical performance of `RandCBP` and `RandCBPside`* on the well-known Apple Tasting (AT) (Helmbold et al., 2000) (further studied in (Raman et al., 2024)) and Label Efficient (LE) (Helmbold et al., 1997) games. AT is a two actions and two outcomes easy game:

$$\mathbf{L} = \begin{matrix} \text{action 1} \\ \text{action 2} \end{matrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{H} = \begin{bmatrix} \bot & \bot \\ \wedge & \odot \end{bmatrix}.$$

LE is a hard game with three actions and two outcomes:

$$\mathbf{L} = \begin{matrix} \text{action 1} \\ \text{action 2} \\ \text{action 3} \end{matrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \mathbf{H} = \begin{bmatrix} \bot & \odot \\ \wedge & \wedge \\ \wedge & \wedge \end{bmatrix}.$$

For reproducibility, we provide in Appendix B a detailed analysis of both games. Code is available at https://github.com/MaxHeuillet/partial-monitoring-algos.

5.1. Evaluation of `RandCBP`

Since both AT and LE admit binary outcomes, the outcome distribution corresponds to $p^\star = [p, 1-p]$ with $p \in [0, 1]$. We consider *imbalanced* and *balanced* instances. Imbalanced instances, where $p \sim \mathcal{U}_{[0,0.2]\cup[0.8,1]}$, are usually solved faster since outcomes have lower variance. Balanced instances, where $p \sim \mathcal{U}_{[0.4,0.6]}$, require more exploration to estimate $p^\star$ with confidence. This leads to four cases: imbalanced/balanced AT and imbalanced/balanced LE. For each of the four cases, we run the experiment 96 times on a $T = 20$k horizon.

Baselines We consider the deterministic `PM-DMED` and `CBP` as baselines, as well as the stochastic `BPM-Least`, `TSPM`, and `TSPM-Gaussian` (in the settings where they have a guarantee). The code is open-sourced for all strategies. Implementation and hyper-parameters details are reported in Appendix A.

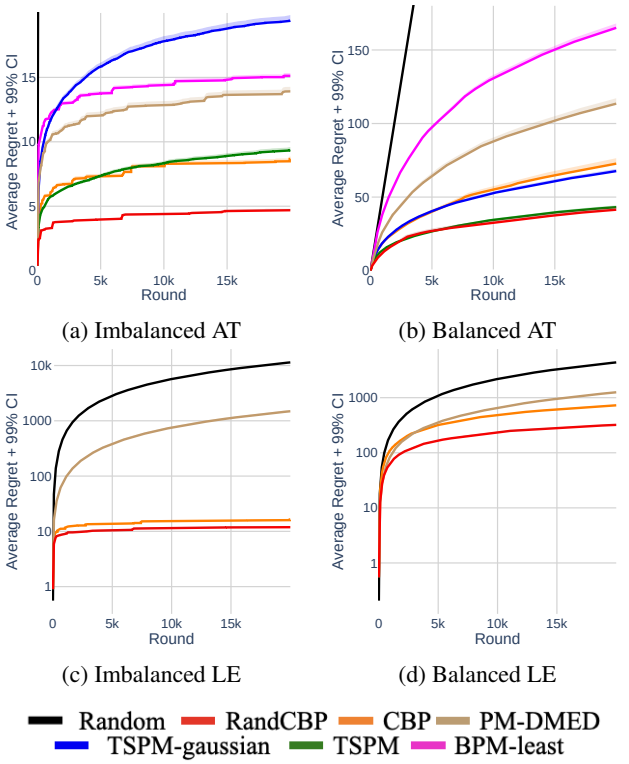
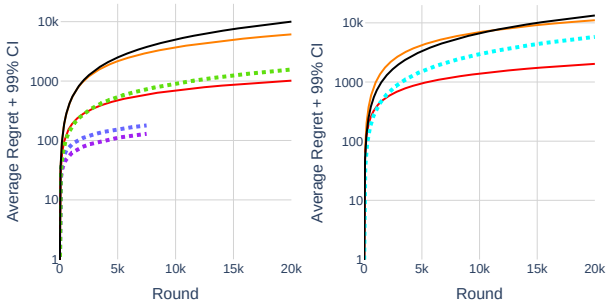


Figure 1: Average regret (with 99% confidence interval above) on non-contextual AT and LE games.

Performance metrics We measure the performance with the average non-contextual cumulative regret (Eq. 1) and the win-count (number of times a strategy achieves the lowest cumulative regret at end of the game). We perform a one sided Welch's t-test to assess if the cumulative regret of `RandCBP` at the end of the game is significantly lower than the baselines' regret.

Results Figure 1 shows the non-contextual cumulative regret for each strategy over the four configurations considered. Numeric details are reported in Table 2 and 3 of Appendix E. In all four cases, `RandCBP` is the best strategy in terms of average regret. `RandCBP` achieves a regret significantly lower (p-value<0.01) than all baselines in three settings (Figures 1a, 1c, and 1d) out of the four considered. In the balanced AT game (Figure 1b), `RandCBP` is not statistically different from `CBP` (p-value=0.055) and `TSPM` (p-value=0.854). For `CBP`, this can be attributed to its high variance (std=138). For `TSPM`, we observe from the win-count that `RandCBP` achieves lowest regret 37 times, against 13 for `TSPM`. Performance similarity between `RandCBP` and `TSPM` reflects the theoretical connections between randomizing confidence bounds and Thompson Sampling (Vaswani et al., 2020), on which `TSPM` is based.



(a) Apple Tasting (AT) (b) Label Efficient (LE)

Random RandCBPside⋆ CBPside⋆
PG-IDS PG-TS STAP CESA

Figure 2: Average regret (with 99% confidence interval above) on contextual games (10-d contexts).

5.2. Evaluation of `RandCBPside`*

Here, the outcome distribution is a linear function of 10-dimensional contexts (sampled uniformly in $[0,1]^{10}$) and a fixed unknown parameter $\theta \in \mathbb{R}^{M\times 10}$ with all values set at 0.1. From the uniform context distribution, we have 0.5 as mean for each context feature. Therefore, the resulting outcome distributions are more *balanced*. We run the experiment 96 times over a $T = 20$k horizon. We report the contextual cumulative regret (Eq. 2), the win-count and Welch's t-test.

Baselines The only PM baseline in this setting is `CBPside`*. We therefore resort to baselines that only apply on specific games. We consider `PG-IDS` (Grant et al., 2021), `PG-TS` (Grant et al., 2021), and `STAP` (Helmbold et al., 2000) for the AT game, and `CESA` (Cesa-Bianchi et al., 2006) for the LE game. Implementation and hyper-parameters details are reported in Appendix A.

Results Figure 2 shows the contextual cumulative regret for each strategy on AT and LE, with dotted-lines indicating game-specific baselines. Numeric details are reported in Table 4 (Appendix E). Over the horizon $T = 20$k, `RandCBPside`* achieves the best regret performance in both settings and significantly improves over `CBPside`*, `STAP`, and `CESA` (p-value<0.01 in AT and LE). In the AT game (Figure 2a), `PG-IDS` achieves the lowest regret on the truncated horizon $T = 7.5$k. However, `PG-IDS` and `PG-TS` scale in cubic time with the number of contexts due to the necessity to sample and invert matrices at each round, whereas `CBPside`* and `RandCBPside`* enjoy lower complexity thanks to the Sherman-Morison update (Sherman et al., 1950), making them practical on long horizons tasks. We emphasize that `PG-IDS`, `PG-TS`, `STAP`, and `CESA` are *game-specific*, unlike `CBPside`* and `RandCBPside`*.

5.3. Discussion on the empirical performance

Recall that the width of the randomized confidence bounds of `RandCBP` and `RandCBPside`* is in expectation smaller than the width of the deterministic confidence bounds of `CBP` and `CBPside`*. As a consequence, the successive elimination criterion (defined in Eq. 4) in charge of populating $\mathcal{U}(t)$ by separating high confidence from low confidence estimates is less restrictive when randomized confidence bounds are used. As a result, a specific loss difference estimate requires less exploration to be considered high confidence (i.e. to be added in $\mathcal{U}(t)$). The additional estimates included in $\mathcal{U}(t)$ bring additional constraints that are used to construct the probability subspace $D(t)$. In conclusion, $D(t)$ becomes tighter around the ground truth solution $p^\star$ or $p^\star(x_t)$ faster, resulting in a faster identification of potentially optimal actions, which translates into a smaller regret.

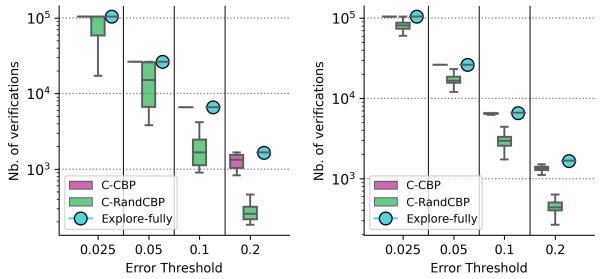
However, one limitation of the proposed approach is that it requires tuning hyper-parameters ϵ, σ and K, whereas hyper-parameters tuning is not easily achievable in online learning.

6. Use-case: Adaptive Monitoring of a Deployed Black-box Classifier

Partial monitoring has a reputation for being a complex framework due to its generality (Kirschner et al., 2023), which can hinder its adoption in real-world problems. Documented applied studies of PM do not emphasize on how to employ the framework towards an application (Singla et al., 2014; Kirschner et al., 2023). Here, we show how to formulate a real-world application as a PM problem, to encourage future applied research.

We consider the problem of cost-efficiently verifying the prediction error rate of a deployed black-box classifier. We assume a streaming setting where, at each round, the classifier receives an input and outputs probabilities to C classes. The index of the highest probability determines the *predicted class*. Each of the C predicted classes has an error rate p_c. The goal is to identify which predicted classes have an error rate greater than a tolerance threshold $\tau \in [0,1]$ while minimizing the number of verifications. In contrast to Kossen et al. (2021), who require a verification budget to be specified, our approach assumes no prior knowledge regarding the number of required verifications.

Problem formulation Everytime class c is predicted, a binary outcome is generated: either the classifier mispredicted (0) or not (1). Thus, the outcome distribution is $p_c^\star = [p_c, 1-p_c]$ where p_c denotes the error rate we aim to estimate for all classes $c \in \{1,\dots,C\}$. We design a PM game, that we name the **τ-detection** game, to estimate the outcome distribution over multiple rounds for a predicted



(a) Case 1 (b) Case 2

Figure 3: Cost-efficiency of the considered approaches (quartiles of number of verifications).

class c:

$$\mathbf{L} = \begin{matrix} \\ \text{verify} \\ \text{pass} \end{matrix} \begin{matrix} \begin{matrix} \text{error} & \text{no error} \end{matrix} \\ \begin{bmatrix} 1 & 1 \\ 1/\tau & 0 \end{bmatrix} \end{matrix}, \mathbf{H} = \begin{matrix} \\ \text{verify} \\ \text{pass} \end{matrix} \begin{matrix} \begin{matrix} \text{error} & \text{no error} \end{matrix} \\ \begin{bmatrix} \wedge & \odot \\ \bot & \bot \end{bmatrix} \end{matrix}.$$

After each classifier prediction (i.e. round in the stream), the PM agent can either require a verification (observation of the true class) or not (pass). The loss matrix is designed such that the optimal action is to pass when $p_c < \tau$ and to verify when $p_c \geq \tau$. The "verify" action is informative about the error rate p_c, but it has a fixed cost no matter the outcome. For reproducibility, we provide in Appendix B.4 the analysis of this game.

Experiment setup We simulate a variety of black-box classifiers by randomly generating confusion matrices with a global error rate lower than 10% (the black-box would not be deployed otherwise). Prediction errors from the black-box can be *uniformly* distributed across the classes or *non-uniformly* distributed. In addition, the distribution of the true classes in the stream can be *balanced* or *imbalanced*. We obtain four configurations (uniform/balanced, uniform/imbalanced, non-uniform/balanced, non-uniform/imbalanced). We consider the two opposite configurations: i) balanced true classes with uniform black-box errors (case 1), and ii) imbalanced true classes with non-uniform black-box errors (case 2).

We run the experiment 96 times. We consider four error tolerance thresholds $\tau \in \{0.025, 0.05, 0.1, 0.2\}$ and a classification task with $C = 10$ classes. We measure the mean and median f1-score to assess how accurate a given approach is at identifying predicted classes whose error rate exceeds τ, and the underlying average number of verifications used by each approach. For validation and comparison purposes, we consider a *maximum number of verifications* that one is willing to spend to estimate accurately each error rate. The maximum number of verifications is derived from Wald's confidence intervals formula (more details in Appendix E.4). The non-adaptive `Explore-fully` baseline consumes entirely the maximum number of verifications. We compare `Explore-fully` against the adaptive strategies `C-RandCBP` and `C-CBP`, which consist of C instances of `RandCBP` (resp. `CBP`) that play the τ-detection game.

Results Tables 5 and 6 (reported in Appendix E) show that `C-RandCBP`, `C-CBP` and `Explore-fully` all have an average f1-score is within the same range, indicating that the three strategies are equally effective in identifying predicted classes that exceed the threshold τ. In case 1, for the smallest threshold $\tau = 0.025$, `Explore-fully` and `C-CBP` have an average f1-score of 0.96 ± 0.15 and `C-RandCBP` of 0.95 ± 0.16. For $\tau = 0.2$, the average f1-score is equal to 1.0 for all strategies. Similar tendencies are observed for the other cases. Figure 3 shows that the number of verifications consumed by `C-RandCBP` to achieve the task is consistently lower than the one of `Explore-fully` and `C-CBP`. In case 1, `C-RandCBP` reduces the verification cost by 15% for a small error threshold ($\tau = 0.025$) and by 73% for $\tau = 0.2$, relatively to `Explore-fully`. In case 2, `C-RandCBP` reduces the verification cost by 18% for a small error threshold ($\tau = 0.025$) and by 62% at $\tau = 0.2$, relatively to `Explore-fully`.

7. Conclusion

This work extends randomization techniques (Kveton et al., 2019; Vaswani et al., 2020) designed for OFU-based methods in the bandit setting to successive elimination strategies in the more general partial monitoring framework. We show that it is possible to randomize `CBP`-based strategies (Bartók et al., 2012b;a), allowing to maintain the regret guarantees while improving significantly their empirical performance. In the contextual PM setting, we propose a correction to the seminal `CBPside`; the resulting `CBPside`* is the first strategy to enjoy regret guarantees on both easy and hard contextual games. Our proposed `RandCBP` and `RandCBPside`* demonstrate competitive performance against state-of-the-art baselines in multiple settings while maintaining regret guarantees. To further bridge the gap between theory and practice, we present a use case on the real-world problem of monitoring the error rate of deployed classifiers. Future research may consist in obtaining tighter regret bounds for `RandCBP` and `RandCBPside`*. Obtaining lower bounds in the contextual setting is another possible future research.

Contributions

Maxime Heuillet: conceptualization, methodology, theory (regret analysis), empirical investigation, visualizations, implementation, writing (original draft, editing), funding acquisition. Ola Ahmad: conceptualization, writing (review, editing), supervision. Audrey Durand: conceptualization, methodology, theory (insights, validation), writing (review, editing), supervision, funding acquisition.

Impact statement

This work aims to improve the practicality and efficiency of partial monitoring agents through randomization. A case study on the cost-efficient verification of deployed classifiers is presented. As such, the work makes a step towards improving the applicability of partial monitoring strategies in the real world. We encourage researchers and practitioners who build upon our work to be cautious of the possible misuse of partial monitoring, which could result in increased dependence on machine inference in decision-making tasks.

Acknowledgments

This work was funded through Mitacs with additional support from CIFAR (CCAI Chair). We thank Alliance Canada and Calcul Quebec for access to computational resources and staff expertise consultation. We would like to thank Junpei Komiyama, Taira Tsuchiya, Ian Lienert, Hastagiri P. Vanchinathan and James A. Grant for answering our technical questions and/or providing total/partial access to private code bases of their approaches. We also acknowledge the library pmlib of Tanguy Urvoy that was helpful to implement PM game environments. We thank Quentin Bertrand and Mathieu Godbout for reading our paper and providing valuable feedback.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *In Proc. NeurIPS*, 24, 2011.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Bartók, G. et al. Minimax regret of finite partial-monitoring games in stochastic environments. In *Proc. CoLT*, pp. 133–154. JMLR Workshop and Conference Proceedings, 2011.

Bartók, G. et al. Partial monitoring with side information. In *Proc. ALT*, 2012a.

Bartók, G. et al. An adaptive algorithm for finite stochastic partial monitoring. *In Proc. ICML*, 2012b.

Bartók, G. et al. Partial monitoring - classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

Cesa-Bianchi, N. et al. *Prediction, learning, and games*. Cambridge University Press, 2006.

Chapelle, O. et al. An empirical evaluation of thompson sampling. *In Proc. NeurIPS*, 24, 2011.

Even-Dar, E., Mannor, S., and Mansour, Y. Pac bounds for multi-armed bandit and markov decision processes. In *Proc. COLT*, 2002.

Grant, J. A. et al. Apple tasting revisited: Bayesian approaches to partially monitored online binary classification, 2021.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL https://www.gurobi.com.

Helmbold, D. et al. Some label efficient learning results. In *Proc. CoLT*, 1997.

Helmbold, D. P. et al. Apple tasting. *Information and Computation*, 161(2):85–139, 2000.

Kirschner, J., Lattimore, T., and Krause, A. Linear partial monitoring for sequential decision-making: Algorithms, regret bounds and applications. *JMLR*, 2023.

Kirschner, J. et al. Information directed sampling for linear partial monitoring. *PMLR*, 2020.

Kleinberg, R. et al. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pp. 594–605. IEEE, 2003.

Komiyama, J. et al. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. *In Proc. NeurIPS*, 28, 2015.

Kossen, J. et al. Active testing: Sample-efficient model evaluation. In *Proc. ICML*, 2021.

Kveton, B. et al. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *Proc. ICML*, 2019.

Lattimore, T. Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. *In Proc. CoLT*, 2022.

Lattimore, T. et al. Exploration by optimisation in partial monitoring. *In Proc. CoLT*, 2020.

Lienert, I. Exploiting side information in partial monitoring games: An empirical study of the cbp-side algorithm with applications to procurement. Master's thesis, Eidgenössische Technische Hochschule Zürich, Department of Computer Science, 2013.

Mitchell, S., OSullivan, M., and Dunning, I. Pulp: a linear programming toolkit for python. *The University of Auckland, Auckland, New Zealand*, 65, 2011.

Piccolboni, A. et al. Discrete prediction games with arbitrary feedback and loss. *In Proc. CoLT*, 2001.

Raman, V., Subedi, U., Raman, A., and Tewari, A. Apple tasting: Combinatorial dimensions and minimax rates. In *Proc. COLT*, 2024.

Sherman, J. et al. Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix. *The Annals of Mathematical Statistics*, 21(1):124 – 127, 1950.

Singla, A. et al. Contextual procurement in online crowdsourcing markets. In *Proc. AAAI*, 2014.

Tsuchiya, T. et al. Analysis and design of thompson sampling for stochastic partial monitoring. *In Proc. NeurIPS*, 33, 2020.

Tsuchiya, T. et al. Best-of-both-worlds algorithms for partial monitoring. In *Proc. ALT*, 2023.

Vanchinathan, H. P. et al. Efficient partial monitoring with prior information. *In Proc. NeurIPS*, 27, 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *In Proc. NeurIPS*, 30, 2017.

Vaswani, S. et al. Old dog learns new tricks: Randomized ucb for bandit problems. *In Proc. AISTATS*, 2020.

Xu, P., Wen, Z., Zhao, H., and Gu, Q. Neural contextual bandits with deep representation and shallow exploration. In *Proc. ICLR*, 2022.

Zhou, D., Li, L., and Gu, Q. Neural contextual bandits with ucb-based exploration. In *In Proc. ICML*, pp. 11492–11502. PMLR, 2020.

Notation	Definition	Observable by the agent?
N	Number of actions	✓
M	Number of outcomes	✓
Σ	Feedback space (space of symbols)	✓
$\mathbf{L} \in [\mathbf{0},\mathbf{1}]^{\mathbf{N}\times\mathbf{M}}$	Loss matrix	✓
L_i	Row i in matrix $\mathbf{L}$ (associated with action i)	✓
$\mathbf{H} \in \mathbf{\Sigma}^{\mathbf{N}\times\mathbf{M}}$	Feedback matrix	✓
H_i	Row i in matrix $\mathbf{H}$ (associated with action i)	✓
σ_i	Number of unique feedback symbols induced by action i (i.e. on row i of H)	✓
Δ_M	Probability simplex of dimension M (i.e. over the outcome space)	✓
Δ_{σ_i}	Probability simplex of dimension σ_i (i.e. over the symbol space induced by action i)	✓
T	Total number of rounds in a game (horizon)	✗
I_t	Action played by the agent at round t	✓
J_t	Outcome at round t	✗
$p^\star \in \Delta_M$	Outcome distribution	✗
$\mathbf{H}[\mathbf{I_t},\mathbf{J_t}]$	Element in matrix $\mathbf{H}$ at row I_t and column J_t (i.e. feedback received at round t)	✓
$\mathbf{L}[\mathbf{I_t},\mathbf{J_t}]$	Element in matrix $\mathbf{L}$ at row I_t and column J_t (i.e. loss incurred at round t)	✗
ℓ_i	Expected loss of action i	✗
$\hat{\ell}_i$	Estimated expected loss of action i (not observable, $p^\star$ can't be computed in practice)	✗
$\mathcal{X}$	Observation space	✗
$x_t \in \mathcal{X}$	Observation received at time t	✓
$X_{i,t} \in \mathbb{R}^{d\times t}$	History of observations for action i up to time t	✓
$Y_{i,t} \in \{0,1\}^{\sigma_i\times t}$	History of one-hot encoded feedback symbols for action i up to time t	✓
$G_{a,t} \in \mathbb{R}^{d\times d}$	Gram matrix for action a up to time t	✓
$\theta_i \in \mathbb{R}^{\sigma_i\times d}$	Parameter of the ridge regression of action i	✓
$p^\star(x_t) \in \Delta_M$	Outcome distribution in the contextual setting	✗
$\mathcal{C}_i \subseteq \Delta_M$	Cell of action i	✓
$S_i \in \{0,1\}^{\sigma_i\times M}$	Signal matrix of action i	✓
$\pi_i \in \Delta_{\sigma_i}$	Distribution for the unique feedback symbols induced by action i	✗
$\hat{\pi}_i \in \Delta_{\sigma_i}$	Estimated distribution for the unique feedback symbols induced by action i	✓
$\delta_{i,j}$	Expected loss difference between action i and j	✗
$\hat{\delta}_{i,j}$	Estimated expected loss difference between action i and j	✓
$n_i(t) \in \mathbb{N}$	Number of times action i was played up to time t	✓
$\nu_i(t) \in \mathbb{N}^{\sigma_i}$	Count for the unique symbols induced by action i up to time t	✓
$\mathcal{P}$	Set of Pareto optimal actions (i.e. set of actions)	✓
$\mathcal{N}$	Set of neighbor action pairs (i.e. set of pairs of actions)	✓
$\mathcal{U}(t)$	Set of confident action pairs (i.e. set of pairs of actions)	✓
$V_{i,j}$	Observer set for pair i,j (i.e. set of actions)	✓
v_{ija}	Observer vector associated with $V_{i,j}$ (index a indicates to which action in $V_{i,j}$ it is associated to)	✓
$c_{i,j}(t)$	Confidence for a pair $\{i,j\}$ at round t	✓
$[A,B]$	Randomization interval (values A and B set by the user)	✓
Z_{ijt}	Value sampled at time t for pair i,j in the discretized interval $[A,B]$ at round t	✓
ϵ	Probability of sampling value B (parameter of the randomization)	✓
K	Number of bins in the discretized distribution (parameter of the randomization)	✓
σ	Variance of the Gaussian distribution (parameter of the randomization)	✓
$c'_{i,j}(t)$	Randomized confidence for a pair $\{i,j\}$ at round t	✓
$D(t) \subseteq \Delta_M$	Sub-space of the simplex based on constraints in $\mathcal{U}(t)$, it includes $p^\star$ with high confidence	✓
$N^+_{i,j}$	Neighbor action set for pair i,j (set of actions)	✓
$\mathcal{V}$	Union of all the observer sets (set of actions)	✓
$\mathcal{P}(t)$	Plausible subset of $\mathcal{P}$ given $D(t)$ (set of actions)	✓
$\mathcal{N}(t)$	Plausible subset of $\mathcal{N}$ given $D(t)$ (set of pairs of actions)	✓
$\mathcal{R}(t)$	Set of underplayed actions at time t (set of actions)	✓
$e(\cdot)$	One hot encoding	✓
$\mathcal{S}(t)$	Final set of actions considered by CBP (set of actions)	✓
$W_a = \max_{\{i,j\}\in\mathcal{N}} \|v_{ija}\|_\infty$	Weight of an action	✓

Table 1: List of notations

A. Implementation Details for `CBP`-based Strategies

A.1. Pseudo-code for `CBPside`* and `RandCBPside`*

Algorithm 3 provides the pseudo-code of `CBPside` as defined by Lienert (2013) and our proposed `RandCBPside`*. Differences are highlighted in purple. The strategies are instantiated with the set of Pareto optimal actions $\mathcal{P}$ (see Definition 2.1), the set of neighbor pairs $\mathcal{N}$ (see Definition 3.1), parameters η_a for each action, the exploration parameter $\alpha > 1$ and the decaying exploring rate $f(t)$.

Remark A.1. Obtaining $\mathcal{P}(t)$ and $\mathcal{N}(t)$ at each round entails solving a computationally expensive optimization problem with evolving constraints. However, by caching the various half-spaces collected over time, the encountered problems can be buffered, significantly enhancing the overall computational complexity of the approach. In practice, Gurobi (Gurobi Optimization, LLC, 2023) or PULP (Mitchell et al., 2011) can be used to solve the optimization problems.

Remark A.2. In the contextual scenario, the update process of the inverse Gram matrix $G_{a,t}$ of action a at time t within `CBPside` and `RandCBPside`* can be efficiently implemented using the Sherman-Morrison update (Sherman et al., 1950) instead of relying on a costly matrix inversion operation.

Algorithm 3 `CBPside` (Lienert, 2013) and `RandCPBside`*

input : $\mathcal{P}, \mathcal{N}, \alpha, f(\cdot), \eta_a, K, \sigma, \varepsilon$
for $t = 1, 2, \ldots, N$ **do**
  Receive side-information x_t
  Play action $I_t = t$
  Observe feedback $\mathbf{H}[I_t, J_t]$
  $X_{i,t} = X_{i,t-1} \cup \{x_t\}$ if $I_t = i$ else $X_{i,t} = X_{i,t-1}, \forall i$
  $Y_{I_t,t} = Y_{I_t,t-1} \cup \{e(\mathbf{H}[I_t, J_t])\}$ if $I_t = i$ else $Y_{i,t} = Y_{i,t-1}, \forall i$
  Compute $G^{-1}_{I_t,t}$ (Sherman-Morrison update, Sherman et al. (1950))
  Update $\hat{\theta}_i(t) = Y_{i,t} X_{i,t}^\top (\lambda I_d + X_{i,t} X_{i,t}^\top)^{-1}$
for $t > N$ **do**
  Receive side-information x_t
  for $a = 1, \ldots, N$ **do**
    $\hat{\pi}_a(x_t) = \hat{\theta}_a x_t$
    ~~$w_a(x_t) = \sigma_a \left(\sqrt{d \log(t) + 2\log(1/t^2)} + \sigma_a \right) \|x_t\|_{G^{-1}_{a,t}}$~~
    $B = \sqrt{d \log(t) + 2 \log(1/t^2)}$
    Sample $Z_{a,t}$, according to Algorithm 2
    $w'_a(x_t) = \sigma_a \left(Z_{a,t} + \sigma_a \right) \|x_t\|_{G^{-1}_{a,t}}$
  for *each neighbor pair* $\{i, j\} \in \mathcal{N}$ **do**
    $\hat{\delta}_{i,j}(x_t) = \sum_{a \in V_{i,j}} v_{ija}^\top \hat{\pi}_a(x_t)$
    ~~$c_{i,j}(x_t) \leftarrow \sum_{a \in V_{i,j}} \|v_{ija}\|_2 w_a(x_t)$~~
    $c'_{ij}(x_t) \leftarrow \sum_{a \in V_{i,j}} \|v_{ija}\|_2 w'_a(x_t)$
    if $|\hat{\delta}_{i,j}(x_t)| >$ ~~$c_{i,j}(x_t)$~~ $c'_{i,j}(x_t)$ **then**
      Add pair $\{i, j\}$ to $\mathcal{U}(t)$
  Compute $D(t)$ based on $\mathcal{U}(t)$
  Get $\mathcal{P}(t)$ and $\mathcal{N}(t)$ from $\mathcal{P}, \mathcal{N}$ and $D(t)$
  $\mathcal{N}^+(t) \leftarrow \bigcup_{ij \in \mathcal{N}(t)} N^+_{ij}$
  $\mathcal{V}(t) \leftarrow \bigcup_{ij \in \mathcal{N}(t)} V_{ij}$
  ~~$\mathcal{R}(t) \leftarrow \{a \in \{1, \ldots, N\} : n_a(t) \leq \eta_a f(t)\}$~~
  $\mathcal{R}(x_t) \leftarrow \{a \in \{1, \ldots, N\} : 1/\|x_t\|^2_{G^{-1}_{a,t}} < \eta_a f(t)\}$
  $\mathcal{S}(t) \leftarrow \mathcal{P}(t) \cup \mathcal{N}^+(t) \cup (\mathcal{V}(t) \cap \mathcal{R}(x_t))$
  ~~Play $I_t = \operatorname{argmax}_{a \in \mathcal{S}(t)} W_a w_a(x_t)$~~
  Play $I_t = \operatorname{argmax}_{a \in \mathcal{S}(t)} W_a w'_a(x_t)$
  Observe feedback $\mathbf{H}[I_t, J_t]$
  $X_{i,t} = X_{i,t-1} \cup \{x_t\}$ if $I_t = i$ else $X_{i,t} = X_{i,t-1}$
  $Y_{I_t,t} = Y_{I_t,t-1} \cup \{e(\mathbf{H}[I_t, J_t])\}$ if $I_t = i$ else $Y_{i,t} = Y_{i,t-1}$
  Compute $G^{-1}_{I_t,t}$ (Sherman-Morrison update, Sherman et al. (1950))
  Update $\hat{\theta}_i(t) = Y_{i,t} X_{i,t}^\top (\lambda I_d + X_{i,t} X_{i,t}^\top)^{-1}$

B. Partial Monitoring Games

In this Appendix, we analyse the Apple Tasting (Helmbold et al., 2000), Label Efficient (Helmbold et al., 1997), and τ-detection games presented in the main paper. The analysis is necessary to implement partial monitoring environments and strategies based on these games.

B.1. Characterizing a partial monitoring game

A game is easy or hard depending on whether it verifies the *global observability* or *local observability* condition. *Easy* games refer to games that are locally observable while *hard* games verify the global observability condition but are not locally observable.

Definition B.1 (Global observability, Piccolboni et al. (2001))**.** A partial-monitoring game with $\mathbf{L}$ and $\mathbf{H}$ admits the *global observability* condition, if all pairs $\{i, j\}$ verify $L_i^\top - L_j^\top \in \oplus_{1\leq a\leq N}\text{Im}(S_a^\top)$.

Definition B.2 (Local observability, Bartók et al. (2012b))**.** A pair of neighbor actions i, j is *locally observable* if $L_i^\top - L_j^\top \in \oplus_{a\in N_{i,j}^+}\text{Im}(S_a^\top)$. 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 neighbor actions is locally observable, i.e., if $\mathcal{L} = \mathcal{N}$.

Remark B.1. When a pair is locally observable, we have $V_{ij} = N_{ij}^+$. For non-locally observable pairs, $V_{ij} = \{1, \dots, N\}$ is always a valid set Bartók et al. (2012b).

B.2. Apple Tasting Game

The Apple Tasting game is defined by the following loss and feedback matrices:

$$\mathbf{L} = \begin{matrix} \\ \text{action 1} \\ \text{action 2} \end{matrix} \begin{matrix} \begin{matrix} \text{A} & \text{B} \end{matrix} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{matrix}, \quad \mathbf{H} = \begin{matrix} \\ \text{action 1} \\ \text{action 2} \end{matrix} \begin{matrix} \begin{matrix} \text{A} & \text{B} \end{matrix} \\ \begin{bmatrix} \bot & \bot \\ \wedge & \odot \end{bmatrix} \end{matrix}.$$

This game has two possible actions and $N = 2$ actions and $M = 2$ outcomes (denoted A and B).

Signal Matrices: Signal matrices are such that $S_1 \in \{0, 1\}^{1\times 2}$ and $S_2 \in \{0, 1\}^{2\times 2}$. The matrices verify:

$$S_1 = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The outcome distribution is denoted $p^\star = [p_A, p_B]^\top$.

- $\pi_1^\star = S_1 p^\star = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} p_A \\ p_B \end{bmatrix} = 1$, there is only one feedback symbol ($\bot$) induced by action 1 therefore the probability of seeing this feedback symbol is always 1.
- $\pi_2^\star = S_2 p^\star = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} p_A \\ p_B \end{bmatrix} = \begin{bmatrix} p_A \\ p_B \end{bmatrix}$, therefore, the probability of seeing feedback $\wedge$ is p_A and the probability of seeing $\odot$ is p_B.

Cells: This game has 2 actions, each associated to a sub-space of the probability simplex:

- For action 1, we have: $\mathcal{C}_1 = \{p \in \Delta_M, \forall j \in \{1, \dots, N\}, (L_1 - L_j)p \leq 0\}$. This probability space corresponds to the following constraints:

$$\begin{bmatrix} L_1 - L_1 \\ L_1 - L_2 \end{bmatrix} p = \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix} p \leq 0$$

  The first constraint $(L_1 - L_1)p \leq 0$ is always verified. The second constraint $(L_1 - L_2)p \leq 0$ implies $p_A - p_B \leq 0$.

- For action 2, we have: $\mathcal{C}_2 = \{p \in \Delta_M, \forall j \in \{1, \dots, N\}, (L_2 - L_j)p \leq 0\}$. This probability space corresponds to the following constraints:

$$\begin{bmatrix} L_2 - L_1 \\ L_2 - L_2 \end{bmatrix} p = \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix} p \leq 0$$

  The second constraint $(L_2 - L_2)p \leq 0$ is always verified. The first constraint $(L_2 - L_1)p \leq 0$ implies $-p_A + p_B \leq 0$.

Pareto optimal actions: The cell respective to each action is neither empty nor included one in another. Therefore, according to Definition 2.1, both actions 1 and 2 are Pareto optimal, i.e. $\mathcal{P} = \{1, 2\}$

Neighbor actions: The space corresponding to $\mathcal{C}_1 \cap \mathcal{C}_2$ includes only one unique point, being $\begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$. Therefore, $\dim(C_1 \cap C_2) = 0 = M - 2$, which satisfies Definition 3.1. This implies that actions 1 and 2 are neighbors, i.e. $\mathcal{N} = \{\{1, 2\}, \}$.

Neighbor action set: This set includes: $N_{12}^+ = N_{21}^+ = [1, 2]$.

Observability of the game: The action pair $\{1, 2\}$ is locally observable because $L_1^\top - L_2^\top$ can be expressed from the set of vectors included in $\text{Im}(S_1^\top) \bigoplus \text{Im}(S_2^\top)$. We can conclude that the game is globally and locally observable. Therefore, it can be classified as an *easy game*.

Observer set: The pair $\{1, 2\}$ is locally observable. According to Definition B.2, we have: $V_{12} = N_{12}^+ = [1, 2]$. The pair of actions 2 and 1 is also locally observable therefore $V_{21} = N_{21}^+ = \{1, 2\}$.

Observer vector: For the pair of actions 1 and 2, we have to find $v_{ija}, a \in V_{ij}$ such that $L_1^\top - L_2^\top = \sum_{a \in V_{ij}} S_i^T v_{ija}$, according to Definition 3.3. Choosing and $v_{121} = 0$ and $v_{122}^\top = \begin{bmatrix} 1 & -1 \end{bmatrix}$ verifies the relation:

$$L_1^\top - L_2^\top = \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \langle \begin{bmatrix} 1 \\ 1 \end{bmatrix}, 0 \rangle + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{7}$$

It suffices to reproduce the same procedure for pair of actions 2 and 1.

B.3. Label Efficient Game

The Label Efficient game (Helmbold et al., 1997) is defined by the following loss and feedback matrices:

$$\mathbf{L} = \begin{array}{c} \\ \text{action 1} \\ \text{action 2} \\ \text{action 3} \end{array} \begin{array}{c} \begin{array}{cc} \text{A} & \text{B} \end{array} \\ \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}, \quad \mathbf{H} = \begin{array}{c} \\ \text{action 1} \\ \text{action 2} \\ \text{action 3} \end{array} \begin{array}{c} \begin{array}{cc} \text{A} & \text{B} \end{array} \\ \begin{bmatrix} \perp & \odot \\ \wedge & \wedge \\ \wedge & \wedge \end{bmatrix} \end{array}.$$

The game includes a set of $N = 3$ possible actions and $M = 2$ possible outcomes (denoted A and B).

Signal Matrices: The dimension of the signal matrices are such that $S_1 \in \{0, 1\}^{2 \times 2}$, $S_2 \in \{0, 1\}^{1 \times 2}$ and $S_3 \in \{0, 1\}^{1 \times 2}$. The matrices verify:

$$S_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

The outcome distribution is noted $p^\star = [p_A, p_B]^\top$.

Cells: Each action can be associated to a sub-space of the probability simplex noted *cell* (see Definition 2.1):

- For action 1, we have: $\mathcal{C}_1 = \{p \in \Delta_M, \forall j \in \{1, \dots, N\}, (L_1 - L_j)p \leq 0\}$. This probability space corresponds to the following constraints:
$$\begin{bmatrix} L_1 - L_1 \\ L_1 - L_2 \\ L_1 - L_3 \end{bmatrix} p = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} p \leq 0$$
The first constraint $(L_1 - L_1)p \leq 0$ is always verified. The second constraint $(L_1 - L_2)p \leq 0$ implies $p_B \leq 0$ and the third constraint $(L_1 - L_3)p \leq 0$ implies $p_A \leq 0$. There exist no probability vector in Δ_M satisfying these three constraints at the same time.

- For action 2, we have: $\mathcal{C}_2 = \{p \in \Delta_M, \forall j \in \{1, \dots, N\}, (L_2 - L_j)p \leq 0\}$. This probability space corresponds to the following constraints:
$$\begin{bmatrix} L_2 - L_1 \\ L_2 - L_2 \\ L_2 - L_3 \end{bmatrix} p = \begin{bmatrix} 0 & -1 \\ 0 & 0 \\ 1 & -1 \end{bmatrix} p \leq 0$$
The second constraint $(L_2 - L_2)p \leq 0$) is always verified. The first constraint $(L_2 - L_1)p \leq 0$ implies $-p_B \leq 0 \iff p_B \geq 0$. The third constraint $(L_2 - L_3)p \leq 0$ implies $p_A - p_B \leq 0 \iff p_A \leq p_B$.

- For action 3, we have: $\mathcal{C}_3 = \{p \in \Delta_M, \forall j \in \{1, \dots, N\}, (L_3 - L_j)p \leq 0\}$. This probability space corresponds to the following constraints:
$$\begin{bmatrix} L_3 - L_1 \\ L_3 - L_2 \\ L_3 - L_3 \end{bmatrix} p = \begin{bmatrix} -1 & 0 \\ -1 & 1 \\ 0 & 0 \end{bmatrix} p \leq 0$$
The third constraint $(L_3 - L_3)p \leq 0$ is always satisfied. The second constraint $(L_3 - L_1)p \leq 0$ implies $-p_A + p_B \leq 0 \iff p_B \geq p_A$. The first constraint $((L_3 - L_1)p \leq 0)$ implies $-p_A \leq 0 \iff p_A \geq 0$.

Pareto optimal actions: From the analysis of the cells, we have $\mathcal{C}_1 = \emptyset$. Therefore, action 1 is dominated, according to Definition 2.1. The remaining actions 2 and 3 are Pareto optimal because their respective cells are not included in one another, i.e. $\mathcal{P} = \{2, 3\}$.

Neighbor actions: In this paragraph, we will determine whether action 2 and 3 are a neighbor pair.

$$\mathcal{C}_1 \cap \mathcal{C}_2 = \begin{cases} p_B \geq 0 \\ p_A \leq p_B \\ p_B \leq p_A \\ p_A \geq 0 \end{cases}$$

The only point in this vector space is $\begin{bmatrix} 0.5 & 0.5 \end{bmatrix}^\top$. Therefore, $\dim(\mathcal{C}_1 \cap \mathcal{C}_2) = 0 = M - 2$ and the pair $\{2, 3\}$ is a neighbor pair, i.e. $\mathcal{N} = \{\{2, 3\}, \}$.

Neighbor action set: This set is defined as $N_{ij}^+ = \{k \in \{1, \dots, N\}, \mathcal{C}_i \cap \mathcal{C}_j \subset \mathcal{C}_k\}$. This yields: $N_{23}^+ = N_{32}^+ = [2, 3]$ because the cell of action 1 is empty.

Observability of the game: The pair $\{2, 3\}$ is not locally observable because it is not possible to express $L_2^\top - L_3^\top$ from $\bigoplus_{i \in N_{23}^+} \text{Im}(S_i^\top)$. On the contrary, it is possible to express $L_2^\top - L_3^\top$ from $\bigoplus_{1 \leq i \leq N} \text{Im}(S_i^\top)$. We can conclude that the game is not locally observable and that the pair $\{2, 3\}$ is globally observable. Therefore, the Label Efficient game belongs to the class of hard games.

Observer set: The pair $\{2, 3\}$ is not locally observable. According to Definition B.2, we have: $V_{23} = \{1, \dots, N\}$ same applies to $V_{32} = \{1, \dots, N\}$.

Observer vector: For the pair $\{2,3\}$, we have to find $v_{ija}, a \in V_{ij}$ such that $L_2^\top - L_3^\top = \sum_{a \in V_{ij}} S_i^T v_{ija}$, according to Definition 3.3. Choosing and $v_{231}^\top = \begin{bmatrix}-1 & 1\end{bmatrix}$, $v_{232} = 0$ and $v_{233} = 0$ verifies the relation:

$$L_2^\top - L_3^\top = \begin{bmatrix}-1\\1\end{bmatrix} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}-1\\1\end{bmatrix} + \begin{bmatrix}1\\1\end{bmatrix}0 + \begin{bmatrix}1\\1\end{bmatrix}0 \tag{8}$$

It suffices to reproduce the same procedure for pair of actions 3 and 2.

B.4. τ-detection Game

Let us consider the τ-detection game, with $\tau \in]0,1[$. The game is defined by the following loss and feedback matrices:

$$\mathbf{L} = \begin{matrix}\\ \text{action 1}\\ \text{action 2}\end{matrix}\begin{matrix}\begin{matrix}\text{A} & \text{B}\end{matrix}\\ \begin{bmatrix}1 & 1\\ 1/\tau & 0\end{bmatrix}\end{matrix}, \mathbf{H} = \begin{matrix}\\ \text{action 1}\\ \text{action 2}\end{matrix}\begin{matrix}\begin{matrix}\text{A} & \text{B}\end{matrix}\\ \begin{bmatrix}\wedge & \odot\\ \bot & \bot\end{bmatrix}\end{matrix}.$$

This game includes a set of $N = 2$ possible actions and $M = 2$ possible outcomes (denoted A and B).

Signal Matrices: The dimension of the signal matrices are such that $S_1 \in \{0,1\}^{2\times 2}$ and $S_2 \in \{0,1\}^{1\times 2}$. The matrices verify:

$$S_1 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, \quad S_2 = \begin{bmatrix}1 & 1\end{bmatrix}$$

Consider a general instance of the problem where the outcome distribution is $p^\star = [p_A, p_B]^\top$.

Cells: This game has two actions, each can be associated to a cell:

- For action 1, we have: $\mathcal{C}_1 = \{p \in \Delta_M, \forall j \in \{1,\dots,N\}, (L_1 - L_j)p \le 0\}$. This probability space corresponds to the following constraints:
$$\begin{bmatrix}L_1 - L_1\\ L_1 - L_2\end{bmatrix} p = \begin{bmatrix}0 & 0\\ 1 - 1/\tau & 1\end{bmatrix} p \le 0$$
The first constraint $(L_1 - L_1)p \le 0$ is always verified. The second constraint $(L_1 - L_2)p \le 0$ implies $p_A(2 - 1/\tau) \le 1$.
- For action 2, we have: $\mathcal{C}_2 = \{p \in \Delta_M, \forall j \in \{1,\dots,N\}, (L_2 - L_j)p \le 0\}$. This probability space corresponds to the following constraints:
$$\begin{bmatrix}L_2 - L_1\\ L_2 - L_2\end{bmatrix} p = \begin{bmatrix}1/\tau - 1 & -1\\ 0 & 0\end{bmatrix} p \le 0$$
The second constraint $(L_2 - L_2)p \le 0)$ is always verified. The first constraint $(L_2 - L_1)p \le 0$ implies $p_A \le \tau$.

Pareto optimal actions: The cell respective to each action is neither empty nor included one in another. Therefore, according to Definition 2.1, both actions 1 and 2 are Pareto optimal, i.e. $\mathcal{P} = \{1,2\}$

Neighbor actions: For values of $\tau \in]0,1[$, $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$. Therefore, $\dim(C_1 \cap C_2) = 0$, which satisfies the definition 3.1. This implies that actions 1 and 2 are neighboring actions, i.e. $\mathcal{N} = \{\{1,2\},\}$.

Neighbor action set: This set is defined as $N_{ij}^+ = \{k \in \{1,\dots,N\}, \mathcal{C}_i \cap \mathcal{C}_j \subset C_k\}$. This yields: $N_{12}^+ = N_{21}^+ = [1,2]$.

Observability of the game: The action pair $\{1,2\}$ is locally observable because $L_1^\top - L_2^\top = \begin{bmatrix}1 - 1/\tau & 1\end{bmatrix}$ can be expressed from the set of basis vectors included in $\text{Im}(S_1) \bigoplus \text{Im}(S_2)$ (see Definition B.2). Since this also applies to the pair $\{2,1\}$, we can conclude that the game is globally and locally observable. Therefore, it can be classified as an *easy game*.

Observer set: The pair $\{1,2\}$ is locally observable. According to the definition 3.2, we have: $V_{12} = N_{12}^+ = [1,2]$. The pair $\{2,1\}$ being also locally observable, we have $V_{21} = N_{21}^+ = \{1,2\}$.

Observer vector: For the pair $\{1,2\}$, we have to find $v_{ija}, a \in V_{ij}$ such that $L_1^\top - L_2^\top = \sum_{a \in V_{ij}} S_i^T v_{ija}$. Choosing and $v_{121} = 0$ and $v_{122}^\top = \begin{bmatrix} 1 & -(1-b_{opt}) \end{bmatrix}$ verifies the relation:

$$L_1^\top - L_2^\top = \begin{bmatrix} 1 - 1/\tau \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} 0 + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -(1-b_{opt}) \end{bmatrix}, \tag{9}$$

where b_{opt} satisfies the constraint $1/b_{opt} - \tau = 0$

C. Regret Analysis of `RandCBP`

In this section, we provide an upper bound on the expected regret of `RandCBP`. The incidence of randomization on *Upper confidence bound* strategies was characterized by Kveton et al. (2019) and Vaswani et al. (2020). CBP-based strategies belong instead to the class of *Successive Elimination* strategies, which utilize both upper and lower confidence bounds.

Let $\delta_i = \max_{1 \le j \le N} \delta_{ij}$ be the sub-optimality gap between the expected loss of action i and the optimal action. Similarly to Bartók et al. (2012a), define g_i as

$$g_i = \max_{\mathcal{P}', \mathcal{N}' \in \Psi, i \in \mathcal{P}'} \min_{\pi \in B_i(\mathcal{N}'), \pi = (i_0, \dots, i_r)} \sum_{s=1}^{r} |V_{i_{s-1}, i_s}| \tag{10}$$

where Ψ corresponds to the set of plausible configurations and $B_i(\mathcal{N}')$ the set of possible paths. The quantity g_i is correlated with the number of actions N.

C.1. Regret decomposition of RandCBP

Assuming action 1 is optimal:

$$\mathbb{E}[R(T)] = \mathbb{E}[\sum_{t=1}^{T} (L_{I_t} - L_1) p^\star] \tag{11}$$

$$= \sum_{k=1}^{N} \mathbb{E}[n_k(T)] \delta_k \tag{12}$$

The goal is to bound $\mathbb{E}[n_k(T)]$. Define the event $\mathcal{E}_t$: "the confidence interval succeeds"[1]. Formally, $\mathcal{E}_t = \{|\hat{\delta}_{i,j}(t) - \delta_{i,j}| \le c_{i,j}(t)\}$. The event $\mathcal{E}_t$ induces the following decomposition:

$$\mathbb{E}[n_k(T)] = \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = k\}}] \tag{13}$$

$$= \underbrace{\mathbb{E}[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = k, \mathcal{E}_t\}}]}_{A_k} + \underbrace{\mathbb{E}[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = k, \mathcal{E}_t^c\}}]}_{B_k} \tag{14}$$

The regret can thus be expressed as:

$$\mathbb{E}[R(T)] = \sum_{k=1}^{N} \delta_k A_k + \delta_k B_k \tag{15}$$

To obtain an upper bound on the regret of `RandCBP`, we need to upper bound the terms A_k and B_k. The bound of A_k is reported in Section C.2. The bound of B_k is reported in Section C.3. The theorem that follows is obtained by combining Eq. 15 and the analyses from Sections Section C.2 and C.3.

Theorem C.1. Consider the randomization over K bins in the interval $[A, B]$, a probability ϵ on the tail and a standard deviation σ. Setting $\eta_a = W_k^{2/3}$, $f(t) = \alpha^{1/3} t^{2/3} \log^{1/3}(t)$ and, with the notations $W = \max_{1 \le a \le N} W_a$, $\mathcal{V} = \bigcup_{i,j \in \mathcal{N}} V_{i,j}$,

[1] We reverse the notation used in Bartók et al. (2012b).

and $N^+ = \bigcup_{i,j \in \mathcal{N}} N^+_{i,j}$, we obtain:

$$\begin{aligned}\mathbb{E}[R_T] \leq \quad & \sum_{1 \leq k \leq N} \left[2(1 + \frac{1}{2\alpha - 2})|\mathcal{V}| + 1 \right] + \sum_{k=1}^{N} \delta_k + \\ & \sum_{k=1, \delta_k > 0}^{N} 4W_k^2 \frac{g_k^2}{\delta_k} \alpha \log(T) + \\ & \sum_{k \in \mathcal{V} \setminus N^+} \delta_k \min(4W_k^2 \frac{g_{l(k)}^2}{\delta_{l(k)}^2} \alpha \log(T), \alpha^{1/3} W_k^{2/3} T^{2/3} \log^{1/3}(T)) + \\ & \sum_{k \in \mathcal{V} \setminus N^+} \delta_k \alpha^{1/3} W_k^{2/3} T^{2/3} \log^{1/3}(T) + 2 g_k \alpha^{1/3} W^{2/3} T^{2/3} \log^{1/3}(T)\end{aligned} \tag{16}$$

On easy games, we have $\mathcal{V} \setminus N^+ = \emptyset$. The theorem implies a bound on the individual regret of `RandCBP` on easy games:

Corollary C.2. Consider an easy game, and the same assumptions as in Theorem C.1. Then:

$$\mathbb{E}[R_T] \leq \quad N \left[2(1 + \frac{1}{2\alpha - 2})|\mathcal{V}| + 1 \right] + \sum_{k=1}^{N} \delta_k + \sum_{k=1, \delta_k > 0}^{N} 4W_k^2 \frac{g_k^2}{\delta_k} \alpha \log(T).$$

Corollary C.2 matches the upper bound on the regret of `CBP` on the time horizon (Bartók et al., 2012b). The first term corresponds to the confidence interval of the failure event. The second term comes from the initialization phase of the algorithm. The third term comes from the exploration-exploitation trade-off achievable on easy games.

Corollary C.3. Consider a hard game and the same assumptions as in Theorem C.1. Then, there exists a constant C_1 and C_2 such that the expected regret can be upper bounded independently of the choice of $p^\star$ as

$$\mathbb{E}[R_T] \leq C_1 N + C_2 T^{2/3} \log^{1/3}(T)$$

The regret bound of `RandCBP` on hard games matches `CBP`'s on hard games on the time horizon (Bartók et al., 2012b). Note that the bound on hard games is problem-independent unlike the bound on easy games.

C.2. Bounding A_k

This part is quite similar to that of Bartók et al. (2012b), except that the underlying Lemma C.4 has been adapted for the randomized confidence bounds. We include the steps for completeness.

The notation I_t corresponds to the action that was effectively played at round t. Define $k(t) = \text{argmax}_{i \in \mathcal{P}(t) \cup V(t)} W_i^2 / n_i(t)$. The event $k(t) \neq I_t$ happens when $k(t) \notin N^+(t)$ and $k(t) \notin \mathcal{R}(t)$, i.e. $k(t)$ is a purely information seeking (exploratory) action which has been sampled frequently. This corresponds to the event $\mathcal{D}_t = \{k(t) \neq I_t\}$ = "the decaying exploration rule is in effect at time t" .

We can decompose:

$$\mathbb{E}[\sum_{t=1}^{T} \mathbb{1}_{\{I_t=k,\mathcal{E}_t\}}]\delta_k \leq \delta_k + \underbrace{\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t,\mathcal{D}_t^c,k\in\mathcal{P}(t)\cup N^+(t),I_t=k\}}]\delta_k}_{A_1} + \underbrace{\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t,\mathcal{D}_t^c,k\notin\mathcal{P}(t)\cup N^+(t),I_t=k\}}]\delta_k}_{A_2} + \underbrace{\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t,\mathcal{D}_t,k\in\mathcal{P}(t)\cup N^+(t),I_t=k\}}]\delta_k}_{A_3} + \underbrace{\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t,\mathcal{D}_t,k\notin\mathcal{P}(t)\cup N^+(t),I_t=k\}}]\delta_k}_{A_4} \tag{17}$$

The first δ_k corresponds to the initialization phase of the algorithm when every action is chosen once. The next paragraphs are devoted to upper bounding the remaining four expressions A_1, A_2, A_3 and A_4, using the results from Lemma C.4. Note that, if action k is optimal, then $\delta_k = 0$, so all the terms are zero. Thus, we can assume from now on that $\delta_k > 0$.

Term A_1: Consider the event $\mathcal{E}_t \cap \mathcal{D}_t^c \cap \{k \in \mathcal{P}(t) \cup N^+(t)\}$. Using case 2 from Lemma C.4 with the choice $k = i$. Thus, from $I_t = i$, we get that $I_t = i = k \in \mathcal{P}(t) \cup N^+(t)$. The result of the lemma gives:

$$n_k(t) \leq A_k(t) = 4W_k^2 \frac{g_k^2}{\delta_k^2} \alpha \log(t)$$

Therefore, we have

$$\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t,\mathcal{D}_t^c,k\in\mathcal{P}(t)\cup N^+(t),I_t=k\}} \tag{18}$$

$$\leq \sum_{t=N+1}^{T} \mathbb{1}_{\{I_t=k,n_k(t)\leq A_k(t)\}} + \sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t,\mathcal{D}_t,k\notin\mathcal{P}(t)\cup N^+(t),I_t=k,n_k(t)>A_k(t)\}} \tag{19}$$

$$= \sum_{t=N+1}^{T} \mathbb{1}_{\{I_t=k,n_k(t)\leq A_k(t)\}} \tag{20}$$

$$\leq A_k(T) = 4W_k^2 \frac{g_k^2}{\delta_k^2} \alpha \log(T) \tag{21}$$

Consequently,

$$\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t,\mathcal{D}_t^c,k\in\mathcal{P}(t)\cup N^+(t),I_t=k\}}\delta_k \leq 4W_k^2 \frac{g_k^2}{\delta_k} \alpha \log(T) \tag{22}$$

Term A_2: Consider the event $\mathcal{E}_t \cap \mathcal{D}_t^c \cap \{k \notin \mathcal{P}(t) \cup N^+(t)\}$. From case 2 of Lemma C.4. The Lemma gives:

$$n_k(t) \leq \min_{j\in\mathcal{P}(t)\cup N^+(t)} 4W_k^2 \frac{g_j^2}{\delta_j} \alpha \log(T)$$

We know that $k \in \mathcal{V}(t) = \bigcup_{k,j \in \mathcal{N}(t)} V_{i,j}$. 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 $l'_{\{i,j\}} = \text{argmax}_{l \in \{i,j\}} \delta_l$ then:

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

If we define $l(k)$ as the action with

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

Then, it follows that:

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

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:

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

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

$$\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t, \mathcal{D}_t^c, k \notin \mathcal{P}(t) \cup N^+(t), I_t = k\}}] \leq \delta_k \min\Big\{4W_k^2 \frac{g^2_{l(k)}}{\delta^2_{l(k)}} \alpha \log(t), \eta_k f(t)\Big\}$$

Term A_3: Consider the event $\mathcal{E}_t \cap D_t \cap \{k \in \mathcal{P}(t) \cup N^+(t)\}$. From Lemma C.4 we have that:

$$\delta_k \leq 2g_k \sqrt{\frac{\alpha \log(T)}{f(t)}} \max_{1 \leq l \leq N} \frac{W_l}{\sqrt{\eta_l}}$$

Thus,

$$\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t, \mathcal{D}_t, k \in \mathcal{P}(t) \cup N^+(t), I_t = k\}}] \leq g_k \sqrt{\frac{\alpha \log(T)}{f(T)}} \max_{1 \leq l \leq N} \frac{W_l}{\sqrt{\eta_l}}$$

Term A_4: Consider the event $\mathcal{E}_t \cap D_t \cap \{k \notin \mathcal{P}(t) \cup N^+(t)\}$ we know that $k \in \mathcal{V}(t) \cap \mathcal{R}(t) \subseteq \mathcal{R}(t)$ and hence $n_k(t) \leq \eta_k f(t)$. With the same argument as in the first and second term, we get that:

$$\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t, \mathcal{D}_t, k \notin \mathcal{P}(t) \cup N^+(t), I_t = k\}}] \leq \delta_k \eta_k f(T)$$

C.3. Bounding term B_k:

In the analysis of B_k, the goal is to upper-bound the probability that the confidence interval fails. For the deterministic `CBP`, this corresponds to Lemma 1 in Bartók et al. (2012b). `RandCBP` uses instead randomized confidence intervals. Following the terminology in Vaswani et al. (2017), we use *uncoupled randomized confidence intervals* because we sample a value for each action pair.

For a pair of actions $\{i,j\} \in \mathcal{N}$, at a time t, note $Q_{ij}(t)$ the probability that the confidence interval of pair $\{i,j\}$ fails:

$$Q_{i,j}(t) = \mathbb{P}_{Z_{ijt}}(\{\delta_{i,j} < \hat{\delta}_{i,j}(t) - c'_{i,j}(t)\} \cup \{\delta_{i,j} > \hat{\delta}_{i,j}(t) + c'_{i,j}(t)\}) \quad (23)$$

$$= \mathbb{P}_{Z_{ijt}}(|\hat{\delta}_{i,j}(t) - \delta_{i,j}| > c'_{i,j}(t)) \quad (24)$$

The event $\mathcal{E}_t^c$ is unlikely to occur when T is large; let

$$\Upsilon_k = \{t \in [T], \forall \{i,j\} \in \mathcal{N}, Q_{i,j}(t) > \frac{1}{T}\}$$

be the set of time steps where the probability of failure is non-negligible, i.e. is higher than $1/T$. Following Kveton et al. (2019), the regret can be decomposed according to Υ_k:

$$\mathbb{E}[\sum_{t=1}^{T} \mathbb{1}_{\{I_t=k, \mathcal{E}_t^c\}}] = \mathbb{E}[\sum_{t \in \Upsilon_k} \mathbb{1}_{\{I_t=k\}}] + \mathbb{E}[\sum_{t \notin \Upsilon_k} \mathbb{1}_{\{\mathcal{E}_t^c\}}] \tag{25}$$

$$\leq \mathbb{E}[\sum_{t=0}^{T} \sum_{\{i,j\} \in \mathcal{N}} \mathbb{1}_{\{Q_{i,j}(t) > \frac{1}{T}\}}] + \mathbb{E}[\sum_{t \notin \Upsilon_k} \frac{1}{T}] \tag{26}$$

$$\leq \sum_{t=0}^{T} \sum_{\{i,j\} \in \mathcal{N}} \mathbb{P}_{\hat{\delta}_{i,j}(t)}(Q_{i,j}(t) > \frac{1}{T}) + 1 \tag{27}$$

For a given pair $\{i,j\} \in \mathcal{N}$, and for a specific time t, define:

$$b_{i,j}(t) = \quad \mathbb{P}_{\hat{\delta}_{i,j}(t)}\left[Q_{i,j}(t) > \frac{1}{T}\right] \tag{28}$$

$$= \quad \mathbb{P}_{\hat{\delta}_{i,j}(t)}\left[\mathbb{P}_{Z_{ijt}}(|\hat{\delta}_{i,j}(t) - \delta_{i,j}| \geq c'_{i,j}(t)) > \frac{1}{T}\right] \tag{29}$$

By definition of Z_{ijt} (that are sampled from a discrete probability distribution) we have:

$$b_{i,j}(t) = \quad \mathbb{P}_{\hat{\delta}_{i,j}(t)}\left[\mathbb{P}_{Z_{ijt}}(|\hat{\delta}_{i,j}(t) - \delta_{i,j}| \geq c'_{i,j}(t)) > \frac{1}{T}\right] \tag{30}$$

$$= \quad \mathbb{P}_{\hat{\delta}_{i,j}(t)}\left[\sum_{k=1}^{K} p_k \mathbb{1}_{\{|\hat{\delta}_{i,j}(t) - \delta_{i,j}| \geq c^k_{i,j}(t)\}} > \frac{1}{T}\right] \tag{31}$$

where $c^k_{i,j}(t)$ denotes the confidence interval associated to the sampled value ρ_k. Since $p_K > \frac{1}{T}$, we have:

$$b_{i,j}(t) = \quad \mathbb{P}_{\hat{\delta}_{i,j}(t)}(|\hat{\delta}_{i,j}(t) - \delta_{i,j}| \geq c^K_{i,j}(t)) \tag{32}$$

$$= \quad \mathbb{P}_{\hat{\delta}_{i,j}(t)}(|\hat{\delta}_{i,j}(t) - \delta_{i,j}| > \sum_{a \in V_{i,j}} ||v_{ija}||_\infty \frac{\rho_K}{\sqrt{n_a(t)}}) \tag{33}$$

$$\leq \quad \sum_{a \in V_{i,j}} \sum_{s=1}^{t} \mathbb{P}_{\hat{\delta}_{i,j}}(\hat{\delta}_{ij}(s) - \delta_{i,j} > ||v_{ija}||_\infty \frac{\rho_K}{\sqrt{l}}) \mathbb{1}_{\{n_a(t)=s\}} \tag{34}$$

$$\leq \quad \sum_{a \in V_{i,j}} \sum_{s=1}^{t} 2\exp(-2s(||v_{ija}||_\infty \frac{\rho_K}{\sqrt{s}})^2) \mathbb{1}_{\{n_a(t)=s\}} \tag{35}$$

$$\leq \quad \sum_{a \in V_{i,j}} \sum_{s=1}^{t} 2\exp(-2||v_{ija}||^2_\infty \rho^2_K) \mathbb{1}_{\{n_a(t)=s\}} \tag{36}$$

$$\leq \quad \sum_{a \in V_{i,j}} 2\exp(-2\rho^2_K) \sum_{s=1}^{t} \mathbb{1}_{\{n_a(t)=s\}} \tag{37}$$

$$\leq \quad \sum_{a \in V_{i,j}} 2\exp(-2\rho^2_K) \tag{38}$$

$$\leq \quad 2|V_{i,j}|\exp(-2\rho^2_K) \tag{39}$$

$$\tag{40}$$

Where the Hoeffding's inequality was used in 35. Therefore,

$$B_k \leq \sum_{t=1}^{T} \sum_{\{i,j\}\in\mathcal{N}} b_{i,j}(t) + 1 \tag{41}$$

$$\leq \sum_{t=1}^{T} \sum_{\{i,j\}\in\mathcal{N}} 2|V_{i,j}|\exp(-2\rho_K^2) + 1 \tag{42}$$

$$\leq 2|\mathcal{V}|\exp(-2\rho_K^2)T + 1 \tag{43}$$

The linear dependency on T is cancelled with $\rho_K = \sqrt{\alpha \log(T)}$ and for $\alpha > 1$, we have:

$$\leq 2(1 + \frac{1}{2\alpha - 2})|\mathcal{V}|+1, \tag{44}$$

where $\mathcal{V} = \bigcup_{i,j\in\mathcal{N}} V_{i,j}$.

C.4. Proofs of lemmas

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

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

$$\delta_i \leq 2g_i\sqrt{\frac{\alpha \log(t)}{f(t)}} \max_{1\leq k\leq N} \frac{W_k}{\sqrt{\eta_k}}$$

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

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

Proof. Observe that for any neighboring action pair $\{i, j\} \in \mathcal{N}(t)$, on $\mathcal{E}_t$, it holds that $\delta_{i,j}(t) \leq 2c'_{i,j}(t)$. Indeed, from $i, j \in \mathcal{N}(t)$ it follows by definition of the algorithm that $\tilde{\delta}_{i,j}(t) \leq c'_{i,j}(t)$. Now, from the definition of $\mathcal{E}_t$, we observe $\delta_{i,j}(t) \leq \tilde{\delta}_{i,j}(t) + c'_{ij}(t)$. Putting together the two inequalities, we get $\delta_{i,j}(t) \leq 2c'_{i,j}(t) \leq 2c_{i,j}(t)$.

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 $\delta_{i'} \geq \delta_i$ and i is in the neighborhood action set of i' and some other Pareto-optimal action. It follows from Bartók et al. (2012b) that i' is well-defined.

Case 1 Consider case 1. Recall that $k(t) = \text{argmax}_{j\in\mathcal{P}(t)\cup\mathcal{V}(t)} \frac{W_j^2}{n_j(t)}$. Thus, $I_t \neq k(t)$. Consequently, $k(t) \notin \mathcal{R}(t)$, i.e. $n_{k(t)}(t) > \eta_{k(t)} f(t)$. 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, we know from Bartók et al. (2012b) that

there exists a path $(i_0, ..., i_r)$ in $\mathcal{N}(t)$ that connects i' to i^* ($i^* \in \mathcal{P}(t)$ holds on $\mathcal{E}_t$). We have that:

$$\begin{aligned}
\delta_i \leq \quad & \delta_{i'} = \sum_{s=1}^{r} \delta_{i_{s-1},i_s} && (45)\\
\leq \quad & 2\sum_{s=1}^{r} c'_{i_{s-1},i_s} && (46)\\
\leq \quad & 2\sum_{s=1}^{r} c_{i_{s-1},i_s} && (47)\\
\leq \quad & 2\sum_{s=1}^{r} \sum_{a\in V_{i_{s-1},i_s}} \|v_{i_{s-1},v_{i_s},a}\|_\infty \sqrt{\frac{\alpha\log(t)}{n_a(t)}} && (48)\\
\leq \quad & 2\sum_{s=1}^{r} \sum_{a\in V_{i_{s-1},i_s}} W_a \sqrt{\frac{\alpha\log(t)}{n_a(t)}} && (49)\\
\leq \quad & 2g_i W_{k(t)} \sqrt{\frac{\alpha\log(t)}{n_{k(t)}(t)}} && (50)\\
\leq \quad & 2g_i W_{k(t)} \sqrt{\frac{\alpha\log(t)}{\eta_{k(t)} f(t)}} && (51)
\end{aligned}$$

Upper bounding $W_{k(t)}/\sqrt{\eta_{k(t)}}$ by $\max_{1\leq k\leq N} W_k/\sqrt{\eta_k}$ we obtain the desired bound.

Case 2: Now, for case 2 take an action k, consider $\mathcal{E}_t \cap \mathcal{D}_t^c$, and assume that $I_t = k$. On the event D_t^c, we have that $I_t = k(t)$. Thus, from $I_t = k$ it follows that $W_k/\sqrt{n_k(t)} \geq W_j/\sqrt{n_j(t)}$ holds true for all $j \in \mathcal{P}(t)$. Let $J_t = \operatorname{argmin}_{j\in\mathcal{P}(t)\cup N^+(t)} \frac{g_j^2}{\delta_j^2}$. Now, similarly to the previous case, there exists a path $(i_0, ..., i_r)$ from the parent action $J_{t'} \in \mathcal{P}(t)$ of J_t to $i^\star \in \mathcal{N}(t)$. Hence,

$$\begin{aligned}
\delta_{J_t} \leq \quad & \delta_{J'_t} = \sum_{s=1}^{r} \delta_{i_{s-1},i_s} && (52)\\
\leq \quad & 2\sum_{s=1}^{r} c'_{i_{s-1},i_s} && (53)\\
\leq \quad & 2\sum_{s=1}^{r} c_{i_{s-1},i_s} && (54)\\
\leq \quad & 2\sum_{s=1}^{r} \sum_{a\in V_{i_{s-1},i_s}} W_a \sqrt{\frac{\alpha\log(t)}{n_a(t)}} && (55)\\
\leq \quad & 2g_{J_t} W_k \sqrt{\frac{\alpha\log(t)}{n_k(t)}} && (56)
\end{aligned}$$

This implies

$$\begin{aligned}
n_k(t-1) \leq \quad & 4W_k^2 \frac{d_{J_t}^2}{\delta_{J_t}^2} \alpha\log(t) && (57)\\
= \quad & 4W_k^2 \min_{j\in\mathcal{P}(t)\cup N^+(t)} \frac{d_j^2}{\delta_j^2} \alpha\log(t) && (58)
\end{aligned}$$

This concludes the proof of the Lemma. □

D. Regret Analysis of `RandCBPside`*

In this Section, we provide an upper bound on the expected regret of `RandCBPside`*. Consider the problem of partial monitoring with linear side information (Bartók et al., 2012a). Let $\delta_i(x) = \max_{1\leq j\leq N} \delta_{i,j}(x)$ be the sub-optimality gap between the expected loss of action i and the optimal action given the context x. Define $\Psi = \max_{1\leq a\leq N} \sigma_a$ as the maximum number of feedback symbols that can be induced by an action in the game.

Similarly to the proof in (Bartók et al., 2012b), consider the events $\mathcal{D}_t$ = *"the decaying exploration rule is in effect at time t"* and $\mathcal{E}_t$ = *"the confidence interval succeeds at time t"* = $\{|\hat{\delta}_{i,j}(x_t) - \delta_{i,j}(x_t)| \leq c_{i,j}(x_t)\}$[2].

D.1. Lemma: the confidence interval succeeds

Lemma D.1. Fix any $t \geq 1$. Take any action i. On the event $\mathcal{E}_t \cap \mathcal{D}_t$, from $i \in \mathcal{P}(t) \cap N^+(t)$ it follows that

$$\delta_i(x_t) \leq \frac{2g_i\Psi\left(\sqrt{d\log(t)} + \Psi\right)}{\sqrt{f(t)}} \max_{1\leq k\leq N} \frac{W_k}{\sqrt{\eta_k}} \tag{59}$$

Proof. We start the proof with the following remarks:

Remark D.2. Observe that for any neighboring action pair $\{i, j\} \in \mathcal{N}(t)$, on $\mathcal{E}_t$, it holds that $\delta_{i,j}(x_t) \leq 2c'_{i,j}(x_t)$. Indeed, from $i, j \in \mathcal{N}(t)$ it follows by definition of the algorithm that $\tilde{\delta}_{i,j}(x_t) \leq c'_{i,j}(x_t)$. Furthermore, we have: $\delta_{i,j}(x_t) \leq \tilde{\delta}_{i,j}(x_t) + c'_{i,j}(x_t)$, by definition of $\mathcal{E}_t$. Putting together the two inequalities, and given that of $c'_{i,j}(x_t) \leq c_{i,j}(x_t)$, we obtain $\delta_{i,j}(x_t) \leq 2c'_{i,j}(x_t) \leq 2c_{i,j}(x_t)$.

Remark D.3. Now, fix some action i that is not *dominated*[3]. We define the *parent action* i' of i as follows: If i is not degenerate then $i' = i$. If i is degenerate then i' is the Pareto-optimal action such that $\delta_{i'}(x_t) \geq \delta_i(x_t)$ and i is in the neighborhood action set of i' and some other Pareto-optimal action. It follows from Lemma 5 in Bartók et al. (2012b) that i' is well-defined.

Define the action $k(t) = \operatorname{argmax}_{j\in\mathcal{P}(t)\cup\mathcal{V}(t)} W_j w_j(t)$. In other words, $k(t)$ represents the action that has the largest confidence width within the set $\mathcal{P}(t) \cup \mathcal{V}(t)$, which corresponds to the exploitation component of the strategy.

Consider $\mathcal{E}_t \cap \mathcal{D}_t$. Due to $\mathcal{D}_t$, the played action I_t is such that $I_t \neq k(t)$. Therefore, $k(t) \notin \mathcal{R}(x_t)$ which implies $\|x_t\|_{G^{-1}_{k(t),t}} \leq \frac{1}{\sqrt{\eta_{k(t)} f(t)}}$ from the definition of $\mathcal{R}(x_t)$ in the contextual setting. 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 $\mathcal{P}(t)$. In any case, there is a path $(i_0, ..., i_r)$ in $\mathcal{N}(t)$ that

[2] The notation is inversed in Bartók et al. (2012b).

[3] see definition 2.1

connects i' to i^*, with $i^* \in \mathcal{P}(t)$ that holds on $\mathcal{E}_t$. We have that:

$$\begin{aligned}
\delta_i(x_t) \leq \quad & \delta_{i'}(x_t) = \sum_{s=1}^{r} \delta_{i_{s-1},i_s}(x_t) && (60)\\
\leq \quad & 2\sum_{s=1}^{r} c'_{i_{s-1},i_s}(x_t) && (61)\\
\leq \quad & 2\sum_{s=1}^{r} c_{i_{s-1},i_s}(x_t) && (62)\\
= \quad & 2\sum_{s=1}^{r} \sum_{a\in V_{i_{s-1},i_s}} \|v_{i_{s-1}i_s a}\|_\infty \sigma_a \left(\sqrt{d\log(t) + 2\log(1/\delta_t)} + \sigma_a\right) \|x_t\|_{G^{-1}_{a,t}} && (63)\\
\leq \quad & 2\sum_{s=1}^{r} \sum_{a\in V_{i_{s-1},i_s}} W_{k(t)}\Psi \left(\sqrt{\Psi\log(t) + 2\log(1/\delta_t)} + \Psi\right) \|x_t\|_{G^{-1}_{k(t),t}} && (64)\\
\leq \quad & 2g_i W_{k(t)}\Psi \left(\sqrt{d\log(t) + 2\log(1/\delta_t)} + \Psi\right) \|x_t\|_{G^{-1}_{k(t),t}} && (65)\\
\leq \quad & 2g_i W_{k(t)}\Psi \left(\sqrt{d\log(t) + 2\log(1/\delta_t)} + \Psi\right) \frac{1}{\sqrt{\eta_{k(t)} f(t)}} && (66)\\
\leq \quad & \frac{2g_i\Psi \left(\sqrt{d\log(t) + 2\log(1/\delta_t)} + \Psi\right)}{\sqrt{f(t)}} \max_{1\leq k\leq N} \frac{W_k}{\sqrt{\eta_k}}, && (67)
\end{aligned}$$

Equation 60 was derived from the definition of a parent action. Equations 61 and 62 follow from remark D.2. In Equation 63, we expand the formula of the confidence bound, defined in Section 4. In Equation 65, we simplify the double sum by using the fact that $\|v_{i_{s-1},i_s,a}\|_\infty$ is upper bounded by $W_{k(t)}$ and that the cardinality of the double sum is g_i. In Equation 66 we use the upper bound on the Gram matrix obtained from the events considered in the Lemma. In Equation 67, we finalize the upper-bound by considering the action in $\{1,\dots,N\}$ that maximizes $\frac{W_k}{\sqrt{\eta_k}}$.

This concludes the proof of the Lemma. □

D.2. Bounding the sum of sub-optimality gaps

The goal of this section is to establish an upper bound for the sum of sub-optimality gaps, specifically under the event denoted as $\mathcal{E}_t$, which signifies the success of the confidence interval.

`CBPside`, as presented by Bartók et al. (2012a), utilizes confidence bounds that are tailored for easy games exclusively. `RandCBPside`* adopts a broader definition of confidence bounds, as originally introduced by Bartók et al. (2012b) and Lienert (2013). This broader definition makes `RandCBPside`* applicable to both easy and hard games.

Lemma D.4. When $\mathcal{E}_t$ holds, the sum of the sub-optimality gaps can be upper-bounded by

$$\sqrt{\sum_{s=1}^{n_k(T)} \delta_k(x_{t_k(s)})^2} \leq 2g_k W_k \Psi d^{1/2} \left(\sqrt{d\log(T) + 2\log(1/\delta_t)} + \Psi\right) \sqrt{2\log(T)},$$

where $n_k(t)$ is the total number of times action k was played up to time t and $t_k(s)$ is the round index where action k was played for the s-th time.

Proof. Recall that $\delta_{I_t}(x_t)$ corresponds to the gap between action k and the optimal action given context x_t. There exist a path of r neighboring actions $I_t = k_0, k_1, ..., k_r = i^\star(x_t)$ between the action played and the optimal action. This sequence always exists thanks to how the algorithm constructs the set of admissible actions [4]. The first step of the proof consists in

[4] for a proof of this statement, refer to Bartók et al. (2012a).

upper-bounding the sub-optimality gap:

$$\delta_k(x_t)^2 \leq \left(\sum_{s=1}^{r} 2c'_{k_{s-1},k_s}(x_t)\right)^2 \tag{68}$$

$$\leq \left(\sum_{s=1}^{r} 2c_{k_{s-1},k_s}(x_t)\right)^2 \tag{69}$$

$$\leq 4\left(\sum_{s=1}^{r}\sum_{a\in V_{k_{s-1},k_s}} \|v_{k_{s-1},k_s,a}\|_\infty \sigma_a \left(\sqrt{d\log(t)+2\log(1/\delta_t)}+\sigma_a\right)\|x_t\|_{G_{a,t}^{-1}}\right)^2 \tag{70}$$

$$\leq 4\left(\sum_{s=1}^{r}\sum_{a\in V_{k_{s-1},k_s}} \|v_{k_{s-1},k_s,a}\|_\infty \sigma_a \left(\sqrt{d\log(t)+2\log(1/\delta_t)}+\sigma_a\right)\|x_t\|_{G_{a,t}^{-1}}\right)^2 \tag{71}$$

$$\leq 4\left(\sum_{s=1}^{r}\sum_{a\in V_{k_{s-1},k_s}} W_k\Psi \left(\sqrt{d\log(t)+2\log(1/\delta_t)}+\Psi\right)\max_{1\leq l\leq N}\|x_t\|_{G_{l,t}^{-1}}\right)^2 \tag{72}$$

$$\leq 4\left(g_k W_k\Psi \left(\sqrt{d\log(t)+2\log(1/\delta_t)}+\Psi\right)\max_{1\leq l\leq N}\|x_t\|_{G_{l,t}^{-1}}\right)^2 \tag{73}$$

$$\leq 4g_k^2W_k^2\Psi^2 \left(\sqrt{d\log(t)+2\log(1/\delta_t)}+\Psi\right)^2\max_{1\leq l\leq N}\|x_t\|^2_{G_{l,t}^{-1}} \tag{74}$$

For the detail between Equation 68 and Equation 72, we refer the reader to the steps described in the proof of Lemma D.1. In Equation 73 we consider the greatest weighted norm over the action space $\{1,\dots,N\}$ to be able to remove it from the double sum.

We now analyse the square root of the sum of the sub-optimality gaps over the time horizon of the action k. We start with the result obtained in Equation 74:

$$\sqrt{\sum_{t=1}^{n_k(T)}\delta_k(x_{t_k(s)})^2} \leq \sqrt{\sum_{s=1}^{n_k(T)}\min(4g_k^2W_k^2\Psi^2\left(\sqrt{\log(t_k(s))+2\log(1/\delta_s)}+\Psi\right)^2\max_{1\leq l\leq N}\|x_{t_k(s)}\|^2_{G^{-1}_{l,t_k(s)}},1)} \tag{75}$$

$$\leq \sqrt{4g_k^2W_k^2\Psi^2\left(\sqrt{d\log(T)+2\log(1/\delta_T)}+\Psi\right)^2\sum_{s=1}^{n_k(T)}\min(\max_{1\leq l\leq N}\|x_t\|^2_{G^{-1}_{l,t_k(s)}},1)} \tag{76}$$

$$\leq 2g_kW_k\Psi\left(\sqrt{d\log(T)+2\log(1/\delta_T)}+\Psi\right)\sqrt{2d\log(1+n_k(T)E^2)} \tag{77}$$

$$\in O(2g_kW_k\Psi d^{1/2}\left(\sqrt{d\log(T)+2\log(1/\delta_T)}+\Psi\right)\sqrt{2\log(T)}) \tag{78}$$

In Equation 77 we have used the upper bound on the sum of weighted norms presented in Lemma 10 of Abbasi-Yadkori et al. (2011), with the assumption $\|x_t\|_2 \leq E$. The difference between Line 78 and Equation 6 in Bartók et al. (2012a) is that a $\sqrt{T}$ term is not appearing. We will see that the $\sqrt{T}$ term appears appears later in the analysis from the Cauchy-Schwartz inequality. □

D.3. Regret analysis of `RandCBPside` using Lemma D.1

In this Section, we analyse the regret rate of `RandCBPside`* on easy and hard games. The initial strategy `CBPside` (Bartók et al., 2012a) has a guarantee restricted to easy games. The key component to obtain the guarantee of `RandCBPside` to hard games is to define underplayed actions in a suitable way for the contextual setting.

Proof. First, we decompose the regret around the event $\mathcal{E}_t$ and its complimentary:

$$\mathbb{E}[R_T] = \sum_{t=1}^{T} \mathbb{E}[\mathbf{L}[I_t, J_t]] - \sum_{t=1}^{T} \mathbb{E}[\mathbf{L}[i^\star(x_t), J_t]] \tag{79}$$

$$= \sum_{t=1}^{T} \mathbb{E}[\delta_{I_t}(x_t)]$$

$$= \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}_{\{\mathcal{E}_t\}} \delta_{I_t}(x_t)] + \mathbb{E}[\mathbb{1}_{\{\mathcal{E}_t^c\}} \delta_{I_t}(x_t)] \tag{80}$$

$$= \underbrace{\sum_{t=1}^{T} \mathbb{E}[\mathbb{1}_{\{\mathcal{E}_t^c\}} \delta_{I_t}(x_t)]}_{A} + \underbrace{\sum_{t=1}^{T} \mathbb{E}[\mathbb{1}_{\{\mathcal{E}_t\}} \delta_{I_t}(x_t)]}_{B} \tag{81}$$

D.4. Term A

In this Section, we will study component A. Assume for each action k at time t, there exist a number such that $p(\mathcal{E}_t^c, I_k = k) \leq \beta_{k,t}$. Therefore, there exist a sequence of numbers $\beta_{k,1}, \beta_{k,2}, \ldots, \beta_{k,T} \in [0, 1]$. These numbers can be seen as some probabilities that $\mathcal{E}^c$ occurs. In the previous analysis (Bartók et al., 2012a) the numbers where action independent. In this work, the numbers are action dependent i.e. we add a dependency on k because the strategy `RandCBPside`* generates a sample $Z_{k,t}$ for each action which influences the value of $\beta_{k,t}$. Define $a(t) = \text{argmax}_{1 \leq k \leq N} \beta_{k,t}$:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[\mathbb{1}_{\{\mathcal{E}_t^c\}} \delta_{I_t}(x_t)] = & \quad \sum_{t=1}^{T} \sum_{k=1}^{N} \mathbb{E}[\mathbb{1}_{\{\mathcal{E}_t^c, k=I_t\}} \delta_k(x_t)] \\
\leq & \quad \sum_{t=1}^{T} \sum_{k=1}^{N} \mathbb{E}[\mathbb{1}_{\{\mathcal{E}_t^c, k=I_t\}}], \text{ because } \delta_k(x_t) \leq 1 \\
= & \quad \sum_{t=1}^{T} \sum_{k=1}^{N} \beta_{k,t} \\
\leq & \quad \sum_{t=1}^{T} N \beta_{a(t),t}
\end{aligned} \tag{82}$$

D.5. Term B:

Consider a specific action k. The regret decomposition is decomposed into multiple components. depending whether $\mathcal{D}_t$ occurs or not. This decomposition was initially presented in Bartók et al. (2012b) in the non-contextual case. Here, we adapt the decomposition to the contextual case, as demonstrated by the presence of contextual sub-optimality gaps $\delta_k(x_t)$.

$$\begin{aligned}\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}_{\{\mathcal{E}_t,I_t=k\}}\delta_k(x_t)] = \quad & \delta_k(x_t) + \\ & \underbrace{\sum_{t=N+1}^{T}\mathbb{E}[\mathbb{1}_{\{\mathcal{E}_t,D_t,k\in P(t)\cup N^+(t),I_t=k\}}]\delta_k(x_t)}_{B_1} + \\ & \underbrace{\sum_{t=N+1}^{T}\mathbb{E}[\mathbb{1}_{\{\mathcal{E}_t,D_t,k\notin P(t)\cup N^+(t),I_t=k\}}]\delta_k(x_t)}_{B_2} + \\ & \underbrace{\sum_{t=N+1}^{T}\mathbb{E}[\mathbb{1}_{\{\mathcal{E}_t,D_t^c,k\in P(t)\cup N^+(t),I_t=k\}}]\delta_k(x_t)}_{B_3} + \\ & \underbrace{\sum_{t=N+1}^{T}\mathbb{E}[\mathbb{1}_{\{\mathcal{E}_t,D_t^c,k\notin P(t)\cup N^+(t),I_t=k\}}]\delta_k(x_t)}_{B_4}\end{aligned} \tag{83}$$

The first term corresponds to the regret suffered at the initialization of the algorithm, where each action is played once. We will now focus on bounding the terms B_1, B_2, B_3, and B_4.

Term B_1: Consider the case $\mathcal{E}_t \cap D_t \cap \{k \in P(t) \cup N^+(t)\}$. From case 1 of Lemma D.1, we have the relation:

$$\delta_k(x_t) \leq \frac{2g_k\Psi\left(\sqrt{d\log(t)+2\log(1/\delta_t)}+\Psi\right)}{\sqrt{f(t)}}\max_{1\leq j\leq N}\frac{W_j}{\sqrt{\eta_j}} \tag{84}$$

$$\begin{aligned}\sum_{t=N+1}^{T}\mathbb{E}[\mathbb{1}_{\{\mathcal{E}_t,D_t,k\in\mathcal{P}(t)\cup N^+(t),I_t=k\}}\delta_k(x_t)] \leq \\ T\frac{2g_k\Psi\left(\sqrt{d\log(T)+2\log(1/\delta_T)}+\Psi\right)}{\sqrt{f(T)}}\max_{1\leq j\leq N}\frac{W_j}{\sqrt{\eta_j}}\end{aligned} \tag{85}$$

Term B_2: Consider the case $\mathcal{E}_t \cap D_t \cap \{k \notin P(t) \cup N^+(t)\}$. It follows that $k \in \mathcal{V}(t) \cap \mathcal{R}(x) \subseteq \mathcal{R}(x)$. Hence, we know by definition of the exploration rule that $1/\|x\|^2_{G^{-1}_{k,t}} < \eta_k f(t)$.

$$\begin{aligned}\sum_{t=N+1}^{T}\mathbb{1}_{\{\mathcal{E}_t,D_t,k\notin\mathcal{P}(t)\cup N^+(t),I_t=k\}} \leq \quad & \sum_{t=N+1}^{T}\mathbb{1}_{\{I_t=k,1/\|x_t\|^2_{G^{-1}_{k,t}}<\eta_k f(t)\}} \\ & + \sum_{t=N+1}^{T}\mathbb{1}_{\{\mathcal{E}_t,D_t,k\notin\mathcal{P}(t)\cup N^+(t),I_t=k,1/\|x_t\|^2_{G^{-1}_{k,t}}\geq\eta_k f(t)\}} \\ \leq \quad & \eta_k f(T)\end{aligned} \tag{86}$$

In Equation 86 there are two antagonist indicators. The second one simplifies to 0 because the inequality is never verified due to D_t. We now apply the Cauchy-Schwartz inequality:

$$\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t, D_t, k\notin\mathcal{P}(t)\cup N^+(t), I_t=k\}}\delta_k(x_t) \leq \sqrt{\left(\sum_{t=N+1}^{T} \mathbb{1}^2_{\{\mathcal{E}_t, D_t, k\notin\mathcal{P}(t)\cup N^+(t), I_t=k\}}\right)\left(\sum_{t=N+1}^{T}\delta_k(x_t)^2\right)} \tag{87}$$

$$\leq \left(\sqrt{\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t, D_t, k\notin\mathcal{P}(t)\cup N^+(t), I_t=k\}}}\right)\left(\sqrt{\sum_{t=N+1}^{T}\delta_k(x_t)^2}\right) \tag{88}$$

$$\leq \sqrt{\eta_k f(T)}\left(\sqrt{\sum_{t=N+1}^{T}\delta_k(x_t)^2}\right) \tag{89}$$

In Equation 87 we use the relation $\left(\sum_i a_i b_i\right)^2 \leq \left(\sum_i a_i^2\right)\left(\sum_i b_i^2\right)$. In Equation 88 we notice that the square of an indicator is equal to the indicator. We also use the relation $\sqrt{ab} \leq \sqrt{a}\sqrt{b}$.

Term B_3: Consider the event $\mathcal{E}_t \cap \mathcal{D}_t^c \cap \{k \in \mathcal{P}(t) \cup N^+(t)\}$. We will use the Cauchy-Schwartz inequality to simplify the expression.

$$\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t, \mathcal{D}_t^c, k\in\mathcal{P}(t)\cup N^+(t), I_t=k\}}\delta_k(x_t) \leq \sqrt{\left(\sum_{t=N+1}^{T} \mathbb{1}^2_{\{\mathcal{E}_t, \mathcal{D}_t^c, k\in\mathcal{P}(t)\cup N^+(t), I_t=k\}}\right)\left(\sum_{t=N+1}^{T}\delta_k(x_t)^2\right)} \tag{90}$$

$$\leq \left(\sqrt{\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t, \mathcal{D}_t^c, k\in\mathcal{P}(t)\cup N^+(t), I_t=k\}}}\right)\left(\sqrt{\sum_{t=N+1}^{T}\delta_k(x_t)^2}\right) \tag{91}$$

$$\leq \sqrt{T}2g_k W_k \Psi d^{1/2}\left(\sqrt{d\log(T) + 2\log(1/\delta_T)} + \Psi\right)\sqrt{2\log(T)} \tag{92}$$

$$\leq \sqrt{T}2g_k W_k \Psi d^{1/2}\left(\sqrt{d\log(T) + 2\log(1/\delta_T)} + \Psi\right)\sqrt{2\log(T)} \tag{93}$$

Term B_4: Consider the event $\{\mathcal{E}_t, \mathcal{D}_t^c, k \notin \mathcal{P}(t) \cup \mathcal{N}^+(t)\}$. Since k is not in $\mathcal{P}(t) \cup N^+(t)$, we also have that $\|x\|^2_{G^{-1}_{k,t}} \leq \frac{1}{\eta_k f(t)} \iff \frac{1}{\|x\|^2_{G^{-1}_{k,t}}} \geq \eta_k f(t)$. We get:

$$\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t, \mathcal{D}_t^c, k\notin\mathcal{P}(t)\cup N^+(t), I_t=k\}}] \leq \min\Big\{T, \eta_k f(T)\Big\} \tag{94}$$

We now use Cauchy-Schwartz,

$$\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t, \mathcal{D}_t^c, k\notin\mathcal{P}(t)\cup N^+(t), I_t=k\}}\delta_k(x_t) \leq \sqrt{\left(\sum_{t=N+1}^{T} \mathbb{1}^2_{\{\mathcal{E}_t, \mathcal{D}_t^c, k\notin\mathcal{P}(t)\cup N^+(t), I_t=k\}}\right)\left(\sum_{t=N+1}^{T}\delta_k(x_t)^2\right)} \tag{95}$$

$$\leq \left(\sqrt{\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t, \mathcal{D}_t^c, k\notin\mathcal{P}(t)\cup N^+(t), I_t=k\}}}\right)\left(\sqrt{\sum_{t=N+1}^{T}\delta_k(x_t)^2}\right) \tag{96}$$

$$\leq \sqrt{\min\Big\{T, \eta_k f(T)\Big\}}\left(\sqrt{\sum_{t=N+1}^{T}\delta_k(x_t)^2}\right) \tag{97}$$

□

D.6. Conclusion:

The following theorem is an individual upper bound on the regret of `RandCBPside`.

Theorem D.5. Consider the interval $[A, B]$, with $B = \sqrt{d\log(t) + 2\log(1/t^2)}$ and $A \leq 0$. Set the randomization over K bins with a probability ϵ on the tail and a standard deviation σ. Let $f(t) = \alpha^{1/3}t^{2/3}\log(t)^{1/3}$, $\eta_a = W_a^{2/3}$ and $\alpha > 1$. Assume $\|x_t\|_2 \leq E$ and positive constants C_1, C_2, C_3, and C_4. Note $W = \max_{1\leq k\leq N} W_k$.

$$\begin{aligned}\mathbb{E}[R(T)] \leq \quad & \sum_{t=1}^{T} N\beta_{a(t),t} + N + \\ & \sum_{1\leq k\leq N}\left(\sqrt{\sum_{s=1}^{n_k(T)}\delta_k(x_{t_k(s)})^2}\right)\sqrt{T} + \\ & \sum_{k\in\mathcal{V}\setminus N^+}\left(\sqrt{\sum_{s=1}^{n_k(T)}\delta_k(x_{t_k(s)})^2}\right)\left(\sqrt{\eta_k f(T)} + \sqrt{\min\left\{T, \eta_k f(T)\right\}}\right) + \\ & \sum_{k\in\mathcal{V}\setminus N^+} T\frac{2g_k\Psi\left(\sqrt{d\log(T) + 2\log(1/\delta_T)} + \Psi\right)}{\sqrt{f(T)}}W^{2/3},\end{aligned} \tag{98}$$

where $\mathcal{V} = \bigcup_{i,j\in\mathcal{N}} V_{i,j}$, and $N^+ = \bigcup_{i,j\in\mathcal{N}} N^+_{i,j}$.

Result on easy games: On easy games, the set $k \in \mathcal{V}\setminus N^+$ is empty which simplifies the expression in Equation 98. The regret rate can be expressed as:

$$\mathbb{E}[R(T)] \leq \quad \sum_{t=1}^{T} N\beta_{a(t),t} + N + N\sqrt{T}2g_kW_k\Psi d^{1/2}\left(\sqrt{d\log(T) + 2\log(1/\delta_T)} + \Psi\right)\sqrt{2\log(T)} \tag{99}$$

Corollary D.6. Consider an easy game and $\delta_t = 1/t^2$, and the same assumptions as in theorem D.5, there exist constants C_1 and C_2 such that the expected regret of `RandCBPside` on this game can be upper bounded independently of the choice of $p^\star$ as:

$$\mathbb{E}[R_T] \leq C_1N + C_2Nd\sqrt{T}\log(T)$$

The guarantee of `CBPside` on easy games proposed in Bartók et al. (2012a) is $C_1N + C_2N^{3/2}d^2\sqrt{T}\log T$. Here, the dependency drops from d^2 to d simply because we corrected the confidence bound formula, but this result should also apply to `CBPside`.

Result on hard games: On hard games, the set $\mathcal{V}\setminus N^+$ is not empty.

We need to study the terms of the regret expression to identify which one dominates. The regret expression is:

$$\begin{aligned}\mathbb{E}[R(T)] \leq & \sum_{t=1}^{T} N\beta_{a(t),t} + N + \\ & \sum_{k\in\underline{\mathbf{N}}}\sqrt{T}2g_kW_kd^{3/2}\left(\sqrt{d\log(T) + 2\log(1/\delta_T)} + \Psi\right)\sqrt{2\log(T)} + \\ & \sum_{k\in\mathcal{V}\setminus N^+}\left(\sqrt{\eta_k f(T)} + \min\left\{\sqrt{T}, \sqrt{\eta_k f(T)}\right\}\right)2g_kW_k\Psi d^{1/2}\left(\sqrt{d\log(T) + 2\log(1/\delta_T)} + \Psi\right)\sqrt{2\log(T)} + \\ & \sum_{k\in\mathcal{V}\setminus N^+} T\frac{2g_k\Psi\left(\sqrt{\Psi\log(T) + 2\log(1/\delta_T)} + \Psi\right)}{\sqrt{f(T)}}W^{2/3}\end{aligned} \tag{100}$$

We will now study the last term in the regret expression. If we choose $\delta_t = 1/t^2$, we can set $f(t) = t^{2/3}\log(t)^{1/3}$ and $\eta_k = W_k^{2/3}$, we have $\sqrt{\eta_k f(T)} + \min\left\{\sqrt{T}, \sqrt{\eta_k f(T)}\right\} \in O(\sqrt{\eta_k f(T)})$

$$\sum_{k\in\mathcal{V}\setminus N^+} T\frac{2g_k\Psi\left(\sqrt{d\log(T)+2\log(1/\delta_T)}+\Psi\right)}{\sqrt{f(T)}}W^{2/3} = \sum_{k\in\mathcal{V}\setminus N^+} T\frac{2g_k\Psi\left(\sqrt{(d+4)\log(T)}+\Psi\right)}{\sqrt{f(T)}}W^{2/3} \tag{101}$$

$$\in O(2g_k\Psi d^{1/2}T\frac{\sqrt{\log(T)}}{\sqrt{f(T)}}W^{2/3}) \tag{102}$$

$$\in O(2g_k\Psi d^{1/2}T^{2/3}\frac{\sqrt{\log(T)}}{\sqrt{\log(T)^{1/3}}}W^{2/3}) \tag{103}$$

$$\in O(2g_k\Psi d^{1/2}T^{2/3}\log(T)^{1/2-1/6}W^{2/3}) \tag{104}$$

$$\in O(2g_k\Psi d^{1/2}T^{2/3}\log(T)^{1/3}W^{2/3}) \tag{105}$$

We will now study the penultimate term in the regret expression. If we choose $\delta_t = 1/t^2$, we can set $f(t) = t^{2/3}\log(t)^{1/3}$ and $\eta_k = W_k^{2/3}$, we have:

$$\sqrt{f(T)\eta_k}\times(\dots) = W_k^{1/3}T^{1/3}\log(T)^{1/6}2g_kW_k\Psi d^{1/2}\left(\sqrt{d\log(T)+2\log(1/\delta_T)}+\Psi\right)\sqrt{2\log(T)} \tag{106}$$

$$= 2g_kW_k^{3/2}\Psi d^{1/2}T^{1/3}\log(T)^{1/6}\left(\sqrt{d\log(T)+2\log(1/\delta_T)}+\Psi\right) \tag{107}$$

$$\in O(2g_kW_k^{3/2}\Psi dT^{1/3}\log(T)^{2/3}) \tag{108}$$

The conclusion is that the last term dominates the penultimate term over time. Therefore, we can conclude:

Corollary D.7. Consider a hard game and $\delta_t = 1/t^2$, and the same assumptions as in theorem D.5. Then, there exist constants C_3 and C_4 such that the expected regret of `RandCBPside` on this game can be upper bounded independently of the choice of $p^\star$ as:

$$\mathbb{E}[R_T] \leq C_3N + C_4\sqrt{d}\log(T)^{1/3}T^{2/3}$$

E. Additional Experiments

E.1. Implementation details and hyper-parameters

Contextual and non-contextual experiments are run on machines with 48 CPUs which justifies why we consider 96 runs rather than 100 ($48\times 2 = 96$ is the optimal allocation).

Non-contextual baselines The stochastic strategies `BPM-Least`, `TSPM` and`TSPM-Gaussian` are initialized with priors $p^\star = [1/M,\dots,1/M]$ as this is the common choice reported in their respective original papers (Vanchinathan et al., 2014; Tsuchiya et al., 2020). The number of samples for `BPM-Least`, `TSPM` and `TSPM-Gaussian` is set to 100. We found that higher values increase drastically the computational complexity of the approaches. The strategies `TSPM` and `TSPM-Gaussian` are set with $\lambda = 0.01$ as reported to be the most competitive value in the original paper (Tsuchiya et al., 2020). The deterministic strategy `PM-DMED` is initialized with $c = 1$ following the value presented in the original paper (Komiyama et al., 2015).

To compare `CBP` and `RandCBP` fairly, both strategies are set with $\alpha = 1.01$. Sampling in `RandCBP` is performed according to the procedure described in Section 3.2 over $K = 5$ bins, with probability $\varepsilon = 10^{-7}$ on the tail and standard deviation $\sigma = 1$. Although this choice is not necessarily the most optimal (see Figures 4 and 5), we find it is the most robust across the different settings considered.

Contextual baselines We run `PG-TS` and `PG-IDS` over a horizon $T = 7.5$k because both strategies scale in cubic time with the number of verifications. For a horizon 20k, on a time budget of 5 hours and a 48-cores machine, less than

Game	Apple Tasting (AT)							
Case	imbalanced				balanced			
Metric	mean	std	pvalue	win count	mean	std	pvalue	win count
RandCBP	4.689	4.07	1.0	82	41.417	78.311	1.0	37
CBP	8.672	8.532	0.0	47	72.748	138.279	0.055	48
PM-DMED	13.915	13.155	0.0	2	113.5	138.047	0.0	3
TSPM	9.359	9.007	0.0	25	43.117	45.53	0.854	13
TSPM-Gaussian	19.417	15.925	0.0	7	67.658	56.203	0.008	3
BPM-Least	15.125	8.063	0.0	0	165.04	111.969	0.0	3

Table 2: Supplement for the non-contextual experiment presented in the main paper (see Figure 1). Imbalanced instances: $p \sim \mathcal{U}_{[0,0.2]\cup[0.8,1]}$. Balanced instances: $p \sim \mathcal{U}_{[0.4,0.6]}$. Mean: average regret at the last step ($T = 20$k). Std: standard deviation at the last step. P-value: Welch's t-test on the distribution of regrets at the last step, with RandCBP as reference (p-value > 0.05 means no statistical difference). Win count: number of times a given strategy achieved the lowest final regret (ties included). Color ● indicates the best; ● indicates second best; ● is the third best.

Game	Label Efficient (LE)							
Case	imbalanced				balanced			
Metric	mean	std	pvalue	win count	mean	std	pvalue	win count
RandCBP	11.887	15.004	1.0	81.0	321.023	353.111	1.0	60.0
CBP	16.47	8.173	0.009	15.0	726.877	643.233	0.0	18.0
PM-DMED	1489.217	2887.675	0.0	0.0	1253.432	1048.542	0.0	18.0

Table 3: Supplement for the non-contextual experiment presented in the main paper (see Figure 1). Imbalanced instances: $p \sim \mathcal{U}_{[0,0.2]\cup[0.8,1]}$. Balanced instances: $p \sim \mathcal{U}_{[0.4,0.6]}$. Mean: average regret at the last step ($T = 20$k). Std: standard deviation at the last step. P-value: Welch's t-test on the distribution of regrets at the last step, with RandCBP as reference (p-value > 0.05 means no statistical difference). Win count: number of times a given strategy achieved the lowest final regret (ties included). Color ● indicates the best; ● indicates second best; ● is the third best.

10 realizations succeed out of the 96 considered. Note that PG-TS and PG-IDS assume a logistic setting while in our experiments we consider a linear setting. The logistic regression still performs well because we consider binary outcome games. For both strategies, we consider 10 Gibbs samples: higher values increase the computational complexity of the approaches. STAP and CESA are hyper-parameter free.

We compare CBPside* to its counterpart RandCBPside* fairly by setting both with $\alpha = 1.01$. Sampling in RandCBPside* is performed according to the randomization procedure described in Section 3.2 with $K = 5$ bins, a probability $\varepsilon = 10^{-7}$ on the tail, and standard deviation $\sigma = 1$. Although this choice is not always the most optimal (see Figures 4 and 5), we find it is the most robust across the various settings considered.

All contextual approaches use a regularization $\lambda = 0.05$.

E.2. Detailed results

Table 2 and 3 provide numeric details to support the non-contextual experiments in the main paper. Table 4 provides numeric details for the contextual experiment presented in the main paper.

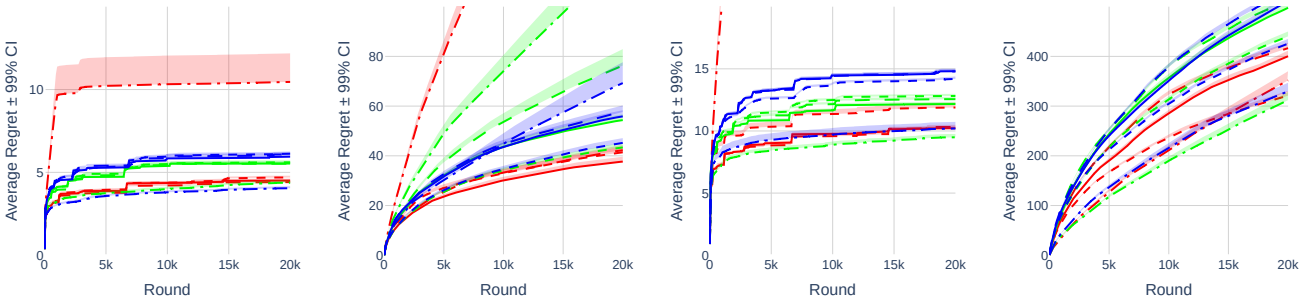
E.3. Sensitivity to hyper-parameters

The goal of this experiment is to illustrate the sensitivity to hyper-parameters of RandCBP and RandCBPside*. We conducted the evaluation for $\varepsilon = 10^{-7}$. Higher values of ϵ imply a higher probability of sampling on the value B in the discretized interval $[A, B]$. We consider the standard deviation values $\sigma \in \{1/2, 1, 2, 10\}$. We consider bin values of K in $\{5, 10, 20\}$. We report averaged regret and upper 99% confidence interval, measured over a 20k horizon and 96 random runs.

Game	Apple Tasting (AT)				Label Efficient (LE)			
Metric	mean	std	pvalue	win count	mean	std	pvalue	win count
RandCBPside*	1016.312	82.151	1.0	96	2026.604	70.161	1.0	96.0
CBPside*	6109.521	86.325	0.0	0	11071.333	86.779	0.0	0
PG-IDS	129.5	12.758	0.0	0				
PG-TS	179.156	15.318	0.0	0				
STAP	1565.917	127.488	0.0	0				
CESA					5792.052	1386.179	0.0	0

Table 4: Numeric detail of the contextual experiment presented in the main paper (see Figure 2). Mean: average regret at the last step ($T = 20$k). Std: standard deviation at the last step. P-value: Welch's t-test on the distribution of regrets at the last step with RandCBPside as reference (p-value> 0.05 means no statistical difference). Win count: number of times a given strategy achieved the lowest final regret (ties included). Color ● indicates the best; ● indicates second best; ● is the third best. Color ● indicates an evaluation on the truncated horizon $T = 7.5$k.

Results (non-contextual case): The experimental setting is described in the main paper. We find the hyper-parameter σ to be the most influential in the performance of RandCBP. Too small values of σ result in a more exploitation, and expose the strategy to a risk of catastrophic failures.

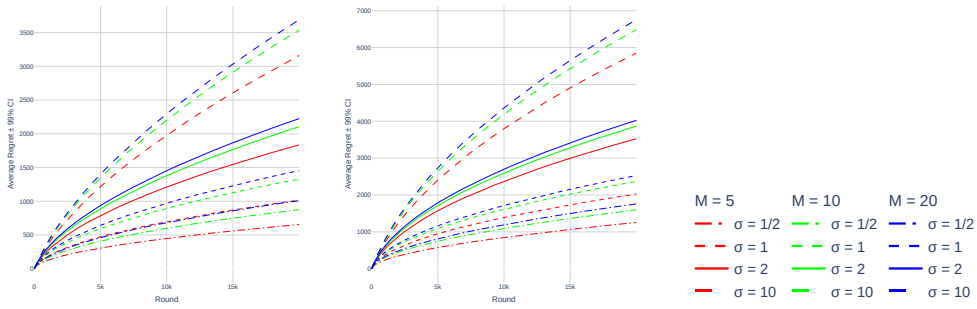


(a) Imbalanced AT (b) Balanced AT (c) Imbalanced LE (d) Balanced LE

(e) Legend

Figure 4: Benchmark of RandCBP on Apple Tasting (AT) and Label Efficient (LE) games, non-contextual case.

Results (contextual case): The experimental setting is described in the main paper. Figure 5 reports multiple hyper-parameter combinations over the Apple Tasting (AT) and Label Efficient (LE) games on linear contexts.



(a) Balanced AT (b) Balanced LE (c) Legend

Figure 5: Benchmark of `RandCBPside`* on Apple Tasting (AT) and Label Efficient (LE) games, contextual case.

E.4. Additional results on the case-study

In this Section, we report additional results for the use-case presented in Section 6. The distribution of true classes in the stream of observations is represented by the vector $B \in [0,1]^N$ and the black-box classifier is represented by its confusion matrix $\mathbf{C} \in [0,1]^{N\times N}$. In each run, B and $\mathbf{C}$ are generated randomly such that the global error rate remains below 10% (the black-box would probably not be deployed otherwise). Each instance of `RandCBP` and `CBP` in the approaches `C-CBP` and `C-RandCBP` is parameterized similarly as in previous experiments.

Maximum number of verification Since the outcomes are binary, the Wald's confidence interval can be used to determine the maximum number of verifications needed to obtain estimates of p_c with a specified level of confidence. We set probability that the confidence interval fails is set to $\zeta = 0.01$ and the acceptable margin length of the Wald interval to $E = \tau/10$. Assuming that the classifier is deployed with a global error rate of at most 10%, a reasonable prior belief per class (noted $\bar{p}_c$) is that the error rate is distributed uniformly across classes $\bar{p}_c = 10/C\%$. For a detection threshold τ, the maximum verification budget for a class c is $n_\tau = \frac{z(1-\zeta/2)^2\bar{p}(1-\bar{p})}{E^2}$, where $z(\cdot)$ is the quantile of the standard normal distribution. In practice, the value $C \times n_\tau$ corresponds to the maximum number of verifications one is willing to use to identify with confidence which of the C predicted classes errors exceed the threshold τ. The goal is to obtain a strategy that performs the task while consuming less verifications than this maximum amount.

Results In the main paper, we reported results for cases: i) the true classes are balanced and the black-box yields uniform mispredictions (case 1), ii) the true classes are imbalanced and the black-box yields non-uniform mispredictions (case 2). Results for the two cases are reported in Tables 5 and 6.

Threshold	Strategy	F1-score (mean)	F1-score (median)	F1-score (std)	Nb. verifs (mean)	Nb. verifs (median)	Nb. verifs (std)
0.025	`Explore-fully`	0.962	1.0	0.15	105120.0	105120.0	0.0
	`C-CBP`	0.962	1.0	0.15	105110.0	105110.0	0.0
	`C-RandCBP`	0.955	1.0	0.163	83818.0	104846.0	32297.0
0.05	`Explore-fully`	0.927	1.0	0.195	26300.0	26300.0	0.0
	`C-CBP`	0.927	1.0	0.195	26290.0	26290.0	0.0
	`C-RandCBP`	0.915	1.0	0.208	15976.0	15091.0	9071.0
0.1	`Explore-fully`	0.907	1.0	0.219	6590.0	6590.0	0.0
	`C-CBP`	0.908	1.0	0.216	6491.0	6580.0	251.0
	`C-RandCBP`	0.91	1.0	0.211	2000.0	1666.0	1094.0
0.2	`Explore-fully`	1.0	1.0	0.0	1670.0	1670.0	0.0
	`C-CBP`	1.0	1.0	0.0	1289.0	1326.0	273.0
	`C-RandCBP`	1.0	1.0	0.0	272.0	255.0	70.0

Table 5: Case 1 - balanced stream and uniform mispredictions

Threshold	Strategy	F1-score (mean)	F1-score (median)	F1-score (std)	Nb. verifs (mean)	Nb. verifs (median)	Nb. verifs (std)
0.025	Explore-fully	0.978	1.0	0.07	103703.0	105120.0	2853.0
	C-CBP	0.978	1.0	0.07	103693.0	105110.0	2853.0
	C-RandCBP	0.976	1.0	0.07	81741.0	81330.0	10921.0
0.05	Explore-fully	0.976	1.0	0.054	26231.0	26300.0	287.0
	C-CBP	0.975	1.0	0.054	26221.0	26290.0	287.0
	C-RandCBP	0.965	1.0	0.063	16901.0	16673.0	2580.0
0.1	Explore-fully	0.953	1.0	0.073	6590.0	6590.0	0.0
	C-CBP	0.953	1.0	0.073	6453.0	6457.0	111.0
	C-RandCBP	0.959	1.0	0.073	2965.0	2961.0	506.0
0.2	Explore-fully	0.927	1.0	0.156	1670.0	1670.0	0.0
	C-CBP	0.927	1.0	0.156	1335.0	1338.0	90.0
	C-RandCBP	0.923	1.0	0.163	447.0	436.0	84.0

Table 6: Case 2 - Imbalanced stream and non-uniform mispredictions