Tracking the Risk of Machine Learning Systems with Partial Monitoring

Maxime Heuillet
Université Laval
Canada
maxime.heuillet.1@ulaval.ca

Audrey Durand
Canada CIFAR AI Chair
Université Laval
Canada

Abstract

Although efficient at performing specific tasks, Machine Learning Systems (MLSs) remain vulnerable to instabilities such as noise or adversarial attacks. In this work, we aim to track the risk exposure of an MLS to these events. We formulate this problem under the stochastic Partial Monitoring (PM) setting. We focus on two instances of partial monitoring, namely the Apple Tasting and Label Efficient games, that are particularly relevant to our problem. Our review of the practicality of existing algorithms motivates RandCBP, a randomized variation of the deterministic algorithm *Confidence Bound* (CBP) inspired by recent theoretical developments in the bandits setting. Our preliminary results indicate that RandCBP enjoys the same regret guarantees as its deterministic counterpart CBP and achieves competitive empirical performance on settings of interest which suggests it could be a suitable candidate for our problem.

1 Introduction

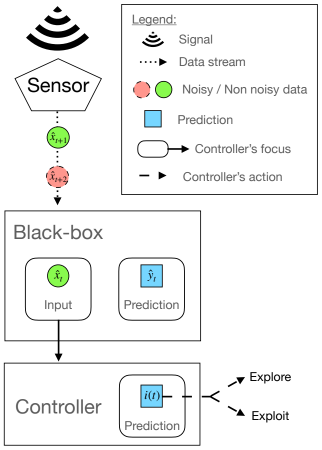


Figure 1: Machine Learning System.

An increasing number of businesses and agencies have started using Machine Learning Systems (MLSs) as they perform well on a wide array of tasks (e.g. decision support in finance Dixon et al. [2020] or healthcare Shailaja et al. [2018]). However, they are vulnerable to instabilities such as noise or adversarial attacks Chakraborty et al. [2021]. In this work, we propose to monitor the risk exposure of a MLS to these instabilities.

We define a MLS as a stream of data points processed sequentially by a *black-box* classifier (see Figure 1). The data points are originate from a *sensor* that can accidentally produce noisy data points that put the black-box at risk of performing a bad prediction. Our goal is to design a *controller* that estimates accurately the proportion of destabilized data points in the stream.

Existing methods Metzen et al. [2017] for the controller emerge from an offline setting (i.e., the controller is trained before its deployment). We propose instead a less studied approach Raginsky et al. [2012] (denoted online setting) where the controller learns progressively

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

over the succession of data points. The difficulty of this task lies in the necessity for the controller to optimally trade-off between exploration and exploitation while only receiving partial feedbacks. Indeed, verifying all the data points is usually impossible due to the amount of data [Henzinger et al., 1998].

We formulate this problem under the partial monitoring (PM) [Bartók et al., 2014] setting as it enables a general expression of the exploration-exploitation trade-off under various feedback scenarios. We focus on two partial monitoring instances called the Label Efficient [Cesa-Bianchi and Lugosi, 2006] and the Apple Tasting [Helmbold et al., 2000] games that are relevant to this problem and that are less studied in the literature. We provide an analysis of the difficulty of each game and review the practicality of existing approaches in these games. This motivates our new algorithm *RandCBP*, a randomized variation of the *Confidence Bound* (CBP) algorithm [Bartok et al., 2012]. We show that *RandCBP* enjoys the same regret guarantees as its deterministic counterpart CBP while demonstrating better empirical performance.

2 Problem setting

Consider a MLS as defined in Figure 1 where a *sensor* produces a sequence of data points that are subject to instabilities assumed to occur stochastically. The idea in this work is to design a controller that samples efficiently the stream to estimate accurately the proportion of perturbed data points.

We formulate this problem as a finite partial monitoring (PM) [Bartók et al., 2014] game played between the controller $\mathcal{C}$ and the sensor. Let K be the number of actions available to the controller and M denote the number of available outcomes that follow a distribution $p^\star$. The game is structured by a loss matrix $L \in \mathbb{R}^{K \times M} \in \{0, 1\}$ and a feedback matrix $H \in \in \Sigma$. Both are known by the controller. Note that Σ is the space of feedbacks is not necessarily numeric. The optimal action of this game is $i^\star = \operatorname{argmin}_i \langle L_i^\top, p^\star \rangle$, where L_i is the i-th column of L. Let $\delta_i = (L_i - L_{i^\star})p^\star$ be the expected loss difference and $N_i(t)$ be the number of times action i was selected before step t. The controller's goal is to explore and exploit in order to minimize the regret: $R(T) = \sum_{t=1}^{T} \delta_{i(t)} = \sum_{i \in N} \delta_i N_i(T+1)$. Therefore, at each step t, the controller aims to incur the smallest loss $L[i(t), j(t)]$ while only observing the feedback $H[i(t), j(t)]$. Notice that a PM game with $L = H$ implies that the controllers observes the loss at each step, which is a bandit game.

Partial monitoring games can have different levels of difficulty. To this end, Bartok [2012] proposed a game classification with 4 categories: *trivial* with zero regret, *easy* with $\tilde{\Theta}(\sqrt{T})$ regret, *hard* with $\Theta(T^{2/3})$ regret and hopeless with linear regret. The difference between *easy* and *hard* games depends on the local observability condition that can be verified for any game (see Appendix A).

3 Tracking the Risk of Machine Learning Systems

In the rest of this work, we focus on two specific games known in the literature as Label Efficient and Apple Tasting games Cesa-Bianchi and Lugosi [2006]. Both are relevant to our application because they admit two outcomes that could be identified as whether the data point is destabilized or not.

3.1 Relevant Partial Monitoring Games

Label Efficient game The Label Efficient game admits 3 actions (flag as anomalous, don't flag, ask an expert) and is structured as follows:

$$L = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, H = \begin{bmatrix} \perp & \odot \\ \wedge & \wedge \\ \wedge & \wedge \end{bmatrix}$$

This game is difficult because the controller receives an informative feedback only when they ask an expert whether a specific data point is destabilized or not. Yet, asking the expert causes a large loss. Hence, the controller must minimize the number of queries to the expert. This dynamic is relevant for high volume streams where not all the data points can not be verified. The Label Efficient game is classified as *hard* and admits a regret lower bound in $\Theta(T^{2/3})$. However, it is a particularly 'harder' game within the class of hard games because the only informative action is also the least optimal one loss wise (see Appendix A.2).

Apple Tasting game The Apple Tasting game differs from the Label Efficient one as it only admits two actions (flag as anomalous, don't flag). The loss and feedback matrices are:

$$L = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, H = \begin{bmatrix} \perp & \perp \\ \perp & \odot \end{bmatrix}$$

In this game, every flagged data point receives a feedback. This ultimately requires more expert resources than the Label Efficient game and is typically relevant for streams with less data volume. This game belongs to the class of *easy* games and thus admits a lower bound in $\Theta(\sqrt{T})$ (see Appendix A.1).

3.2 Practicality of Existing Approaches

Existing stochastic partial monitoring approaches are either deterministic or bayesian. On the bayesian side, BPM-Least [Vanchinathan et al., 2014] admits a regret guarantee in $O(\sqrt{\log(T)})$ for easy games and TSPM Tsuchiya et al. [2020] achieves a $O(\sqrt{T})$ regret rate on easy games only as well. Therefore, both BPM-Least and TSPM fail in hard games such as the Label Efficient one. On the deterministic side, CBP [Bartok et al., 2012] achieves a $O(\sqrt{T})$ regret rate on easy games and $O(T^{2/3})$ rate on hard games. However, CBP achieves a $O(\sqrt{T})$ regret rate for some hard games under a favorable outcome distribution[1]. PM-DMED [Komiyama et al., 2015] achieves an asymptotic $O(\log T)$ regret rate yet, the empirical performance of PM-DMED relies on the tuning of a distribution dependent parameter c.

4 Method: a Randomized Confidence Bound Algorithm

In this section, we present RandCBP, a randomized variation of the algorithm CBP. RandCBP is inspired from recent theoretical developments in confidence based approaches in the bandit setting Vaswani et al. [2020]. The aim is to bridge these developments to the partial monitoring setting. Implementation details are available in Appendix B.

CBP In short, the algorithm works as follows. The algorithm estimates for all action pairs in $\mathcal{N}$ the loss differences $\tilde{\delta}_{i,j}$ and maintains a deterministic confidence bound $c_{i,j}$ over these estimates . When there is enough confidence, the sign of the loss difference estimation indicates which action in the pair is better allowing the elimination of sub-optimal actions. This procedure is repeated at each time t, and yields a set of promising actions. However, in order to guarantee sufficient exploration the algorithm must also select a few times sub-optimal actions for information seeking purpose. Therefore, a set of rarely sampled actions $\mathcal{R}(t)$ is computed at each round and is added to the set of promising actions. Finally, the selected action is the one that potentially reduces the confidence bound the most within this larger set.

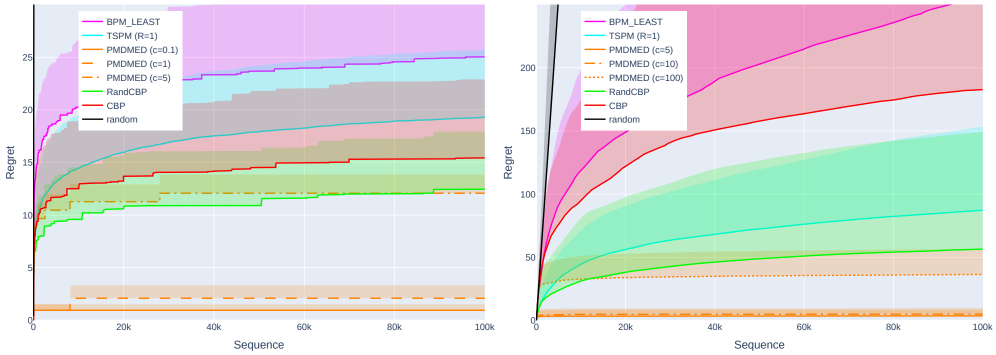
RandCBP We derive a randomized version of the deterministic algorithm CBP. We propose to use stochastic confidence bounds $z_{i,j}$ instead of deterministic ones. More precisely, $z_{i,j}$ are proportional to a value Z_t sampled at each time $t \leq T$ from from a Gaussian $\mathcal{N}(0, \sigma)$ truncated over $[0, \alpha log(t)]$. This approach brings additional variability to trade off exploration and exploitation.

Regret guarantee of RandCBP In Appendix B.3, we show that RandCBP enjoys the same regret guarantees as CBP. The change in the definition of the confidence intervals impacts the two lemmas at the core of CBP's regret proof [Bartok, 2012]. Our approach differs from the path used in Vaswani et al. [2020] as we observe that $z_{i,j}(t) \leq c_{i,j}(t)$ is verified anytime whereas they reduce their proof to a known results from the linear bandits literature.

5 Experiments

We track the risk of a MLS on synthetic streams. To measure the performance, we average the expected regret over 100 sequences of $100k$ elements. A lower regret implies a better estimation of the MLS risk. We consider *simple* instances with $p^\star = [p_{anomaly}, 1 - p_{anomaly}]$ is such that

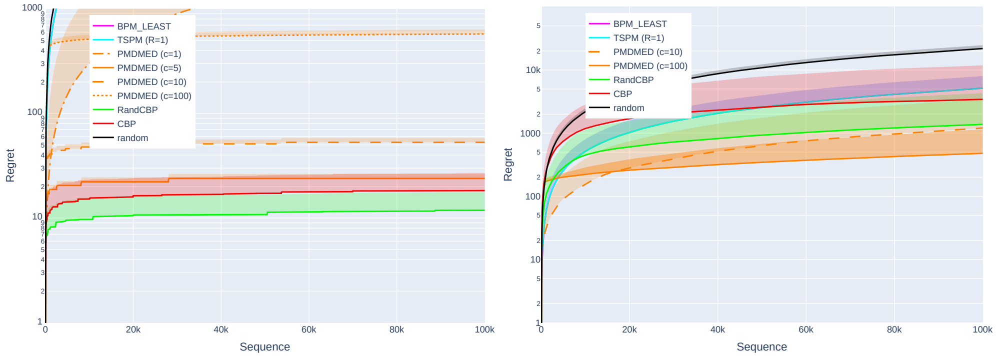
[1] This is occurs when $p^\star$ verifies the *hilly property* defined in Bartok [2012]. In the Label Efficient and Apple Tasting games, the *hilly property* is verified when $p_{spam} \leq 0.4$.



(a) Simple instances

(b) Difficult instances

Figure 2: Apple Tasting game



(a) Simple instances

(b) Difficult instances

Figure 3: Label Efficient game

$p_{anomaly} \in \mathcal{U}[0, 0.2]$ which is coherent with our scenario. We also consider *difficult* instances where $p_{anomaly} \in \mathcal{U}([0.4, 0.5])$. See footnote 1 for a justification of the probability thresholds. A study of RandCBP hyper-parameters is available in C.2.

Baselines We rely on 5 baselines. We consider a *random* selection policy, the bayesian approaches BPM-Least and TSPM as well as the deterministic algorithms PM-DMED and CBP. These four approaches represent all existing approaches in stochastic PM.

Results on the Apple Tasting game From Figure 2, we observe that RandCBP achieves a lower regret than CBP on both simple and difficult instances on the Apple Tasting game. Over 100 trials, RandCBP achieved or lower or equal regret than CBP 74 (resp. 93) times on simple (resp. difficult) instances.A well tuned PM-DMED is the best strategy on difficult and easy apple tasting games. However, on simple games, the performance gap between PM-DMED and RandCBP is narrower and the greater time and implementation complexity of PM-DMED over RandCBP should be taken into consideration.

Results on the Label Efficient game As expected, bayesian approaches such as TSPM and BPM-Least fail in the Label Efficient game. Again, the RandCBP's regret is lower CBP's one on both simple and difficult instances of the Label Efficient game. Over 100 trials, RandCBP achieved or lower or equal regret than CBP 90 (resp. 72) times on simple (resp. difficult) instances. In the simple case, the best realizations of PM-DMED (with $c = 5$) under-perform RandCBP but are narrow. In the difficult case, PM-DMED (with $c = 100$) is the most efficient strategy. However, as can be seen in C.3, this is due to PM-DMED randomly selecting actions over the 1000 first data points which corresponds to an *explore-the-commit* behavior that could not be appropriate in some applications whereas RandCBP explores over a longer time span.

6 Conclusion

Our preliminary results show that RandCBP enjoys the same regret guarantees as CBP and has better empirical performance. RandCBP is also the best approach in simple Label Efficient games which are particularly relevant for our application. Our next steps are to extend the approach to the contextual setting where the controller has access to the data points contained in the stream to inform its actions which makes sense in practice. We also plan to investigate the adversarial setting to generalize to the presence of state-of-the-art adversaries Mladenovic et al. [2022].

References

G. Bartok. The role of information in online learning (ph.d. thesis, 2012.

G. Bartok, N. Zolghadr, and C. Szepesvari. An adaptive algorithm for finite stochastic partial monitoring, 2012. URL `https://arxiv.org/abs/1206.6487`.

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

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

A. Chakraborty et al. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45, 2021.

M. F. Dixon, I. Halperin, and P. Bilokon. *Machine learning in Finance*, volume 1406. Springer, 2020.

J. A. Grant and D. S. Leslie. Apple tasting revisited: Bayesian approaches to partially monitored online binary classification, 2021. URL `https://arxiv.org/abs/2109.14412`.

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

M. R. Henzinger et al. Computing on data streams. *External memory algorithms*, 50:107–118, 1998.

J. Komiyama, J. Honda, and H. Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. *Advances in Neural Information Processing Systems*, 28, 2015.

T. Lattimore and C. Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. *CoRR*, abs/1902.00470, 2019. URL `http://arxiv.org/abs/1902.00470`.

J. H. Metzen et al. On detecting adversarial perturbations. *In Proc. ICLR*, 2017.

A. Mladenovic et al. Online adversarial attacks. *In Proc. ICLR*, 2022.

M. Raginsky, R. M. Willett, C. Horn, J. Silva, and R. F. Marcia. Sequential anomaly detection in the presence of noise and limited feedback. *IEEE Transactions on Information Theory*, 58(8): 5544–5562, 2012.

K. Shailaja, B. Seetharamulu, and M. Jabbar. Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, pages 910–914. IEEE, 2018.

T. Tsuchiya, J. Honda, and M. Sugiyama. Analysis and design of thompson sampling for stochastic partial monitoring. *Advances in Neural Information Processing Systems*, 33:8861–8871, 2020.

H. P. Vanchinathan, G. Bartók, and A. Krause. Efficient partial monitoring with prior information. *Advances in Neural Information Processing Systems*, 27, 2014.

S. Vaswani, A. Mehrabian, A. Durand, and B. Kveton. Old dog learns new tricks: Randomized ucb for bandit problems. *In Proc. AISTATS*, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [No] **The code and the data are proprietary.**

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] **We don't see any negative societal impact in a technology that aims to identify possible instabilities in a stream of data points.**
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [No]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] **Our work does not use existing assets.**
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] **Our work does not use crowdsourcing or human subjects.**

A Analysis of Partial Monitoring games

In this Section, we provide a detailed analysis of the Apple Tasting and Label Efficient games, using the theoretical framework developed in Bartok [2012]. These in depth analysis are valuable for the implementation of partial monitoring algorithms and understanding the difficulty of these games. We point the reader to the appendix of Grant and Leslie [2021] and Lattimore and Szepesvári [2019] for higher level analyses.

A.1 Analysis of the Apple Tasting game

Let us consider the Apple Tasting problem defined by the following loss (L) and feedback (H) matrices:

$$L = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, H = \begin{bmatrix} \perp & \perp \\ \perp & \odot \end{bmatrix}$$

This problem includes a set of 2 possible actions and 2 possible outcomes (denoted A and B).

Signal Matrices: The signal matrices as defined in Bartok [2012] associated to this problem are such that $S_1 \in \{0,1\}^{1\times 2}$ and $S_2 \in \{0,1\}^{2\times 2}$. More precisely, they verify:

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

To illustrate the properties of the signal matrices, let us consider an instance of the problem where the outcome distribution $p^\star$ is such that $p^\star(A) = 0.8$ and $p^\star(B) = 0.2$. Let us define $\langle \cdot, \cdot \rangle$ as the scalar product between matrices. We express the probability of obtaining each symbol of H given a specific action:

- $\langle S_1, p \rangle = 1$, indeed there is only one observation induced by action one therefore the probability of seeing this observation can only be one.
- $\langle S_2, p \rangle = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix}$, therefore the probability of seeing outcome B is 0.8 and the probability of seeing outcome A is 0.2.

Cells: This game has 2 actions, we can characterize these actions by a sub-space of the probability simplex noted *cell* in Bartok [2012]:

- For action 1, we have: $C_1 = \{p \in \Delta_M \| \forall j \in N, (\ell_1 - \ell_j)^\top p \leq 0\}$. This probability space corresponds to the following constraints:
$$C_1 = \begin{bmatrix} \ell_1 - \ell_1 \\ \ell_1 - \ell_2 \end{bmatrix}^\top p = \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}^\top p \leq 0$$
The first constraint ($(\ell_1 - \ell_1)^\top p \leq 0$) is always true. The second constraint ($(\ell_1 - \ell_2)^\top p \leq 0$) implies $p_1 - p_2 \leq 0$.
- For action 2, we have: $C_2 = \{p \in \Delta_M \| \forall j \in N(\ell_2 - \ell_j)^\top p \leq 0\}$. This probability space corresponds to the following constraints:
$$C_2 = \begin{bmatrix} \ell_2 - \ell_1 \\ \ell_2 - \ell_2 \end{bmatrix}^\top p = \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix}^\top p \leq 0$$
The second constraint ($(\ell_2 - \ell_2)^\top p \leq 0$) is always true. The first constraint ($(\ell_2 - \ell_1)^\top p \leq 0$) implies $p_2 - p_1 \leq 0$.

Therefore, for action 1 to be optimal the outcome distribution needs to verify $p_1 - p_2 \leq 0$ whereas it is the opposite for action 2 to be optimal.

Pareto optimal actions: From the analysis of the cells, we have that both actions are Pareto optimal according to the definition in Bartok [2012] because their respective cells are neither empty, neither included one in another.

Neighboring actions: In this paragraph, we will determine whether action 1 and 2 constitute a neighboring pair, i.e if $dim(\mathcal{C}_1 \cap \mathcal{C}_2) = M - 2 = 0$.

$$\mathcal{C}_1 \cap \mathcal{C}_2 = \begin{cases} p_1 - p_2 \leq 0 \\ p_2 - p_1 \leq 0 \end{cases}$$

The only point in this vector space is $[0.5 \quad 0.5]$. Therefore, $dim(C_1 \cap C_2) = 0$. Therefore, action 1 and action 2 are neighboring action pairs according to the definition in Bartok [2012].

Observability of the game: In this paragraph we will determine whether this game is globally and/or locally observable according to the definition in Bartok [2012]. Let's calculate $Im(S_1)$ and $Im(S_2)$:

- For $Im(S_1)$ we have: $\begin{bmatrix}1\\1\end{bmatrix}[x] = \begin{bmatrix}x\\x\end{bmatrix} = x\begin{bmatrix}1\\0\end{bmatrix} + x\begin{bmatrix}0\\1\end{bmatrix}$
- For $Im(S_2)$ we have: $\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}x\\y\end{bmatrix} = x\begin{bmatrix}1\\0\end{bmatrix} + y\begin{bmatrix}0\\1\end{bmatrix}$

We have $Im(S_1) = Im(S_2) = Span(\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix})$. In conclusion, we have: $Im(S_1) \bigoplus Im(S_2) = Span(\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix})$

Finally, the action pair $\{1, 2\}$ is locally observable because $\ell_1 - \ell_2 = \begin{bmatrix}1\\-1\end{bmatrix}$ can be expressed as the sum of the canonic vectors in $Im(S_1) \bigoplus Im(S_2)$. Since this also applies to the action pair $\{2, 1\}$, we can conclude that the game is globally and locally observable. Therefore, it can be classified as an *easy game* with a bound on the regret in $\tilde{\Theta}(\sqrt{T})$.

A.2 Analysis of the Label Efficient game

Let us consider the Label Efficient game [Cesa-Bianchi and Lugosi, 2006] defined by the following loss and feedback matrices:

$$L = \begin{bmatrix}1 & 1\\1 & 0\\0 & 1\end{bmatrix}, H = \begin{bmatrix}\perp & \odot\\ \wedge & \wedge\\ \wedge & \wedge\end{bmatrix}$$

This problem includes a set of 3 possible actions and 2 possible outcomes (denoted A and B).

Signal Matrices: The signal matrices [Bartok, 2012] associated to this problem are such that $S_1 \in \{0, 1\}^{2\times2}$, $S_2 \in \{0, 1\}^{1\times2}$ and $S_3 \in \{0, 1\}^{1\times2}$. We have:

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

Cells: This game has 3 actions, each associated to a sub-space of the probability simplex called cell. We will now characterize the cell of each action in the Label Efficient game:

- For action 1, we have: $C_1 = \{p \in \Delta_M, \forall j \in N, (\ell_1 - \ell_j)^\top p \leq 0\}$. This probability space corresponds to the following constraints:

$$C_1 = \begin{bmatrix}\ell_1 - \ell_1\\ \ell_1 - \ell_2\\ \ell_1 - \ell_3\end{bmatrix}^\top p = \begin{bmatrix}0 & 0\\0 & 1\\1 & 0\end{bmatrix}^\top p \leq 0$$

  The first constraint ($(\ell_1 - \ell_1)^\top p \leq 0$) is always true. The second constraint ($(\ell_1 - \ell_2)^\top p \leq 0$) implies $p_2 \leq 0$ and the third constraint ($(\ell_1 - \ell_3)^\top p \leq 0$) implies $p_1 \leq 0$. Therefore, there exist no probability vector in Δ_M satisfying this constraint.

- For action 2, we have: $C_2 = \{p \in \Delta_M, \forall j \in N(\ell_2 - \ell_j)^\top p \leq 0\}$. This probability space corresponds to the following constraints:

$$C_2 = \begin{bmatrix}\ell_2 - \ell_1\\ \ell_2 - \ell_2\\ \ell_2 - \ell_3\end{bmatrix}^\top p = \begin{bmatrix}0 & -1\\0 & 0\\1 & -1\end{bmatrix}^\top p \leq 0$$

The second constraint ($(\ell_2 - \ell_2)^\top p \leq 0$) is always true. The first constraint ($(\ell_2 - \ell_1)^\top p \leq 0$) implies $-p_2 \leq 0 \iff p_2 \geq 0$. The third constrain ($(\ell_2 - \ell_3)^\top p \leq 0$) implies $p_1 - p_2 \leq 0 \iff p_1 \leq p_2$.

- For action 3, we have: $C_3 = \{p \in \Delta_M, \forall j \in N(\ell_3 - \ell_j)^\top p \leq 0\}$. This probability space corresponds to the following constraints:

$$C_3 = \begin{bmatrix} \ell_3 - \ell_1 \\ \ell_3 - \ell_2 \\ \ell_3 - \ell_3 \end{bmatrix}^\top p = \begin{bmatrix} -1 & 0 \\ -1 & 1 \\ 0 & 0 \end{bmatrix}^\top p \leq 0$$

The third constraint ($(\ell_3 - \ell_3)^\top p \leq 0$) is always true. The second constraint ($(\ell_3 - \ell_1)^\top p \leq 0$) implies $-p_1 + p_2 \leq 0 \iff p_2 \geq p_1$. The first constraint ($(\ell_3 - \ell_1)^\top p \leq 0$) implies $-p_1 \leq 0 \iff p_1 \geq 0$.

Pareto optimal actions: From the analysis of the cells, we have $\mathcal{C}_1 = \emptyset$. Therefore, action 1 is dominated according to the definition in [Bartok, 2012]. The remaining action 2 and 3 are Pareto optimal because their respective cells are not included in one another.

Neighboring actions: In this paragraph, we will determine whether action 2 and 3 constitute a neighboring pair, i.e if $dim(\mathcal{C}_2 \cap \mathcal{C}_3) = M - 2 = 0$.

$$\mathcal{C}_1 \cap \mathcal{C}_2 = \begin{cases} p_2 \geq 0 \\ p_1 \leq p_2 \\ p_2 \leq p_1 \\ p_1 \geq 0 \end{cases}$$

The only point in this vector space is $[0.5 \quad 0.5]$. Therefore, $dim(C_1 \cap C_2) = 0$. Therefore, $\{2, 3\}$ is a neighboring action pair.

Observability of the game: In this paragraph we will determine whether this game is globally and/or locally observable. Let's calculate $Im(S_1)$, $Im(S_2)$ and $Im(S_3)$:

- For $Im(S_1)$ we have: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} = x \begin{bmatrix} 1 \\ 0 \end{bmatrix} + y \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
- For $Im(S_2) = In(S_3)$ we have: $\begin{bmatrix} 1 \\ 1 \end{bmatrix} [x] = \begin{bmatrix} x \\ x \end{bmatrix} = x \begin{bmatrix} 1 \\ 0 \end{bmatrix} + x \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

We have $Im(S_1) = Im(S_2) = Span(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix})$. In conclusion, we have: $Im(S_1) \bigoplus Im(S_2) = Span(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix})$

The action pair (2,3) is not locally observable. We can conclude that the game is not locally observable. Howver, the action pair (1,2) is globally observable. Therefore, it can be classified as an *hard game* with a bound on the regret in $T^{2/3}$.

A.3 Instances of Label Efficient game with variable difficulty

Note that there exist instances of label efficient game where the informative action is not dominated. Let us consider the expected loss of each action obtained from a general label efficient game:

$$\langle \begin{bmatrix} a & a \\ b & 0 \\ 0 & c \end{bmatrix}, p \rangle = \begin{bmatrix} p_1 a + p_2 a \\ p_2 b \\ p_1 c \end{bmatrix}$$

Therefore, it would take the cost a to satisfy $p_1 a + p_2 a \leq p_2 b \Rightarrow a \leq p_2 b$. This would imply that the expected cost of playing the informative action is lower than the expected loss of action 2. The other option would be to satisfy $p_1 a + p_2 a \leq p_1 c \Rightarrow a \leq p_1 c$. This implies that the expected loss for playing the informative action is lower than the expected loss of playing action 3.

In practice, such combinations of the problem are not very realistic, since the goal is to play the least possible times the costly yet informative action. Therefore, such instances are less susceptible to be found in real world applications.

B More details on CBP and RandCBP

In this Section, we provide more details and code[2] on the proposed RandCBP. RandCBP is a randomized variation of the algorithm CBP. The development of RandCBP is inspired from a recent breakthrough on confidence-based approaches in the bandits setting [Vaswani et al., 2020]. The aim is to bridge these theoretical developments and empirical improvements to the partial monitoring setting and make sure that the optimality guarantees hold in this setting as well.

B.1 Instanciation of CBP and RandCBP

The difference between RandCBP and CBP is a change in the confidence interval definition. In CBP, the confidence interval is deterministic and defined as $c_{i,j}(t) = \|v_{ijk}\|_\infty \sqrt{\frac{\alpha \log(t)}{n_k(t-1)}}$. On the other hand, in RandCBP, the confidence interval is stochastic and defined as $z_{i,j}(t) = \|v_{ijk}\|_\infty Z_t \sqrt{\frac{1}{n_k(t-1)}}$ where $Z_t \sim \mathcal{N}(0, \sigma)$ and is truncated between $[0, \sqrt{\alpha \log(t)}]$. The differences between both algorithms are highlighted in purple in Algorithm 1. For both algorithms, we set $\alpha = 1.01$ and the parameter $\eta = W^{2/3}$. The set of rarely sampled actions is defined as $\mathcal{R}(t) = \{k \in \bar{N} : n_k(t) \leq f(t)\}$ with $f(t) = \alpha^{1/3} t^{2/3} \log(t)^{1/3}$. More details can be found in Bartok et al. [2012] and Bartok [2012].

B.2 The sampling distribution

We consider a discrete distribution for Z on the interval $[0, \sqrt{\alpha \log(t)}]$, supported on M points. The samples $Z_1, Z_2, ... Z_T$ are i.i.d and have the same distribution as Z. Let $\rho_1 = 0, ..., \rho_M = \sqrt{\alpha \log(t)}$ denote M equally spaced points. If M = 1 and the interval is a single point $\sqrt{\alpha \log(t)}$, then we recover CBP algorithm. If the interval is instead $[0, \sqrt{\alpha \log(t)}]$ split over M points, then we obtain RandCBP. Let $p_1, ..., p_M$ denote the probability of sampling each $\rho_1, ..., \rho_M$ value. The p_m values correspond to Gaussian distribution truncated in the $[0, \sqrt{\alpha \log(t)}]$ interval and has tunable hyper-parameters $\epsilon, \sigma > 0$. The former is the constant probability to be put on the highest point: $\rho_M = \sqrt{\alpha \log(t)}$ with $p_M = \epsilon$. For the remaining $M - 1$ points, we use a discretized Gaussian distribution; formally, for $1 \leq m \leq M - 1$, let $p_m := \exp(-\rho_m^2 / 2\sigma^2)$ and let p_m denote the normalized probabilities, that is, $p_m := (1 - \epsilon) p_m / (\sum_m p_m)$. The above choice can be viewed as a truncated (between 0 and $\alpha \log(t)$) and discretized (into M points) Gaussian distribution. We refer the reader to Vaswani et al. [2020] for more details on the sampling procedure.

B.3 Regret of RandCBP

The key to extend to the stochastic case is to observe that $z_{i,j}(t) \leq c_{i,j}(t)$ is verified anytime. We include the complete proof for completeness and write in the purple the changes made to the original proof. The modifications are located in the two lemmas at the core of the proof Bartok [2012].

In this section we provide individual and minimax upper bounds on the expected regret of RandCBP. The first theorem is an individual upper bound on the regret.

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

[2] `https://anonymous.4open.science/r/attack-detection-DF70`

```
Algorithm 1: CBP and RandCPB
input : $\mathcal{P}, \mathcal{N}, \alpha, \eta, f(\cdot), M, \sigma, \epsilon$
$n \leftarrow 0, \nu \leftarrow 0$
$N \leftarrow$ length(A)
for $t \leq N$ do
    Choose $I_t = t$, observe $Y_t$, $n_{I_t} = 1$, $\nu_{I_t} = Y_t$
for $t > N$ do
    $Z_t \sim \mathcal{N}(0, \sigma)$ , truncated between $[0, \alpha log(t)]$ according to Section B.2
    for each $\{i, j\} \in \mathcal{N}$ do
        $\tilde{\delta}_{i,j} \leftarrow \sum_{k \in V_{i,j}} v_{i,j,k}^{\top} \frac{\nu_k}{n_k}$
        ~~$c_{i,j} \leftarrow \|v_{i,j,k}\|_{\infty} \sqrt{\frac{\alpha log(t)}{n_k}}$~~
        $z_{i,j} \leftarrow \|v_{i,j,k}\|_{\infty} Z_t \sqrt{\frac{1}{n_k}}$

        if $|\tilde{\delta}_{i,j}| \leq$ ~~$c_{i,j}$~~ $z_{i,j}$ then
            Halfspace$(i, j) \leftarrow$ sign$(\tilde{\delta}_{i,j})$
        else
            Halfspace$(i, j) \leftarrow 0$
$\mathcal{P}(t), \mathcal{N}(t) \leftarrow$ GetPolytope$(\mathcal{P}, \mathcal{N},$ Halfspace$)$
$\mathcal{N}^{+}(t) \leftarrow \bigcup_{i,j \in \mathcal{N}(t)} \mathcal{N}_{i,j}^{+}$
$\mathcal{V}(t) \leftarrow \bigcup_{i,j \in \mathcal{N}(t)} \mathcal{V}_{i,j}$
$\mathcal{R}(t) \leftarrow \{k \in \bar{N} : n_k(t) \leq f(t)\}$
$\mathcal{S}(t) \leftarrow \mathcal{P}(t) \cup \mathcal{N}^{+}(t) \cup (\mathcal{V}(t) \cap \mathcal{R}(t))$
Choose $I_t = \text{argmax}_i \frac{W_i^2}{n_i}$, observe $Y_t$
$n_{I_t} = n_{I_t} + 1, \nu_{I_t} = \nu_{I_t} + Y_t$
```

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

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

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

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

Lemma 1. For any $\{i, j\} \in \mathcal{N}, t \geq 1$, we have that:

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

If for some t, i, j, the event whose probability is upper-bounded in Lemma 1 happens, we say that the *confidence interval fails*. Let $\mathcal{E}_t$ be the event that some confidence interval fails in time step t. An immediate corollary of Lemma 1 is that the sum of the probabilities that some confidence interval fails is small:

$$\sum_{t=1}^{T} \mathbb{P}(\mathcal{E}_t) \leq \sum_{t=1}^{T} \sum_{\{i,j\} \in \mathcal{N}} 2|V_i, j| t^{-2\alpha} \leq \sum_{\{i,j\} \in \mathcal{N}} 2|V_i, j|(1 + \frac{1}{2\alpha - 2})$$

Let $k(t) = \mathrm{argmax}_{i \in \mathcal{P}(t) \cup V(t)} W_i^2 / n_i(t-1)$. When $k(t) \neq I_t$, this happens because $k(t) \notin N^+(t)$ and $k(t) \notin \mathcal{R}(t)$ i.e. the action $k(t)$ is a purely information seeking action that has been sampled frequently. When this holds we say that *the decaying exploration rule is in effect at time step t*. The corresponding event is denoted by $\mathcal{D}_t = \{k(t) \neq I_t\}$. Using the notation in Bartok [2012], we can recycle the definition of d_i; and redefine these values using observer sets instead of neighborhood action sets:

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

Now we can state the following lemma:

Lemma 2. Fix any $t \geq 1$.

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

$$\delta_i \leq 2 d_i \sqrt{\frac{\alpha \log(t)}{f(t)}} \max \frac{W_k}{\sqrt{\eta_k}}$$

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

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

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

$$\begin{aligned}
\mathbb{E}[R_T] = {} & \mathbb{E}[\sum_{t=1}^{T} L[I_t, J_t]] - \sum_{t=1}^{T} \mathbb{E}[L[i^\star, J_1]] && (2) \\
= {} & \sum_{k=1}^{N} \mathbb{E}[n_k(T)] \delta_i && (3) \\
= {} & \sum_{k=1}^{N} \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = k\}}] \delta_k && (4) \\
= {} & \sum_{k=1}^{N} \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = k, \mathcal{E}_t\}}] \delta_k + \sum_{k=1}^{N} \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = k, \mathcal{E}_t^c\}}] \delta_k && (5) \\
& && (6)
\end{aligned}$$

Now, because $\delta_k \leq 1$,

$$
\begin{aligned}
\sum_{k=1}^{N}\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=k,\mathcal{E}_t\}}]\delta_k \leq &\quad \sum_{k=1}^{N}\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=k,\mathcal{E}_t\}}] \qquad (7)\\
= &\quad \mathbb{E}[\sum_{t=1}^{T}\sum_{k=1}^{N}\mathbb{1}_{\{I_t=k,\mathcal{E}_t\}}] \qquad (8)\\
= &\quad \mathbb{E}[\sum_{t=1}^{T}\mathbb{1}_{\{\mathcal{E}_t\}}] \qquad (9)\\
= &\quad \sum_{t=1}^{T}\mathbb{P}(\mathcal{E}_t) \qquad (10)\\
& \qquad (11)
\end{aligned}
$$

Hence,

$$
\mathbb{E}[R_T] \leq \sum_{t=1}^{T}\mathbb{P}(\mathcal{E}_t) + \sum_{k=1}^{N}\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}_{\{I_t,\mathcal{E}_t^c\}}]\delta_k
$$

Here, the first term can be bounded using the result from Lemma 1. Let us thus consider the elements of the second sum:

$$
\begin{aligned}
\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=k,\mathcal{E}_t^c\}}]\delta_k \leq \delta_k + &\\
&\mathbb{E}[\sum_{t=N+1}^{T}\mathbb{1}_{\{\mathcal{E}_t^c,\mathcal{D}_t^c,k\in\mathcal{P}(t)\cup N^+(t),I_t=k\}}]\delta_k +\\
&\mathbb{E}[\sum_{t=N+1}^{T}\mathbb{1}_{\{\mathcal{E}_t^c,\mathcal{D}_t^c,k\notin\mathcal{P}(t)\cup N^+(t),I_t=k\}}]\delta_k + \qquad (12)\\
&\mathbb{E}[\sum_{t=N+1}^{T}\mathbb{1}_{\{\mathcal{E}_t^c,\mathcal{D}_t,k\in\mathcal{P}(t)\cup N^+(t),I_t=k\}}]\delta_k +\\
&\mathbb{E}[\sum_{t=N+1}^{T}\mathbb{1}_{\{\mathcal{E}_t^c,\mathcal{D}_t,k\notin\mathcal{P}(t)\cup N^+(t),I_t=k\}}]\delta_k
\end{aligned}
$$

First term: 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. Note that, if action k is optimal, that is, if $\delta_k = 0$ then all the terms are zero. Thus, we can assume from now on that $\delta_k > 0$.

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

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

Therefore, we have

$$\sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t^c, \mathcal{D}_t^c, k \in \mathcal{P}(t) \cup N^+(t), I_t = k\}} \tag{13}$$

$$\leq \sum_{t=N+1}^{T} \mathbb{1}_{\{I_t = k, n_k(t-1) \leq A_k(t)\}} + \sum_{t=N+1}^{T} \mathbb{1}_{\{\mathcal{E}_t^c, \mathcal{D}_t, k \notin \mathcal{P}(t) \cup N^+(t), I_t = k, n_k(t-1) > A_k(t)\}} \tag{14}$$

$$= \sum_{t=N+1}^{T} \mathbb{1}_{\{I_t = k, n_k(t-1) \leq A_k(t)\}} \tag{15}$$

$$\leq A_k(T) = 4W_k^2 \frac{d_k^2}{\delta_k^2} \alpha \log(T) \tag{16}$$

$$\leq 4W_k^2 \frac{d_k^2}{\delta_k} \alpha \log(T) \tag{17}$$

Third term: Consider the event $\mathcal{E}_t^c \cap \mathcal{D}_t^c \cap \{k \notin \mathcal{P}(t) \cup N^+(t)\}$. We use case 2 of Lemma 2. The Lemma gives that:

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

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

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

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

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

Then, it follows that:

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

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 $n_k(t-1) \leq \mu_k f(t)$. Define $A_k(t)$ as:

$$A_k(t) = \delta_k \min(4W_k^2 \frac{d_{l(k)}^2}{\delta_{l(k)}^2} \alpha \log(t), \eta_k f(t))$$

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

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

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

Fourth term: Consider the event $\mathcal{E}_t^c \cap D_t \cap \{k \in \mathcal{P}(t) \cup N^+(t)\}$. From Lemma 2 we have that

$$\delta_k \leq 2d_k \sqrt{\frac{\alpha \log(T)}{f(t)}} \max_{j \in N} \frac{W_j}{\sqrt{\nu_j}}$$

. Thus,

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

Fifth term: Consider the event $\mathcal{E}_t^c \cap D_t \cap \{k \notin \mathcal{P}(t) \cup N^+(t)\}$ we know that $k \in \mathcal{V}(t) \cap \mathcal{R}(t) \subseteq \mathcal{R}(t)$ and hence $n_k(t-1) \leq \mu_k f(t)$. With the same argument as in the first and second term, we get that:

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

To conclude the proof of Theorem 1 we set $\nu_k = W_k^{2/3}$, $f(t) = \alpha^{1/3} t^{2/3} \log^{1/3}(t)$ and, with the notation $W = \max_{k \in N} W_k$, $\mathcal{V} = \bigcup_{i,j \in \mathcal{N}} V_{i,j}$, $N^+ = \bigcup_{i,j \in \mathcal{N}} N_{i,j}^+$, we write

$$\begin{aligned} \mathbb{E}[\mathcal{R}_T] \leq \quad & \sum_{\{i,j\} \in \mathcal{N}} 2|V_i, j|(1 + \frac{1}{2\alpha - 2}) + \sum_{k=1}^{N} \delta_k + \sum_{k=1, \delta_k > 0}^{N} 4W_k^2 \frac{d_k^2}{\delta_k^2} \alpha \log(T) + \\ & \sum_{k \in \mathcal{V} N^+} \delta_k \min(4W_k^2 \frac{d_{l(k)}^2}{\delta_{l(k)}^2} \alpha \log(T), \alpha^{1/3} W_k^{2/3} T^{2/3} \log^{1/3}(T)) + \\ & \sum_{k \in \mathcal{V} N^+} \delta_k \alpha^{1/3} W_k^{2/3} T^{2/3} \log^{1/3}(T) + 2 d_k \alpha^{1/3} W_k^{2/3} T^{2/3} \log^{1/3}(T) \end{aligned} \tag{18}$$

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

Corollary 1. Let G be a globally observable game. Then there exists a constant c such that the excpected regret can be upper bounded independently of the choice of $p^\star$ as

$$\mathbb{E}[\mathcal{R}_T] \leq c T^{2/3} \log^{1/3}(T)$$

□

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

$$\mathbb{E}[R_T] \leq c d_{pmax} \sqrt{bT \log T}$$

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

Proof. To prove this theorem, we use a scheme similar to the proof of Theorem 1. Repeating that proof, we arrive at the same expression

$$\begin{aligned} \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}_{I_t = k, \mathcal{E}_t^c}] \delta_k \leq \delta_k + & \\ & \mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\mathcal{E}_t^c, \mathcal{D}_t^c, k \in \mathcal{P}(t) \cup N^+(t), I_t = k}] \delta_k + \\ & \mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\mathcal{E}_t^c, \mathcal{D}_t^c, k \notin \mathcal{P}(t) \cup N^+(t), I_t = k}] \delta_k + \\ & \mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\mathcal{E}_t^c, \mathcal{D}_t, k \in \mathcal{P}(t) \cup N^+(t), I_t = k}] \delta_k + \\ & \mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\mathcal{E}_t^c, \mathcal{D}_t, k \notin \mathcal{P}(t) \cup N^+(t), I_t = k}] \delta_k \end{aligned} \tag{19}$$

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

From the condition of Δ' we have that there exists a positive constant ρ_1 such that for every neighboring action pair $\{i, j\} \in \mathcal{NL}, \max(\delta_i, \delta_j) \leq 4N\sqrt{\frac{\alpha \log(t)}{f(t)}} \max(W_{k'}/\sqrt{\nu'_k}) = g(t)$. It follows that if $t > g^{-1}(\rho_1)$ then the decaying exploration rule can not be in effect. Therefore, third and fourth terms can be upper bounded by $g^{-1}(\rho_1)$. With the value ρ_1 defined in the previous paragraph, we have that for any action $k \in \mathcal{V}N^+, l(k) \geq \rho_1$ holds. Therefore, third term can be upper bounded by

$$\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\mathcal{E}_t^c, \mathcal{D}_t, k \in \mathcal{P}(t) \cup N^+(t), I_t = k}] \leq 4W_k^2 \frac{4N^2}{\rho_1^2} \alpha \log(t)$$

using that d_k, defined in the proof of Theorem 1, is at most $2N$. It remains to carefully upper bound second term. For that, we first need a definition and a lemma. Let $A_\rho = \{i \in N, \delta_i \leq \rho\}$

Lemma 3. Let G = (L,H) be a finite partial-monitoring game and $p \in \Delta_M$ an opponent strategy. There exists a $\rho_2 > 0$ such that A_{ρ_2} is a point-local game in G.

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

$$\begin{aligned}
\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\mathcal{E}_t^c, \mathcal{D}_t^c, k \notin \mathcal{P}(t) \cup N^+(t), I_t = k}] = \mathbb{E}[\sum_{t=N+1}^{T} \mathbb{1}_{\mathcal{E}_t^c, \mathcal{D}_t^c, k \in P(t) \cup N^+(t), T_t = k}] \\
\leq \quad \mathbb{1}_{\delta_k < \gamma} n_k(T) \delta_k + \mathbb{1}_{k \in A_{\rho_2}, \delta_k \geq \gamma} 4W_k^2 \frac{d_k^2}{\delta_k^2} \alpha \log(T) \\
+ \mathbb{1}_{k \notin A_{\rho_2}} 4W_k^2 \frac{8N^2}{\rho_2} \alpha \log(T) \\
\leq \quad \mathbb{1}_{\delta_k < \gamma} n_k(T) \delta_k \\
+ |A_{\rho_2}| 4W^2 \frac{d_{pmax}^2}{\gamma} \alpha \log(T) + 4NW^2 \frac{8N^2}{\rho_2} \alpha \log(T)
\end{aligned} \tag{20}$$

Now we can choose γ to be

$$\gamma = 2W d_{pmax} \sqrt{\frac{b\alpha \log(T)}{T}}$$

and we get

$$\mathbb{E}[R_T] = c_1 + c_2 \log(T) + 4W d_{pmax} \sqrt{b\alpha T \log(T)}$$

□

B.4 Proofs of Lemmas

Lemma 4. For any $\{i, j\} \in \mathcal{N}, t \geq 1$, we have that:

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

Proof. Recall that the confidence $z_{i,j}(t)$ is a random variable defined as

$$z_{i,j}(t) = \sum_{k \in V_{i,j}^+} \|v_{i,j,k}\|_\infty Z_t \sqrt{\frac{1}{n_k(t-1)}}$$

where $Z_t \sim \mathcal{N}(0, \sigma)$ truncated over $[0, \sqrt{\alpha \log(t)}]$.

First, we use the triangle inequality and the union bound and the definition of $z_{i,j}(t)$:

$$\mathbb{P}(|\ \tilde{\delta}_{i,j} - \delta_{i,j}\ | \geq z_{i,j}(t)) \leq \sum_{k\in V^{+}_{i,j}} \mathbb{P}(|\ v_{ijk}^{\top} \frac{\nu_k(t)}{n_k(t)} - v_{ijk} S_k p^{\star}\ | \geq \|v_{ijk}\|_{\infty} Z_t \sqrt{\frac{1}{n_k(t-1)}}) \tag{21}$$

$$\leq \sum_{k\in V^{+}_{i,j}} \sum_{s=1}^{t-1} \mathbb{1}_{\{n_k(t-1)=s\}} \mathbb{P}(|\ v_{ijk}^{\top} \frac{\nu_k(t)}{s} - v_{ijk} S_k p^{\star}\ | \geq \|v_{ijk}\|_{\infty} Z_t \sqrt{\frac{1}{s}}) \tag{22}$$

Using Hoeffding's inequality, we obtain:

$$\leq \sum_{k\in V^{+}_{i,j}} \sum_{s=1}^{t-1} \mathbb{1}_{\{n_k(t-1)=s\}} 2\exp(\frac{-2s\|v_{ijk}\|_{\infty}^{2} Z_t^2 \sqrt{\frac{1}{s}}^{2}}{a^2}) \tag{23}$$

$$\leq \sum_{k\in V^{+}_{i,j}} \sum_{s=1}^{t-1} \mathbb{1}_{\{n_k(t-1)=s\}} 2\exp(\frac{-2\|v_{ijk}\|_{\infty}^{2} Z_t^2}{a^2}) \tag{24}$$

Choosing $a = \|v_{ijk}\|_{\infty}$, we simplify:

$$\leq \sum_{k\in V^{+}_{i,j}} \sum_{s=1}^{t-1} \mathbb{1}_{\{n_k(t-1)=s\}} 2\exp(-2Z_t^2) \tag{25}$$

Since $Z_t \leq \sqrt{\alpha \log(t)}$ we have $Z_t^2 \leq \alpha \log(t)$:

$$\leq \sum_{k\in V^{+}_{i,j}} \sum_{s=1}^{t-1} \mathbb{1}_{\{n_k(t-1)=s\}} 2\exp(-2\alpha \log(t)) \tag{26}$$

$$\leq \sum_{k\in V^{+}_{i,j}} \sum_{s=1}^{t-1} \mathbb{1}_{\{n_k(t-1)=s\}} 2\exp(\log(t^{-2\alpha})) \tag{27}$$

$$\leq \sum_{k\in V^{+}_{i,j}} 2t^{1-2\alpha} \tag{28}$$

$$\leq 2\ |\ V^{+}_{i,j}\ |\ t^{1-2\alpha} \tag{29}$$

$$\tag{30}$$

□

Lemma 5. Fix any $t \geq 1$.

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

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

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

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

Proof. Recall that the confidence $z_{i,j}(t)$ is a random variable defined as

$$z_{i,j}(t) = \sum_{k \in V_{i,j}^{+}} \|v_{i,j,k}\|_\infty Z_t \sqrt{\frac{1}{n_k(t-1)}}$$

where $Z_t \sim \mathcal{N}(0, \sigma)$ truncated over $[0, \sqrt{\alpha \log(t)}]$. Let $c_{i,j}(t)$ be the deterministic upper-bound of $z_{i,j}(t)$ such that

$$c_{i,j}(t) = \|v_{,i,j,k}\|_\infty Z_t \sqrt{\frac{\alpha \log(t)}{n_k(t-1)}}$$

First we observe that for any neighboring action pair $\{i, j\} \in \mathcal{N}(t)$, on $\mathcal{E}_t^c$, 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 $\tilde{\delta}_{i,j}(t) \leq z_{i,j}(t)$. Now, from the definition of $\mathcal{E}_t^c$, we observe $\delta_{i,j}(t) \leq \tilde{\delta}_{i,j}(t) + z_{i,j}(t)$. Putting together the two inequalities, we get $\delta_{i,j}(t) \leq 2z_{i,j}(t) \leq 2c_{i,j}(t)$.

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 Bartok [2012] that i' is well-defined.

Consider case 1. Thus, $I_t \neq k(t) = \text{argmax}_{j \in \mathcal{P}(t) \cup \mathcal{V}(t)} \frac{W_j^2}{n_j(t-1)}$. Therefore, $k(t) \notin \mathcal{R}(t)$, i.e. $n_{k(t)}(t-1) > \nu_{k(t)} f(t)$. 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, there is a path $(i_0, ..., i_r)$ in $\mathcal{N}(t)$ that connects i' to i^* ($i^* \in \mathcal{P}(t)$ holds on $\mathcal{E}_t^c$). We have that:

$$\begin{aligned} \delta_i \leq \quad & \delta_{i'} = \sum_{s=1}^{r} \delta_{i_{s-1}, i_s} && (31) \\ \leq \quad & 2 \sum_{s=1}^{r} z_{i_{s-1}, i_s} && (32) \end{aligned}$$

Using the deterministic upper-bound $c_{i,j}(t)$ on the random variable $z_{i,j}(t)$

$$\begin{aligned} \leq \quad & 2 \sum_{s=1}^{r} c_{i_{s-1}, i_s} && (33) \\ \leq \quad & 2 \sum_{s=1}^{r} \sum_{j \in V_{i_{s-1}, i_s}} \|v_{i_{s-1}, v_{i_s}, j}\|_\infty \sqrt{\frac{\alpha \log(t)}{n_j(t-1)}} && (34) \\ \leq \quad & 2 \sum_{s=1}^{r} \sum_{j \in V_{i_{s-1}, i_s}} W_j \sqrt{\frac{\alpha \log(t)}{n_j(t-1)}} && (35) \\ \leq \quad & 2 d_i W_{k(t)} \sqrt{\frac{\alpha \log(t)}{n_{k(t)}(t-1)}} && (36) \\ \leq \quad & 2 d_i W_{k(t)} \sqrt{\frac{\alpha \log(t)}{\nu_{k(t)} f(t)}} && (37) \\ & && (38) \end{aligned}$$

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

Now, for case 2 take an action k, consider $\mathcal{E}_t^c \cap \mathcal{D}_t^c$, and assume that $I_t = k$. On $D_t^c, I_t = k(t)$. Thus, from $I_t = k$ it follows that $W_k / \sqrt{n_k(t-1)}$ holds for all $j \in \mathcal{P}(t)$. Let $J_t =$

$\text{argmin}_{j \in \mathcal{P}(t) \cup N^+(t)} \frac{d_j^2}{\delta_j^2}$. Now, similarly to the previous case, there exists a path $(i_0, ..., i_r)$ from the parent action $J_{t'} \in \mathcal{P}(t)$ of J_t to $i^\star \in \mathcal{N}(t)$. Hence,

$$\delta_{J_t} \leq \quad \delta_{J_t'} = \sum_{s=1}^{r} \delta_{i_{s-1}, i_s} \tag{39}$$

$$\leq \quad 2 \sum_{s=1}^{r} z_{i_{s-1}, i_s} \tag{40}$$

Using the deterministic upper-bound $c_{i,j}(t)$ on the random variable $z_{i,j}(t)$

$$\leq \quad 2 \sum_{s=1}^{r} c_{i_{s-1}, i_s} \tag{41}$$

$$\leq \quad 2 \sum_{s=1}^{r} \sum_{j \in V_{i_{s-1}, i_s}} W_j \sqrt{\frac{\alpha \log(t)}{n_j(t-1)}} \tag{42}$$

$$\leq \quad 2 d_{J_t} W_k \sqrt{\frac{\alpha \log(t)}{n_k(t-1)}} \tag{43}$$

$$\tag{44}$$

This concludes the proof of the Lemma. □

C Additional results

C.1 Experiment details

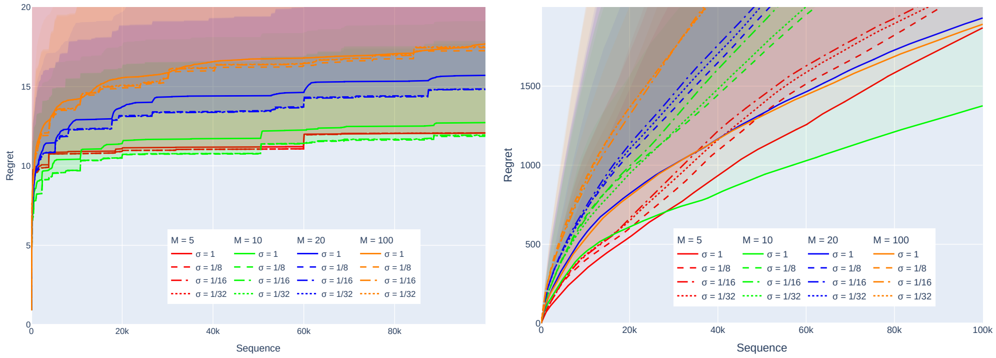
The experiments were run on a 8 cores GPU engine. The evaluation of RandCBP and all the baselines takes approximately 48 hours. The most computational baselines are BPM-Least and PM-DMED. The evaluation of PM-DMED over a set of 5 different hyper-parameters ($c \in \{0.1, 1, 5, 10, 100\}$) required a much longer time budget than evaluating RandCBP over 62 different hyper-parameter configurations although we spent an equal amount of time optimizing both algorithms for fast computation.

C.2 Hyper-parameters of RandCBP

In this Section, we show that the set of optimal hyper-parameters identified in Vaswani et al. [2020] for randomizing bandit confidence bound algorithms is close to be optimal in the partial monitoring setting. In both Figures 4 and 5 we evaluated RandCBP over sequences of $100k$ elements and averaged the regret over 100 random seeds. We conducted the evaluation for $\epsilon \in \{10^{-7}, 0.1\}$ but we only show results from $\epsilon = 10^{-7}$ because we observed the performance was much better and would provide more informative insights. This ϵ value corresponds to the one used in Vaswani et al. [2020]. Similarly, we observed that the range of optimal values for hyper-parameter M is close to the one studied in Vaswani et al. [2020]. As for hyper-parameter σ we considered the values $\sigma \in \{1/8, 1/16, 1/32\}$ which are the same as in Vaswani et al. [2020] but also considered $\sigma = 1$. This latter value achieved the best regret performance (i.e, lowest regret and smallest variance) in difficult instances of Apple Tasting and Label Efficient games with $M = 10$. For simple instances of Apple Tasting and Label Efficient, the best hyper-parameters configuration was obtained with $M = 10$ and $\sigma = 1/8$.

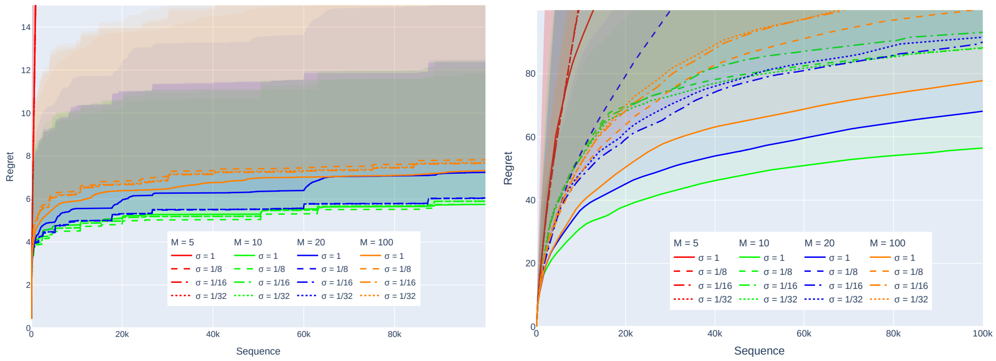
C.3 Logarithmic views of the experiments

In this Section, we provide a log-scale view of the experiments on the Label Efficient game. This allows to better visualize how different or similar the evaluated approaches are to the random selection strategy.



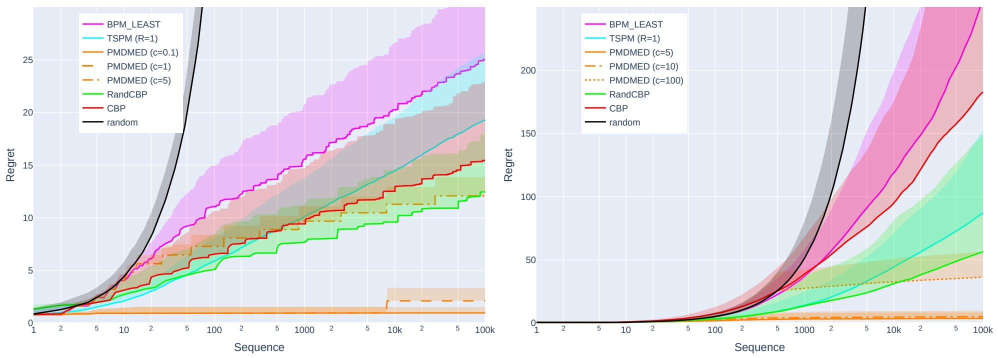
(a) Simple instances (b) Difficult instances

Figure 4: Benchmarking of RandCBP on the Label Efficient game



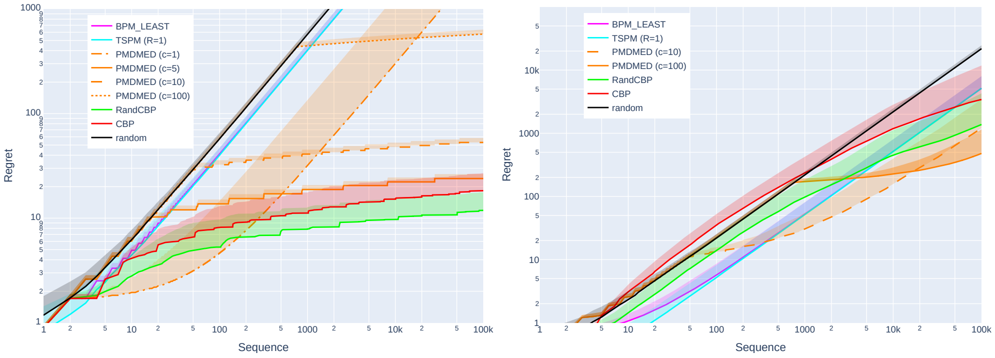
(a) Simple instances (b) Difficult instances

Figure 5: Benchmarking of RandCBP on the Apple Tasting game



(a) Simple instances (b) Difficult instances

Figure 6: Apple Tasting game, log-scales



(a) Simple instances

(b) Difficult instances

Figure 7: Label Efficient game, log-scales