Non-Asymptotic Pure Exploration by Solving Games

Rémy Degenne, Wouter M. Koolen, Pierre Ménard

Non-Asymptotic Pure Exploration by Solving Games

Rémy Degenne
Centrum Wiskunde & Informatica
Science Park 123, 1098 XG Amsterdam
remy.degenne@cwi.nl

Wouter M. Koolen
Centrum Wiskunde & Informatica
Science Park 123, 1098 XG Amsterdam
wmkoolen@cwi.nl

Pierre Ménard
Inria Lille
40 Avenue Halley, 59650 Villeneuve-d'Ascq
pierre.menard@inria.fr

Abstract

Pure exploration (aka active testing) is the fundamental task of sequentially gathering information to answer a query about a stochastic environment. Good algorithms make few mistakes and take few samples.

Lower bounds (for multi-armed bandit models with arms in an exponential family) reveal that the sample complexity is determined by the solution to an optimisation problem. The existing state of the art algorithms achieve asymptotic optimality by solving a plug-in estimate of that optimisation problem at each step.

We interpret the optimisation problem as an unknown game, and propose sampling rules based on iterative strategies to estimate and converge to its saddle point. We apply no-regret learners to obtain the first finite confidence guarantees that are adapted to the exponential family and which apply to any pure exploration query and bandit structure. Moreover, our algorithms only use a best response oracle instead of fully solving the optimisation problem.

1 Introduction

We study fundamental trade-offs arising in sequential interactive learning. We adopt the framework of Pure Exploration, in which the learning system interacts with its environment by performing a sequence of experiments, with the goal of maximising information gain. We aim to design general, efficient systems that can answer a given query with few experiments yet few mistakes.

As usual, we model the environment by a multi-armed bandit model with exponential family arms, and work in the fixed confidence (δ-PAC) setting. Information-theoretic lower bounds [13] show that a certain number of samples is unavoidable to reach a certain confidence. Moreover, algorithms are developed [13] that match these lower bounds asymptotically, in the small confidence $\delta \to 0$ regime.

Our contribution is a framework for obtaining efficient algorithms with *non-asymptotic guarantees*. The main object of study is the "Pure Exploration Game" [9], a two-player zero-sum game that is central to lower bounds as well as to the widely used GLRT-based stopping rules. We develop iterative methods that provably converge to saddle-point behaviour. The game itself is not known to the learner, and has to be explored and estimated on the fly. Our methods are based on pairs of low-regret algorithms, combined with optimism and tracking. We prove sample complexity guarantees for several combinations of algorithms, and discuss their computational and statistical trade-offs.

The rest of the introduction provides more detail on pure exploration problems, the pure exploration game, the connection between them, and expands on our contribution. We also review related work.

Preprint. Under review.

Our model for the environment is a K-armed bandit, i.e. distributions $(\nu_1, \dots, \nu_K)$ on $\mathbb{R}$. We assume throughout that these distributions come from a one-dimensional exponential family, and we denote by $d(\mu, \lambda)$ the relative entropy (Kullback-Leibler divergence) from the distribution with mean μ to that with mean λ. A pure exploration problem is parameterised by a set $\mathcal{M}$ of K-armed bandit models (the possible environments), a finite set $\mathcal{I}$ of candidate answers and a correct-answer function $i^* : \mathcal{M} \to \mathcal{I}$. We focus on *Best Arm Identification*, for which $i^*(\boldsymbol{\mu}) = \operatorname{argmax}_i \mu_i$ and the *Minimum Threshold* problem, which is defined for any fixed threshold γ by $i^*(\boldsymbol{\mu}) = \mathbf{1}_{\{\min_i \mu_i < \gamma\}}$. The goal of the learner is to learn $i^*(\boldsymbol{\mu})$ confidently and efficiently by means of sequentially sampling from the arms of $\boldsymbol{\mu}$, no matter which $\boldsymbol{\mu} \in \mathcal{M}$ it faces. When an algorithm sequentially interacts with $\boldsymbol{\mu}$, we denote by N_t^k and $\hat{\mu}_t^k$ the sample count and empirical mean estimate (these form a sufficient statistic) for arm k after t rounds. We write τ_δ for the time at which the algorithm stops and $\hat{\imath}$ for the answer it recommends. The algorithm is correct (on a particular run) if it recommends $\hat{\imath} = i^*(\boldsymbol{\mu})$ the correct answer for $\boldsymbol{\mu}$. An algorithm is δ-PAC (or δ-correct) if $\mathbb{P}_{\boldsymbol{\mu}}(\hat{\imath} \neq i^*(\boldsymbol{\mu})) \leq \delta$ for each $\boldsymbol{\mu} \in \mathcal{M}$. Among δ-PAC algorithms, we are interested in those minimising the sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$. As it turns out, what can be achieved, and how, is captured by a certain game.

For each $\boldsymbol{\mu} \in \mathcal{M}$, [9] define the two-player zero-sum simultaneous-move *Pure Exploration Game*: MAX plays an arm $k \in [K]$, MIN plays an "alternative" bandit model $\boldsymbol{\lambda} \in \mathcal{M}$ with a different correct answer $i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})$. We denote the set of such alternatives to answer i by $\neg i = \{\boldsymbol{\lambda} \in \mathcal{M} : i^*(\boldsymbol{\lambda}) \neq i\}$. MAX then receives payoff $d(\mu^k, \lambda^k)$ from MIN. As the payoff is neither concave in k (since discrete) nor convex in $\boldsymbol{\lambda}$ (both domain and divergence are problematic), we will analyse the game by sequencing the moves and considering a mixed strategy for the player moving first. With MAX moving first and playing a mixed strategy $k \sim \boldsymbol{w} \in \triangle_K$ (we identify distributions over $[K]$ and the simplex $\triangle_K$), the value of the game is

$$D_{\boldsymbol{\mu}} \coloneqq \sup_{\boldsymbol{w} \in \triangle_K} D_{\boldsymbol{\mu}}(\boldsymbol{w}) \qquad \text{where} \qquad D_{\boldsymbol{\mu}}(\boldsymbol{w}) \coloneqq \inf_{\boldsymbol{\lambda} \in \mathcal{M} : i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})} \sum_{k=1}^{K} w^k d(\mu^k, \lambda^k). \quad (1)$$

We denote a minimiser of $D_{\boldsymbol{\mu}}$ by $\boldsymbol{w}^*(\boldsymbol{\mu})$ and call it an *oracle allocation*. The analogue where MIN plays first using a mixed strategy $\boldsymbol{\lambda} \sim \boldsymbol{q} \in \mathbb{P}(\neg i^*(\boldsymbol{\mu}))$ (distributions over that set) is proposed and analysed in [9]. Despite the baroque domain of $\boldsymbol{\lambda}$ in (1), there always exist minimax $\boldsymbol{q}$ supported on $\leq K$ points due to dimension constraints.

The Pure Exploration Game is essential to both characterising the complexity of learning, and also to algorithm design. Namely, first, any δ-correct algorithm has sample complexity for each bandit $\boldsymbol{\mu} \in \mathcal{M}$ at least $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq \operatorname{kl}(\delta, 1-\delta)/D_{\boldsymbol{\mu}} \approx \ln \frac{1}{\delta} / D_{\boldsymbol{\mu}}$, and matching this rate requires sampling proportions $\mathbb{E}_{\boldsymbol{\mu}}[\boldsymbol{N}_{\tau_\delta}]/\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ converging to $\boldsymbol{w}^*(\boldsymbol{\mu})$ [see 13]. Moreover, second, the general approach for obtaining δ-correct algorithms is based on the Generalised Likelihood Ratio Test (GLRT) statistic $Z_t \coloneqq t D_{\hat{\boldsymbol{\mu}}_t}(\boldsymbol{N}_t/t)$. There are universal thresholds $\beta(t, \delta) \approx \ln \frac{1}{\delta} + \frac{K}{2} \ln \ln \frac{t}{\delta}$ [see e.g. 12, 13, 19, 23] such that $\mathbb{P}_{\boldsymbol{\mu}}\{\exists t : Z_t \geq \beta(\delta, t)\} \leq \delta$ for any $\boldsymbol{\mu} \in \mathcal{M}$. Hence stopping when $Z_t \geq \beta(t, \delta)$ and recommending $\hat{\imath} = i^*(\hat{\boldsymbol{\mu}}_t)$ is δ-correct for any sampling rule. Maximising the GLRT to stop as early as possible is achieved by the sampling proportions $\boldsymbol{N}_t/t = \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}_t)$.

These considerations show that any successful Pure Exploration agent needs to (approximately) solve the Pure Exploration Game $D_{\boldsymbol{\mu}}$. The Track-and-Stop approach, pioneered by [13], ensures that $\hat{\boldsymbol{\mu}}_t \to \boldsymbol{\mu}$ using *forced exploration*, and $\boldsymbol{N}_t/t \to \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}_t)$ using *tracking*. Continuity of $\boldsymbol{w}^*$ and $D_{\boldsymbol{\mu}}$ then yields that $Z_t \approx t D_{\boldsymbol{\mu}}(\boldsymbol{w}^*(\boldsymbol{\mu})) = t D_{\boldsymbol{\mu}}$. The GLRT stopping rule triggers when $t = \beta(\delta, t)/D_{\boldsymbol{\mu}} \approx \ln \frac{1}{\delta} / D_{\boldsymbol{\mu}}$, meeting the lower bound in the asymptotic regime $\delta \to 0$.

Our contributions. We explore methods to solve the Pure Exploration game $D_{\boldsymbol{\mu}}$ associated with the unknown bandit model $\boldsymbol{\mu}$, and discusses their statistical and computational trade-offs. We look at solving the game iteratively, by instantiating a low-regret online learners for each player. In particular for the k-player we use a self-tuning instance of Exponentiated Gradient called AdaHedge [8]. The $\boldsymbol{\lambda}$-player needs to play a distribution to deal with non-convexity; we consider Follow the Perturbed Leader as well as an ensemble of Online Gradient Descent experts. We show how a combination of optimistic gradient estimates, concentration of measure arguments and regret guarantees combine to deliver the first non-asymptotic sample complexity guarantees (which retain asymptotic optimality for $\delta \to 0$). The advantage of this approach is that it only requires a best response oracle (1, right) instead of a computationally more costly max-min oracle (1, left) employed by Track-and-Stop. Going the other extreme, we also develop Optimistic Track-and-Stop based on a max-max-min oracle (the outer

max implementing optimism over a confidence region for $\boldsymbol{\mu}$), which trades increased computation for tighter sample complexity guarantees with simpler proofs.

Our cocktail sheds new light on the trade-offs involved in the design of pure exploration algorithms. We show how "big-hammer" forced exploration can be refined using problem-adapted optimism. We show how tracking is unnecessary when the k player goes second. We show how computational complexity can be traded off using oracles of various sophistication. And finally, we validate our approach empirically in benchmark experiments at practical δ, and find that our algorithms are either competitive with Track-and-Stop (dense $\boldsymbol{w}^*$) or dominate it (sparse $\boldsymbol{w}^*$).

Related work Besides maximising information gain, there is a vast literature on maximising reward in multi-armed bandit models for which a good starting point is [21]. The canonical Pure Exploration problem is Best Arm Identification [10, 3], which is actively studied in the fixed confidence, fixed budget and simple regret settings [21, Ch. 33]. Its sample complexity as a function of the confidence level δ has been analysed very thoroughly in the (sub)-Gaussian case, where we have a rather complete picture, even including lower order terms [5]. [18] initiated the quest for correct instance-dependent constants for arms from any exponential family. [26] stresses the importance of the "moderate confidence" regime $\delta \gg 0$. Although it is not the focus here, we do believe that it is crucial to obtain the right problem dependence not only in $\ln \frac{1}{\delta}$ but also in K and other structural parameters, as the latter may in practice dominate the sample complexity.

Pure Exploration queries beyond Best Arm include Top-M [15], Thresholding [22], Minimum Threshold [20], Combinatorial Bandits [6], pure-strategy Nash equilibria [29] and Monte-Carlo Tree Search [27]. There is also significant interest in these problems in structured bandit models, including Rank-one [17], Lipschitz [23], Monotonic [14], Unimodal [7] and Unit-Sum [26]. Our framework applies to all these cases. Problems with multiple correct answers were recently considered by [9]. Existing learning strategies do not work unmodified; some fail and others need to be generalised.

Optimism is ubiquitous in bandit optimisation since [1], and was adapted to pure exploration by [16]. We are not aware of optimism being used to solve unknown min-max problems. Optimism was employed in the UCB Frank-Wolfe method by [2] for maximising an unknown smooth function faster. We do not currently know how to make use of such fast rate results. For games the best response value is a non-smooth function of the action.

Using a pair of independent no-regret learners to solve a fixed and known game goes back to [11]. More recently game dynamics were used to explain (Nesterov) acceleration in offline optimisation [28]. Ensuring faster convergence with coordinating learners is an active area of research [25]. Unfortunately, we currently do not know how to obtain an advantage in this way, as our main learning overhead comes from concentration, not regret.

2 Algorithms with finite confidence sample complexity bounds

We introduce a family of algorithms, presented as Algorithm 1, with sample complexity bounds for non-asymptotic confidence. It uses the following ingredients: the GLRT stopping rule, a saddle point algorithm (possibly formed by two regret minimization algorithms) and optimistic loss estimates.

2.1 Model and assumption: sub-Gaussian exponential families.

We suppose that the distributions belong to a known one-parameter exponential family. That is, there is a reference measure ν_0 and parameters $\eta_1, \dots, \eta_K \in \mathbb{R}$ such that the distribution of arm $k \in [K]$ is defined by $d\nu_k/d\nu_0(x) \propto e^{\eta_k x}$. Examples include Gaussians with a given variance or Bernoulli with means in $(0, 1)$. All results can be extended to arms each in a possibly different known exponential family. Let Θ be the open interval of possible means of such distributions. A distribution ν is said to be σ^2-sub-Gaussian if for all $u \in \mathbb{R}$, $\log \mathbb{E}_{X\sim\nu} e^{u(X-\mathbb{E}_{X\sim\nu}[X])} \leq \frac{\sigma^2}{2}u^2$. An exponential family has all distributions sub-Gaussian with constant σ^2 iff for all $\mu, \lambda \in \Theta$, it verifies $d(\mu, \lambda) \geq \frac{1}{2\sigma^2}(\mu - \lambda)^2$.

Assumption 1. The arm distributions belong to sub-Gaussian exponential families with constant σ^2.

Assumption 2. There exists a closed interval $[\mu_{\min}, \mu_{\max}] \subset \Theta$ such that $\mathcal{M} \subseteq [\mu_{\min}, \mu_{\max}]^K$.

As a consequence of Assumption 2, there exists $L, D > 0$ such that for all $y \in [\mu_{\min}, \mu_{\max}]$, the function $x \mapsto d(x, y)$ is L-Lipschitz on $[\mu_{\min}, \mu_{\max}]$ and $d(x, y) \leq D$. Assumption 1 is implied by

Algorithm 1 Pure exploration meta-algorithm.

Require: Algorithms $\mathcal{A}^k$ and $\mathcal{A}^{\boldsymbol{\lambda}}$, stopping threshold $\beta(t,\delta)$ and exploration bonus $f(t)$.

Sample each arm once and form estimate $\hat{\boldsymbol{\mu}}_K$.
for $t = K+1, \dots$ **do**
  For $k \in [K]$, let $[\alpha_t^k, \beta_t^k] = \{\xi : N_{t-1}^k d(\hat{\mu}_{t-1}^k, \xi) \le f(t-1)\}$. ▷ KL confidence intervals
  Let $\tilde{\boldsymbol{\mu}}_{t-1} = \operatorname{argmin}_{\boldsymbol{\lambda} \in \mathcal{M} \cap \times_{k=1}^K [\alpha_t^k, \beta_t^k]} \sum_{k=1}^K N_{t-1}^k d(\hat{\mu}_{t-1}^k, \lambda^k)$. ▷ $= \hat{\boldsymbol{\mu}}_{t-1}$ if $\hat{\boldsymbol{\mu}}_{t-1} \in \mathcal{M}$
  Let $i_t = i^*(\tilde{\boldsymbol{\mu}}_{t-1})$.
  Stop and output $\hat{\imath} = i_t$ **if** $\inf_{\boldsymbol{\lambda} \in \neg i_t} \sum_k N_{t-1}^k d(\hat{\mu}_{t-1}^k, \lambda^k) > \beta(t,\delta)$. ▷ GLRT Stopping rule
  Get $\boldsymbol{w}_t$ and $\boldsymbol{q}_t$ from $\mathcal{A}_{i_t}^k$ and $\mathcal{A}_{i_t}^{\boldsymbol{\lambda}}$.
  For $k \in [K]$, let $U_t^k = \max\left\{ f(t-1)/N_{t-1}^k, \max_{\xi \in \{\alpha_t^k, \beta_t^k\}} \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_t} d(\xi, \lambda^k) \right\}$. ▷ Optimism
  Feed $\mathcal{A}_{i_t}^k$ the loss $\ell_t^{\boldsymbol{w}}(\boldsymbol{w}) = -\sum_{k=1}^K w^k U_t^k$.
  Feed $\mathcal{A}_{i_t}^{\boldsymbol{\lambda}}$ the loss $\ell_t^{\boldsymbol{\lambda}}(\boldsymbol{q}) = \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}} \sum_{k=1}^K w_t^k d(\hat{\mu}_{t-1}^k, \lambda^k)$.
  Pick arm $k_t = \operatorname{argmin}_k N_{t-1}^k / \sum_{s=1}^t w_s^k$. ▷ Cumulative tracking
  Observe sample $X_t \sim \nu_{k_t}$. Update $\hat{\boldsymbol{\mu}}_t$.
end for

Assumption 2. Both are discussed in Appendix F. In particular, Assumption 2 can often be relaxed. L and D will appear in the sample complexity bounds but none of our algorithms use them explicitly.

Everywhere below, $\hat{\boldsymbol{\mu}}_t$ denotes the orthogonal projection of the empirical mean onto $[\mu_{\min}, \mu_{\max}]^K$, with one possible exception: the GLRT stopping rule may use it either projected or not, indifferently.

2.2 Algorithmic ingredients

Stopping and recommendation rules. The algorithm stops if any one of $|\mathcal{I}|$ GLRT tests succeeds [13]. Let $\mathcal{L}_{\boldsymbol{\mu}}$ denote the likelihood under the model parametrized by $\boldsymbol{\mu}$. The generalized log-likelihood ratio between a set Λ and the whole parameter space Θ^K is

$$\mathrm{GLR}_t^{\Theta^K}(\Lambda) = \log \frac{\sup_{\tilde{\boldsymbol{\mu}} \in \Theta^K} \mathcal{L}_{\tilde{\boldsymbol{\mu}}}(X_1, \dots, X_t)}{\sup_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}_{\boldsymbol{\lambda}}(X_1, \dots, X_t)} = \inf_{\boldsymbol{\lambda} \in \Lambda} \sum_{k \in [K]} N_t^k d(\hat{\mu}_t^k, \lambda^k) \,.$$

By concentration of measure arguments, we may find $\beta(t,\delta)$ such that with probability greater than $1-\delta$, for all $t \in \mathbb{N}$, $\mathrm{GLR}_t^{\Theta^K}(\{\boldsymbol{\mu}\}) \le \beta(t,\delta)$ [see 12, 13, 19, 23]. Test $i \in \mathcal{I}$ succeeds if $\mathrm{GLR}_t^{\Theta^K}(\neg i) > \beta(t,\delta)$. If the algorithm stops because of test i, recommend $\hat{\imath} = i$. If several tests succeed at the same time, choose arbitrarily among these.

Theorem 1. *Any algorithm using the GLRT stopping and recommendation rules with threshold $\beta(t,\delta)$ such that $\mathbb{P}_{\boldsymbol{\mu}}\{GLR_t^{\Theta^K}(\{\boldsymbol{\mu}\}) > \beta(t,\delta)\} \le \delta$ is δ-correct.*

A game with two players An algorithm is unable to stop at time t if the stopping condition is not met, i.e.

$$\beta(t,\delta) \ge \inf_{\boldsymbol{\lambda} \in \neg i^*(\hat{\boldsymbol{\mu}}_t)} \sum_{k \in [K]} N_t^k d(\hat{\mu}_t^k, \lambda^k) \,.$$

In order to stop early, the right hand side has to be maximized, i.e. made close to $t \sup_{\boldsymbol{w} \in \triangle_K} \inf_{\boldsymbol{\lambda} \in \neg i^*(\hat{\boldsymbol{\mu}}_t)} \sum_{k \in [K]} w_t^k d(\hat{\mu}_t^k, \lambda^k) = t D_{\hat{\boldsymbol{\mu}}_t} \approx t D_{\boldsymbol{\mu}}$. Then with $\beta(t,\delta) \approx \log 1/\delta + o(t)$ we obtain $t \le \log(1/\delta)/D_{\boldsymbol{\mu}}$ up to lower order terms, i.e. the stopping time is close to optimality.

We propose to approach that max-min saddle-point by implementing two iterative algorithms, $\mathcal{A}^k$ and $\mathcal{A}^{\boldsymbol{\lambda}}$, for the k-player and a $\boldsymbol{\lambda}$-player. Our sample complexity bound is a function of two quantities R_t^k and $R_t^{\boldsymbol{\lambda}}$, regret bounds of algorithms $\mathcal{A}^k$ and $\mathcal{A}^{\boldsymbol{\lambda}}$ when used for t steps on appropriate losses.

One player of our choice goes first. The second player can see the action of the first, see the corresponding loss function and use an algorithm with zero regret (e.g. Best-Response or Be-The-Leader). One of the players has to play distributions on its action set. We have one of the following:

1. $\boldsymbol{\lambda}$-player plays first and uses a distribution in $\mathbb{P}(\neg i_t)$. The k-player plays $k_t \in [K]$.
2. k-player plays first and uses $\boldsymbol{w}_t \in \triangle_K$ (distribution over $[K]$). The $\boldsymbol{\lambda}$-player plays $\boldsymbol{\lambda}_t \in \neg i_t$.
3. Both players play distributions and go in any order, or concurrently.

Algorithm 1 presents two players playing concurrently but can be modified: if for example $\boldsymbol{\lambda}$ plays second, then it gets to see $\ell_t^{\boldsymbol{\lambda}}(\boldsymbol{q})$ before computing $\boldsymbol{q}_t$.

The sampling rule at stage t first computes the most likely answer i_t for $\hat{\boldsymbol{\mu}}_{t-1}$. If the set over which the algorithm optimizes at line 4 is empty, i_t is arbitrary. The k-player plays $\boldsymbol{w}_t$ coming from $\mathcal{A}_{i_t}^k$, an instance of $\mathcal{A}^k$ running only on the rounds on which the selected answer is that i_t. The $\boldsymbol{\lambda}$-player similarly uses an instance $\mathcal{A}_{i_t}^{\boldsymbol{\lambda}}$ of $\mathcal{A}^{\boldsymbol{\lambda}}$.

Tracking. Since a single arm has to be pulled, if the k-player plays $\boldsymbol{w} \in \triangle_K$ an additional procedure is needed to translate that play into a sampling rule. We use a so-called tracking procedure, $k_t = \operatorname{argmin}_{k\in[K]} N_{t-1}^k / \sum_{s=1}^t w_s^k$, which ensures that $\sum_{s=1}^t w_s^k - (K-1) \leq N_t^k \leq \sum_{s=1}^t w_s^k$.

Optimism in face of uncertainty. Existing algorithms for general pure exploration use forced exploration to ensure convergence of $\hat{\boldsymbol{\mu}}_t$ to $\boldsymbol{\mu}$, making sure that every arm is sampled more than e.g. $\sqrt{t}$ times. We replace that method by the "optimism in face of uncertainty" principle, which gives a more adaptive exploration scheme. While that heuristic is widely used in the bandit literature, this work is its first successful implementation for general pure exploration. In Algorithm 1, the k-player algorithm gets an optimistic loss depending on $\boldsymbol{w}_t$ and $\boldsymbol{q}_t$. The $\boldsymbol{\lambda}$-player gets a non-optimistic loss.

2.3 Proof scheme and sample complexity result

In order to bound the sample complexity, we introduce a sequence of concentration events $\mathcal{E}_t = \{\forall s \leq t, \forall k \in [K],\ d(\hat{\mu}_s^k, \mu^k) \leq \frac{\widehat{W}((1+a)\log(t))}{N_s^k}\}$ for $a > 0$ and $\widehat{W}(x) = x + \log x + 1/2$. It verifies $\sum_{t=3}^{+\infty} \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_t^c) \leq 2eK/a^2$ (see Appendix B for a proof). The concentration intervals used in Algorihtm 1 are a function of $f(t) = \widehat{W}((1+a)(1+b)\log t)$ for $b > 0$.

Lemma 1. *Let $\mathcal{E}_t$ be an event and $T_0(\delta) \in \mathbb{N}$ be such that for $t \geq T_0(\delta)$, $\mathcal{E}_t \subseteq \{\tau_\delta \leq t\}$. Then*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] = \sum_{t=1}^{+\infty} \mathbb{P}\{\tau_\delta > t\} \leq T_0(\delta) + \sum_{t=T_0(\delta)}^{+\infty} \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_t^c)\,.$$

We now present briefly the steps of the proof for the stopping time upper bound before stating our main theorem on the sample complexity of Algorithm 1. These steps are inexact and should be regarded as a guideline and not as rigorous computations. A full proof of our results can be found in the appendices (Appendix B for concentration results, C for tracking and D for the main sample complexity proof). We simplify the presentation by supposing that $i_t = i^*(\boldsymbol{\mu})$ throughout (the main proof will show this may fail only $o(t)$ rounds). For $t < \tau_\delta$, under concentration event $\mathcal{E}_t$,

$$\begin{aligned}
\beta(t,\delta) &\geq \inf_{\boldsymbol{\lambda}\in\neg i^*(\boldsymbol{\mu})} \sum_{k\in[K]} N_t^k d(\hat{\mu}_t^k, \lambda^k) && \text{(stopping condition)}\\
&\geq \inf_{\boldsymbol{\lambda}\in\neg i^*(\boldsymbol{\mu})} \sum_{s\in[t]} \sum_{k\in[K]} w_s^k d(\hat{\mu}_t^k, \lambda^k) - KD && \text{(tracking)}\\
&\geq \inf_{\boldsymbol{\lambda}\in\neg i^*(\boldsymbol{\mu})} \sum_{s\in[t]} \sum_{k\in[K]} w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) - \mathcal{O}(\sqrt{t\log(t)})\,. && \text{(concentration)}
\end{aligned}$$

The first term is now the infimum of a sum of losses, $\inf_{\boldsymbol{\lambda}\in\neg i^*(\boldsymbol{\mu})} \sum_{s\in[t]} \ell_s^{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$. We use the regret property of the $\boldsymbol{\lambda}$-player's algorithm on those losses, then we introduce optimistic values U_s^k such

that for $\xi^k \in \{\mu^k, \hat{\mu}^k_{s-1}\}$ we have $\mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s}\, d(\xi^k, \lambda^k) \le U^k_s \le \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s}\, d(\xi^k, \lambda^k) + \mathcal{O}(\sqrt{1/s})$.

$$\begin{aligned}
\inf_{\boldsymbol{\lambda}\in\neg i^*(\boldsymbol{\mu})} \sum_{s\in[t]}\sum_{k\in[K]} w^k_s d(\hat{\mu}^k_{s-1}, \lambda^k) &\ge \sum_{s\in[t]} \mathop{\mathbb{E}}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} \sum_{k\in[K]} w^k_s d(\hat{\mu}^k_{s-1}, \lambda^k) - R^{\boldsymbol{\lambda}}_t && \text{(regret } \boldsymbol{\lambda}\text{)} \\
&\ge \sum_{s\in[t]}\sum_{k\in[K]} w^k_s U^k_s - \mathcal{O}(\sqrt{t}) - R^{\boldsymbol{\lambda}}_t && \text{(optimism)} \\
&\ge \max_{k\in[K]} \sum_{s\in[t]} U^k_s - R^k_t - \mathcal{O}(\sqrt{t}) - R^{\boldsymbol{\lambda}}_t && \text{(regret } \boldsymbol{w}\text{)} \\
&\ge \max_{k\in[K]} \sum_{s\in[t]} \mathop{\mathbb{E}}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} d(\mu^k, \lambda^k) - R^k_t - \mathcal{O}(\sqrt{t}) - R^{\boldsymbol{\lambda}}_t && \text{(optimism)}
\end{aligned}$$

Finally, $1/t \sum_{s\in[t]} \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s}$ is itself the expectation of another distribution on $\mathbb{P}(\neg i^*(\boldsymbol{\mu}))$. Hence

$$\max_{k\in[K]} \sum_{s\in[t]} \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s}\, d(\mu^k, \lambda^k) \ge t \inf_{\boldsymbol{q}} \max_k \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}}\, d(\mu^k, \lambda^k) = tD_{\boldsymbol{\mu}}\,.$$

Putting these inequalities together, we get finally an inequality on such a $t < \tau_\delta$. The exact result we obtain is the following Theorem, proved in Appendix D.

Theorem 2. *Under Assumption 2, the sample complexity of Algorithm 1 on model $\boldsymbol{\mu} \in \mathcal{M}$ is*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \le T_0(\delta) + \frac{2eK}{a^2} \quad \textit{with} \quad T_0(\delta) = \max\{t \in \mathbb{N} : t \le \frac{\beta(t,\delta)}{D_{\boldsymbol{\mu}}} + C_{\boldsymbol{\mu}}(R^{\boldsymbol{\lambda}}_t + R^k_t + h(t))\}\,,$$

where $C_{\boldsymbol{\mu}}$ depends on $\boldsymbol{\mu}$ and $\mathcal{M}$ and $h(t) = \mathcal{O}(\sqrt{t\log(t)})$. See Appendix D for an exact definition.

The forms of $h(t)$ and of $T_0(\delta)$ depend on the particular algorithm but we now show how an inequality of that type translates into $T_0(\delta)$. The next lemma is a consequence of the concavity of $t \mapsto \sqrt{t\log t}$.

Lemma 2. *Suppose that $t \in \mathbb{R}$ verifies the equation $t - C\sqrt{t\log t} \le \frac{\log 1/\delta}{D_{\boldsymbol{\mu}}}$. Then for $T^*_\delta = \frac{\log 1/\delta}{D_{\boldsymbol{\mu}}}$,*

$$t \le \frac{\log 1/\delta}{D_{\boldsymbol{\mu}}}\left(1 + C\sqrt{\frac{\log T^*_\delta}{T^*_\delta}}\,\frac{1}{1 - C\frac{1+\log T^*_\delta}{2\sqrt{T^*_\delta \log T^*_\delta}}}\right).$$

3 Practical Implementations

Next we discuss instantiating no-regret learners. We consider a hierarchy of computational oracles:

1. Min aka Best-Response oracle: obtain for any $i \in \mathcal{I}$, $\boldsymbol{w} \in \triangle_K$ and $\boldsymbol{\xi} \in \Theta^K$ a minimizer in $\neg i$ of $\boldsymbol{\lambda} \mapsto \sum_{k\in[K]} w^k d(\xi^k, \lambda^k)$.
2. Max-min aka Game-Solving oracle: obtain for any $i \in \mathcal{I}$ and $\boldsymbol{\xi} \in \Theta^K$ a vector $\boldsymbol{w}^* \in \triangle_K$ such that there is a Nash equilibrium $(\boldsymbol{w}^*, \boldsymbol{q}^*) \in \triangle_K \times \mathbb{P}(\neg i)$ for the zero-sum game with reward $d(\xi^k, \lambda^k)$ with the k-player using the mixed strategy $\boldsymbol{w}^*$.
3. Max-max-min oracle: for any confidence region $\mathcal{C} = [a_1, b_1] \times \ldots \times [a_K, b_K]$, obtain $(\boldsymbol{\mu}^+, i^+, \boldsymbol{w}^+)$ with $(\boldsymbol{\mu}^+, i^+) = \operatorname{argmax}_{\boldsymbol{\xi}\in\mathcal{C}, i\in\mathcal{I}} \sup_{\boldsymbol{w}\in\triangle_K} \inf_{\boldsymbol{\lambda}\in\neg i} \sum_{k=1}^K w^k d(\xi^k, \lambda^k)$ and $\boldsymbol{w}^+$ a k-player strategy of a Nash equilibrium of the game with reward $d(\mu^{+k}, \lambda^k)$.

For Minimum Threshold all oracles can be evaluated in closed form in $O(K)$ time, and the same is true for Best Response in Best Arm Identification. Max-min for Best Arm requires binary search [13] and Max-max-min requires $O(K)$ max-min calls. See [24] for run-time data on Track-and-Stop (max-min oracle) and gradient ascent (min oracle) for Best Arm. Our approach also extends naturally to min-max and max-min-max oracles, which we plan to incorporate in full detail in our future work.

3.1 A Learning Algorithm for the k-Player vs Best-Response for the $\boldsymbol{\lambda}$-Player

In this section the k-player plays first, employing a regret minimization algorithm for linear losses on the simplex to produce $\boldsymbol{w}_t \in \triangle_K$ at time t. We pick AdaHedge of [8], which runs in $O(K)$ per round and adapts to the scale of the losses. The $\boldsymbol{\lambda}$-player goes second and can use a zero-regret algorithm: Best-Response. It plays $\boldsymbol{q}_t$, a Dirac at $\boldsymbol{\lambda}_t \in \operatorname{argmin}_{\boldsymbol{\lambda}\in\neg i_t} \sum_{k\in[K]} w^k_t d(\hat{\mu}^k_{t-1}, \lambda^k)$.

Lemma 3. *AdaHedge has regret* $R_t^k \leq \sqrt{\sum_{s\leq t} b_s^2 \ln K} + \max_{s\leq t} b_s(\frac{4}{3}\ln K + 2)$ *where* $b_s = \max_k U_s^k - \min_k U_s^k \leq \max\{D, f(s)\}$ *is the loss scale in round s, so that* $R_t^k = \mathcal{O}(\sqrt{t \ln K}\ln t)$. *Best-Response has no regret,* $R_t^{\boldsymbol{\lambda}} \leq 0$. *The sample complexity is bounded per Theorem 2.*

We expect that in practice the scale converges to $b_s \to D_{\boldsymbol{\mu}}$ after a transitory startup phase.

Computational complexity: one best-response oracle call per time step.

3.2 Learning Algorithms for the $\boldsymbol{\lambda}$-Player vs Best Response for the k-Player

Using a learner for the $\boldsymbol{\lambda}$-player removes the need for a tracking procedure. In this section the k-player goes second and uses Best-Response, with zero regret, i.e. $k_t = \operatorname{argmax}_{k\in[K]} U_t^k$ (see Algorithm 1). After playing $\boldsymbol{q}_t \in \mathbb{P}(\neg i_t)$, the $\boldsymbol{\lambda}$-player suffers loss $\mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_t} d(\hat{\mu}_{t-1}^{k_t}, \lambda^{k_t})$.

Most existing regret minimization algorithms do not apply since the function $\lambda \mapsto d(\mu, \lambda)$ is not convex in general and the action set $\neg i_t$ is also not convex. The challenge is to come up with an algorithm able to play distributions with only access to a best-response oracle.

Follow-The-Perturbed-Leader. Follow-The-Perturbed-Leader can sample points from a distribution on $\mathbb{P}(\neg i)$ by only using best-response oracle calls on $\neg i$. The version we use here incorporates all the information available to the $\boldsymbol{\lambda}$-player: the loss of $\boldsymbol{\lambda} \in \neg i_t$ will be $d(\hat{\mu}_{t-1}^{k_t}, \lambda^{k_t})$ where the only unknown quantity is k_t. Let $\sigma_t \in \mathbb{R}^K$ be a random vector with independent exponentially distributed coordinates. The idea is that the distribution $\boldsymbol{q}_t$ played by the $\boldsymbol{\lambda}$-player should be the distribution of

$$\operatorname*{argmin}_{\boldsymbol{\lambda}\in\neg i_t} \sum_{s=1}^{t-1} d(\hat{\mu}_{s-1}^{k_s}, \lambda^{k_s}) + \sum_{k=1}^{K} \sigma_t^k d(\hat{\mu}_{t-1}^k, \lambda^k)\,.$$

We show in Appendix E.2 that this argmin can be computed by a single best-response oracle call. However, the k-player has to be able to compute the best response to $\boldsymbol{q}_t$. Since we cannot get the above distribution exactly, we instead take for $\boldsymbol{q}_t$ an empirical distribution from t samples. A regret bound $R_t^{\boldsymbol{\lambda}} = \mathcal{O}(\sqrt{t\log t})$ for that algorithm is in Appendix E.2. The sample complexity is then bounded by Theorem 2.

Computational complexity: t best-response oracle calls at time step t.

Online Gradient Descent. While the learning problem for $\boldsymbol{\lambda}$ is hard in general, in several common cases the sets $\neg i$ have a simple structure. If these sets are unions of a finite number J of convex sets and $\lambda \mapsto d(\mu, \lambda)$ is convex (i.e. for Gaussian or Bernoulli arm distributions), then we can use off-the-shelf regret algorithms. One gradient descent learner can be used on each convex set, and these J experts are then aggregated by an exponential weights algorithm. This procedure would have $\mathcal{O}(\sqrt{t})$ regret. The computational complexity is J (convex) best-response oracle calls per time step.

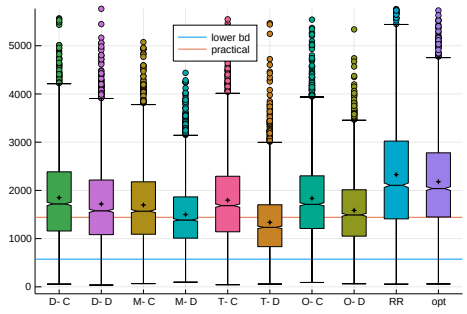
3.3 Optimistic Track-and-Stop.

At stage t, this algorithm computes $(\boldsymbol{\mu}^+, i_t) = \operatorname{argmax}_{\boldsymbol{\xi},i} \sup_{\boldsymbol{w}\in\triangle_K} \inf_{\boldsymbol{\lambda}\in\neg i} \sum_{k=1}^K w^k d(\xi^k, \lambda^k)$ where $\boldsymbol{\xi}$ ranges over all points in Θ^K in a confidence region around $\hat{\boldsymbol{\mu}}_{t-1}$ and $i \in \mathcal{I}$. Then, the k-player plays $\boldsymbol{w}_t$ such that there exists a Nash equilibrium $(\boldsymbol{w}_t, \boldsymbol{q}_t)$ of the game with reward $d(\mu^{+k}, \lambda^k)$. The proof of its sample complexity bound proceeds slightly differently from the sketch of part 2.3, although the ingredients are still the GLRT, concentration, optimism and game-solving. The proof of the following lemma can be found in appendix E.2.

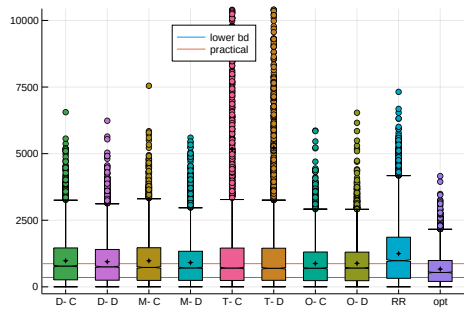
Lemma 4. *Take* $b = 1$ *in the definition of* $f(t)$. *Let* $h(t) = 2\sqrt{t}D_{\boldsymbol{\mu}} + 3L\sqrt{2\sigma^2 f(t)}(K^2 + (2\sqrt{2} + 1/3)\sqrt{Kt}) + f(t)(K^2 + 2K\log(t/K)) + KD$. *Then the expected sample complexity is at most* $T_0(\delta) + \frac{2eK}{a^2}$, *where* $T_0(\delta)$ *is the maximal* $t \in \mathbb{N}$ *such that* $t \leq (\beta(t,\delta) + h(t))/D_{\boldsymbol{\mu}}$.

Note: the K^2 factors are due to the tracking. We conjecture that they should be $K\log K$ instead.

Computational complexity: one max-max-min oracle call per time step.



(a) Best Arm for Bernoulli bandit model $\boldsymbol{\mu} = (0.3, 0.21, 0.2, 0.19, 0.18)$. The oracle weights are $\boldsymbol{w}^* = (0.34, 0.25, 0.18, 0.13, 0.10)$.



(b) Minimum Threshold for Gaussian bandit model $\boldsymbol{\mu} = (0.5, 0.6)$ with threshold $\gamma = 0.6$, $\boldsymbol{w}^* = \boldsymbol{e}_1$. Note the excessive sample complexity of T-C/ T-D.

Figure 1: Selected experiments. In both cases $\delta = 0.1$. Plots based on 3000 and 5000 runs.

This algorithm is the most computationally expensive but has the best sample complexity upper bound, has a simpler proof and works well in experiments where computing the max-max-min oracle is feasible, like the Best Arm and Minimum Threshold problems (see section 4).

4 Experiments

The goal of our experiments is to empirically validate Algorithm 1 on benchmark problems for practical δ. We use stylised stopping threshold $\beta(\delta, t) = \ln \frac{1+\ln t}{\delta}$ and exploration bonus $f(t) = \ln t$. Both are unlicensed by theory yet conservative in practise (the error frequency is way below δ). We use the following letter coding to designate sampling rules: **D** for AdaHedge vs Best-Response as advocated in Section 3.1, **T** for Track-and-Stop of [13], **M** for the Gradient Ascent algorithm of [24], **O** for Optimistic Track-and-Stop from Section 3.3, **RR** for uniform, and **opt** for following the oracle proportions $\boldsymbol{w}^*(\boldsymbol{\mu})$. We also ran all our experiments on a simplification of **D** that uses a single learner instead of partitioning the rounds according to i_t. We omit it from the results, as it was always within a few percent of **D**. We append **-C** or **-D** to indicate whether cumulative ($\boldsymbol{N}_t \rightsquigarrow \sum_{s \le t} \boldsymbol{w}_s$) or direct ($\boldsymbol{N}_t \rightsquigarrow t\boldsymbol{w}_t$) tracking [13] is employed. We finally note that we tune the learning rate of **M** in terms of (the unknown) $D_{\boldsymbol{\mu}}$.

We perform two series of experiments, one on Best Arm instances from [13, 24], and one on Minimum Threshold instances from [20]. Two selected experiments are shown in Figure 1, the others are included in Appendix G. We contrast the empirical sample complexity with the lower bound $\mathrm{kl}(\delta, 1-\delta)/D_{\boldsymbol{\mu}}$, and with a more "practical" version, which indicates the time t for which $t = \beta(t, \delta)/D_{\boldsymbol{\mu}}$, which is, approximately, the first time at which the GLRT stopping rule crosses the threshold β.

We see in Figures 1(a) and 1(b) that direct tracking **-D** has the advantage over cumulative tracking **-C** across the board, and that uniform sampling **RR** is sub-optimal as expected. In Figure 1(a) we see that **T** performs best, closely followed by **M** and **O**. Sampling from the oracle weights **opt** performs surprisingly poorly (as also observed in [26, Table 1]). The main message of Figure 1(b) is that **T** can be highly sub-optimal. We comment on the reason in Appendix G.2. Asymptotic optimality of **T** implies that this effect disappears as $\delta \to 0$. However, for this example this kicks in excruciatingly slowly. Figure 5 shows that **T** is still not competitive at $\delta = 10^{-20}$. On the other hand, **O** performs best, closely followed by **M** and then **D**. Practically, we recommend using **O** if its computational cost is acceptable, **M** if an estimate of the problem scale is available for tuning, and **D** otherwise.

The gap between **opt** and **T** (or **O**) shows that Track-and-Stop outperforms its design motivation. It is an exciting open problem to understand exactly why, and to optimise for stopping early ($\boldsymbol{N}_t/t \approx \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}_t)$) while ensuring optimality ($\mathbb{E}_{\boldsymbol{\mu}}[\boldsymbol{N}_\tau]/\mathbb{E}_{\boldsymbol{\mu}}[\tau] \approx \boldsymbol{w}^*(\boldsymbol{\mu})$).

5 Conclusion

We leveraged the game point of view of the pure exploration problem, together with the use of the optimism principle, to derive algorithms with sample complexity guarantees for non-asymptotic confidence. Varying the flavours of optimism and saddle-point strategies leads to procedures with diverse tradeoffs between sample and computational complexities. Our sample complexity bounds attain asymptotic optimality while offering guarantees for moderate confidence and the obtained algorithms are empirically sound. Our bounds however most probably do not depend optimally on the problem parameters, like the number of arms K. For BAI and the Top-K arms problems, lower bounds with lower order terms as well as matching algorithms were derived by [26]. A generalization of such lower bounds to the general pure exploration problem could shed light upon the optimal complexity across the full confidence spectrum.

The richness of existing saddle-point iterative algorithms may bring improved performance over our relatively simple choices. A smart algorithm could possibly take advantage of the stochastic nature of the losses instead of treating them as completely adversarial.

Acknowledgements

We are grateful to Zakaria Mhammedi and Emilie Kaufmann for multiple generous discussions. Travel funding was provided by INRIA Associate Team ^{6}PAC. The experiments were carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

References

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

[2] Quentin Berthet and Vianney Perchet. Fast rates for bandit optimization with upper-confidence Frank-Wolfe. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2222–2231, 2017.

[3] S. Bubeck, R. Munos, and G. Stoltz. Pure Exploration in Finitely Armed and Continuous Armed Bandits. *Theoretical Computer Science 412, 1832-1852*, 412:1832–1852, 2011.

[4] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[5] Lijie Chen, Jian Li, and Mingda Qiao. Towards instance optimal bounds for best arm identification. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 535–592, Amsterdam, Netherlands, July 2017. PMLR.

[6] S. Chen, T. Lin, I. King, M. Lyu, and W. Chen. Combinatorial Pure Exploration of Multi-Armed Bandits. In *Advances in Neural Information Processing Systems*, 2014.

[7] Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning*, pages 521–529, 2014.

[8] Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, Hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316, April 2014.

[9] Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems (NeurIPS)*, December 2019.

[10] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and markov decision processes. In *15th Annual Conference on Learning Theory (COLT)*, volume 2375 of *Lecture Notes in Computer Science*, pages 255–270. Springer, 2002.

[11] Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

[12] Aurélien Garivier. Informational confidence bounds for self-normalized averages and applications. In *2013 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2013.

[13] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027, 2016.

[14] Aurélien Garivier, Pierre Ménard, and Laurent Rossi. Thresholding bandit for dose-ranging: The impact of monotonicity. *arXiv preprint arXiv:1711.04454*, 2017.

[15] S. Kalyanakrishnan and P. Stone. Efficient Selection in Multiple Bandit Arms: Theory and Practice. In *International Conference on Machine Learning (ICML)*, 2010.

[16] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning (ICML)*, 2012.

[17] Sumeet Katariya, Branislav Kveton, Csaba Szepesvári, Claire Vernade, and Zheng Wen. Stochastic rank-1 bandits. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 392–401. PMLR, 2017.

[18] E. Kaufmann, O. Cappé, and A. Garivier. On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

[19] Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. Preprint, October 2018.

[20] Emilie Kaufmann, Wouter M. Koolen, and Aurélien Garivier. Sequential test for the lowest mean: From Thompson to Murphy sampling. In *Advances in Neural Information Processing Systems (NeurIPS) 31*, pages 6333–6343, December 2018.

[21] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2019.

[22] Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1690–1698. JMLR.org, 2016.

[23] Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pages 975–999, 2014.

[24] Pierre Ménard. Gradient ascent for active exploration in bandit problems. *arXiv preprint arXiv:1905.08165*, 2019.

[25] Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3066–3074, 2013.

[26] Max Simchowitz, Kevin Jamieson, and Benjamin Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. In *Conference on Learning Theory*, pages 1794–1834, 2017.

[27] Kazuki Teraoka, Kohei Hatano, and Eiji Takimoto. Efficient sampling method for Monte Carlo tree search problem. *IEICE Transactions*, 97-D(3):392–398, 2014.

[28] Jun-Kun Wang and Jacob D. Abernethy. Acceleration through optimistic no-regret dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3828–3838, 2018.

[29] Yichi Zhou, Jialian Li, and Jun Zhu. Identify the Nash equilibrium in static games with random payoffs. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4160–4169, International Convention Centre, Sydney, Australia, August 2017. PMLR.

A Likelihood Ratio and Exponential Families

A.1 Canonical one-parameter exponential families

We suppose that all arms have distributions in a canonical one-parameter exponential family. That is, there is a reference measure ν_0 and parameters $\eta_1, \dots, \eta_K \in \mathbb{R}$ such that the distribution of arm $k \in [K]$ is defined by

$$d\nu_k/d\nu_0(x) \propto e^{\eta_k x - \psi(\eta_k)} \qquad \text{with} \qquad \psi(\eta) = \log \mathbb{E}_{X\sim\nu_0} e^{\eta x} .$$

Let ϕ be the convex conjugate of ψ, i.e. $\phi(x) = \sup_{y\in \text{dom }\psi}(xy - \psi(y))$. Let $\Theta \subset \mathbb{R}$ be the open interval on which the first derivative ϕ' is defined. The Kullback-Leibler divergence between the distributions of the exponential family with means μ and λ in Θ is

$$d(\mu, \lambda) = \phi(\mu) - \phi(\lambda) - (\mu - \lambda)\phi'(\lambda) .$$

A distribution ν is said to be σ^2-sub-Gaussian if for all $u \in \mathbb{R}$, $\log \mathbb{E}_{X\sim\nu} e^{u(X - \mathbb{E}_{X\sim\nu}[X])} \leq \frac{\sigma^2}{2}u^2$. A canonical one-parameter exponential family has all distributions sub-Gaussian with constant σ^2 iff for all $\mu, \lambda \in \Theta$, it verifies $d(\mu, \lambda) \geq \frac{1}{2\sigma^2}(\mu - \lambda)^2$.

A.2 The Generalized log-likelihood ratio

The generalized log-likelihood ratio between the whole model space $\mathcal{M}$ and a subset $\Lambda \subseteq \mathcal{M}$ is

$$\text{GLR}_t^{\mathcal{M}}(\Lambda) = \log \frac{\sup_{\tilde{\boldsymbol{\mu}}\in\mathcal{M}} \mathcal{L}_{\tilde{\boldsymbol{\mu}}}(X_1, \dots, X_t)}{\sup_{\boldsymbol{\lambda}\in\Lambda} \mathcal{L}_{\boldsymbol{\lambda}}(X_1, \dots, X_t)} .$$

In the case of a canonical one-parameter exponential family, the likelihood of the model with means $\boldsymbol{\mu}$ is

$$\mathcal{L}_{\boldsymbol{\mu}}(X_1, \dots, X_t) = \prod_{s=1}^{t} \exp(\phi'(\mu^{k_s})(X_s - \mu^{k_s}) + \phi(\mu^{k_s}))d\nu_0(X_s)$$

For $\boldsymbol{\xi}, \boldsymbol{\lambda} \in \mathcal{M}$ two mean vectors,

$$\log \frac{\mathcal{L}_{\boldsymbol{\xi}}(X_1, \dots, X_t)}{\mathcal{L}_{\boldsymbol{\lambda}}(X_1, \dots, X_t)} = \sum_{s=1}^{t} d(X_s, \lambda^{k_s}) - d(X_s, \xi^{k_s}) = \sum_{k=1}^{K} N_t^k [d(\hat{\mu}_t^k, \lambda^k) - d(\hat{\mu}_t^k, \xi^k)] .$$

The maximum likelihood estimator $\tilde{\boldsymbol{\mu}}_t$ corresponding to the data $X_1, \dots, X_t$ is

$$\tilde{\boldsymbol{\mu}}_t = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\mathcal{M}} \sum_{k=1}^{K} N_t^k d(\hat{\mu}_t^k, \lambda^k) .$$

The GLR for set Λ is

$$\begin{aligned}
\text{GLR}_t^{\mathcal{M}}(\Lambda) &= \operatorname*{argmin}_{\boldsymbol{\lambda}\in\Lambda} \sum_{k=1}^{K} N_t^k d(\hat{\mu}_t^k, \lambda^k) - \operatorname*{argmin}_{\boldsymbol{\lambda}\in\mathcal{M}} \sum_{k=1}^{K} N_t^k d(\hat{\mu}_t^k, \lambda^k) \\
&= \operatorname*{argmin}_{\boldsymbol{\lambda}\in\Lambda} \sum_{k=1}^{K} N_t^k d(\hat{\mu}_t^k, \lambda^k) - \sum_{k=1}^{K} N_t^k d(\hat{\mu}_t^k, \tilde{\mu}_t^k) .
\end{aligned}$$

B Concentration Lemmas

B.1 Concentration bounds

For $x > 0$, let $\widehat{W}(x) = x + \log(x)$. Let W_{-1} be the negative branch of the Lambert W function and for $x \geq 1$, $\overline{W}(x) = -W_{-1}(-e^{-x})$. Then

- For $x, y \geq 1$, $x - \log x \geq y \Leftrightarrow x \geq \overline{W}(y)$.

- For $x > 1$, $\widehat{W}(x) \leq \overline{W}(x) \leq \widehat{W}(x) + \min\{\frac{1}{2}, \frac{1}{\sqrt{x}}\}$.

Lemma 5 ([12])**.** *Let $Y_1^k, \dots, Y_t^k$ be i.i.d. random variables in a canonical one-parameter exponential family with mean μ^k. Then for $\alpha > 0$,*

$$\mathbb{P}_{\boldsymbol{\mu}} \left\{ \exists s \leq t, sd(\frac{1}{s}\sum_{r=1}^{s} Y_r^k, \mu^k) \geq \alpha \right\} \leq 2e \log(t) e^{-(\alpha - \log \alpha)} .$$

Remark that for $s \leq t$, the number of pulls verifies $N_s^k \leq t$. For $t > e$ and $\alpha = \overline{W}((1+a)\log t)$ with $a > 0$, the lemma above implies

$$\mathbb{P}_{\boldsymbol{\mu}} \left\{ \exists s \leq t,\ N_s^k d(\hat{\mu}_s^k, \mu^k) \geq \overline{W}((1+a)\log t) \right\} \leq 2e \frac{\log t}{t^{1+a}} .$$

Definition 1. Let $f(s) = \overline{W}((1+a)(1+b)\log s)$.

For $s \geq t^{1/(1+b)}$, $\overline{W}((1+a)(1+b)\log s) \geq \overline{W}((1+a)\log t)$, and when the event above happens,

$$d(\hat{\mu}_s^k, \mu^k) \leq \frac{f(s)}{N_s^k} .$$

B.2 Main concentration event

Concentration event for $t \geq 3$:

$$\mathcal{E}_t = \left\{ \forall s \leq t, \forall k \in [K]\ N_s^k d(\hat{\mu}_s^k, \mu^k) \leq \overline{W}((1+a)\log t) \right\}$$

Lemma 6.

$$\forall t \geq 3,\ \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_t^c) \leq 2eK \frac{\log t}{t^{1+a}} , \qquad \sum_{t=3}^{+\infty} \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_t^c) \leq \frac{2eK}{a^2} .$$

Proof.

$$\sum_{t=3}^{+\infty} \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_t^c) \leq 2eK \sum_{t=3}^{+\infty} \frac{\log t}{t^{1+a}} \leq 2eK \int_{x=1}^{+\infty} \frac{\log x}{x^{1+a}} dx = \frac{2eK}{a^2} .$$

□

C Tracking

Lemma 7. *Let $(\boldsymbol{w}_s)_{s\in\mathbb{N}} \in \triangle_K^{\mathbb{N}}$ be vectors in the simplex with $\boldsymbol{w}_1, \dots, \boldsymbol{w}_K$ equal to the basis vectors. We recursively define for $t \in \mathbb{N}$,*

$$\forall k \in [K],\ N_K^k = 1 ,$$

$$\forall t \geq K+1,\ k_t = \operatorname*{argmin}_k \frac{N_{t-1}^k}{\sum_{s=1}^t w_s^k} , \qquad \forall k \in [K],\ N_t^k = \sum_{s=1}^{t} \mathbb{I}\{k_s = k\} .$$

The tie-breaking for the argmin *is arbitrary. Then for all $t \geq K$, all $k \in [K]$,*

$$\sum_{s=1}^{t} w_s^k - (K-1) \leq N_t^k \leq \sum_{s=1}^{t} w_s^k + 1 .$$

Proof. Let $\Sigma_t^k = \sum_{s=1}^t w_s^k$. We start by proving the inequality on the right by induction. At $t = K$, for all k, $N_K^k = \Sigma_K^k = 1$.

Suppose now that $N_s^i \leq \Sigma_s^i + 1$ for all $i \in [K]$ and all $s \leq t-1$. We prove that it also holds for t.

If $i \neq k_t$, by the induction hypothesis, $N_{t-1}^i \leq \Sigma_{t-1}^i + 1$. We obtain $N_t^i = N_{t-1}^i \leq \Sigma_{t-1}^i + 1 \leq \Sigma_t^i + 1$.

If $i = k_t$, we use that $\sum_{j=1}^K N_{t-1}^j = t-1$ and $\sum_{j=1}^K \Sigma_t^j = t$ to say that $\min_j \frac{N_{t-1}^j}{\Sigma_t^j} \leq \frac{t-1}{t} \leq 1$. Since k_t realizes that minimum, we have

$$\frac{N_t^{k_t}}{\Sigma_t^{k_t}} = \frac{N_{t-1}^{k_t}}{\Sigma_t^{k_t}} + \frac{1}{\Sigma_t^{k_t}} \leq 1 + \frac{1}{\Sigma_t^{k_t}} \,.$$

The inequality is proved for all $k \in [K]$ at t.

The lower bound for N_t^i follows from the fact that $\sum_{i=1}^K N_t^i = \sum_{i=1}^K \Sigma_t^i = t$.

$$N_t^i = t - \sum_{j \neq i} N_t^j \geq t - \sum_{j \neq i} (\Sigma_t^j + 1) = \Sigma_t^i - (K-1) \,.$$

□

Lemma 8. *For $t \geq t_0 \geq 1$ and $(x_s)_{s \in [t]}$ non-negative real numbers such that $\sum_{s=1}^{t_0-1} x_s > 0$,*

$$\sum_{s=t_0}^{t} \frac{x_s}{\sqrt{\sum_{r=1}^s x_r}} \leq 2\sqrt{\sum_{s=1}^{t} x_s} - 2\sqrt{\sum_{s=1}^{t_0-1} x_s} \,.$$

$$\sum_{s=t_0}^{t} \frac{x_s}{\sum_{r=1}^s x_r} \leq \log(\sum_{s=1}^{t} x_s) - \log(\sum_{s=1}^{t_0-1} x_s) \,.$$

Proof. By concavity of $x \mapsto \sqrt{x}$, we have $\sqrt{x} \leq \sqrt{x+y} - \frac{y}{2\sqrt{x+y}}$. We obtain $\frac{x_s}{\sqrt{\sum_{r=1}^s x_r}} \leq 2(\sqrt{\sum_{r=1}^s x_r} - \sqrt{\sum_{r=1}^{s-1} x_r})$. The sum is then telescopic. The second result uses the concavity of $x \mapsto \log(x)$. □

Lemma 9. *Let $(\boldsymbol{w}_s)_{s \in \mathbb{N}} \in \triangle_K^{\mathbb{N}}$ be vectors in the simplex. Let N_t be defined as in Lemma 7. Then*

$$\sum_{k=1}^K \sum_{s=K}^{t} \frac{w_s^k}{\sqrt{N_s^k}} \leq K^2 + 2\sqrt{Kt} \quad \textit{and} \quad \sum_{k=1}^K \sum_{s=K+1}^{t} \frac{w_s^k}{\sqrt{N_{s-1}^k}} \leq K^2 + 2\sqrt{2Kt} \,.$$

Proof. We first prove the inequality on the left. Let t_0^k be the first time such that $\sum_{r=1}^{t_0^k-1} w_r^k > K-1$. Then

$$\sum_{s=K}^{t} \frac{w_s^k}{\sqrt{N_s^k}} = \sum_{s=K}^{t_0^k-1} \frac{w_s^k}{\sqrt{N_s^k}} + \sum_{s=t_0^k}^{t} \frac{w_s^k}{\sqrt{N_s^k}} \leq \sum_{s=K}^{t_0^k-1} w_s^k + \sum_{s=t_0^k}^{t} \frac{w_s^k}{\sqrt{N_s^k}} \leq K + \sum_{s=t_0^k}^{t} \frac{w_s^k}{\sqrt{N_s^k}} \,.$$

By the tracking property of Lemma 7,

$$\sum_{s=t_0^k}^{t} \frac{w_s^k}{\sqrt{N_s^k}} \leq \sum_{s=t_0^k}^{t} \frac{w_s^k}{\sqrt{\sum_{r=1}^s w_r^k - (K-1)}} \,.$$

By Lemma 8,

$$\sum_{s=t_0^k}^{t} \frac{w_s^k}{\sqrt{\sum_{r=1}^s w_r^k - (K-1)}} \leq 2\sqrt{\sum_{s=1}^{t} w_s^k - (K-1)} - 2\sqrt{\sum_{s=1}^{t_0^k} w_s^k - (K-1)} \leq 2\sqrt{\sum_{s=1}^{t} w_s^k} \,.$$

Putting all these computations together, we obtain

$$\sum_{k=1}^K \sum_{s=K}^{t} \frac{w_s^k}{\sqrt{N_s^k}} \leq K^2 + 2\sum_{k=1}^K \sqrt{\sum_{s=1}^{t} w_s^k} \leq K^2 + 2\sqrt{Kt} \,.$$

We now prove the inequality on the right. For s such that $N_{s-1}^k \geq 1$, we have $N_{s-1}^k \geq \frac{1}{2} N_s^k$. We remark that this is true for all $s \geq K$, apply it to the sum starting from t_0^k, and obtain the wanted inequality. □

D Sample complexity proof

D.1 Upper confidence bounds

At stage t, we compute the empirical mean vector $\hat{\boldsymbol{\mu}}_{t-1}$ and the mixed strategies of the two players $\boldsymbol{w}_t$ and $\boldsymbol{q}_t$. A concentration event ensures that for all $k \in [K]$, both μ^k and $\hat{\mu}^k_{t-1}$ belong to an interval $[a^k_t, b^k_t]$. We introduce two types of coordinate-wise upper confidence bounds (UCB). The first type is a vector $U_t \in \mathbb{R}^K$ such that

$$\forall k \in [K], \forall \xi^k \in [a^k_t, b^k_t],\ U^k_t \geq \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_t}\, d(\xi^k, \lambda^k)\,.$$

The second type is a function of $\boldsymbol{\lambda}$, $U^k_t(\boldsymbol{\lambda})$ such that

$$\forall k \in [K], \forall \boldsymbol{\lambda} \in \mathcal{M}, \forall \xi^k \in [a^k_t, b^k_t],\ U^k_t(\boldsymbol{\lambda}) \geq d(\xi^k, \lambda^k)\,.$$

Let $[\alpha^k_t, \beta^k_t]$ be the intersection of $[\mu_{\min}, \mu_{\max}]$ and the interval $\{\xi \in \Theta\ :\ d(\hat{\mu}^k_{t-1}, \xi) \leq \frac{f(t-1)}{N^k_{t-1}}\}$, where f is defined in Definition 1 in section B.

Let $[a^t_t, b^k_t] = [\mu_{\min}, \mu_{\max}] \cap [\hat{\mu}^k_{t-1} - \sqrt{2\sigma^2 \frac{f(t-1)}{N^k_{t-1}}}, \hat{\mu}^k_{t-1} + \sqrt{2\sigma^2 \frac{f(t-1)}{N^k_{t-1}}}]$.

We consider the following UCBs.

1. $U_t^{k\,(1)} = \max\left\{\frac{f(t-1)}{N^k_{t-1}}, \max_{\xi\in[\alpha^k_t,\beta^k_t]} \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_t}\, d(\xi,\lambda^k)\right\}$.
2. $U_t^{k\,(2)} = \max\left\{\frac{f(t-1)}{N^k_{t-1}}, \max_{\xi\in[a^k_t,b^k_t]} \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_t}\, d(\xi,\lambda^k)\right\}$.
3. $U_t^{k\,(1)}(\boldsymbol{\lambda}) = \max\left\{\frac{f(t-1)}{N^k_{t-1}}, \max_{\xi\in[\alpha^k_t,\beta^k_t]} d(\xi,\lambda^k)\right\}$.
4. $U_t^{k\,(2)}(\boldsymbol{\lambda}) = \max\left\{\frac{f(t-1)}{N^k_{t-1}}, \max_{\xi\in[a^k_t,b^k_t]} d(\xi,\lambda^k)\right\}$.

The UCBs indexed by (2) are larger but potentially easier to compute that the ones indexed by (1), since a^k_t and b^k_t are easier to compute than α^k_t and β^k_t. The next lemma simplifies the computation of the UCBs.

Lemma 10. *In all the UCBs introduced, the maximum over the interval is attained at one of the two extremal points.*

Proof. We need to prove that a function of the form $\xi \mapsto \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}}\, d(\xi, \lambda^k)$ attains its maximum at an extremity of any interval. That function has derivative equal to $\phi'(\xi) - \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}}\, \phi'(\lambda^k)$. Since ϕ' is increasing, that derivative is negative below a point and positive afterwards. Hence the function is decreasing then increasing. We obtain that its maximum is indeed attained on an extremity of the interval. □

Lemma 11. *For all $k \in [K]$, all $t \in \mathbb{N}$, $U_t^{k\,(1)} \leq U_t^{k\,(2)}$. Furthermore for all $\boldsymbol{\lambda} \in \mathcal{M}$, $U_t^{k\,(1)}(\boldsymbol{\lambda}) \leq U_t^{k\,(2)}(\boldsymbol{\lambda})$.*

Proof. By the sub-Gaussian assumption 1, $[\alpha^k_t, \beta^k_t] \subseteq [a^k_t, b^k_t]$. □

Lemma 12. *$U_t^{k\,(1)}$ and $U_t^{k\,(2)}$ verify $\forall \xi \in [\alpha^k_t, \beta^k_t],\ U^k_t \geq \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_t}\, d(\xi, \lambda^k)$. $U_t^{k\,(1)}(\boldsymbol{\lambda})$ and $U_t^{k\,(2)}(\boldsymbol{\lambda})$ verify $\forall \xi \in [\alpha^k_t, \beta^k_t],\ U^k_t \geq d(\xi, \lambda^k)$.*

Proof. It is true for $U_t^{k\,(1)}$ and $U_t^{k\,(1)}(\boldsymbol{\lambda})$ by definition and true for $U_t^{k\,(2)}$ and $U_t^{k\,(2)}(\boldsymbol{\lambda})$ by Lemma 11. □

The analysis will proceed identically with $U_t^{k\,(1)}$ or $U_t^{k\,(2)}$ (resp. $U_t^{k\,(1)}(\boldsymbol{\lambda})$ or $U_t^{k\,(2)}(\boldsymbol{\lambda})$), which will be denoted simply by U^k_t (resp. $U^k_t(\boldsymbol{\lambda})$). The following lemma is an immediate consequence of the definition.

Lemma 13. *All UCBs presented verify* $U_t^k \geq \frac{f(t-1)}{N_{t-1}^k}$ *(resp.* $U_t^k(\boldsymbol{\lambda}) \geq \frac{f(t-1)}{N_{t-1}^k}$*).*

This lower bound is the reason the UCBs are computed as the maximum of some expression and $\frac{f(t-1)}{N_{t-1}^k}$. But for $U_t^{k\,(2)}$ and $U_t^{k\,(2)}(\boldsymbol{\lambda})$, that lower bound is also obtained automatically as soon as $[a_t^k, b_t^k] \subseteq [\mu_{\min}, \mu_{\max}]$. Indeed in that case

$$U_t^{k\,(2)} \geq \min\{d(\hat{\mu}_{t-1}^k - \sqrt{2\sigma^2 \tfrac{f(t-1)}{N_{t-1}^k}}, \hat{\mu}_{t-1}^k), d(\hat{\mu}_{t-1}^k + \sqrt{2\sigma^2 \tfrac{f(t-1)}{N_{t-1}^k}}, \hat{\mu}_{t-1}^k)\} \; .$$

From the sub-Gaussian assumption, they are both bigger than $\frac{f(t-1)}{N_{t-1}^k}$.

D.2 Saddle point algorithms

Let Λ be a subset of $\mathcal{M}$.

Definition 2. In the context of this proof, an algorithm playing sequences $(\boldsymbol{w}_s, \boldsymbol{q}_s)_{s\leq t} \in (\triangle_K \times \mathbb{P}(\Lambda))^{[t]}$ is said to be an approximate optimistic saddle point algorithm with slack x_t if

$$\inf_{\boldsymbol{\lambda}\in\Lambda} \sum_{s=1}^t \sum_{k=1}^K w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \geq \max_k \sum_{s=1}^t \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} U_s^k(\boldsymbol{\lambda}) - x_t \; ,$$

$$\text{or} \qquad \inf_{\boldsymbol{\lambda}\in\Lambda} \sum_{s=1}^t \sum_{k=1}^K w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \geq \max_k \sum_{s=1}^t U_s^k - x_t \; .$$

We now show two ways to prove that a procedure is an approximate optimistic saddle point algorithm, introducing either upper bounds $U_t^k(\boldsymbol{\lambda})$ or U_t^k.

Introduce UCBs, then use a saddle point property. We can start by replacing $d(\hat{\mu}_{s-1}^k, \lambda^k)$ by an UCB $U_s^k(\boldsymbol{\lambda})$. Let $C_s^k = \sup_{\boldsymbol{\lambda}\in\Lambda}(U_s^k(\boldsymbol{\lambda}) - d(\hat{\mu}_{s-1}^k, \lambda^k))$.

$$\inf_{\boldsymbol{\lambda}\in\Lambda} \sum_{s=1}^t \sum_{k=1}^K w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \geq \inf_{\boldsymbol{\lambda}\in\Lambda} \sum_{s=1}^t \sum_{k=1}^K w_s^k U_s^k(\boldsymbol{\lambda}) - \sum_{s=1}^t \sum_{k=1}^K w_s^k C_s^k \; .$$

Consider the following "optimistic" zero-sum games, indexed by $t \in \mathbb{N}$: to actions $(k, \boldsymbol{\lambda}) \in [K] \times \Lambda$ corresponds a reward of $U_t^k(\boldsymbol{\lambda})$ for the k-player.

An iterative saddle point algorithm attains an $(R_t^{\boldsymbol{\lambda}}, R_t^k)$ equilibrium at time t on that sequence if

$$\inf_{\boldsymbol{\lambda}\in\Lambda} \sum_{s=1}^t \sum_{k=1}^K w_s^k U_s^k(\boldsymbol{\lambda}) + R_t^{\boldsymbol{\lambda}} \geq \sum_{s=1}^t \sum_{k=1}^K w_s^k \, \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} U_s^k(\boldsymbol{\lambda}) \geq \max_{k\in[K]} \sum_{s=1}^t \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} U_s^k(\boldsymbol{\lambda}) - R_t^k \; .$$

The notations $R_t^{\boldsymbol{\lambda}}$ and R_t^k reflect a common strategy to attain such an equilibrium: instantiate two regret minimization algorithms for the two players, with linear losses $\ell_t^{\boldsymbol{w}}(\boldsymbol{w}) = -\mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_t} \sum_{k=1}^K w^k U_t^k(\boldsymbol{\lambda})$ and $\ell_t^{\boldsymbol{\lambda}}(\boldsymbol{q}) = \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}} \sum_{k=1}^K w_t^k U_t^k(\boldsymbol{\lambda})$. If we do so, the left and right inequalities are the regret properties of the algorithm for $\boldsymbol{\lambda}$ and k respectively. At that point we have

$$\inf_{\boldsymbol{\lambda}\in\Lambda} \sum_{s=1}^t \sum_{k=1}^K w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \geq \max_{k\in[K]} \sum_{s=1}^t \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} U_s^k(\boldsymbol{\lambda}) - R_t^{\boldsymbol{\lambda}} - R_t^k - \sum_{s=1}^t \sum_{k=1}^K w_s^k C_s^k \; .$$

We obtain the desired property with $x_t = R_t^{\boldsymbol{\lambda}} + R_t^k + \sum_{s=1}^t \sum_{k=1}^K w_s^k C_s^k$.

Use a regret property for $\boldsymbol{\lambda}$, then introduce UCBs, then use a regret property for k. We take here for the $\boldsymbol{\lambda}$-player a regret minimization algorithm for the loss $\ell_t^{\boldsymbol{\lambda}}(\boldsymbol{q}) = \sum_{k=1}^K w_t^k \, \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}} \, d(\hat{\mu}_{t-1}^k, \lambda^k)$. It verifies

$$\inf_{\boldsymbol{\lambda}\in\Lambda} \sum_{s=1}^t \sum_{k=1}^K w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \geq \sum_{s=1}^t \sum_{k=1}^K w_s^k \, \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} \, d(\hat{\mu}_{s-1}^k, \lambda^k) - R_t^{\boldsymbol{\lambda}} \; .$$

We now introduce UCBs U_s^k,

$$\sum_{s=1}^{t}\sum_{k=1}^{K} w_s^k \, \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} \, d(\hat{\mu}_{s-1}^k, \lambda^k) \geq \sum_{s=1}^{t}\sum_{k=1}^{K} w_s^k U_s^k - \sum_{s=1}^{t}\sum_{k=1}^{K} w_s^k C_s^k \; .$$

The k-player uses a regret minimization algorithm for the loss $\ell_t^k(\boldsymbol{w}) = -\sum_{k=1}^{K} w_s^k U_s^k$, with regret R_t^k. Let $C_s^k = U_s^k - \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} \, d(\hat{\mu}_{t-1}^k, \lambda^k)$.

$$\sum_{s=1}^{t}\sum_{k=1}^{K} w_s^k U_s^k \geq \max_k \sum_{s=1}^{t} U_s^k - R_t^k \; .$$

We obtain the desired property with $x_t = R_t^{\boldsymbol{\lambda}} + R_t^k + \sum_{s=1}^{t}\sum_{k=1}^{K} w_s^k C_s^k$.

D.3 Concentration arguments

Concentration event: $\mathcal{E}_t = \left\{ \forall s \leq t, \forall k \in [K], \; d(\hat{\mu}_s^k, \mu^k) \leq \frac{f(t^{1/(1+b)})}{N_s^k} \right\}$.

Lemma 14. *Under the event $\mathcal{E}_t$, for all $s \in [t]$, $k \in [K]$ and $\boldsymbol{\lambda} \in \mathcal{M}$,*

$$|d(\mu^k, \lambda^k) - d(\hat{\mu}_{s-1}^k, \lambda^k)| \leq L\sqrt{2\sigma^2 \frac{f(t^{1/(1+b)})}{N_{s-1}^k}} \; .$$

Proof. Use the Lipschitz property of $x \mapsto d(x, y)$, then the sub-Gaussian assumption and finally the definition of $\mathcal{E}_t$. □

Lemma 15. *Let $C_s^k = \max\left\{ 2L\sqrt{2\sigma^2 \frac{f(\max\{s-1, t^{1/(1+b)}\})}{N_{s-1}^k}}, \frac{f(\max\{s-1, t^{1/(1+b)}\})}{N_{s-1}^k} \right\}$. Let α_s^k and β_s^k be defined as in section D.1. Under the event $\mathcal{E}_t$, for all $s \in [t]$, all $\boldsymbol{\lambda} \in \mathcal{M}$,*

$$\sup_{\xi \in [\alpha_s^k, \beta_s^k]} (U_s^k(\boldsymbol{\lambda}) - d(\xi, \lambda^k)) \leq C_s^k \; ,$$
$$\sup_{\xi \in [\alpha_s^k, \beta_s^k]} (U_s^k - \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} \, d(\xi, \lambda^k)) \leq C_s^k \; .$$

Proof. Let $u_s^k = \hat{\mu}_{s-1}^k - \sqrt{2\sigma^2 \frac{f(\max\{s-1, t^{1/(1+b)}\})}{N_{s-1}^k}}$ and $v_s^k = \hat{\mu}_{s-1}^k + \sqrt{2\sigma^2 \frac{f(\max\{s-1, t^{1/(1+b)}\})}{N_{s-1}^k}}$.

Under the event $\mathcal{E}_t$, for all $s \in [t]$, we have $\hat{\mu}_{s-1}^k, \mu^k \in [u_s^k, v_s^k]$. U_s^k is defined as the maximum of $\frac{f(s-1)}{N_{s-1}^k}$ and a maximum over an interval which is contained in $[u_s^k, v_s^k]$. If U_s^k is equal to the latter,

$$\begin{aligned}\sup_{\xi \in [\alpha_s^k, \beta_s^k]} (U_s^k - \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} \, d(\xi, \lambda^k)) &\leq \sup_{\eta, \xi \in [u_s^k, v_s^k]} |\, \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} \, d(\eta, \lambda^k) - \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} \, d(\xi, \lambda^k)| \\ &\leq L|u_s^k - v_s^k| \leq 2L\sqrt{2\sigma^2 \frac{f(\max\{s-1, t^{1/(1+b)}\})}{N_{s-1}^k}} \; .\end{aligned}$$

If $U_s^k = \frac{f(s-1)}{N_{s-1}^k}$, then $\sup_{\xi \in [\alpha_s^k, \beta_s^k]} (U_s^k - \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s} \, d(\xi, \lambda^k)) \leq U_s^k = \frac{f(s-1)}{N_{s-1}^k}$.

Same computations for $U_s^k(\boldsymbol{\lambda})$, without expectations. □

Lemma 16.

$$\sum_{s=K+1}^{t}\sum_{k=1}^{K} w_s^k C_s^k \leq 2L\sqrt{2\sigma^2 f(t)}(K^2 + 2\sqrt{2Kt}) + f(t)(K^2 + 2K\log(t/K)) \; .$$

Proof. Since C_s^k is the maximum of two quantities, it is smaller than their sum. By Lemma 9,

$$\sum_{s=K+1}^{t}\sum_{k=1}^{K} w_s^k 2L\sqrt{2\sigma^2\frac{f(\max\{s-1,t^{1/(1+b)}\})}{N_{s-1}^k}} \le 2L\sqrt{2\sigma^2 f(t)}\sum_{s=K+1}^{t}\sum_{k=1}^{K}\frac{w_s^k}{\sqrt{N_{s-1}^k}}$$
$$\le 2L\sqrt{2\sigma^2 f(t)}(K^2+2\sqrt{2Kt})\,,$$

Similarly,

$$\sum_{s=K+1}^{t}\sum_{k=1}^{K} w_s^k \frac{f(\max\{s-1,t^{1/(1+b)}\})}{N_{s-1}^k} \le f(t)\sum_{s=K+1}^{t}\sum_{k=1}^{K}\frac{w_s^k}{N_{s-1}^k}$$
$$\le f(t)(K^2+2K\log(t/K))\,,$$

□

Lemma 17. *Under $\mathcal{E}_t$, for any $\boldsymbol{\lambda}\in\mathcal{M}$,*

$$\sum_{k=1}^{K} N_t^k d(\hat{\mu}_t^k,\lambda^k) \ge \sum_{k=1}^{K} N_t^k d(\mu^k,\lambda^k) - L\sqrt{2\sigma^2 K t f(t)}\,.$$

Proof. By the Lipschitzness assumption,

$$\sum_{k=1}^{K} N_t^k d(\hat{\mu}_t^k,\lambda^k) \ge \sum_{k=1}^{K} N_t^k d(\mu^k,\lambda^k) - L\sum_{k=1}^{K} N_t^k|\hat{\mu}_t^k-\mu^k|\,.$$

Using the sub-Gaussian hypothesis, under $\mathcal{E}_t$, $|\hat{\mu}_t^k-\mu^k|\le\sqrt{2\sigma^2 d(\hat{\mu}_t^k,\mu^)}\le\sqrt{2\sigma^2\frac{f(t)}{N_t^k}}$.

$$\sum_{k=1}^{K} N_t^k d(\hat{\mu}_t^k,\lambda^k) \ge \sum_{k=1}^{K} N_t^k d(\mu^k,\lambda^k) - L\sum_{k=1}^{K} N_t^k\sqrt{2\sigma^2\frac{f(t)}{N_t^k}}$$
$$= \sum_{k=1}^{K} N_t^k d(\mu^k,\lambda^k) - L\sqrt{2\sigma^2 f(t)}\sum_{k=1}^{K}\sqrt{N_t^k}$$
$$\ge \sum_{k=1}^{K} N_t^k d(\mu^k,\lambda^k) - L\sqrt{2\sigma^2 K t f(t)}\,.$$

□

D.4 The candidate answer

The data seen before time t is summarized in the vector $\hat{\boldsymbol{\mu}}_{t-1}\in\Theta^K$. That vector does not in general belong to $\mathcal{M}$.

Our algorithm finds any point in the intersection of $\mathcal{M}$ and the confidence box around $\hat{\boldsymbol{\mu}}_{t-1}$. The point obtained is denoted by $\boldsymbol{\mu}_{t-1}^{\mathcal{M}}$ and verifies that for all $k\in[K]$, $d(\hat{\mu}_{t-1}^k,\mu_{t-1}^{\mathcal{M}k})\le\frac{f(t-1)}{N_{t-1}^k}$. The candidate answer used at time t is then $i_t=i^*(\boldsymbol{\mu}_{t-1}^{\mathcal{M}})$.

D.5 When the candidate answer is not the correct answer

Chernoff information. For $x,y\in\Theta$, let $\mathrm{ch}(x,y)=\inf_{u\in\Theta}(d(u,x)+d(u,y))$ be the Chernoff information between x and y.

Assumption 3. There exists $\varepsilon>0$ such that for all $\boldsymbol{\lambda}\in\neg i^*(\boldsymbol{\mu})$, there exists $k\in[K]$ such that $\mathrm{ch}(\lambda^k,\mu^k)\ge\varepsilon$.

If the distributions are sub-Gaussian with parameter σ^2, then $\mathrm{ch}(x,y)\ge\frac{(x-y)^2}{8\sigma^2}$ and that assumption is true for every $\boldsymbol{\mu}\in\mathcal{M}$ with $D_{\boldsymbol{\mu}}>0$. i.e. Assumption 1 implies Assumption 3.

Lemma 18. *Suppose that Assumption 3 holds for $\boldsymbol{\mu} \in \mathcal{M}$ and that for all $k \in [K]$, $d(\hat{\mu}^k_{t-1}, \mu^k) \leq \frac{\log(t-1)}{N^k_{t-1}}$. If $i^*(\boldsymbol{\mu}^{\mathcal{M}}_{t-1}) \neq i^*(\boldsymbol{\mu})$ then there exists $j \in [K]$ such that $\frac{f(t-1)}{N^j_{t-1}} \geq \frac{\varepsilon}{2}$.*

Proof. If $i^*(\tilde{\boldsymbol{\mu}}_{t-1}) \neq i^*(\boldsymbol{\mu})$ then $\boldsymbol{\mu}^{\mathcal{M}}_{t-1}$ belongs to the set $\neg i^*(\boldsymbol{\mu})$.

By Assumption 3, there exists $j \in [K]$ such that $\mathrm{ch}(\mu^k, {\mu^{\mathcal{M}}_{t-1}}^k) \geq \varepsilon$. By definition of ch as an infimum over Θ, it is smaller than $d(\hat{\mu}^j_{t-1}, \mu^j) + d(\hat{\mu}^j_{t-1}, {\mu^{\mathcal{M}}_{t-1}}^j)$. That sum is then bigger than ε, with consequence that either $d(\hat{\mu}^j_{t-1}, \mu^j) \geq \varepsilon/2$ or $d(\hat{\mu}^j_{t-1}, {\mu^{\mathcal{M}}_{t-1}}^j) \geq \varepsilon/2$.

If $d(\hat{\mu}^j_{t-1}, \mu^j) \geq \varepsilon/2$, then by hypothesis, $\frac{f(t-1)}{N^j_{t-1}} \geq d(\hat{\mu}^j_{t-1}, \mu^j) \geq \varepsilon/2$.

Otherwise $d(\hat{\mu}^j_{s-1}, {\mu^{\mathcal{M}}_{t-1}}^j) \geq \varepsilon/2$. By definition of $\mu^{\mathcal{M}}_{t-1}$, $\frac{f(t-1)}{N^j_{t-1}} \geq d(\hat{\mu}^j_{t-1}, {\mu^{\mathcal{M}}_{t-1}}^j)$. We proved that $\frac{f(t-1)}{N^j_{t-1}} \geq \frac{\varepsilon}{2}$. □

Linear increase in information. For $i \in \mathcal{I}$, let $n_i(t)$ be the number of stages $s \leq t$ in which $i_s = i$. To shorten notations, let $i^* = i^*(\boldsymbol{\mu})$. The goal of this section is to find a lower bound for $n_{i^*}(t)$. We do it by showing that when the answer i_s is not the correct one, a quantity is linearly increasing, while at the same time being $O(\sqrt{t})$ by a concentration argument. Hence the number of time steps this can happen is also $O(\sqrt{t})$.

Using that $\boldsymbol{\mu} \in \neg i_s$,

$$\sum_{s \leq t, i_s \neq i^*} \sum_{k=1}^{K} w^k_s d(\hat{\mu}^k_{s-1}, \mu^k) \geq \sum_{i \in \mathcal{I} \setminus \{i^*\}} \inf_{\boldsymbol{\lambda} \in \neg i} \sum_{s \leq t, i_s = i} \sum_{k=1}^{K} w^k_s d(\hat{\mu}^k_{s-1}, \lambda^k) \,.$$

Let ε_t be the quantity on the left, which will be small by a concentration argument.

The algorithm used when $i_s = i$ is an optimistic approximate saddle point algorithm with slack $R^k_{n_i(t)} + R^{\boldsymbol{\lambda}}_{n_i(t)} + \sum_{s \leq t, i_s = i} \sum_{k=1}^{K} w^k_s C^k_s$. Hence we have

$$\varepsilon_t \geq \sum_{i \in \mathcal{I} \setminus \{i^*\}} \max_k \sum_{s \leq t, i_s = i} U^k_s - \sum_{i \in \mathcal{I} \setminus \{i^*\}} (R^k_{n_i(t)} + R^{\boldsymbol{\lambda}}_{n_i(t)}) - \sum_{s \leq t, i_s \neq i^*} \sum_{k=1}^{K} w^k_s C^k_s \,.$$

Note: if UCBs of the form $U^k_s(\boldsymbol{\lambda})$ are used instead of U^k_s, replace U^k_s by $\mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_s} U^k_s(\boldsymbol{\lambda})$ here and in the following expressions.

For fixed $i \in \mathcal{I} \setminus \{i^*\}$, we now shos that the quantity $\max_k \sum_{s \leq t, i_s = i} U^k_s$ increases linearly with the number of terms of the sum, $n_i(t)$. We proved in Lemma 13 that for all $s \in \mathbb{N}$ and $k \in [K]$, $U^k_s \geq \frac{f(s-1)}{N^k_{s-1}}$. When the event $\mathcal{E}_t$ holds, for all $s \in [t^{1/(1+b)}, t]$ with $i_s \neq i^*$, there is a $j_s \in [K]$ such that $U^{j_s}_s \geq \varepsilon/2$ by Lemma 18.

Let t' be the last term of the sum and suppose that $t' > \sqrt{t}$. Let j be such that $U^j_{t'} \geq \varepsilon/2$. Then for all $s \in [\lceil t^{1/(1+b)} \rceil, t']$,

$$\frac{f(s-1)}{N^j_{s-1}} \geq \frac{f(s-1)}{N^j_{t'-1}} = \frac{f(s-1)}{f(t'-1)} \frac{f(t'-1)}{N^j_{t'-1}} \geq \frac{f(t^{1/(1+b)})}{f(t)} \varepsilon/2 \,.$$

For $t > e$, $\frac{f(t^{1/(1+b)})}{f(t)} \geq \frac{1}{3(1+b)}$. Let $C_b = 1/(3(1+b))$.

Hence for that arm j, $\sum_{s \leq t, i_s = i} U^j_s \geq C_b \varepsilon (n_i(t) - n_i(t^{1/(1+b)}))/2$.

We conclude that the maximum over k of the sums is also bigger than this quantity. We have shown

$$\begin{aligned}\varepsilon_t &\geq \sum_{i\in\mathcal{I}\setminus\{i^*\}} \frac{C_b\varepsilon}{2}(n_i(t) - n_i(t^{1/(1+b)})) - \sum_{i\in\mathcal{I}\setminus\{i^*\}} (R^k_{n_i(t)} + R^{\boldsymbol{\lambda}}_{n_i(t)}) - \sum_{s=K+1}^{t}\sum_{k=1}^{K} w_s^k C_s^k \\ &\geq \frac{C_b\varepsilon}{2}(t - t^{1/(1+b)} - n_{i^*}(t)) - (|\mathcal{I}|-1)(R_t^k + R_t^{\boldsymbol{\lambda}}) - \sum_{s=K+1}^{t}\sum_{k=1}^{K} w_s^k C_s^k \,.\end{aligned}$$

If $n \mapsto R_n^k$ and $n \mapsto R_n^{\boldsymbol{\lambda}}$ are concave (for example regret proportional to $\sqrt{n}$), the regret term has the form $(|\mathcal{I}|-1)(R^k_{(t-n_{i^*})/(|I|-1)} + R^{\boldsymbol{\lambda}}_{(t-n_{i^*})/(|I|-1)})$.

By concentration,

$$\varepsilon_t \leq f(t^{1/(1+b)})\sum_{s=1}^{t}\sum_{k=1}^{K}\frac{w_s^k}{N_{s-1}^t} \leq f(t)(K^2 + 2K\log(t/K)) \,.$$

We proved

$$n_{i^*}(t) \geq t - t^{1/(1+b)} - \frac{2}{C_b\varepsilon}\left((|\mathcal{I}|-1)(R_t^k + R_t^{\boldsymbol{\lambda}}) + f(t)(K^2 + 2K\log(t/K)) + \sum_{s=K+1}^{t}\sum_{k=1}^{K} w_s^k C_s^k\right) \tag{2}$$

D.6 When the candidate answer is the correct answer

Let $t' \leq t$ be the last round in which $i_{t'} = i^*$ before the algorithm stops. Then $t' \geq n_{i^*}(t)$, we have $i_{t'} = i^*$ and $n_{i^*}(t') = n_{i^*}(t)$.

$$\begin{aligned}\beta(t,\delta) \geq \beta(t',\delta) &\geq \inf_{\boldsymbol{\lambda}\in\neg i_{t'}} \sum_{k=1}^{K} N_{t'}^k d(\hat{\mu}_{t'}^k, \lambda^k) = \inf_{\boldsymbol{\lambda}\in\neg i^*} \sum_{k=1}^{K} N_{t'}^k d(\hat{\mu}_{t'}^k, \lambda^k) \\ &\geq \inf_{\boldsymbol{\lambda}\in\neg i^*} \sum_{k=1}^{K} N_{t'}^k d(\mu^k, \lambda^k) - L\sqrt{2\sigma^2 K t f(t)} \,.\end{aligned}$$

Using the tracking Lemma 7, then concentration Lemma 14,

$$\begin{aligned}\beta(t,\delta) &\geq \inf_{\boldsymbol{\lambda}\in\neg i^*} \sum_{s=1}^{t'}\sum_{k=1}^{K} w_s^k d(\mu^k, \lambda^k) - KD - L\sqrt{2\sigma^2 K t f(t)} \\ &\geq \inf_{\boldsymbol{\lambda}\in\neg i^*} \sum_{s=K+1}^{t'}\sum_{k=1}^{K} w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \\ &\qquad - L\sqrt{2\sigma^2 f(t)}\sum_{s=K+1}^{t}\sum_{k=1}^{K}\frac{w_s^k}{\sqrt{N_{s-1}^k}} - KD - L\sqrt{2\sigma^2 K t f(t)} \\ &\geq \inf_{\boldsymbol{\lambda}\in\neg i^*} \sum_{s=K+1}^{t'}\sum_{k=1}^{K} w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \\ &\qquad - 2L\sqrt{2\sigma^2 f(t)}(K^2 + 2\sqrt{2Kt}) - KD - L\sqrt{2\sigma^2 K t f(t)}\end{aligned}$$

We drop the rounds in which $i_s \neq i^*$.

$$\begin{aligned}\beta(t,\delta) \geq \inf_{\boldsymbol{\lambda}\in\neg i^*} &\sum_{K+1\leq s\leq t', i_s = i^*}\sum_{k=1}^{K} w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \\ &- 2L\sqrt{2\sigma^2 f(t)}(K^2 + 2\sqrt{2Kt}) - KD - L\sqrt{2\sigma^2 K t f(t)} \,.\end{aligned}$$

The algorithm used is an optimistic approximate saddle point algorithm with slack $R_t^{\boldsymbol{\lambda}} + R_t^k + \sum_{s=1}^t \sum_{k=K+1}^K w_s^k C_s^k$:

$$\inf_{\boldsymbol{\lambda} \in \neg i^*} \sum_{s \leq t', i_s = i^*} \sum_{k=1}^K w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \geq \max_k \sum_{s \leq t', i_s = i^*} U_s^k - (R_t^{\boldsymbol{\lambda}} + R_t^k + \sum_{s=K+1}^t \sum_{k=1}^K w_s^k C_s^k) \, .$$

Let $A_t = \sum_{s=K+1}^t \sum_{k=1}^K w_s^k C_s^k + 2L\sqrt{2\sigma^2 f(t)}(K^2 + 2\sqrt{2Kt}) + KD + L\sqrt{2\sigma^2 Ktf(t)}$. We obtain

$$\beta(t, \delta) \geq \max_k \sum_{K < s \leq t', i_s = i^*} U_s^k - R_t^k - R_t^{\boldsymbol{\lambda}} - A_t \, .$$

Let $t_b = t^{1/(1+b)}$. Since U_t is a coordinate-wise upper confidence bound when concentration holds (for $s \geq t^{1/(1+b)}$), we have

$$\begin{aligned}
\beta(t, \delta) &\geq \max_k \sum_{t_b \leq s \leq t', i_s = i^*} \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_s} \, d(\mu^k, \lambda^k) - R_t^k - R_t^{\boldsymbol{\lambda}} - A_t \\
&= (n_{i^*}(t') - t_b) \max_k \frac{1}{(n_{i^*}(t') - t_b)} \sum_{t^{1/(1+b)} \leq s \leq t', i_s = i^*} \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_s} \, d(\mu^k, \lambda^k) - R_t^k - R_t^{\boldsymbol{\lambda}} - A_t \\
&\geq (n_{i^*}(t') - t_b) \inf_{\boldsymbol{q} \in \mathbb{P}(\neg i^*)} \max_k \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}} \, d(\mu^k, \lambda^k) - R_t^k - R_t^{\boldsymbol{\lambda}} - A_t \\
&= (n_{i^*}(t') - t^{1/(1+b)}) D_{\boldsymbol{\mu}} - R_t^k - R_t^{\boldsymbol{\lambda}} - A_t \, .
\end{aligned}$$

t' is such that $n_{i^*}(t') = n_{i^*}(t)$. Combining that result and the lower bound on $n_{i^*}(t)$ of equation (2), we have

$$\frac{\beta(t, \delta) + A_t + R_t^k + R_t^{\boldsymbol{\lambda}}}{D_{\boldsymbol{\mu}}} \tag{3}$$

$$\geq t - 2t^{1/(1+b)} - \frac{2}{C_b \varepsilon} \left((|\mathcal{I}| - 1)(R_t^k + R_t^{\boldsymbol{\lambda}}) + f(t)(K^2 + 2K \log(t/K)) + \sum_{s=K+1}^t \sum_{k=1}^K w_s^k C_s^k \right) . \tag{4}$$

D.7 Stopping time upper bound

We can solve equation (3) to find an upper bound for t such that the algorithm does not stop. Suppose that there exists $R > 0$ such that $R_t^k + R_t^{\boldsymbol{\lambda}} \leq R\sqrt{Kt}$. Take $b = 1$. By Lemma 16,

$\sum_{s=K+1}^t \sum_{k=1}^K w_s^k C_s^k \leq 2L\sqrt{2\sigma^2 f(t)}(K^2 + 2\sqrt{2Kt}) + f(t)(K^2 + 2K \log(t/K))$.

$A_t \leq 4L\sqrt{2\sigma^2 f(t)}(K^2 + 2\sqrt{2Kt}) + KD + L\sqrt{2\sigma^2 Ktf(t)} + f(t)(K^2 + 2K \log(t/K))$.

We now define

$$\begin{aligned}
h(t) = 2\sqrt{t} &+ \frac{A_t + R\sqrt{Kt}}{D_{\boldsymbol{\mu}}} \\
&+ \frac{2}{C_b \varepsilon} \left((|\mathcal{I}| - 1) R\sqrt{Kt} + 2f(t)(K^2 + 2K \log(t/K)) + 2L\sqrt{2\sigma^2 f(t)}(K^2 + 2\sqrt{2Kt}) \right) \, .
\end{aligned}$$

We have that $h(t) = \mathcal{O}(\sqrt{t \log t})$ and we obtained that if $t < \tau_\delta$ then

$$t - h(t) \leq \frac{\beta(t, \delta)}{D_{\boldsymbol{\mu}}} \, .$$

E Algorithms

E.1 Optimistic Track and Stop

We prove that under the concentration event $\mathcal{E}_t$, there is an upper bound on t such that $t < \tau_\delta$.

Let $\mathcal{C}_s = \{\boldsymbol{\xi} \in \Theta^K : \forall k \in [K], d(\hat{\mu}_{s-1}^k, \xi^k) \leq \frac{f(s-1)}{N_{s-1}^k}\}$ be a confidence region around $\hat{\boldsymbol{\mu}}_{s-1}$.

When $i_t \neq i^*(\boldsymbol{\mu})$. Let $i \in \mathcal{I} \setminus \{i^*(\boldsymbol{\mu})\}$. Since $i_s \neg i^*(\boldsymbol{\mu})$ implies that $\boldsymbol{\mu} \in \neg i_s$,

$$\sum_{s \leq t, i_s = i} \sum_{k=1}^{K} w_s^k d(\hat{\mu}_{s-1}^k, \mu^k) \geq \inf_{\boldsymbol{\lambda} \in \neg i} \sum_{s \leq t, i_s = i} \sum_{k=1}^{K} w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \, .$$

Let ε_t^i be the left hand side of that inequality. Since $\hat{\boldsymbol{\mu}}_{s-1}$ and $\boldsymbol{\mu}_s^+$ both belong to $\mathcal{C}_s$, we have

$$\sum_{s \leq t, i_s = i} \sum_{k=1}^{K} w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \geq \sum_{s \leq t, i_s = i} \sum_{k=1}^{K} w_s^k d(\mu_s^{+k}, \lambda^k) - L\sqrt{2\sigma^2 f(t)} \sum_{s \leq t, i_s = i} \frac{w_s^k}{\sqrt{N_{s-1}^k}} \, .$$

By definition of $\boldsymbol{\mu}_s^+$,

$$\inf_{\boldsymbol{\lambda} \in \neg i} \sum_{s \leq t, i_s = i} \sum_{k=1}^{K} w_s^k d(\mu_s^{+k}, \lambda^k) \geq \sum_{s \leq t, i_s = i} \inf_{\boldsymbol{\lambda} \in \neg i} \sum_{k=1}^{K} w_s^k d(\mu_s^{+k}, \lambda^k) = \sum_{s \leq t, i_s = i} D_{\boldsymbol{\mu}_s^+} \, .$$

For $s \geq t^{1/(1+b)}$, $\boldsymbol{\mu} \in \mathcal{C}_s$ and by definition of $\boldsymbol{\mu}_s^+$, $D_{\boldsymbol{\mu}_s^+} \geq D_{\boldsymbol{\mu}}$. We obtain, with $n_i(t)$ the number of times with $i_s = i$ until t,

$$\begin{aligned}
\sum_{i \in \mathcal{I} \setminus \{i^*(\boldsymbol{\mu})\}} \varepsilon_t^i &\geq (t - n_{i^*(\boldsymbol{\mu})}(t) - t^{1/(1+b)}) D_{\boldsymbol{\mu}} - L\sqrt{2\sigma^2 f(t)} \sum_{s \leq t} \frac{w_s^k}{\sqrt{N_{s-1}^k}} \\
&\geq (t - n_{i^*(\boldsymbol{\mu})}(t) - t^{1/(1+b)}) D_{\boldsymbol{\mu}} - L\sqrt{2\sigma^2 f(t)} (K^2 + 2\sqrt{2Kt}) \, .
\end{aligned}$$

See Lemma 9 for that last inequality. By concentration,

$$\sum_{i \in \mathcal{I} \setminus \{i^*(\boldsymbol{\mu})\}} \varepsilon_t^i \leq f(t) \sum_{s=1}^{t} \sum_{k=1}^{K} \frac{w_s^k}{N_{s-1}^k} \leq f(t)(K^2 + 2K \log(t/K)) \, .$$

Finally,

$$n_{i^*(\boldsymbol{\mu})}(t) \geq t - t^{1/(1+b)} - \frac{1}{D_{\boldsymbol{\mu}}} \left(L\sqrt{2\sigma^2 f(t)} (K^2 + 2\sqrt{2Kt}) + f(t)(K^2 + 2K \log(t/K)) \right) \, .$$

When $i_t = i^*(\boldsymbol{\mu})$. Let $t' \geq n_{i^*(\boldsymbol{\mu})}(t)$ be such that $i_{t'} = i^*(\boldsymbol{\mu})$ and $n_{i^*(\boldsymbol{\mu})}(t') = n_{i^*(\boldsymbol{\mu})}(t)$. Using concentration and tracking properties, as in the main sample complexity proof of Appendix D.6,

$$\begin{aligned}
\beta(t', \delta) &\geq \inf_{\boldsymbol{\lambda} \in \neg i^*(\boldsymbol{\mu})} \sum_{k=1}^{K} N_t^k d(\hat{\mu}_{t'}^k, \lambda^k) \\
&\geq \inf_{\boldsymbol{\lambda} \in \neg i^*(\boldsymbol{\mu})} \sum_{k=1}^{K} N_t^k d(\mu^k, \lambda^k) - L\sqrt{2\sigma^2 K t f(t)} \\
&\geq \inf_{\boldsymbol{\lambda} \in \neg i^*(\boldsymbol{\mu})} \sum_{s=1}^{t'} \sum_{k=1}^{K} w_s^k d(\mu^k, \lambda^k) - KD - L\sqrt{2\sigma^2 K t f(t)} \\
&\geq \inf_{\boldsymbol{\lambda} \in \neg i^*(\boldsymbol{\mu})} \sum_{s=1}^{t'} \sum_{k=1}^{K} w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \\
&\qquad - L\sqrt{2\sigma^2 f(t)} (K^2 + 2\sqrt{2Kt}) - KD - L\sqrt{2\sigma^2 K t f(t)}
\end{aligned}$$

Since $\hat{\boldsymbol{\mu}}_{s-1}$ and $\boldsymbol{\mu}_s^+$ both belong to $\mathcal{C}_s$, we have

$$\inf_{\boldsymbol{\lambda} \in \neg i^*(\boldsymbol{\mu})} \sum_{s=1}^{t'} \sum_{k=1}^{K} w_s^k d(\hat{\mu}_{s-1}^k, \lambda^k) \geq \inf_{\boldsymbol{\lambda} \in \neg i^*(\boldsymbol{\mu})} \sum_{s=1}^{t'} \sum_{k=1}^{K} w_s^k d(\mu_s^{+k}, \lambda^k) - L\sqrt{2\sigma^2 f(t)} (K^2 + 2\sqrt{2Kt}) \, .$$

Let $B_t = 2L\sqrt{2\sigma^2 f(t)}(K^2 + 2\sqrt{2Kt}) + KD + L\sqrt{2\sigma^2 Ktf(t)}$.

$$\begin{aligned}\beta(t,\delta) &\geq \inf_{\boldsymbol{\lambda}\in\neg i^*(\boldsymbol{\mu})} \sum_{s\leq t', i_s = i^*(\boldsymbol{\mu})} \sum_{k=1}^{K} w_s^k d(\mu_s^{+k}, \lambda^k) - B_t \\ &\geq \sum_{s\leq t', i_s = i^*(\boldsymbol{\mu})} \inf_{\boldsymbol{\lambda}\in\neg i_t} \sum_{k=1}^{K} w_s^k d(\mu_s^{+k}, \lambda^k) - B_t \\ &= \sum_{s\leq t', i_s = i^*(\boldsymbol{\mu})} D_{\boldsymbol{\mu}_s^+} - B_t \ .\end{aligned}$$

For $s \geq t^{1/(1+b)}$, $\boldsymbol{\mu} \in \mathcal{C}_s$. Then by definition of $\boldsymbol{\mu}_s^+$, $D_{\boldsymbol{\mu}_s^+} \geq D_{\boldsymbol{\mu}}$.

$$\begin{aligned}\beta(t,\delta) &\geq \sum_{t^{1/(1+b)}\leq s\leq t', i_s = i^*(\boldsymbol{\mu})} D_{\boldsymbol{\mu}} - B_t \\ &= (n_{i^*(\boldsymbol{\mu})}(t) - t^{1/(1+b)}) D_{\boldsymbol{\mu}} - B_t \ .\end{aligned}$$

Putting things together. Let $h(t) = 3L\sqrt{2\sigma^2 f(t)}(K^2 + 2\sqrt{2Kt}) + f(t)(K^2 + 2K\log(t/K)) + KD + L\sqrt{2\sigma^2 Ktf(t)}$. When the concentration event $\mathcal{E}_t$ holds, if $t < \tau_\delta$ then

$$\frac{\beta(t,\delta) + h(t)}{D_{\boldsymbol{\mu}}} \geq t - 2t^{1/(1+b)} \ .$$

Let $T_0(\delta)$ be the maximal t verifying this inequality. Then the expected sample complexity is lower than $T_0(\delta) + \frac{2eK}{a^2}$. Note that $f(t)$ depends on a and b.

E.2 Follow The Perturbed Leader

In this section, we suppose that the rewards are bounded and we define $C > 0$ such that for all times s and $k \in [K]$, $|X_s^k - \hat{\mu}_{s-1}^k| \leq C$.

At stage t, the loss of a vector $\boldsymbol{\lambda}$ is $\ell_t(\boldsymbol{\lambda}) = d(\hat{\mu}_{t-1}^{k_t}, \lambda^{k_t})$. The only unknown quantity for the $\boldsymbol{\lambda}$-player is k_t. We will use the form of that loss in the way we perturb the leader. For $\boldsymbol{\sigma} \in \mathbb{R}_+^K$ and $\boldsymbol{\xi} \in \Theta^K$ we define

$$\boldsymbol{\lambda}_t(\boldsymbol{\sigma}, \boldsymbol{\xi}) = \operatorname*{argmin}_{\boldsymbol{\lambda}} \sum_{s=1}^{t-1} \ell_s(\boldsymbol{\lambda}) + \sum_{k=1}^{K} \sigma^k d(\xi^k, \lambda^k) \ .$$

We study the expected regret of an algorithm playing $\boldsymbol{\lambda}_t(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})$ with exponentially distributed perturbations $\boldsymbol{\sigma}_t$. Let $\boldsymbol{q}_t$ be the distribution of $\boldsymbol{\lambda}_t(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})$. Let $\tilde{\mu}_{t-1}^k = \frac{1}{N_{t-1}^k} \sum_{s=1}^{t-1} \hat{\mu}_{s-1}^k \mathbb{I}\{k_s = k\}$. We show in the following lemma that the point $\boldsymbol{\lambda}_t(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})$ can be computed by the best-response oracle, as

$$\operatorname*{argmin}_{\boldsymbol{\lambda}\in\Lambda} \sum_{k=1}^{K} (N_{t-1}^k + \sigma_t^k) d\left(\frac{N_{t-1}^k}{N_{t-1}^k + \sigma_t^k} \tilde{\mu}_{t-1}^k + \frac{\sigma_t^k}{N_{t-1}^k + \sigma_t^k} \hat{\mu}_{t-1}^k, \lambda^k \right) \ .$$

Lemma 19. *Let $(\boldsymbol{\mu}_s)_{s\in[t]}$ be t points in Θ^K. Then*

$$\operatorname*{argmin}_{\boldsymbol{\lambda}\in\Lambda} \sum_{s=1}^{t} d(\mu_s^{k_s}, \lambda^{k_s}) = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\Lambda} \sum_{k=1}^{K} N_t^k d(\frac{\sum_{s=1}^{t} \mu_s^k \mathbb{I}\{k_s = k\}}{N_t^k}, \lambda^k) \ .$$

Proof. This is an extension of the following property:

$$\operatorname*{argmin}_{\lambda} d(\mu_1, \lambda) + d(\mu_2, \lambda) = \operatorname*{argmin}_{\lambda} d(\frac{\mu_1 + \mu_2}{2}, \lambda) \ .$$

Indeed we can observe that fact by developing the divergence in terms of ϕ and observing that the terms depending on λ are the same up to a multiplicative factor.

$$d(\mu_1,\lambda)+d(\mu_2,\lambda)=\phi(\mu_1)+\phi(\mu_2)-2\left(\phi(\lambda)+\phi'(\lambda)(\frac{\mu_1+\mu_2}{2}-\lambda)\right),$$
$$d(\frac{\mu_1+\mu_2}{2},\lambda)=\phi(\frac{\mu_1+\mu_2}{2})-\left(\phi(\lambda)+\phi'(\lambda)(\frac{\mu_1+\mu_2}{2}-\lambda)\right).$$

□

Theorem 3. *The expected regret of the FTPL procedure introduced above against an oblivious adversary, with perturbations $\sigma_t^k=\eta_t^k\sigma_1^k$ with $\eta_t^k=\sqrt{N_{t-1}^k}$ and σ_1^k exponential with parameter η is*

$$\sum_{s=1}^t \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s}\,\ell_s(\boldsymbol{\lambda})-\inf_{\boldsymbol{\lambda}\in\Lambda}\sum_{s=1}^t\ell_s(\boldsymbol{\lambda})\le R_t=\sqrt{Kt}\left(\frac{D+2CL}{\eta}+2D\eta\right).$$

The expected regret of the FTPL algorithm in which the noises are independent in time and σ_t^k is exponential with parameter η/η_t^k is the same.

For non-oblivious adversaries, the quantity $\sum_{s=1}^t \mathbb{E}_{\boldsymbol{\lambda}\sim\boldsymbol{q}_s}\,\ell_s(\boldsymbol{\lambda})-\inf_{\boldsymbol{\lambda}\in\Lambda}\sum_{s=1}^t\ell_s(\boldsymbol{\lambda})$ is also bounded by the same R_t, according to Lemma 4.1 of [4].

Proof of Theorem 3. Regret decomposition: for any u, the regret compared to u is

$$\sum_{s=1}^t\ell_s(\boldsymbol{\lambda}_s(\boldsymbol{\sigma}_s,\hat{\boldsymbol{\mu}}_{s-1}))-\sum_{s=1}^t\ell_s(u)\le\sum_{s=1}^t\ell_s(\boldsymbol{\lambda}_s(\boldsymbol{\sigma}_s,\hat{\boldsymbol{\mu}}_{s-1}))-\ell_s(\boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s,\hat{\boldsymbol{\mu}}_{s-1}))$$
$$+\sum_{s=1}^t\ell_s(\boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s,\hat{\boldsymbol{\mu}}_{s-1}))-\sum_{s=1}^t\ell_s(u)$$

Second term of the regret. We are analysing here the regret of a noisy Be-The-Leader. We first show by induction that

$$\sum_{s=1}^t\ell_s(\boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s,\hat{\boldsymbol{\mu}}_{s-1}))-\sum_{s=1}^t\ell_s(u)$$
$$\le\boldsymbol{\sigma}_t^\intercal d(\hat{\boldsymbol{\mu}}_{t-1},u)+\sum_{s=1}^t\boldsymbol{\sigma}_s^\intercal d(\hat{\boldsymbol{\mu}}_{s-1},\boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s,\hat{\boldsymbol{\mu}}_{s-1}))-\boldsymbol{\sigma}_{s-1}^\intercal d(\hat{\boldsymbol{\mu}}_{s-2},\lambda_{s+1}(\boldsymbol{\sigma}_s,\hat{\boldsymbol{\mu}}_{s-1}))$$

Initialization: for all $u\in\Lambda$,

$$\ell_1(\boldsymbol{\lambda}_2(\boldsymbol{\sigma}_1,\hat{\boldsymbol{\mu}}_0))=\ell_1(\boldsymbol{\lambda}_2(\boldsymbol{\sigma}_1,\hat{\boldsymbol{\mu}}_0))+\boldsymbol{\sigma}_1^\intercal d(\hat{\boldsymbol{\mu}}_0,\boldsymbol{\lambda}_2(\boldsymbol{\sigma}_1,\hat{\boldsymbol{\mu}}_0))-\boldsymbol{\sigma}_1^\intercal d(\hat{\boldsymbol{\mu}}_0,\boldsymbol{\lambda}_2(\boldsymbol{\sigma}_1,\hat{\boldsymbol{\mu}}_0))$$
$$\le\ell_1(u)+\boldsymbol{\sigma}_1^\intercal d(\hat{\boldsymbol{\mu}}_0,u)-\boldsymbol{\sigma}_1^\intercal d(\hat{\boldsymbol{\mu}}_0,\boldsymbol{\lambda}_2(\boldsymbol{\sigma}_1,\hat{\boldsymbol{\mu}}_0)).$$

Let $A_1(u)=\boldsymbol{\sigma}_1^\intercal d(\hat{\boldsymbol{\mu}}_0,u)-\boldsymbol{\sigma}_1^\intercal d(\hat{\boldsymbol{\mu}}_0,\boldsymbol{\lambda}_2(\boldsymbol{\sigma}_1,\hat{\boldsymbol{\mu}}_0))$. Then for all $u\in\Lambda$, $\ell_1(\boldsymbol{\lambda}_2(\boldsymbol{\sigma}_1,\hat{\boldsymbol{\mu}}_0))-\ell_1(u)\le A_1(u)$.

Induction: suppose that for all $u\in\Lambda$, $\sum_{s=1}^{t-1}\ell_s(\boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s,\hat{\boldsymbol{\mu}}_{s-1}))\le\sum_{s=1}^{t-1}\ell_s(u)+A_{t-1}(u)$, with

$$A_{t-1}(u)=\boldsymbol{\sigma}_{t-1}^\intercal d(\hat{\boldsymbol{\mu}}_{t-2},u)+\sum_{s=1}^{t-1}\boldsymbol{\sigma}_{s-1}^\intercal d(\hat{\boldsymbol{\mu}}_{s-2},\boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s,\hat{\boldsymbol{\mu}}_{s-1}))-\boldsymbol{\sigma}_s^\intercal d(\hat{\boldsymbol{\mu}}_{s-1},\boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s,\hat{\boldsymbol{\mu}}_{s-1}))$$

where $\boldsymbol{\sigma}_0 = 0$. Apply it to $u = \boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})$.

$$\begin{aligned}
\sum_{s=1}^{t} \ell_s(\boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s, \hat{\boldsymbol{\mu}}_{s-1})) &\le \sum_{s=1}^{t-1} \ell_s(\boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) + \ell_t(\boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) \\
&\quad + A_{t-1}(\boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) \\
&= \sum_{s=1}^{t} \ell_s(\boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) + \boldsymbol{\sigma}_t^\intercal d(\hat{\boldsymbol{\mu}}_{t-1}, \boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) \\
&\quad - \boldsymbol{\sigma}_t^\intercal d(\hat{\boldsymbol{\mu}}_{t-1}, \boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) + A_{t-1}(\boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) \\
&\le \sum_{s=1}^{t} \ell_s(u) + \boldsymbol{\sigma}_t^\intercal d(\hat{\boldsymbol{\mu}}_{t-1}, u) \\
&\quad - \boldsymbol{\sigma}_t^\intercal d(\hat{\boldsymbol{\mu}}_{t-1}, \boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) + A_{t-1}(\boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1}))\,.
\end{aligned}$$

We obtain

$$\begin{aligned}
A_t(u) - \boldsymbol{\sigma}_t^\intercal d(\hat{\boldsymbol{\mu}}_{t-1}, u) &= A_{t-1}(\boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) - \boldsymbol{\sigma}_t^\intercal d(\hat{\boldsymbol{\mu}}_{t-1}, \boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) \\
&= \sum_{s=1}^{t} \boldsymbol{\sigma}_{s-1}^\intercal d(\hat{\boldsymbol{\mu}}_{s-2}, \boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s, \hat{\boldsymbol{\mu}}_{s-1})) - \boldsymbol{\sigma}_s^\intercal d(\hat{\boldsymbol{\mu}}_{s-1}, \boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s, \hat{\boldsymbol{\mu}}_{s-1}))\,.
\end{aligned}$$

End of the induction proof.

We now bound $A_t(u)$. First we write

$$\begin{aligned}
A_t(u) - \boldsymbol{\sigma}_t^\intercal d(\hat{\boldsymbol{\mu}}_{t-1}, u) &= \sum_{s=1}^{t} \boldsymbol{\sigma}_{s-1}^\intercal d(\hat{\boldsymbol{\mu}}_{s-2}, \boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s, \hat{\boldsymbol{\mu}}_{s-1})) - \boldsymbol{\sigma}_s^\intercal d(\hat{\boldsymbol{\mu}}_{s-1}, \boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s, \hat{\boldsymbol{\mu}}_{s-1})) \\
&= \sum_{s=1}^{t} \boldsymbol{\sigma}_s^\intercal [d(\hat{\boldsymbol{\mu}}_{s-2}, \boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s, \hat{\boldsymbol{\mu}}_{s-1})) - d(\hat{\boldsymbol{\mu}}_{s-1}, \boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s, \hat{\boldsymbol{\mu}}_{s-1}))] \\
&\quad + \sum_{s=1}^{t} (\boldsymbol{\sigma}_{s-1} - \boldsymbol{\sigma}_s)^\intercal d(\hat{\boldsymbol{\mu}}_{s-2}, \boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s, \hat{\boldsymbol{\mu}}_{s-1}))\,.
\end{aligned}$$

We now bound separately the two sums. The first one uses the Lipschitz-continuity of d and the fact that successive $\hat{\boldsymbol{\mu}}_t$ are not far from each other.

$$\begin{aligned}
&\mathbb{E} \sum_{s=1}^{t} \boldsymbol{\sigma}_s^\intercal [d(\hat{\boldsymbol{\mu}}_{s-2}, \boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s, \hat{\boldsymbol{\mu}}_{s-1})) - d(\hat{\boldsymbol{\mu}}_{s-1}, \boldsymbol{\lambda}_{s+1}(\boldsymbol{\sigma}_s, \hat{\boldsymbol{\mu}}_{s-1}))] \\
&\quad = \mathbb{E} \sum_{s=1}^{t} \boldsymbol{\sigma}_s^{k_{s-1}} [d(\hat{\mu}_{s-2}^{k_{s-1}}, \lambda_{s+1}^{k_{s-1}}(\boldsymbol{\sigma}_s, \hat{\boldsymbol{\mu}}_{s-1})) - d(\hat{\mu}_{s-1}^{k_{s-1}}, \lambda_{s+1}^{k_{s-1}}(\boldsymbol{\sigma}_s, \hat{\boldsymbol{\mu}}_{s-1}))] \\
&\quad \le \sum_{s=1}^{t} \mathbb{E}[\sigma_s^{k_{s-1}}] L |\hat{\mu}_{s-1}^{k_{s-1}} - \hat{\mu}_{s-2}^{k_{s-1}}| \le CL \sum_{s=1}^{t} \mathbb{E}[\sigma_s^{k_{s-1}}] \frac{1}{N_{s-1}^{k_{s-1}}} \le \frac{CL}{\eta} \sum_{s=1}^{t} \frac{\eta_s^{k_{s-1}}}{N_{s-1}^{k_{s-1}}}\,.
\end{aligned}$$

For η_t^k non-decreasing in t, $\boldsymbol{\sigma}_{s-1} - \boldsymbol{\sigma}_s$ has non-positive coordinates and the second sum is negative.

We obtain

$$\mathbb{E}\, A_t(u) \le \mathbb{E}\, \boldsymbol{\sigma}_t^\intercal d(\hat{\boldsymbol{\mu}}_{t-1}, u) + \frac{CL}{\eta} \sum_{s=1}^{t} \frac{\eta_s^{k_{s-1}}}{N_{s-1}^{k_{s-1}}} \le \frac{D\|\eta_t\|_1}{\eta} + \frac{CL}{\eta} \sum_{s=1}^{t} \frac{\eta_s^{k_{s-1}}}{N_{s-1}^{k_{s-1}}}\,.$$

First term of the regret. Remark that $\boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}, \hat{\boldsymbol{\mu}}_{t-1}) = \boldsymbol{\lambda}_t(\boldsymbol{\sigma} + e_{k_t}, \hat{\boldsymbol{\mu}}_{t-1})$. Let f be the density of the distribution of σ_t. In expectation, the first term of the regret is

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}_t}[\ell_t(\boldsymbol{\lambda}_t(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) - \ell_t(\boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1}))] \\
&\quad = \int_{\boldsymbol{\sigma}_t} [\ell_t(\boldsymbol{\lambda}_t(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) - \ell_t(\boldsymbol{\lambda}_t(\boldsymbol{\sigma}_t + e_{k_t}, \hat{\boldsymbol{\mu}}_{t-1}))] f(\boldsymbol{\sigma}_t) d\boldsymbol{\sigma}_t \\
&\quad = \int_{\boldsymbol{\sigma}_t} \ell_t(\boldsymbol{\lambda}_t(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1}))(f(\boldsymbol{\sigma}_t) - f(\boldsymbol{\sigma}_t - e_{k_t})) d\boldsymbol{\sigma}_t
\end{aligned}$$

By positivity of ℓ_t (since it is a divergence),

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}_t}[\ell_t(\boldsymbol{\lambda}_t(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) - \ell_t(\boldsymbol{\lambda}_{t+1}(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1}))] \\
&\quad \leq \int_{\boldsymbol{\sigma}_t} \ell_t(\boldsymbol{\lambda}_t(\boldsymbol{\sigma}_t, \hat{\boldsymbol{\mu}}_{t-1})) \mathbb{I}\{f(\boldsymbol{\sigma}_t) - f(\boldsymbol{\sigma}_t - e_{k_t}) > 0\}(f(\boldsymbol{\sigma}_t) - f(\boldsymbol{\sigma}_t - e_{k_t})) d\boldsymbol{\sigma}_t \\
&\quad \leq D \int_{\boldsymbol{\sigma}_t} \mathbb{I}\{f(\boldsymbol{\sigma}_t) - f(\boldsymbol{\sigma}_t - e_{k_t}) > 0\}(f(\boldsymbol{\sigma}_t) - f(\boldsymbol{\sigma}_t - e_{k_t})) d\boldsymbol{\sigma}_t \\
&\quad \leq D \int_{\boldsymbol{\sigma}_t} \mathbb{I}\{f(\boldsymbol{\sigma}_t) - f(\boldsymbol{\sigma}_t - e_{k_t}) > 0\} f(\boldsymbol{\sigma}_t) d\boldsymbol{\sigma}_t \\
&\quad = D \int_{\sigma_t^{k_t} \leq 1} f(\boldsymbol{\sigma}_t) d\boldsymbol{\sigma}_t \\
&\quad = D(1 - e^{-\eta/\eta_t^{k_t}}) \\
&\quad \leq D\eta/\eta_t^{k_t} \, .
\end{aligned}$$

Putting things together. Choose $\eta_t^k = \sqrt{N_{t-1}^k}$.

$$\mathbb{E}\, R_t \leq D\frac{\|\eta_t\|_1}{\eta} + \frac{CL}{\eta} \sum_{s=1}^{t} \frac{\eta_s^{k_{s-1}}}{N_{s-1}^{k_{s-1}}} + D\eta \sum_{s=1}^{t} \frac{1}{\eta_s^{k_s}} \leq \sqrt{Kt} \left(\frac{D + 2CL}{\eta} + 2D\eta \right) \, .$$

□

Approximation of $\boldsymbol{q}_t$ by an empirical distribution. We want the k-player to use optimistic best-response to $\boldsymbol{q}_t$. This requires the computation of

$$\operatorname*{argmax}_{k \in [K]} U_t^k \qquad \text{with } U_t^k = \max_{\xi \in \{a_t^k, b_t^k\}} \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_t}\, d(\xi, \lambda^k) \, .$$

for some values a_t^k, b_t^k.

Since we cannot compute an expectation under $\boldsymbol{q}_t$ exactly, we compute instead the expectation under an empirical distribution based on t samples $\boldsymbol{\lambda}_t^{(1)}, \ldots, \boldsymbol{\lambda}_t^{(t)}$ of $\boldsymbol{q}_t$. For all ξ, $d(\xi, \lambda^k)$ is bounded by D. Hence, by Hoeffding's inequality,

$$\mathbb{P}\left\{ \frac{1}{t} \sum_{j=1}^{t} d(\xi, \lambda_t^{(j)k}) - \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_t}\, d(\xi, \lambda^k) \geq \sqrt{\frac{3D^2 \log(t)}{2t}} \right\} \leq \frac{1}{t^3} \, .$$

In the concentration analysis of the algorithm, we replace $\mathcal{E}_t$ by $\mathcal{E}_t \cap \mathcal{E}'_t$ with

$$\mathcal{E}'_t = \left\{ \forall k \in [K], \forall s \leq t, \forall \xi \in \{a_s^k, b_s^k\}\ \frac{1}{s} \sum_{j=1}^{s} d(\xi, \lambda_s^{(j)k}) - \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_s}\, d(\xi, \lambda^k) \leq D\sqrt{\frac{3\log(t)}{2t}} \right\}$$

It verifies $\sum_{t=1}^{+\infty} \mathbb{P}(\mathcal{E}_t'^c) \leq 2K \sum_{t=1}^{+\infty} 1/t^2 \leq K\pi^2/3$.

Under the event $\mathcal{E}'_t$,

$$\sum_{s=1}^{t} \frac{1}{s} \sum_{j=1}^{s} d(\xi, \lambda_s^{(j)k}) - \sum_{s=1}^{t} \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_s}\, d(\xi, \lambda^k) \leq D\sqrt{\frac{3}{2} t \log(t)} \, .$$

We obtain that the procedure based on these empirical distributions has $\mathcal{O}(\sqrt{t \log t})$ regret.

F On the statistical assumptions

F.1 The sub-Gaussian assumption

The natural coordinate-wise concentration events for exponential families have the form $N_t^k d(\hat{\mu}_t^k, \mu^k) \leq c$ for some constant $c > 0$. In our proofs, we need then to relate $d(\hat{\mu}_t^k, \lambda^k)$ and $d(\mu^k, \lambda^k)$ for a given λ^k under such a concentration constraint. However, we now show that for some convex function ϕ (such that d is the associated Bregman divergence), these two quantities could be very far apart even under the constraint $d(\hat{\mu}_t^k, \mu^k) = 0$.

If $d(\hat{\mu}_t^k, \mu^k) = 0$, we have the equalities

$$\begin{aligned} d(\hat{\mu}_t^k, \lambda^k) - d(\mu^k, \lambda^k) &= d(\hat{\mu}_t^k, \mu^k) + (\hat{\mu}_t^k - \mu^k)(\phi'(\mu^k) - \phi'(\lambda^k)) \\ &= (\hat{\mu}_t^k - \mu^k)(\phi'(\mu^k) - \phi'(\lambda^k)) \,. \end{aligned}$$

Let $\phi : \mathbb{R} \to \mathbb{R}$ be defined by $\phi(x) = \max\{0, x\}$. Let $\lambda^k = 1$, $\mu^k = -1$ and $\hat{\mu}_t^k < -1$. Then

$$\begin{aligned} d(\hat{\mu}_t^k, \mu^k) &= 0 \,, \\ d(\hat{\mu}_t^k, \lambda^k) - d(\mu^k, \lambda^k) &= |\hat{\mu}_t^k - \mu^k| \,. \end{aligned}$$

In that example, the constraint on $d(\hat{\mu}_t^k, \mu^k)$ is not sufficient to bound $d(\hat{\mu}_t^k, \lambda^k) - d(\mu^k, \lambda^k)$.

The example exploits the piecewise linearity of ϕ. Such a function ϕ cannot arise from an exponential family. Indeed, for an exponential family ϕ is the convex conjugate of a cumulant generating function. In particular, ϕ is strictly convex. But it could still have very low curvature (for example for an exponential distribution with high mean). The sub-Gaussian assumption ensures that ϕ is strongly convex.

Our work and previous parametric pure exploration papers treat d as a general Bregman divergence. The present example shows that either we need to also use more specific properties of d due to the fact that it is a Kullback-Leibler divergence, or we need to impose additional assumptions like sub-Gaussianity.

F.2 The upper bound assumption

A first way to relax the assumption that $\mathcal{M} \subseteq [\mu_{\min}, \mu_{\max}]^K$ is to remark that we do not need to bound $d(\mu, \lambda)$ for any μ and λ.

For $\boldsymbol{\mu} \in \mathcal{M}$ and $\boldsymbol{w} \in \triangle_K$, let $\boldsymbol{\lambda}(\boldsymbol{\mu}, \boldsymbol{w}) = \operatorname{argmin}_{\boldsymbol{\lambda} \in \neg i} \sum_{k=1}^K w^k d(\mu^k, \lambda^k)$. Our proofs are valid for example under the following assumption.

Assumption 4. There exists $D > 0$ and $L > 0$ such that for all $\boldsymbol{w} \in \triangle_K$, for all $\boldsymbol{\mu} \in \mathcal{M}$, $\|d(\boldsymbol{\mu}, \boldsymbol{\lambda}(\boldsymbol{\mu}, \boldsymbol{w}))\|_\infty \leq D$ and $\|\phi'(\boldsymbol{\mu}) - \phi'(\boldsymbol{\lambda}(\boldsymbol{\mu}, \boldsymbol{w}))\|_\infty \leq L$.

We could also use the concentration events to replace it with weaker hypotheses. Under event $\mathcal{E}_t$ and with Assumption 1, for all $s \leq t$, $\|d(\hat{\boldsymbol{\mu}}_s, \boldsymbol{\mu})\|_\infty \leq f(t)$ and $\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|_\infty \leq \sqrt{2\sigma^2 f(t)}$. That is, we get from concentration only, without assumptions, that $\hat{\boldsymbol{\mu}}_t$ is in a bounded set around $\boldsymbol{\mu}$. We can then quantify L and D on that set.

Let $L_{\boldsymbol{\mu}} = \sup_{\boldsymbol{w} \in \triangle_k} \max_k |\phi'(\mu^k) - \phi'(\lambda(\boldsymbol{\mu}, \boldsymbol{w})^k)|$.

Assumption 5. For all $\boldsymbol{\mu} \in \mathcal{M}$, $L_{\boldsymbol{\mu}}$ is finite.

This is true for BAI, where $L_{\boldsymbol{\mu}} \leq \phi'(\max_k \mu^k) - \phi'(\min_k \mu^k)$.

Assumption 6. There exists $M > 0$ such that $\boldsymbol{\mu} \mapsto L_{\boldsymbol{\mu}}$ is M-Lipschitz for the ℓ^∞ norm.

This is true for BAI on sets on which ϕ' is Lipschitz. For example, it is true on $\mathbb{R}$ for Gaussian arm distributions, but is still only true in intervals of the form $[\varepsilon, 1 - \varepsilon]$ for Bernoulli distributions.

Then for any $\boldsymbol{\mu}, \boldsymbol{\xi}$ and $\boldsymbol{\lambda}_{\boldsymbol{\mu}}$ minimal point for $\boldsymbol{\mu}$, for any coordinate $k \in [K]$ (omitted in the computations),

$$\begin{aligned} d(\mu, \lambda_\mu) &= d(\xi, \lambda_\mu) + (\mu - \xi)(\phi'(\mu) - \phi'(\lambda_\mu)) - d(\xi, \mu) \\ &\geq d(\xi, \lambda_\mu) - |\mu - \xi| L_\mu - d(\xi, \mu) \\ &\geq d(\xi, \lambda_\mu) - |\mu - \xi| L_\xi - M(\mu - \xi)^2 - d(\xi, \mu) \\ &\geq d(\xi, \lambda_\mu) - L_\xi \sqrt{2\sigma^2 \min\{d(\mu, \xi), d(\xi, \mu)\}} - 2\sigma^2 M \min\{d(\mu, \xi), d(\xi, \mu)\} - d(\xi, \mu) \end{aligned}$$

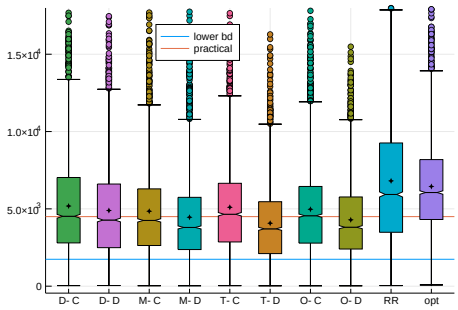
Examples for the quantities used in the proofs:

$$\begin{aligned} d(\mu, \lambda_\mu) &\geq d(\hat{\mu}_{s-1}, \lambda_\mu) - L_\mu \sqrt{2\sigma^2 d(\hat{\mu}_{s-1}, \mu)} - d(\hat{\mu}_{s-1}, \mu) \\ d(\hat{\mu}_t, \lambda_{\hat{\mu}_t}) &\geq d(\mu, \lambda_{\hat{\mu}_t}) - L_\mu \sqrt{2\sigma^2 d(\hat{\mu}_t, \mu)} - 2\sigma^2 M d(\hat{\mu}_t, \mu) - d(\mu, \hat{\mu}_t) \end{aligned}$$

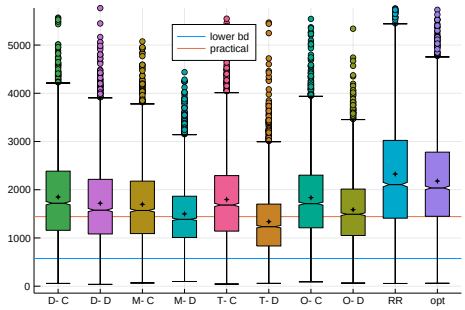
The proofs must then be adapted to account for the additional terms in these inequalities.

G Numerical Experiments

G.1 Best Arm

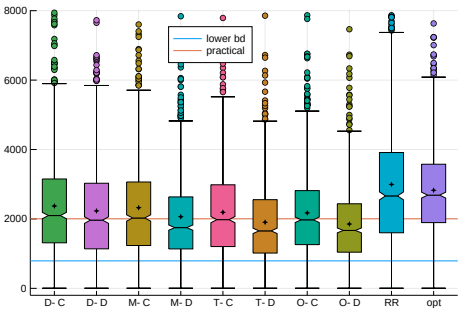


(a) Bernoulli bandit $\boldsymbol{\mu} = (0.5, 0.45, 0.43, 0.4)$, $\boldsymbol{w}^* = (0.42, 0.39, 0.14, 0.06)$

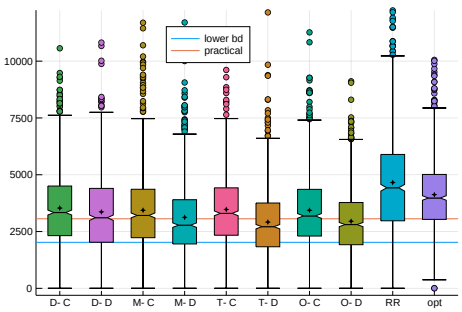


(b) Bernoulli bandit $\boldsymbol{\mu} = (0.3, 0.21, 0.2, 0.19, 0.18)$, $\boldsymbol{w}^* = (0.34, 0.25, 0.18, 0.13, 0.10)$

Figure 2: Best Arm experiments from [13]. In both cases $\delta = 0.1$. Plots show 3000 runs.



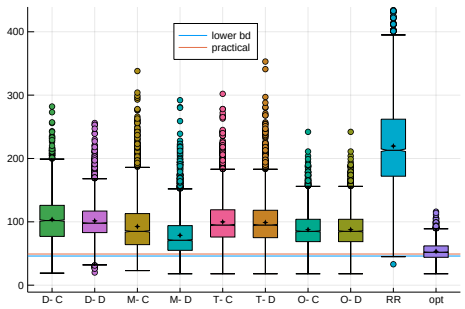
(a) $\delta = 0.1$



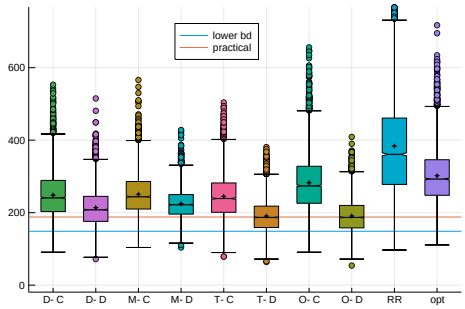
(b) $\delta = 0.01$

Figure 3: Best Arm experiment from [24]. Gaussian bandit $\boldsymbol{\mu} = (1., 0.85, 0.8, 0.7)$, $\boldsymbol{w}^* = (0.41, 0.38, 0.15, 0.06)$. Plots show 3000 runs.

G.2 Minimum Threshold

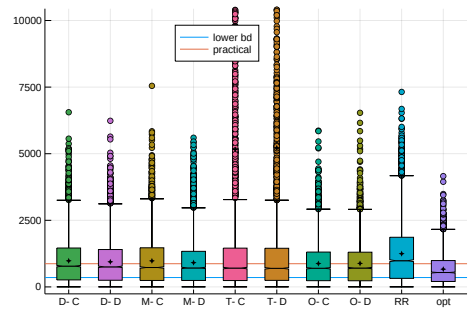


(a) Gaussian bandit $\boldsymbol{\mu} = (-1, \ldots, 1)$ with $K = 10$ arms and $\delta = e^{-23}$, $\boldsymbol{w}^* = \boldsymbol{e}_1$

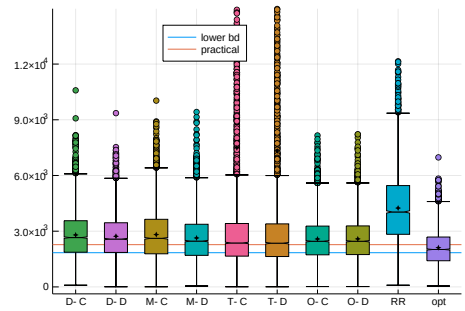


(b) Gaussian bandit $\boldsymbol{\mu} = (0.5, \ldots, 1)$ with $K = 5$ arms and $\delta = e^{-7}$, $\boldsymbol{w}^* = (0.38, 0.24, 0.17, 0.12, 0.09)$

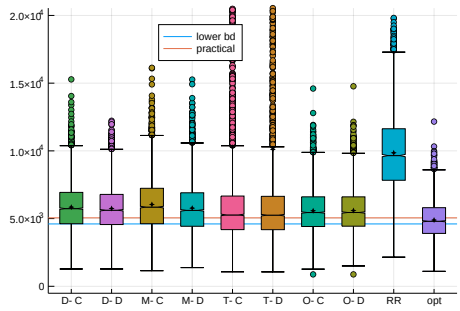
Figure 4: Minimum Threshold experiments from [20] with threshold $\gamma = 0$. Plots show 5000 runs.



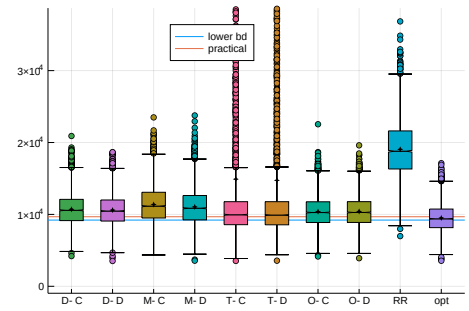
(a) $\delta = 0.1$



(b) $\delta = 0.0001$



(c) $\delta = 10^{-10}$



(d) $\delta = 10^{-20}$

Figure 5: Minimum Threshold experiment (new): Gaussian bandit $\boldsymbol{\mu} = (0.5, 0.6)$ with threshold $\gamma = 0.6$, $\boldsymbol{w}^* = \boldsymbol{e}_1$. Note the excessive sample complexity of Track-and-Stop (T-C and T-D). Plots show 5000 runs.

The reason for the bad performance of Track-and-Stop in Figure 5 is that with small but non-negligible probability the algorithm finds $\hat{\mu}_t^1 \gg \gamma$ estimated too high at some early t. In this situation $\boldsymbol{w}^*(\boldsymbol{\mu}_t)$ will be $\boldsymbol{e}_2$ (exactly if $\hat{\mu}_t^2 \leq \gamma$, approximately if $\hat{\mu}_t^2 > \gamma$), and constantly pulling arm 2 will not correct the estimate of arm 1. **T** relies on forced exploration to correct the estimate.