Efficient Pure Exploration for Combinatorial Bandits with Semi-Bandit Feedback

Marc Jourdan MARC.JOURDAN@INF.ETHZ.CH
Mojmír Mutný MOJMIR.MUTNY@INF.ETHZ.CH
Johannes Kirschner JKIRSCHNER@INF.ETHZ.CH
Andreas Krause KRAUSEA@ETHZ.CH
Department of Computer Science
ETH Zürich, Switzerland

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

Combinatorial bandits with semi-bandit feedback generalize multi-armed bandits, where the agent chooses sets of arms and observes a noisy reward for each arm contained in the chosen set. The action set satisfies a given structure such as forming a base of a matroid or a path in a graph. We focus on the pure-exploration problem of identifying the best arm with fixed confidence, as well as a more general setting, where the structure of the answer set differs from the one of the action set. Using the recently popularized game framework, we interpret this problem as a sequential zero-sum game and develop a CombGame meta-algorithm whose instances are asymptotically optimal algorithms with finite time guarantees. In addition to comparing two families of learners to instantiate our meta-algorithm, the main contribution of our work is a specific oracle efficient instance for best-arm identification with combinatorial actions. Based on a projection-free online learning algorithm for convex polytopes, it is the first computationally efficient algorithm which is asymptotically optimal and has competitive empirical performance.

Keywords: Combinatorial Bandits, Pure Exploration, Best-Arm Identification

1. Introduction

The multi-armed bandit (MAB) setting is an extensively studied problem in statistics and machine learning (Robbins, 1952; Lattimore and Szepesvári, 2020). The environment consists of a set of arms, each characterized by an unknown reward distribution. An agent interacts with it by playing the arms sequentially in order to identify the arm with the highest expected reward.

Combinatorial bandits (Cesa-Bianchi and Lugosi, 2012; Chen et al., 2013) are a natural extension of the standard framework. The agent chooses actions (or super arms) which are defined by *sets of arms* satisfying certain constraints. The most studied families of actions stem from matroid theory (Kveton et al., 2014; Perrault et al., 2019). Matroids encompass the batch setting where actions are sets of size k (Jun et al., 2016; Kuroki et al., 2020; Rejwan and Mansour, 2020) and graph-based structures where arms are edges and actions are spanning trees or matching trees. This formulation can model various application-specific structures such as paths taken in routing problems (Talebi et al., 2018). Another example is protein design, where experimental constraints force the agent to evaluate specific sequences of proteins. Instead of inducing a single mutation, a range of localized mutations are performed at once. The main challenge in combinatorial bandits is to cope with the exponential size of the action set. This renders standard approaches for the bandit setting computationally inefficient and also – without further assumptions like linearity – statistically inefficient.

To overcome this hurdle, existing approaches assume the reward is linear over the set of arms, and leverage an efficient oracle which solves a linear optimization problem over the combinatorial set of feasible actions. Efficient combinatorial oracles are known for many constraint families such as matroid polytopes, intersections of matroids and path polytopes. Combinatorial bandit strategies vary depending on the received feedback. We consider *semi-bandit* feedback where the agent *observes a reward for each selected arm*. Moreover, we assume that the reward for each arm is *independent*.

We focus on the *pure-exploration* framework, in which the agent aims at maximizing the information gathered to answer a given query and disregards the accumulated cost. Two major theoretical frameworks exist (Gabillon et al., 2012, 2016; Jun et al., 2016; Kaufmann et al., 2016): the *fixed-budget* setting and the *fixed-confidence* setting. In the fixed-budget setting, the goal is to minimize the probability of misidentifying the correct answer given a fixed number of pulls. We consider the fixed-confidence setting where the objective is to minimize the number of pulls necessary to identify the correct answer with a given confidence $1 - \delta$. The most studied problems are best-arm identification (BAI) (Karnin et al., 2013; Jamieson et al., 2014; Zaki et al., 2020) and top-k identification (Gabillon et al., 2011; Kalyanakrishnan et al., 2012; Bubeck et al., 2013; Scarlett et al., 2019).

In the spirit of transductive bandits (Fiez et al., 2019) we consider a more general setting where answers are sets of arms. The set of actions and the set of answers can be different. For example, in a routing or transportation network the objective might be to identify a weak link in order to fix it. The agent evaluates a path (action) in the network and gets access to time-stamped data for each link (answer) of a played path. Similarly, in protein design, researchers often generate many mutant proteins in one experiment, but the goal is to identify the best mutant.

We adopt the recently popularized game approach of Degenne et al. (2019). The idea is to consider a sequential zero-sum game between two players. This game approximates the optimal allocation given by the lower bound (Kaufmann et al., 2016). The objective of our work is to design asymptotically optimal algorithms with finite-time guarantees. They should have computationally efficient implementations as long as the offline combinatorial problem can be solved efficiently.

Contributions (1) We use the game framework for pure exploration to study combinatorial bandits with semi-bandit feedback. The action and answer sets are arbitrary and the feedback is independent across arms. Despite its increasing popularity, the game framework has not yet been used in combinatorial bandits or in the transductive setting. (2) We develop a pure-exploration CombGame meta-algorithm whose instances are asymptotically optimal algorithms with finite time guarantees. The family of algorithms directly adapts the pure-exploration meta-algorithm of Degenne et al. (2019) to the combinatorial nature of the problem allowing for tractable implementation of the game framework. (3) To overcome the limitation of prior work, we employ the projection-free algorithm over convex polyhedral sets of Garber and Hazan (2013). This approach is the first computationally efficient algorithm which is asymptotically optimal and has competitive empirical performance.

1.1. Related Work

Combinatorial bandits have been introduced by Cesa-Bianchi and Lugosi (2012) and Chen et al. (2013). The emblematic examples of combinatorial actions are the basis of a matroid (Perrault et al., 2019) and the paths in a graph (Talebi et al., 2018). Semi-bandit feedback is extensively studied (Kveton et al., 2015; Wen et al., 2015). Other works have considered the *bandit* feedback where the agent observes an aggregated reward (Combes et al., 2015). Generalizing them both, the partial linear monitoring feedback has been studied for cumulative regret minimization (Kirschner et al., 2020) and for pure exploration (Chen et al., 2020). Combinatorial bandits have also been used to denote a different setting where the agent plays arms to identify the best action (Chen et al., 2014, 2016, 2017a; Cao and Krishnamurthy, 2019). Combinatorial bandits have also been generalized to consider submodular reward functions (Hazan and Kale, 2012a; Chen et al., 2017b).

Before the game approach was introduced, Jamieson and Nowak (2014) highlighted three important types of algorithms to solve BAI. They were based on action elimination (Karnin et al., 2013), upper confidence bound (UCB) (Audibert et al., 2010) or lower UCB (Kalyanakrishnan et al., 2012). Bayesian strategies have also been proposed with Thompson sampling like algorithms (Russo, 2016; Kaufmann et al., 2018; Shang et al., 2020). Generalizing the BAI problem to the identification of the k best arms, top-k identification has been studied for an agent playing arms (Gabillon et al., 2011; Kalyanakrishnan et al., 2012; Scarlett et al., 2019) or batches of arms (Jun et al., 2016; Kuroki et al., 2020; Rejwan and Mansour, 2020). The pure-exploration framework encompasses more complex queries such as maximin (Garivier et al., 2016) or minimum threshold (Degenne et al., 2019). Some problems admit multiple correct answers (Degenne and Koolen, 2019).

In the fixed-confidence pure-exploration setting the first known lower bounds on the sample complexity involve a characteristic time whose inverse is a complexity measure (Kaufmann et al., 2016). Those setting-dependent lower bounds have motivated the search for algorithms with matching upper bound, both in finite-time (Simchowitz et al., 2017) and asymptotic regime (Garivier and Kaufmann, 2016). Unfortunately, existing algorithms often require an expensive oracle to compute the optimal allocation weights which are used for sampling, such as Track-and-Stop (Garivier and Kaufmann, 2016) or RAGE (Fiez et al., 2019). Degenne et al. (2019) introduces the game framework which interprets the optimization problem as a zero-sum game between two players. In particular, it proposes a pure-exploration meta-algorithm which uses a cheaper best-response oracle. The game framework has inspired recent algorithms for linear bandits, such as PELEG (Zaki et al., 2020) or LinGame(-C) (Degenne et al., 2020a). PELEG extends the phased-elimination algorithm of Fiez et al. (2019). The idea has also been adapted to cumulative regret in Degenne et al. (2020b).

2. Preliminaries

In this section we formally define *pure exploration for combinatorial bandits with semi-bandit feedback*, and prove a lower bound on the sample complexity. We then use the lower bound to determine sampling strategies for our algorithm.

2.1. Problem Formulation

Suppose the environment consists of d arms (or base arms). Each arm $a \in [d] := \{1, \cdots, d\}$ is associated with a probability distribution from the exponential family ν_a characterized by the unknown mean μ_a. Given known σ_a, we consider two cases, in which: (a) ν_a is σ_a^2-sub-Gaussian and (b) ν_a is Gaussian $\mathcal{N}(\mu_a, \sigma_a^2)$. An exponential family ν_a is σ_a^2-sub-Gaussian if and only if for all (μ_a, λ_a) the KL divergence satisfies $d_{\mathrm{KL}}(\mu_a, \lambda_a) \geq \frac{(\mu_a - \lambda_a)^2}{2\sigma_a^2}$. The independent joint distribution of the arms is denoted by ν and defined uniquely by $\mu := (\mu_a)_{a \in [d]} \in \mathcal{M}$. The set of possible parameters $\mathcal{M} \subset \mathbb{R}^d$ is known to the agent. Similarly to earlier work on bandits, $\mathcal{M}$ is assumed to be bounded. As proven in Appendix H, this assumption is immaterial for Gaussian distributions. The component-wise KL divergence between the true parameter μ and a different parameter $\lambda \in \mathcal{M}$ is denoted by a vector $d_{\mathrm{KL}}(\mu, \lambda) := (d_{\mathrm{KL}}(\mu_a, \lambda_a))_{a \in [d]}$.

We define the action set $\mathcal{A} \subset 2^{[d]}$ as a collection of sets of arms (a subset of the power set of arms). The agent can only play actions. In the literature, *super arms* or *multiple arms* are used to denote actions. As a special case, the action set could be the singletons (arms) $\mathcal{A} = \{\{a\}\}_{a\in[d]}$. Let $K := \max_{A\in\mathcal{A}} |A|$ be the maximum size of an action. At each round $t \geq 1$, the agent chooses an action $A_t \in \mathcal{A}$ and observes a noisy semi-bandit feedback $Y_{t,A_t} := \left(Y_{t,a}\mathbf{1}_{(a\in A_t)}\right)_{a\in[d]}$ where $\mathbf{1}_S := \left(\mathbf{1}_{(a\in S)}\right)_{a\in[d]}$ is the indicator vector for $S \subset [d]$ and $Y_t \sim \nu$ is the observation vector in $\mathbb{R}^d$.

We define the answer set $\mathcal{I} \subset 2^{[d]}$ as a collection of sets of arms, possibly different from the set of actions $\mathcal{A}$. The setting where $\mathcal{I}$ and $\mathcal{A}$ differ is also known as the transductive bandit setting. Given a parameter λ, the reward of an answer $I \in \mathcal{I}$ is the sum of the rewards of each arm $\langle \lambda, \mathbf{1}_I\rangle := \sum_{a\in[d]} \lambda_a \mathbf{1}_{(a\in I)}$. The *correct answer* is given by the function $I^* : \mathcal{M} \mapsto \mathcal{I}$ defined as $I^*(\lambda) := \operatorname{argmax}_{I\in\mathcal{I}}\langle \lambda, \mathbf{1}_I\rangle$. For simplicity, we assume that $I^*(\lambda)$ is unique for all $\lambda \in \mathcal{M}$. A more careful analysis would allow to relax this assumption to: $I^*(\mu)$ is unique for the unknown μ characterizing the bandit ν. The goal of the agent is to identify the correct answer $I^*(\mu)$ by interacting with the environment. BAI is a special case where $\mathcal{I} = \{\{a\}\}_{a\in[d]}$. Best-action identification is obtained for $\mathcal{I} = \mathcal{A}$.

We assume that the agent has access to efficient oracles[1] to solve the offline linear optimization problems $\operatorname{argmax}_{A\in\mathcal{A}}\langle \mathbf{1}_A, c\rangle$ and $\operatorname{argmax}_{I\in\mathcal{I}}\langle \mathbf{1}_I, c\rangle$ for a given linear objective $c \in \mathbb{R}^d$. This assumption is commonly made for semi-bandits (Cao and Krishnamurthy, 2019; Kuroki et al., 2020; Perrault et al., 2020). It is crucial, since the offline problem cannot be efficiently solved without this oracle.

Policies The history $\mathcal{F}_t := \sigma(A_1, Y_{1,A_1}, \cdots, A_t, Y_{t,A_t})$ contains all the information available to the agent at step $t+1$. In the fixed-confidence setting a strategy is described by three rules: a *sampling rule* $(A_t)_{t\geq 1}$ where $A_t \in \mathcal{A}$ is $\mathcal{F}_{t-1}$-measurable, a *stopping rule*, τ_δ being the stopping time with respect to the filtration $(\mathcal{F}_t)_{t\geq 1}$, and a *recommendation rule* I_{τ_δ} which is $\mathcal{F}_{\tau_\delta}$-measurable.

While the sampling rule can be randomized, we consider only deterministic strategies in our work. In the fixed-confidence setting, the learner is given a confidence parameter $\delta \in (0,1)$. The strategy is said to be δ-PAC if it terminates and recommends the correct answer with probability at least $1-\delta$: $\mathbb{P}_\nu\left[\tau_\delta = \infty \vee I_{\tau_\delta} \neq I^*(\mu)\right] \leq \delta$. Among δ-PAC algorithms, the objective is to minimize the expected number of samples required to terminate $\mathbb{E}_\nu[\tau_\delta]$, also known as the *sample complexity*.

2.2. Sample Complexity Lower Bound

Given an answer $I \in \mathcal{I}$, the *cell* Θ_I is the set of parameters for which the correct answer is I, $\Theta_I := \{\lambda \in \mathcal{M} : I^*(\lambda) = I\}$. The *alternative to* I is the set of parameters for which I is not the correct answer, $\Theta_I^\complement$. It is also equal to the set of parameters for which there exists an answer $J \neq I$ having a higher reward, $\Theta_I^\complement = \bigcup_{J\in\mathcal{I}\setminus\{I\}} \bar{\Theta}_J^I$ where $\bar{\Theta}_J^I := \{\lambda \in \mathcal{M} : \langle \mathbf{1}_J - \mathbf{1}_I, \lambda\rangle \geq 0\}$. The *neighbors to* I is the set of answers whose cells' boundaries intersect the boundary of the cell I, $N(I) := \{J \in \mathcal{I} : \partial\Theta_I \cap \partial\Theta_J \neq \emptyset\}$.

The *transformed simplex* $\mathcal{S}_\mathcal{A} := \{W_\mathcal{A} w : w \in \Delta_{|\mathcal{A}|}\} \subset \mathbb{R}^d$ is the image of the $|\mathcal{A}|$-dimensional probability simplex $\Delta_{|\mathcal{A}|} := \left\{w \in \mathbb{R}^{|\mathcal{A}|} : w \geq 0 \wedge \sum_{A\in\mathcal{A}} w_A = 1\right\}$ by $W_\mathcal{A} := \left[\mathbf{1}_{A_1} \ldots \mathbf{1}_{A_{|\mathcal{A}|}}\right] \in \mathbb{R}^{d\times|\mathcal{A}|}$. The matrix $W_\mathcal{A}$ collects the action incidence vectors. For a distribution over actions $w \in$

[1] In practice, such an oracle may be an efficient algorithm tailored to the combinatorial constraints (e.g., Kruskal's algorithm for minimum spanning trees etc.), or a search strategy given by a Mixed Integer Programming solver.

$\Delta_{|\mathcal{A}|}$, $W_{\mathcal{A}}w$ represents the effect at the base arm level when sampling actions according to w. The probability of sampling the arm $a \in [d]$ is $\tilde{w}_a$, where $\tilde{\cdot} := W_{\mathcal{A}}\cdot$ denotes implicitly the operator $W_{\mathcal{A}}$.

Lower bound Given any δ-PAC strategy, Theorem 1 gives a finite-time and asymptotic lower bound on the sample complexity, see Appendix C for a proof. This result is a technical extension of previous work, see Theorem 1 in Garivier and Kaufmann (2016).

Theorem 1 *For any δ-PAC strategy and any bandit ν characterized by μ,*

$$\frac{\mathbb{E}_\nu[\tau_\delta]}{\ln(1/(2.4\delta))} \geq D_\nu^{-1} \quad \textit{and} \quad \limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\ln(1/\delta)} \geq D_\nu^{-1}$$

where the complexity *D_ν is the inverse of the characteristic time, defined by*

$$D_\nu := \max_{\tilde{w} \in \mathcal{S}_{\mathcal{A}}} \inf_{\lambda \in \Theta^{\complement}_{I^*(\mu)}} \langle \tilde{w}, d_{KL}(\mu, \lambda) \rangle$$

Similar bounds were already proven for other settings (Garivier and Kaufmann, 2016; Degenne and Koolen, 2019). The technical difference is that we sample actions. A δ-PAC strategy is said to be *asymptotically optimal* if the bound is tight, meaning that for any ν, $\limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\ln(1/\delta)} \leq D_\nu^{-1}$.

The *set of optimal allocations* is $w^*(\mu) := \left\{ w \in \Delta_{|\mathcal{A}|} : \inf_{\lambda \in \Theta^{\complement}_{I^*(\mu)}} \langle W_{\mathcal{A}}w, d_{\mathrm{KL}}(\mu, \lambda) \rangle = D_\nu \right\}$. It is non-empty since $\tilde{w} \mapsto \inf_{\lambda \in \Theta^{\complement}_{I^*(\mu)}} \langle \tilde{w}, d_{\mathrm{KL}}(\mu, \lambda) \rangle$ is concave on the compact $\mathcal{S}_{\mathcal{A}}$. Moreover, $w^*(\mu)$ contains multiple optimal allocations, except for specific choice of $\mathcal{A}$. Computing an element of $w^*(\mu)$ is a difficult minmax optimization even for a known μ. To the best of our knowledge, there are no theoretical results on the hardness of this specific optimization problem.

3. Algorithms

After introducing the game approach, we discuss two asymptotically optimal families of algorithms which instantiate our proposed pure-exploration CombGame meta-algorithm, see Algorithm 1. The learners used to instantiating it are either on $\Delta_{|\mathcal{A}|}$ or on $\mathcal{S}_{\mathcal{A}}$.

3.1. Game Approach

At round $t \geq 1$, the agent computes a distribution over actions $w_t \in \Delta_{|\mathcal{A}|}$ which is converted into a deterministic action A_t by tracking (Garivier and Kaufmann, 2016), as explained below. Since we observe semi-bandit feedback, w_t corresponds to $\tilde{w}_t = W_{\mathcal{A}}w_t$ at the base arms level. Importantly, due to the independence assumption and the linearity of the considered operators, all computations on $\Delta_{|\mathcal{A}|}$ can be done on $\mathcal{S}_{\mathcal{A}}$.

Since $\mathcal{S}_{\mathcal{A}} = \mathrm{conv}\left(\{\mathbf{1}_A\}_{A \in \mathcal{A}}\right)$, the transformed simplex is a 0-1 polytope in $\mathbb{R}^d$. A pulling proportion $w \in \Delta_{|\mathcal{A}|}$ is said to be *sparse* if its support is small, $\mathrm{supp}(w) \ll |\mathcal{A}|$. A simple application of Carathéodory's theorem yields that for all $w \in \Delta_{|\mathcal{A}|}$ there exists a sparse $w_0 \in \Delta_{|\mathcal{A}|}$ with $|\mathrm{supp}(w_0)| \leq d + 1$ such that both w and w_0 have the same allocation over arms, $W_{\mathcal{A}}w = W_{\mathcal{A}}w_0$.

Two-player, minimax approach As noted in the early work by Chernoff (1959) and extended in the recent papers using gamification (Degenne et al., 2019, 2020a), the complexity D_ν is the value of a fictitious zero-sum game between two players. The agent chooses a pulling proportion over arms, $\tilde{w} \in \mathcal{S}_\mathcal{A}$. The nature plays the most confusing alternative with respect to the KL divergence in order to fool the agent into predicting an incorrect answer, $\lambda \in \Theta^{\complement}_{I^*(\mu)}$.

Allowing nature to play distributions over alternatives and using Sion's minimax theorem, we can invert the order of the players to obtain the dual formulation of the complexity D_ν,

$$D_\nu = \max_{\tilde{w} \in \mathcal{S}_\mathcal{A}} \inf_{\lambda \in \Theta^{\complement}_{I^*(\mu)}} \langle \tilde{w}, d_{\mathrm{KL}}(\mu, \lambda) \rangle = \inf_{q \in \mathcal{P}\left(\Theta^{\complement}_{I^*(\mu)}\right)} \max_{A \in \mathcal{A}} \mathbb{E}_{\lambda \sim q} \left[\langle \mathbf{1}_A, d_{\mathrm{KL}}(\mu, \lambda) \rangle \right]$$

where $\mathcal{P}\left(\Theta^{\complement}_{I^*(\mu)}\right)$ denotes the set of probability distributions over $\Theta^{\complement}_{I^*(\mu)}$.

In our work we focus on a sequential game where the agent, or A-player, plays first and nature, or the λ-player, is second. The A-player uses a learner that minimizes the cumulative regret. The λ-player has access to a best-response oracle that has no regret. This combination ensures a saddle-point property required to derive the finite-time upper bound on the sample complexity. Alternatively the order could be reversed, or they could play simultaneously (Degenne et al., 2019).

3.2. CombGame Meta-Algorithm

First, we briefly introduce the estimator, stopping and recommendation rules, which define the pure-exploration algorithm. Since μ (and the best answer $I^*(\mu)$) is unknown, we use the maximum likelihood estimator (MLE) μ_t as a plug-in estimator. The recommendation and the stopping rules are frequentist and use the value of μ_t. Based on μ_t, the sampling rule corresponds to playing an optimistic sequential game. Both the sample complexity and the computational efficiency depend on the learner used to approximate this game.

Estimator Let $N_{t-1} \in \mathbb{R}^{|\mathcal{A}|}$ be the count of sampled actions at the beginning of round t and $\tilde{N}_{t-1} = W_\mathcal{A} N_{t-1}$ its counterpart at the base arms level. The MLE, $\mu_{t-1,a} := \frac{1}{\tilde{N}_{t-1,a}} \sum_{s=1}^{t-1} \mathbf{1}_{(a \in A_s)} Y_{s,a}$ for all $a \in [d]$, is associated with the confidence hyperbox for the exploration bonus f, $\mathcal{C}_t := \bigtimes_{a \in [d]} [\alpha_{t,a}, \beta_{t,a}]$ where $[\alpha_{t,a}, \beta_{t,a}] := \{\lambda : \tilde{N}_{t-1,a} d_{\mathrm{KL}}(\mu_{t-1,a}, \lambda) \leq f(t-1)\}$. As in Degenne et al. (2019), the exploration bonus is chosen as $f(t) = \overline{W}((1+c)(1+b)\ln(t))$ where $c > 0$, $b > 0$ and $\overline{W}(x) \approx x + \ln(x)$, see Appendix G.1 for an exact definition.

When $\mu_{t-1} \notin \mathcal{M}$, we consider $\tilde{\mu}_{t-1} \in \operatorname{argmin}_{\lambda \in \mathcal{M} \cap \mathcal{C}_t} \langle \tilde{N}_{t-1}, d_{\mathrm{KL}}(\mu_{t-1}, \lambda) \rangle$, the projection of μ_{t-1} on $\mathcal{M} \cap \mathcal{C}_t$. $\tilde{\mu}_{t-1}$ is chosen randomly when $\mathcal{M} \cap \mathcal{C}_t = \emptyset$. When all arms are sampled an infinite number of times, we have $\lim_\infty \mu_t = \mu \in \mathcal{M}$: there exists T_0 such that for all $t \geq T_0$, $\mu_{t-1} \in \mathcal{M}$.

Stopping and recommendation rules We will use the recommendation and stopping rules based on a frequentist estimator μ_{t-1}. Given the feasible $\tilde{\mu}_{t-1}$, we recommend the unique best answer $I_t := \operatorname{argmax}_{I \in \mathcal{I}} \langle \mathbf{1}_I, \tilde{\mu}_{t-1} \rangle$. I_t can be computed with the efficient oracle. We stop as soon as the generalized likelihood ratio is above a stopping threshold $\beta(t-1, \delta)$:

$$\tau_\delta := \inf \left\{ t \in \mathbb{N} : \min_{J \in N(I_t)} \inf_{\lambda \in \bar{\Theta}^{I_t}_J} \langle \tilde{N}_{t-1}, d_{\mathrm{KL}}(\mu_{t-1}, \lambda) \rangle > \beta(t-1, \delta) \right\}$$

Given any sampling rule, this pair of rules is sufficient to obtain a δ-PAC strategy, see Theorem 2. The proof leverages the concentration inequalities of Kaufmann and Koolen (2018) (Appendix D).

Algorithm 1: CombGame meta-algorithm

```
Input: Learner $\mathcal{A}^A$ with associated init, stopping threshold $\beta(t-1,\delta)$, exploration bonus $f(t)$
Output: Answer $I_t$
$(w_{n_0}, \tilde{w}_{n_0}, B_{n_0}) = \text{INIT(init)}$ ;                                  ▷initialization
for $t = n_0 + 1, \cdots$ do
    $I_t = \text{argmax}_{I \in \mathcal{I}} \langle \mathbf{1}_I, \tilde{\mu}_{t-1} \rangle$ ;                          ▷recommendation rule
    If $\min_{J \in N(I_t)} \inf_{\lambda \in \bar{\Theta}_J^{I_t}} \langle \tilde{N}_{t-1}, d_{\text{KL}}(\mu_{t-1}, \lambda) \rangle > \beta(t-1,\delta)$ then return $I_t$ ;   ▷stopping rule
    Get $(w_t, \tilde{w}_t, B_t)$ from $\mathcal{A}^A_{I_t}$;
    $A_t \in \text{argmin}_{A \in B_t} \frac{N_{t-1,A}}{\sum_{s=1}^t w_{s,A}}$ ;                              ▷sparse C-Tracking
    $(\cdot, \lambda_t) \in \text{argmin}_{J \in N(I_t), \lambda \in \Theta_J^{I_t}} \langle \tilde{w}_t, d_{\text{KL}}(\mu_{t-1}, \lambda) \rangle$ ;              ▷λ-player
    $\forall a \in [d], \quad r_{t,a} = \max \left\{ \frac{f(t-1)}{\tilde{N}_{t-1,a}}, \max_{\phi \in \{\alpha_{t,a}, \beta_{t,a}\}} d_{\text{KL}}(\phi, \lambda_{t,a}) \right\}$ ;   ▷optimism
    Feed $\mathcal{A}^A_{I_t}$ with the reward $r_t$;
    Observe a sample $Y_{t,A_t}$ and update $(\mu_t, N_t, \tilde{\mu}_t)$ ;
end
```

Theorem 2 *Let $\mathcal{M}$ be bounded. Regardless of the sampling rule, a strategy using the frequentist recommendation/stopping pair with the stopping threshold:*

$$\beta(t,\delta) := \begin{cases} 3d_0 \ln\left(1 + \ln\left(\frac{tK}{d_0}\right)\right) + d_0 \mathcal{T}\left(\frac{\ln\left(\frac{|\mathcal{I}|-1}{\delta}\right)}{d_0}\right) & \textit{for (a)} \\ 2d_0 \ln\left(4 + \ln\left(\frac{tK}{d_0}\right)\right) + d_0 \mathcal{C}^{gG}\left(\frac{\ln\left(\frac{|\mathcal{I}|-1}{\delta}\right)}{d_0}\right) & \textit{for (b)} \end{cases}$$

is δ-PAC. In the above, $d_0 := \max_{I,J \in \mathcal{I}, J \neq I} |(I \setminus J) \cup (J \setminus I)|$, $\mathcal{T}$ and $\mathcal{C}^{gG}$ are the functions defined in Kaufmann and Koolen (2018), $\mathcal{C}^{gG}(x) \approx x + \ln(x)$ and $\mathcal{T}(x) \approx x + 4\ln(1 + x + \sqrt{2x})$ for $x \geq 5$.

3.2.1. SAMPLING RULE

The challenge is to define the sampling rule in order to satisfy the stopping criterion as soon as possible. Based on the definition of τ_δ, $\min_{J \in N(I_t)} \inf_{\lambda \in \bar{\Theta}_J^I} \langle \tilde{N}_t, d_{\text{KL}}(\mu_t, \lambda) \rangle$ should be maximized. We will achieve the desired saddle-point property by combining $|\mathcal{I}|$ learners for the A-player, one per candidate answer $\mathcal{A}^A_{I_t}$, and one best-response oracle for the λ-player.

We present two categories of learners, both aiming at minimizing the cumulative regret $R_t^A := \max_{A \in \mathcal{A}} \sum_{s=1}^t \langle \mathbf{1}_A, r_s \rangle - \sum_{s=1}^t \langle \tilde{w}_s, r_s \rangle$. r_t is the optimistic reward at time t as defined in the paragraph below. Learners on the simplex update w_t and need a *full* initialization where each action is sampled once. To overcome the computational inefficiency of those learners, we also consider learners on the transformed simplex which update $\tilde{w}_t$. By leveraging the sparse support when tracking, they only require a *covering* initialization where each arm is observed at least once. The length of the initialization is denoted by n_0. We compare the different learners in Table 1.

By knowing w_t used by the A-player, the λ-player can adopt the most confusing parameter in $\Theta_{I_t}^{\complement}$: $(\cdot, \lambda_t) \in \text{argmin}_{J \in N(I_t), \lambda \in \bar{\Theta}_J^{I_t}} \langle \tilde{w}_t, d_{\text{KL}}(\mu_{t-1}, \lambda) \rangle$.

Optimism Since the estimator is not exact but associated to a confidence region, following the exact sequential game for μ_t cannot lead to sufficient exploration. Degenne et al. (2019) overcome this hurdle by using the optimism principle. Since $\mu \in \mathcal{C}_t$ with high probability, the optimistic reward r_t is the upper bound on the gain of the agent given the λ-player's response, λ_t: for all $a \in [d]$, $r_{t,a} := \max\left\{\frac{f(t-1)}{\tilde{N}_{t-1,a}}, \max_{\phi \in \{\alpha_{t,a}, \beta_{t,a}\}} d_{\mathrm{KL}}(\phi, \lambda_{t,a})\right\}$ where $\frac{f(t-1)}{\tilde{N}_{t-1,a}}$ fosters exploration. The clipping is due to non-symmetric d_{KL}. It disappears for Gaussian as shown in Lemma 8.

Tracking Since a learner plays pulling proportion over actions w_t, we need to convert it into an action choice A_t. Introduced in Garivier and Kaufmann (2016), C-Tracking and D-Tracking allow to deterministically convert weights into pulls. Due to the non-uniqueness of the optimal allocation of weights, we consider C-Tracking, which ensures $1 - |\mathcal{A}| \leq N_{t,A} - \sum_{s=1}^{t} w_{t,A} \leq 1$, see Appendix G.5.1. We obtain a sparse tracking procedure by limiting the choice of A_t to the incremental support $B_t := \operatorname{supp}\left(\sum_{s=1}^{t} w_s\right)$: $A_t \in \operatorname{argmin}_{A \in B_t} \frac{N_{t-1,A}}{\sum_{s=1}^{t} w_{s,A}}$. Alternatives include D-Tracking or the rounding procedure in Fiez et al. (2019). For a non-deterministic algorithm we can directly sample the next action, $A_t \sim w_t$.

3.3. Learners on the Simplex

Since we are playing pulling proportion over actions, the immediate approach is to consider Hedge-type algorithms. They constitute a family of learners on the probability simplex $\Delta_{|\mathcal{A}|}$. As examples from this family, we will use Hedge (Cesa-Bianchi et al., 2005) and the adaptive version AdaHedge (Rooij et al., 2014). An algorithm is said to be *anytime* if it is independent of the horizon T. Those learners require computations at the actions level to obtain a reward vector U_t: for all $A \in \mathcal{A}$, $U_{t,A} := \langle \mathbf{1}_A, r_t \rangle$. For both learners the update of w_t is: for all $A \in \mathcal{A}$, $w_{t,A} = \frac{w_{n_0,A} \exp\left(-\eta_t L_{t-1,A}\right)}{\sum_{A' \in \mathcal{A}} w_{n_0,A'} \exp\left(-\eta_t L_{t-1,A'}\right)}$ where $L_{t-1,A} = -\sum_{s=1}^{t-1} U_{s,A}$ is the cumulative loss, η_t is the learning rate and $w_{n_0} = \frac{1}{|\mathcal{A}|}\mathbf{1}$ is the sampling parameter for a full initialization. In Hedge, η_t is a constant depending on T. While in AdaHedge, η_t is decreasing and defined as a function of a cumulative mixability gap. As shown in Lemmas 9 and 10, both Hedge and AdaHedge have optimal cumulative regret, $O\left(\ln(t)\sqrt{t}\right)$. The additional $\ln(t)$-factor originates from the unbounded losses.

Due to the potentially exponential number of actions, a closer examination of those learners reveals the *computational inefficiency* of three steps. First, we initialize by sampling all the actions once. Second, at each round the update step requires the computation of $U_t \in \mathbb{R}^{|\mathcal{A}|}$ and $w_t \in \Delta_{|\mathcal{A}|}$. Third, C-Tracking is equivalent to finding the minimum of $|\mathcal{A}|$ values since w_t is dense. This motivates considering the second family of algorithms, which defines the learner directly on $\mathcal{S}_{\mathcal{A}}$.

3.4. Learners on the Transformed Simplex

To circumvent the shortcomings of the learners on $\Delta_{|\mathcal{A}|}$, we introduce a second family of learners for which we update $\tilde{w}_t \in \mathcal{S}_{\mathcal{A}}$ by using r_t. Note that the loss for the A-learner is linear, $f_t(x) = -\langle x, r_t \rangle$ for $x \in \mathcal{S}_{\mathcal{A}}$. The online convex optimization (OCO) literature provides algorithms achieving optimal cumulative regret guarantees for adversarial linear losses. Since we want a computationally efficient algorithm, the learner should satisfy three additional requirements. First, it should be projection-free, since projections onto $\mathcal{S}_{\mathcal{A}}$ require a solution to a costly quadratic optimization problem. Second, the learner should access at most one efficient linear optimization oracle per round. Third, the algorithm should maintain efficiently an incrementally sparse representation in the simplex, which is used for sparse tracking. Projection-free algorithms have been extensively

	Sparse support	Computational cost	Anytime	Cumulative regret
Hedge	✗	$O\left(\lvert\mathcal{A}\rvert\right)$	✗	$O\left(\ln(t)\sqrt{t}\right)$
AdaHedge	✗	$O\left(\lvert\mathcal{A}\rvert\right)$	✓	$O\left(\ln(t)\sqrt{t}\right)$
OFW	✓	$O\left(\lvert B_t\rvert\right)$	✓	$O\left(\ln(t)^2 t^{3/4}\right)$
LLOO	✓	$O\left(\lvert B_t\rvert(d+\ln(\lvert B_t\rvert))\right)$	✗	$O\left(\ln(t)\sqrt{t}\right)$

Table 1: Comparison of the relevant properties of the learners used to instantiate CombGame. For cumulative regret, the notation $O(\cdot)$ hides parameters independent of t, see Appendix F. For the computational cost, $O(\cdot)$ hides constant values, small compared to $|B_t|$ and $|\mathcal{A}|$.

studied since they are computationally efficient, as long as the linear optimization oracle is computationally efficient and increase support incrementally. They are often based on the Frank-Wolfe approach (Frank and Wolfe, 1956; Jaggi, 2013; Lacoste-Julien and Jaggi, 2015).

The anytime Online Frank-Wolfe (OFW) (Hazan and Kale, 2012b) and Local Linear Optimization Oracle-based OCO (LLOO) (Garber and Hazan, 2013) satisfy those requirements. LLOO is tailored to convex polyhedral sets, see Appendix I for details. Therefore, the assumptions of LLOO are satisfied in our setting. Both use a single call per round to the linear optimization oracle in order to compute the best vertex $\mathbf{1}_{\tilde{A}_t}$ of the polytope with respect to the gradient of a regularized cumulative loss F_t: $\tilde{A}_t \in \operatorname{argmin}_{A\in\mathcal{A}}\langle \mathbf{1}_A, \nabla F_t(\tilde{w}_t)\rangle$. While $F_t(x) = \frac{1}{t}\sum_{s=1}^{t} \frac{s^{-1/4}}{\text{diam}(\mathcal{S}_{\mathcal{A}})}\|x - \tilde{w}_{n_0}\|_2^2 - \langle x, r_s\rangle$ for OFW, where diam$(\mathcal{S}_{\mathcal{A}})$ denotes the diameter of $\mathcal{S}_{\mathcal{A}}$, we have $F_t(x) = \|x - \tilde{w}_{n_0}\|_2^2 - \eta_{\mathcal{A},T}\sum_{s=1}^{t}\langle x, r_s\rangle$ for LLOO. OFW simply moves on the segment connecting $\tilde{w}_t$ and $\mathbf{1}_{\tilde{A}_t}$, $\tilde{w}_{t+1} = \tilde{w}_t + t^{-1/4}\left(\mathbf{1}_{\tilde{A}_t} - \tilde{w}_t\right) \in \mathcal{S}_{\mathcal{A}}$. LLOO adopts a more sophisticated strategy whose parameters $\eta_{\mathcal{A},T}$, $\gamma_{\mathcal{A}}$ and $M_{\mathcal{A},T}$ depend on the horizon T, see Lemma 12 for explicit formulas. LLOO simultaneously moves towards the best corner $\mathbf{1}_{\tilde{A}_t}$ and away from the ordered worst corners by using several pairwise Frank-Wolfe steps, see Lacoste-Julien and Jaggi (2015). The corresponding update is $\tilde{w}_{t+1} = \tilde{w}_t + \gamma_{\mathcal{A}}\left(M_{\mathcal{A},T}\mathbf{1}_{\tilde{A}_t} - \tilde{w}_{t,-}\right)$, where $(\tilde{w}_{t,-}, w_{t,-}) = \mathcal{A}^{\text{reduce}}(w_t, B_t, M_{\mathcal{A},T}, \nabla F_t(\tilde{w}_t))$ and $\mathcal{A}^{\text{reduce}}$ is detailed in Algorithm 2. The computations of $\mathcal{A}^{\text{reduce}}$ are dominated by the cost of sorting $|B_t|$ inner-products in $\mathbb{R}^d$, $O\left(|B_t|(d + \ln(|B_t|))\right)$. Since $W_{\mathcal{A}}$ is a linear map, both variants of the convex-combination update of $\tilde{w}_{t+1}$ are propagated to the simplex to obtain w_{t+1} by using w_t, $\delta_{\tilde{A}_t}$ (dirac function in $\tilde{A}_t$) and $w_{t,-}$ instead of $\tilde{w}_t$, $\mathbf{1}_{\tilde{A}_t}$ and $\tilde{w}_{t,-}$. The corresponding support is incrementally sparse, $B_{t+1} \setminus B_t \subset \{\tilde{A}_t\}$, and unchanged when $\tilde{A}_t$ is already included in B_t. Lemma 11 shows that OFW has an upper bound on the cumulative regret in $O\left(\ln(t)^2 t^{3/4}\right)$, which is in general suboptimal for the online linear optimization setting. Thanks to these extra computations, Lemma 12 yields that LLOO has optimal cumulative regret, $O\left(\ln(t)\sqrt{t}\right)$. Those results are obtained by modifying existing ones (Hazan and Kale, 2012b; Garber and Hazan, 2013) to account for an unbounded reward and modified parameters for OFW. Since R_t^A appears in the finite-time upper bound on the sample complexity (Theorem 3), optimal cumulative regret is a desirable property if we strive for low sample complexity. This is validated by our experimental results.

4. Sample Complexity Upper Bound

In this section we present and sketch the proof of the finite-time upper bound on the sample complexity of our instantiated CombGame meta-algorithm.

Algorithm 2: LLOO's $\mathcal{A}^{\text{reduce}}$

```
Input: $w \in \Delta_{|\mathcal{A}|}$ with sparse support $B$, probability mass $M \in \mathbb{R}$ and cost vector $c \in \mathbb{R}^d$
$\forall A \in B, \quad l_A = \langle \mathbf{1}_A, c \rangle$;
Let $i_1, \cdots, i_{|B|}$ be a permutation such that $l_{A_{i_1}} \geq \cdots \geq l_{A_{i_{|B|}}}$;
Let $k$ be the smallest integer such that $\sum_{j=1}^{k} w_{A_{i_j}} \geq M$ ;
$(\tilde{w}_-, w_-) = \sum_{j=1}^{k-1} w_{A_{i_j}} \left(\mathbf{1}_{A_{i_j}}, \delta_{A_{i_j}}\right) + \left(M - \sum_{j=1}^{k-1} w_{A_{i_j}}\right) \left(\mathbf{1}_{A_{i_k}}, \delta_{A_{i_k}}\right)$ ;
Return $(\tilde{w}_-, w_-)$;
```

4.1. Finite-time Upper Bound

Given a learner with sub-linear cumulative regret $R_t^A = o(t)$, Theorem 3 shows that the instances of Algorithm 1, the CombGame meta-algorithm, satisfy a finite-time upper bound on the sample complexity. The upper bound involves the complexity D_ν. The leading constant is optimal in the asymptotic regime $\delta \to 0$. Those results and their proofs are inspired from Theorem 2 in Degenne et al. (2019). It also bares similarity with Theorem 2 of Degenne et al. (2020a).

Theorem 3 *Let $\mathcal{M}$ be bounded. The sample complexity of the instantiated CombGame meta-algorithm on bandit $\mu \in \mathcal{M}$ satisfies:*

$$\mathbb{E}_\nu[\tau_\delta] \leq T_0(\delta) + \frac{2ed}{c^2} \quad \textit{with} \quad T_0(\delta) := \max\left\{ t \in \mathbb{N} : t \leq \frac{\beta(t,\delta)}{D_\nu} + C_\nu(R_t^A + h(t)) \right\}$$

where $c > 0$ is the parameter of the exploration bonus $f(t)$ when taking $b = 1$. The reminder terms are: the approximation error $h(t) = O\left(\sqrt{t \ln(t)}\right)$, the learner's cumulative regret R_t^A and a constant C_ν depending on the distribution.

Moreover, the instantiated CombGame meta-algorithm is an asymptotically optimal algorithm.

Even though the upper bound in Theorem 3 holds for finite-time, it is an asymptotic result by nature. The additive term, which is independent of δ, can't be neglected in finite-time, and is likely to be loose due to the analysis. Therefore, we won't compare the upper bounds of different learners.

Proof Scheme Detailed in Appendix G, the proof of Theorem 3 uses Lemma 4, which is an adaptation of Lemma 1 in Degenne et al. (2019) with the same exploration bonus.

Lemma 4 *Let $(\mathcal{E}_t)_{t \geq 1}$ be a sequence of concentrations events for the exploration bonus f with parameters $c > 0$ and $b > 0$: $\mathcal{E}_t := \left\{ \forall s \leq t, \forall a \in [d], \quad \tilde{N}_{s,a} d_{KL}(\mu_{s,a}, \mu_a) \leq f\left(t^{1/(1+b)}\right) \right\}$ for all $t \geq 1$. Suppose that there exists $T_0(\delta) \in \mathbb{N}$ such that for all $t > T_0(\delta)$, $\mathcal{E}_t \subset \{\tau_\delta \leq t\}$. Then*

$$\mathbb{E}_\nu[\tau_\delta] \leq T_0(\delta) + \sum_{t > T_0(\delta)} \mathbb{P}_\nu\left[\mathcal{E}_t^{\complement}\right] \quad \textit{where} \quad \sum_{t > T_0(\delta)} \mathbb{P}_\nu\left[\mathcal{E}_t^{\complement}\right] \leq \frac{2ed}{c^2}$$

The challenging part of the proof is the characterization of $T_0(\delta)$ with an equation involving the complexity D_ν, similarly to Appendix D in Degenne et al. (2019). We need to exhibit an upper bound $T_0(\delta)$ such that for $t \geq T_0(\delta)$, if $\mathcal{E}_t$ holds then the algorithm has already stopped, $\tau_\delta \leq t$. In contrast to Degenne et al. (2019), the particularity of our proof is to consider computations on $\mathcal{S}_\mathcal{A}$

and not on the simplex. Even though the idea of the proof is identical, we need different technical arguments such as the tracking and concentration results in Appendices G.5.1 and G.5.2. For sake of simplicity we suppose that $I_t = I^*(\mu)$ in the following informal exposition. This fails only for $o(t)$ rounds as shown in Appendix G.3.1. Using C-Tacking, we obtain that as long as the stopping criterion is not satisfied, under the concentration event $\mathcal{E}_{t-1}$,

$$\beta(t-1, \delta) \geq \inf_{\lambda \in \Theta^{\complement}_{I^*(\mu)}} \langle \tilde{N}_{t-1}, d_{\mathrm{KL}}(\mu_{t-1}, \lambda) \rangle \geq \inf_{\lambda \in \Theta^{\complement}_{I^*(\mu)}} \sum_{s=1}^{t-1} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda) \rangle - O\left(\sqrt{t \ln(t)}\right)$$

Then, we leverage the approximate saddle-point property of the CombGame meta-algorithm. This property is obtained by combining the optimism, the no-regret λ-player and the cumulative regret of the A-player, see Appendix G.2:

$$\inf_{\lambda \in \Theta^{\complement}_{I^*(\mu)}} \sum_{s=1}^{t-1} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda) \rangle \geq \max_{A \in \mathcal{A}} \sum_{s=1}^{t-1} \langle \mathbf{1}_A, r_s \rangle - O\left(\sqrt{t}\right) - R_t^A$$

Under the concentration event $\mathcal{E}_{t-1}$, the optimism implies $r_s \geq d_{\mathrm{KL}}(\mu, \lambda_s)$ for $s \leq t-1$. Combining the dual formulation of D_ν and the average of diracs, $\frac{1}{t-1} \sum_{s=1}^{t-1} \delta_{\lambda_s} \in \mathcal{P}\left(\Theta^{\complement}_{I^*(\mu)}\right)$, yields:

$$\max_{A \in \mathcal{A}} \sum_{s=1}^{t-1} \langle \mathbf{1}_A, d_{\mathrm{KL}}(\mu, \lambda_s) \rangle \geq t \inf_{q \in \mathcal{P}\left(\Theta^{\complement}_{I^*(\mu)}\right)} \max_{A \in \mathcal{A}} \mathbb{E}_{\lambda \sim q} \left[\langle \mathbf{1}_A, d_{\mathrm{KL}}(\mu, \lambda) \rangle \right] = t D_\nu$$

Combining all inequalities justifies the definition of $T_0(\delta)$ as the largest time such that the following inequality is satisfied: $T_0(\delta) = \max\left\{ t \in \mathbb{N} : t \leq \frac{\beta(t,\delta)}{D_\nu} + C_\nu(R_t^A + h(t)) \right\}$. Taking the limit $\delta \to 0$ yields that the instances of CombGame are asymptotically optimal.

5. Experiments

The goal of our experiments is to validate the sample effectiveness and computational efficiency of CombGame's instances for the finite-time regime, $\delta = 0.1$. We will compare the sample complexity of our learners, the uniform sampling and GCB-PE (Chen et al., 2020). To our knowledge, GCB-PE is the only algorithm which can be used to solve the pure-exploration problem for combinatorial bandits with semi-bandit feedback. Other works consider bandit feedback, cumulative regret or MAB. In addition, we demonstrate that learners on $\mathcal{S}_{\mathcal{A}}$ have an *exponentially smaller computational cost* compared to the learners on $\Delta_{|\mathcal{A}|}$. As an illustrative example, we use the best-arm identification with batch size k for a Gaussian bandit, $\nu = \mathcal{N}(\mu, \sigma^2 I_d)$. In BAI the *informative* actions are the ones containing the best arm I^*, $\mathcal{A}^* := \{A \in \mathcal{A} : I^* \subset A\}$. They provide direct feedback on the best arm, while other actions are sampled to answer indirectly to our query. The batch setting is used in real-world applications and admits an efficient oracle, the greedy algorithm. The number of actions is $|\mathcal{A}| = \binom{d}{k}$ and the ratio of informative actions is $\frac{|\mathcal{A}^*|}{|\mathcal{A}|} = \frac{k}{d}$. By increasing the dimension d, we observe the effect of an exponential increase of $|\mathcal{A}|$ while the ratio of informative actions $\frac{|\mathcal{A}^*|}{|\mathcal{A}|}$ is decreasing harmonically. In Appendix I.1, additional experiments include BAI by playing paths in a graph (Figures 4 and 5).

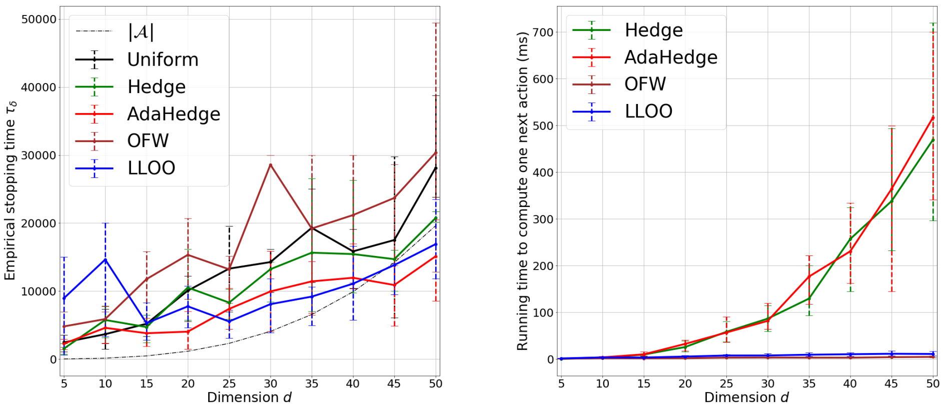

Figure 1: Uniform matroid, $k = 3$, where $\nu = \mathcal{N}(\mu, \sigma^2 I_d)$ with μ as in Appendix I.1.1 and $\sigma = 0.1$. (a) Empirical stopping time τ_δ. (b) Average running time to compute the next action. Note LLOO's competitive sample complexity for a low and constant computational cost.

As described in Appendix I, the empirical results of CombGame's instances are similar in behavior if we adopt D-Tracking instead of C-Tracking, one learner $\mathcal{A}^A$ instead of $|\mathcal{I}|$ learners, stylized stopping threshold $\beta(t,\delta) = \ln\left(\frac{1+\ln(t)}{\delta}\right)$ and exploration bonus $f(t) = \ln(t)$ instead of the ones licensed by theory. Doubling trick is used for Hedge and LLOO. The results over 750 runs are summarized in Figure 1 by plotting the mean of the empirical stopping time τ_δ and the average running time to compute the next action. The error bars correspond to the first and third quartiles.

In Figure 1(a), we observe that the sample complexity of Hedge, AdaHedge and LLOO is similar and increases proportionally to the number of actions. They perform better than uniform sampling, which still works reasonably well thanks to the high number of informative actions when $k \ll d$, $|\mathcal{A}^*| = \binom{d-1}{k-1}$. OFW's sample complexity is significantly higher than previous algorithms. This highlights the importance of cumulative regret's guarantees in order to have competitive empirical performance. In Figure 3(c) in Appendix I.1.1, we empirically show that on this example, the sample complexity of GCB-PE is about an order of magnitude higher compared to the other sampling rules: the mean over 750 runs of τ_δ is $\{19914, 85314, 166179, 316552\}$ for $d \in \{5, 10, 15, 20\}$.

In Figure 1(b), the computational efficiency of the learners on the transformed simplex is striking when compared to the learners on the simplex. While the computational cost increases exponentially for Hedge and AdaHedge, it remains almost constant for OFW and LLOO. Uniform sampling has constant run time per round. As detailed in Appendix I, the computational cost of GCB-PE is dominated by solving an NP-hard binary quadratic program in $\mathbb{R}^{m_{n_0}}$ with $m_{n_0} = \sum_{A \in B_{n_0}} |A| \geq d$. Since it has no efficient solver to our knowledge, the algorithm cannot run when d is high. Therefore, we were unable to perform further experiments on GCB-PE.

Despite the fact that there is no clear-cut ranking between all algorithms in Figure 1(a), Figure 1(b) highlights that, with a greatly lower computational cost, we obtain similar sample complexity.

6. Conclusion

In this paper we designed the first *computationally efficient and asymptotically optimal* algorithm to solve best-arm identification with combinatorial actions and semi-bandit feedback.

We highlight two directions to improve on our work. First, due to the learner's central role in the empirical performance, a more thorough benchmark of the existing learners should be made. An interesting choice is SFTPL from Hazan and Minasyan (2020) which meets our requirements. Second, the best-reponse oracle used by the λ-player is not computationally efficient for combinatorial *answer sets*, as in best-action identification, since the computations per round scale with $|N(I_t)|$ (which is usually lower than $|\mathcal{I}|$). In the spirit of Fiez et al. (2019); Zaki et al. (2020), this flaw could be mitigated by considering a phase-based algorithm discarding suboptimal answers.

Addressing a richer bandit structure where the arms are correlated is yet another avenue. Extending our approach to correlated Gaussian with known covariance matrix Σ is straightforward. The *correlated transformed simplex* is a subset of the cone of symmetric positive semi-definite matrices: $\mathrm{conv}(\{V_{\delta_A}\}_{A \in \mathcal{A}})$ where $V_{\delta_A} := S_A^\intercal \left(S_A \Sigma S_A^\intercal\right)^{-1} S_A$ and $S_A := \left(\mathbf{1}_{(\tilde{a}=a)}\right)_{\tilde{a} \in A, a \in [d]}$. Unfortunately, the oracle has the form $\mathrm{argmin}_{A \in \mathcal{A}} \mathrm{Tr}(V_{\delta_A}^\intercal C)$ for a cost matrix $C \in \mathbb{R}^{d \times d}$. To our knowledge, there is no computationally efficient oracle for this linear optimization over matrices. Therefore, it is not clear how and to what extent we can conserve the computational efficiency of our sampling rules.

Finally, as already noted in Degenne et al. (2020a), we observed that the stopping threshold is the major bottleneck in terms of finite-time empirical sample complexity. Using thresholds guarantying δ-PAC algorithms is too conservative since empirical error rates are orders of magnitude below the theoretical confidence error δ.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program grant agreement No 815943. It was also supported by the Swiss National Science Foundation through the NCCR Catalysis.

References

Jean-Yves Audibert, Sébastien Bubeck, and Remi Munos. Best arm identification in multi-armed bandits. In *COLT 2010 - The 23rd Conference on Learning Theory*, pages 41–53, November 2010.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Séebastian Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple Identifications in Multi-Armed Bandits. In *International Conference on Machine Learning*, pages 258–265, February 2013.

Tongyi Cao and Akshay Krishnamurthy. Disagreement-based combinatorial pure exploration: Sample complexity bounds and an efficient algorithm. volume 99 of *Proceedings of Machine Learning Research*, pages 558–588, Phoenix, USA, 25–28 Jun 2019. PMLR.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404 – 1422, 2012.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved Second-Order Bounds for Prediction with Expert Advice. In *Learning Theory*, pages 217–232, Berlin, Heidelberg, 2005.

Lijie Chen, Anupam Gupta, and Jian Li. Pure exploration of multi-armed bandit under matroid constraints. volume 49 of *Proceedings of Machine Learning Research*, pages 647–669, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

Lijie Chen, Anupam Gupta, Jian Li, Mingda Qiao, and Ruosong Wang. Nearly Optimal Sampling Algorithms for Combinatorial Pure Exploration. In *Conference on Learning Theory*, pages 482–534, June 2017a.

Lin Chen, Andreas Krause, and Amin Karbasi. Interactive Submodular Bandit. In *Advances in Neural Information Processing Systems 30*, pages 141–152. 2017b.

Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial Pure Exploration of Multi-Armed Bandits. In *Advances in Neural Information Processing Systems 27*, pages 379–387. 2014.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. volume 28 of *Proceedings of Machine Learning Research*, pages 151–159, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Wei Chen, Yihan Du, and Yuko Kuroki. Combinatorial pure exploration with partial or full-bandit linear feedback. *arXiv preprint arXiv:2006.07905*, 2020.

Herman Chernoff. Sequential design of experiments. *Annals of Mathematical Statistics*, 30(3): 755–770, 09 1959.

Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, and Marc Lelarge. Combinatorial Bandits Revisited. In *Advances in Neural Information Processing Systems 28*, pages 2116–2124. 2015.

Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, 2020a.

Rémy Degenne, Han Shao, and Wouter M Koolen. Structure adaptive algorithms for stochastic bandits. In *International Conference on Machine Learning*, 2020b.

Rémy Degenne and Wouter M Koolen. Pure Exploration with Multiple Correct Answers. In *Advances in Neural Information Processing Systems 32*, pages 14591–14600. 2019.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-Asymptotic Pure Exploration by Solving Games. In *Advances in Neural Information Processing Systems 32*, pages 14492–14501. 2019.

Tanner Fiez, L. Jain, K. Jamieson, and L. Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems 32*, pages 10667–10677, 2019.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck. Multi-Bandit Best Arm Identification. In *Advances in Neural Information Processing Systems 24*, pages 2222–2230. 2011.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence. In *Advances in Neural Information Processing Systems 25*, pages 3212–3220. 2012.

Victor Gabillon, Alessandro Lazaric, Mohammad Ghavamzadeh, Ronald Ortner, and Peter Bartlett. Improved Learning Complexity in Combinatorial Pure Exploration Bandits. In *Artificial Intelligence and Statistics*, pages 1004–1012, May 2016.

Dan Garber and Elad Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *SIAM Journal on Optimization*, 26, January 2013.

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

Aurélien Garivier, Emilie Kaufmann, and Wouter M. Koolen. Maximin Action Identification: A New Bandit Framework for Games. In *Conference on Learning Theory*, pages 1028–1050, June 2016.

Elad Hazan and Satyen Kale. Online Submodular Minimization. *Journal of Machine Learning Research*, 13(93):2903–2922, 2012a.

Elad Hazan and Satyen Kale. Projection-free online learning. In *Proceedings of the 29th International Conference on Machine Learning*, page 1843–1850, Madison, WI, USA, 2012b.

Elad Hazan and Edgar Minasyan. Faster projection-free online learning. In *Conference on Learning Theory*, 2020.

Martin Jaggi. Revisiting Frank-Wolfe: projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, pages I–427–I–435, Atlanta, GA, USA, June 2013.

K. Jamieson and R. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, 2014.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' ucb : An optimal exploration algorithm for multi-armed bandits. volume 35 of *Proceedings of Machine Learning Research*, pages 423–439, Barcelona, Spain, 13–15 Jun 2014. PMLR.

Kwang-Sung Jun, Kevin G. Jamieson, Robert D. Nowak, and Xiaojin Zhu. Top Arm Identification in Multi-Armed Bandits with Batch Arm Pulls. In *Artificial Intelligence and Statistics*, pages 139–148, Cadiz, Spain, 09–11 May 2016.

Shivaram Kalyanakrishnan, Ambuj Tewari, P. Auer, and P. Stone. Pac subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*, page 227–234, 2012.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, pages 1238–1246, Atlanta, GA, USA, June 2013.

Emilie Kaufmann and Wouter Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *arXiv preprint arXiv:1811.11419*, 2018.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, January 2016.

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 31*, pages 6332–6342. 2018.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. volume 125 of *Proceedings of Machine Learning Research*, pages 2328–2369. PMLR, 09–12 Jul 2020.

Yuko Kuroki, Liyuan Xu, Atsushi Miyauchi, Junya Honda, and Masashi Sugiyama. Polynomial-time algorithms for multiple-arm identification with full-bandit feedback. *Neural Computation*, 32(9):1733–1773, 2020.

Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, page 420–429, Arlington, Virginia, USA, 2014.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits. volume 38 of *Proceedings of Machine Learning Research*, pages 535–543, San Diego, California, USA, 09–12 May 2015. PMLR.

S. Lacoste-Julien and M. Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, page 496–504, 2015.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Pierre Perrault, Vianney Perchet, and Michal Valko. Exploiting structure of uncertainty for efficient matroid semi-bandits. volume 97 of *Proceedings of Machine Learning Research*, pages 5123–5132, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Pierre Perrault, Etienne Boursier, Vianney Perchet, and Michal Valko. Statistical efficiency of thompson sampling for combinatorial semi-bandits. *arXiv preprint arXiv:2006.06613*, 2020.

Idan Rejwan and Yishay Mansour. Top-k combinatorial bandits with full-bandit feedback. volume 117 of *Proceedings of Machine Learning Research*, pages 752–776, San Diego, California, USA, 08 Feb–11 Feb 2020. PMLR.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952.

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(37):1281–1316, 2014.

Daniel Russo. Simple bayesian algorithms for best arm identification. volume 49 of *Proceedings of Machine Learning Research*, pages 1417–1418, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

J. Scarlett, I. Bogunovic, and V. Cevher. Overlapping multi-bandit best arm identification. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2544–2548, 2019.

Xuedong Shang, Rianne de Heide, Pierre Menard, Emilie Kaufmann, and Michal Valko. Fixed-confidence guarantees for bayesian best-arm identification. volume 108 of *Proceedings of Machine Learning Research*, pages 1823–1832, Online, 26–28 Aug 2020. PMLR.

Max Simchowitz, Kevin Jamieson, and Benjamin Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. volume 65 of *Proceedings of Machine Learning Research*, pages 1794–1834, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

M. S. Talebi, Z. Zou, R. Combes, A. Proutiere, and M. Johansson. Stochastic online shortest path routing: The value of feedback. *IEEE Transactions on Automatic Control*, 63(4):915–930, 2018.

Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, and Alessandro Lazaric. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, page 1113–1122, 2015.

Mohammadi Zaki, Avi Mohan, and Aditya Gopalan. Explicit best arm identification in linear bandits using no-regret learners. *arXiv preprint arXiv:2006.07562*, 2020.

Appendix A. Notation

We recall some commonly used notations: the set of base arms $a \in [d] := \{1, \cdots, d\}$, the indicator vector $\mathbf{1}_S := (\mathbf{1}_{(a \in S)})_{a \in [d]}$ for a subset $S \subset [d]$, the symmetric difference of two sets $A \triangle B := (A \setminus B) \cup (B \setminus A)$, the euclidean inner-product $\langle x, y \rangle := \sum_{a \in [d]} x_a y_a$, the support of a vector $\text{supp}(x) := \{a \in [d] : x_a \neq 0\}$, the set of probability distributions $\mathcal{P}(\mathcal{X})$ over $\mathcal{X}$ and the n-dimensional probability simplex $\Delta_n := \{x \in \mathbb{R}^n : x \geq 0 \wedge \langle 1_d, x \rangle = 1\}$. In Table 2, we summarize problem-specific notations.

Notation	Meaning
$(\nu_a)_{a \in [d]}$	distributions of base arms
μ	bandit mean parameter, $(\mu_a)_{a \in [d]}$
$\mathcal{M}$	set of possible parameters
$d_{\text{KL}}(\mu, \lambda)$	component-wise KL divergence between two parameters, $(d_{\text{KL}}(\mu_a, \lambda_a))_{a \in [d]}$
K	maximum size of an action, $\max_{A \in \mathcal{A}} \lvert A \rvert$
A_t	sampled action at time t
Y_{t,A_t}	semi-bandit feedback at time t, $(Y_{t,a})_{a \in A_t}$
$\mathcal{F}_t$	history up to time t, $\sigma(A_1, Y_{1,A_1}, \cdots, A_t, Y_{t,A_t})$
μ_t	maximum likelihood estimator, $\left(\frac{1}{\tilde{N}_{t-1,a}} \sum_{s=1}^{t-1} \mathbf{1}_{(a \in A_s)} Y_{s,a}\right)_{a \in [d]}$
$\mathcal{C}_t$	confidence region associated to μ_t
$\tilde{\mu}_t$	projection of μ_t onto $\mathcal{M}$
$I^*(\lambda)$	unique correct answer for parameter λ, $\text{argmax}_{I \in \mathcal{I}} \langle \lambda, \mathbf{1}_I \rangle$
I_t	recommended answer at time t
Θ_I	cell I, $\{\lambda \in \mathcal{M} : I^*(\lambda) = I\}$
$\bar{\Theta}^I_J$	set of parameters for which J outperforms I, $\{\lambda \in \mathcal{M} : \langle \mathbf{1}_J - \mathbf{1}_I, \lambda \rangle \geq 0\}$
$N(I)$	neighbors to I, $\{J \in \mathcal{I} : \partial\Theta_I \cap \partial\Theta_J \neq \emptyset\}$
τ_δ	stopping time for confidence δ
$\mathcal{S}_{\mathcal{A}}$	transformed simplex
D_ν	complexity for the bandit ν
$N_t, \tilde{N}_t$	empirical count of sampled actions and its equivalent for arms, $W_{\mathcal{A}} N_t$
$w_t, \tilde{w}_t$	pulling distribution over actions and its equivalent for arms, $W_{\mathcal{A}} w_t$
$\beta(t, \delta)$	stopping threshold at time t for confidence δ
$f(t)$	exploration bonus at time t
r_t	optimistic reward
U_t	extended optimistic reward
R^A_t	cumulative regret of the learner of the A-player

Table 2: Table of notations

Appendix B. Outline

The appendices are organized as follows:

- The proof of Theorem 1 is detailed in Appendix C.

- The proof of Theorem 2 is detailed in Appendix D.
- The results concerning the optimistic reward are detailed in Appendix E: bounds on $\|r_t\|_\infty$ and explicit formulas for Gaussian bandit.
- The upper bounds on the learners' cumulative regret are proven in Appendix F.
- The full proof of Theorem 3 is detailed in Appendix G.
- In Appendix H, we sketch why the boundedness assumption is immaterial for Gaussian bandit.
- The implementation details for the experiments are presented in Appendix I. Additional empirical results are also displayed.

Appendix C. Proof of Theorem 1

Let $\mathrm{kl}(x, y)$ be the KL divergence of a Bernoulli distribution. Let ν and ν' be two bandit models such that for all $a \in [d]$ the distributions ν_a and ν'_a are mutually absolutely continuous. The associated density are denoted f_{ν_a} and $f_{\nu'_a}$. Given the history up to time t, the log-likelihood ratio of the independent observations is:

$$L_t = L_t(A_1, Y_{1,A_1}, \cdots, A_t, Y_{t,A_t}) := \sum_{a\in[d]} \sum_{s\in[t]:a\in A_s} \ln\left(\frac{f_{\nu_a}(Y_{s,a})}{f_{\nu'_a}(Y_{s,a})}\right)$$

The proof of Theorem 1 is an adaptation of the proof of Theorem 1 in Garivier and Kaufmann (2016) to our setting. We use Lemma 19 of Kaufmann et al. (2016), which shows a lower bound on the expectation of the log-likelihood ratio. Combined with Wald's lemma, we obtain the transportation inequality of Lemma 5, which replaces the Lemma 1 in Kaufmann et al. (2016).

Lemma *(Lemma 19 in Kaufmann et al. (2016)) Let τ be the almost-surely finite stopping time with respect to the filtration $(\mathcal{F}_t)_{t\geq 1}$. For every event $\mathcal{E} \in \mathcal{F}_\tau$,*

$$\mathbb{E}_\nu\left[L_\tau\right] \geq kl(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E}))$$

Lemma 5 *Let ν and ν' be two bandit models with independent arms. For any almost-surely finite stopping time τ with respect to the filtration $(\mathcal{F}_t)_{t\geq 1}$,*

$$\sum_{a\in[d]} \mathbb{E}_\nu[\tilde{N}_{\tau,a}] d_{KL}(\nu_a, \nu'_a) \geq \sup_{\mathcal{E}\in\mathcal{F}_\tau} kl(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E}))$$

Proof For all $a \in [d]$, we denote $(Y_{s,a})_{s\in[t]:a\in A_s}$ the sequence of $\tilde{N}_{t,a}$ i.i.d. samples observed for the arm a. By definition of L_τ, the fact that $d_{\mathrm{KL}}(\nu_a, \nu'_a) := \mathbb{E}_\nu\left[\ln\left(\frac{f_{\nu_a}(Y_{s,a})}{f_{\nu'_a}(Y_{s,a})}\right)\right]$ and applying Wald's lemma to L_τ, we obtain that: for all $a \in [d]$, $\mathbb{E}_\nu\left[L_\tau\right] = \sum_{a\in[d]} \mathbb{E}_\nu[\tilde{N}_{\tau,a}] d_{\mathrm{KL}}(\nu_a, \nu'_a)$. Combining this equation with Lemma 19 of Kaufmann et al. (2016) yields the desired result. ■

Theorem *For any δ-PAC strategy and any bandit ν,*

$$\frac{\mathbb{E}_\nu[\tau_\delta]}{\ln(1/(2.4\delta))} \geq D_\nu^{-1} \quad \text{and} \quad \limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\ln(1/\delta)} \geq D_\nu^{-1}$$

where the complexity D_ν*, inverse of a characteristic time, is defined by*

$$D_\nu := \max_{\tilde{w}\in\mathcal{S}_\mathcal{A}} \inf_{\lambda\in\Theta^{\complement}_{I^*(\mu)}} \langle \tilde{w}, d_{KL}(\mu,\lambda)\rangle$$

Proof Let $\delta \in (0,1)$, ν a bandit with parameter $\mu \in \mathcal{M}$ and consider a δ-PAC strategy. Let $\lambda \in \Theta^{\complement}_{I^*(\mu)}$ be the parameter of a bandit ν' with a unique correct answer $J \neq I^*(\mu)$. Let $\mathcal{E}_J := \{I_{\tau_\delta} = J\} \in \mathcal{F}_{\tau_\delta}$ be the event in which we recommend J instead of $I^*(\mu)$ at the stopping time. Since the strategy is δ-PAC, we have: $\mathbb{P}_\nu(\mathcal{E}_J) \leq \delta$ and $\mathbb{P}_{\nu'}(\mathcal{E}_J) \geq 1-\delta$. Therefore, we have:

$$\sup_{\mathcal{E}\in\mathcal{F}_{\tau_\delta}} \mathrm{kl}(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})) \geq \mathrm{kl}(\mathbb{P}_\nu(\mathcal{E}_J), \mathbb{P}_{\nu'}(\mathcal{E}_J)) \geq \mathrm{kl}(\delta, 1-\delta) \geq \ln(1/(2.4\delta))$$

The last inequality $\mathrm{kl}(\delta, 1-\delta) = \delta \ln\left(\frac{\delta}{1-\delta}\right) + (1-\delta)\ln\left(\frac{1-\delta}{\delta}\right) \geq \ln(1/(2.4\delta))$ was shown in Kaufmann et al. (2016). Combined with Lemma 5, we obtain: $\sum_{a\in[d]} \mathbb{E}_\nu[\tilde{N}_{\tau_\delta,a}] d_{\mathrm{KL}}(\mu_a, \lambda_a) \geq \ln(1/(2.4\delta))$ for all $\lambda \in \Theta^{\complement}_{I^*(\mu)}$. By construction, we have $\frac{\mathbb{E}_\nu[\tilde{N}_{\tau_\delta,a}]}{\mathbb{E}_\nu[\tau_\delta]} \in \mathcal{S}_\mathcal{A}$. Instead of considering a specific alternative bandit minimizing the lower bound, we combine all the inequalities. Taking the infimum:

$$\begin{aligned} \ln(1/(2.4\delta)) &\leq \mathbb{E}_\nu[\tau_\delta] \inf_{\lambda\in\Theta^{\complement}_{I^*(\mu)}} \sum_{a\in[d]} \frac{\mathbb{E}_\nu[\tilde{N}_{\tau_\delta,a}]}{\mathbb{E}_\nu[\tau_\delta]} d_{\mathrm{KL}}(\mu_a,\lambda_a) \\ &\leq \mathbb{E}_\nu[\tau_\delta] \sup_{\tilde{w}\in\mathcal{S}_\mathcal{A}} \inf_{\lambda\in\Theta^{\complement}_{I^*(\mu)}} \sum_{a\in[d]} \tilde{w}_a d_{\mathrm{KL}}(\mu_a,\lambda_a) = \mathbb{E}_\nu[\tau_\delta] D_\nu \end{aligned}$$

This concludes the proof of the finite-time lower bound. Taking the limit $\delta \to 0$ in the previous lower bound yields directly the asymptotic lower bound. ■

Appendix D. Proof of Theorem 2

The proof of Theorem 2 uses the deviation inequalities of Kaufmann and Koolen (2018), see Appendix D.1. The idea of the proof is similar to the proof of Proposition 21 in Kaufmann and Koolen (2018), as well as Theorem 2 in Shang et al. (2020).

Theorem *Let $\mathcal{M}$ be bounded. Regardless of the sampling rule, a strategy using the frequentist recommendation/stopping pair with the stopping threshold:*

$$\beta(t,\delta) := \begin{cases} 3d_0 \ln\left(1+\ln\left(\frac{tK}{d_0}\right)\right) + d_0\mathcal{T}\left(\frac{\ln\left(\frac{|\mathcal{I}|-1}{\delta}\right)}{d_0}\right) & \text{for (a)} \\ 2d_0 \ln\left(4+\ln\left(\frac{tK}{d_0}\right)\right) + d_0\mathcal{C}^{gG}\left(\frac{\ln\left(\frac{|\mathcal{I}|-1}{\delta}\right)}{d_0}\right) & \text{for (b)} \end{cases}$$

is δ-PAC. In the above, $d_0 := \max_{I,J\in\mathcal{I},J\neq I} |(I\setminus J)\cup(J\setminus I)|$, $\mathcal{T}$ and $\mathcal{C}^{gG}$ are the functions defined in Kaufmann and Koolen (2018), $\mathcal{C}^{gG}(x) \approx x+\ln(x)$ and $\mathcal{T}(x) \approx x+4\ln(1+x+\sqrt{2x})$ for $x \geq 5$.

Proof First let's show that $\tau_\delta < \infty$. We recall the following expressions: $I_t = \operatorname{argmax}_{I \in \mathcal{I}} \langle \mathbf{1}_I, \tilde{\mu}_{t-1} \rangle$ where $\tilde{\mu}_{t-1} \in \operatorname{argmin}_{\lambda \in \mathcal{M} \cap \mathcal{C}_t} \langle \tilde{N}_{t-1}, d_{\mathrm{KL}}(\mu_{t-1}, \lambda) \rangle$ and

$$\tau_\delta = \inf \left\{ t \in \mathbb{N} : \min_{J \in N(I_t)} \inf_{\lambda \in \bar{\Theta}^{I_t}_J} \langle \tilde{N}_{t-1}, d_{\mathrm{KL}}(\mu_{t-1}, \lambda) \rangle > \beta(t-1, \delta) \right\}$$

Let an arbitrary sampling rule, the set of arms sampled only a finite time, $\mathcal{U} := \{a \in [d] : \lim_{t \to \infty} \tilde{N}_{t,a} < +\infty\}$, and the limit of the empirical sampling rate, $\tilde{w}_\infty := \left(\lim_\infty \frac{\tilde{N}_{t,a}}{t} \right)_{a \in [d]}$. For all $a \in \mathcal{U}^{\complement}$, the law of large number proves that $\mu_t \to_\infty \mu_a$, while for all $a \in \mathcal{U}$, $\mu_t \to_\infty \tilde{\mu}_a \neq \mu_a$. Since it is a basic requirement for a sampling rule to predict the unique correct answer, we consider only sampling rules satisfying $\lim_\infty I_t = I^*(\mu)$. Since $\mu_t \to \mu$ implies $\tilde{\mu}_t \to \mu \in \mathcal{M}$ and $\lim_\infty I_t = I^*(\mu)$, this condition is weaker than assuming the convergence of the parameter μ_t towards the true parameter, which happens if $\mathcal{U} = \emptyset$. Let $T_0 \in \mathbb{N}$ such that for all $t \geq T_0$, $I_t = I^*(\mu)$. For $t \geq T_0$, the stopping condition rewrites as: $\inf_{\lambda \in \Theta^{\complement}_{I^*(\mu)}} \langle \frac{\tilde{N}_{t-1}}{t}, d_{\mathrm{KL}}(\mu_{t-1}, \lambda) \rangle > \frac{\beta(t-1,\delta)}{t}$. By continuity, dominated convergence (to invert $\lim$ and $\inf$ for $\mathcal{M}$ bounded), and using that $\beta(t, \cdot) \sim_\infty c_d \ln(\ln(t))$, taking the limit on both side yields: $\inf_{\lambda \in \Theta^{\complement}_{I^*(\mu)}} \sum_{a \in \mathcal{U}^{\complement}} \tilde{w}_{\infty,a} d_{\mathrm{KL}}(\mu_a, \lambda_a) \geq 0$.

By construction, we have $\tilde{w}_\infty \in \mathcal{S}_{\mathcal{A}}$, hence the left term is strictly positive if: for all $a \in [d]$ such that $\tilde{w}_{\infty,a} \neq 0$, we have $d_{\mathrm{KL}}(\mu_a, \lambda_a) \neq 0$. Since $d_{\mathrm{KL}}(\mu_a, \lambda_a) = 0$ if and only if $\mu_a = \lambda_a$, the fact that $\lambda \in \Theta^{\complement}_{I^*(\mu)}$ allows us to conclude that the inequality is strict. Therefore, there exists a finite time such that the stopping condition is met: $\tau_\delta < \infty$.

Second, let's show that $\mathbb{P}_\nu \left[I_{\tau_\delta} \neq I^*(\mu) \right] \leq \delta$. Let $\beta(t, \delta)$ be an arbitrary stopping threshold. Since $\{I_{\tau_\delta} \neq I^*(\mu)\} = \bigcup_{I \neq I^*(\mu)} \{\exists t \in \mathbb{N} : I_{t+1} = I \wedge \inf_{\lambda \in \Theta^{\complement}_{I_{t+1}}} \langle \tilde{N}_t, d_{\mathrm{KL}}(\mu_t, \lambda) \rangle > \beta(t, \delta)\}$, the union bound yields:

$$\mathbb{P}_\nu \left[I_{\tau_\delta} \neq I^*(\mu) \right] \leq \sum_{I \neq I^*(\mu)} \mathbb{P}_\nu \left[\exists t \in \mathbb{N} : \inf_{\lambda \in \Theta^{\complement}_I} \langle \tilde{N}_t, d_{\mathrm{KL}}(\mu_t, \lambda) \rangle > \beta(t, \delta) \right]$$

Let $I \neq I^*(\mu)$. Since $\Theta^{\complement}_I = \bigcup_{J \neq I} \bar{\Theta}^I_J$, we have $\inf_{\lambda \in \Theta^{\complement}_I} h(\lambda) = \min_{J \neq I} \inf_{\lambda \in \bar{\Theta}^I_J} h(\lambda) \leq \inf_{\lambda \in \bar{\Theta}^I_{I^*(\mu)}} h(\lambda)$. Using $h(\lambda) = \langle \tilde{N}_t, d_{\mathrm{KL}}(\mu_t, \lambda) \rangle$, we obtain:

$$\inf_{\lambda \in \Theta^{\complement}_I} \langle \tilde{N}_t, d_{\mathrm{KL}}(\mu_t, \lambda) \rangle \leq \inf_{\lambda \in \bar{\Theta}^I_{I^*(\mu)}} \langle \tilde{N}_t, d_{\mathrm{KL}}(\mu_t, \lambda) \rangle \leq \sum_{a \in I^*(\mu) \triangle I} \tilde{N}_{t,a} d_{\mathrm{KL}}(\mu_{t,a}, \mu_a)$$

The last inequality is obtained by considering a parameter λ defined as: $\lambda_a = \mu_a$ when $a \in I^*(\mu) \triangle I$ and $\lambda_a = \mu_{t,a}$ else. Since $\mu \in \bar{\Theta}^I_{I^*(\mu)}$, we have that $\langle \mathbf{1}_{I^*(\mu)} - \mathbf{1}_I, \lambda_a \rangle = \langle \mathbf{1}_{I^*(\mu)} - \mathbf{1}_I, \mu_a \rangle \geq 0$. Therefore $\lambda \in \bar{\Theta}^I_{I^*(\mu)}$, hence it is a valid parameter. The upper bound rewrites as:

$$\mathbb{P}_\nu \left[I_{\tau_\delta} \neq I^*(\mu) \right] \leq \sum_{I \neq I^*(\mu)} \mathbb{P}_\nu \left[\exists t \in \mathbb{N} : \sum_{a \in I^*(\mu) \triangle I} \tilde{N}_{t,a} d_{\mathrm{KL}}(\mu_{t,a}, \mu_a) > \beta(t, \delta) \right]$$

To conclude, we need to control the deviation of the *self-normalized sums*, $\sum_{a \in S} \tilde{N}_{t,a} d_{\mathrm{KL}}(\mu_{t,a}, \mu_a)$ for $S \subset [d]$, thanks to concentration inequalities, which are uniform in time. The concentration inequality depends on the setting (a) sub-Gaussian bandit or (b) Gaussian bandit. Moreover, we want an expression for $\beta(t, \delta)$ which doesn't depend on the answer I or the empirical count $\tilde{N}_t$. Let $d_0 := \max_{I,J \in \mathcal{I}, J \neq I} |I \triangle J|$. Combining the concavity of $x \mapsto \ln(c + \ln(x))$ and the fact that $\sum_{a \in I^*(\mu) \triangle I*} \tilde{N}_{t,a} \leq \sum_{a \in [d]} \tilde{N}_{t,a} \leq tK$, we obtain:

$$\sum_{a \in I^*(\mu) \triangle I} \ln(c + \ln(\tilde{N}_{t,a})) \leq |I^*(\mu) \triangle I| \ln \left(c + \ln \left(\frac{\sum_{a \in I^*(\mu) \triangle I} \tilde{N}_{t,a}}{|I^*(\mu) \triangle I|} \right) \right)$$
$$\leq |I^*(\mu) \triangle I| \ln \left(c + \ln \left(\frac{tK}{|I^*(\mu) \triangle I|} \right) \right) \leq d_0 \ln \left(c + \ln \left(\frac{tK}{d_0} \right) \right)$$

The last inequality is due to the fact that $h_t(x) = x \ln \left(c_0 + \ln \left(\frac{t}{x} \right) \right)$ is increasing on $]0, d]$ when $t \gg d$. The higher t is, the longer h_t is increasing. Numerically, h_1 is increasing till 424. Let $\mathcal{T}$ and $\mathcal{C}^{gG}$ the functions defined in Kaufmann and Koolen (2018). Since $x \mapsto x\mathcal{T}\left(\frac{c}{x}\right)$ and $x \mapsto x\mathcal{C}^{gG}\left(\frac{c}{x}\right)$ are increasing (Appendix D.1), we obtain:

$$|I^*(\mu) \triangle I| \mathcal{T} \left(\frac{\ln \left(\frac{|\mathcal{I}|-1}{\delta} \right)}{|I^*(\mu) \triangle I|} \right) \leq d_0 \mathcal{T} \left(\frac{\ln((|\mathcal{I}| - 1)/\delta)}{d_0} \right)$$
$$|I^*(\mu) \triangle I| \mathcal{C}^{gG} \left(\frac{\ln \left(\frac{|\mathcal{I}|-1}{\delta} \right)}{|I^*(\mu) \triangle I|} \right) \leq d_0 \mathcal{C}^{gG} \left(\frac{\ln((|\mathcal{I}| - 1)/\delta)}{d_0} \right)$$

Combining those inequalities with $c = 1$ for (a) and $c = 4$ for (b), we obtain that:

$$3 \sum_{a \in I^*(\mu) \triangle I} \ln(1 + \ln(\tilde{N}_{t,a})) + |I^*(\mu) \triangle I| \mathcal{T} \left(\frac{\ln \left(\frac{|\mathcal{I}|-1}{\delta} \right)}{|I^*(\mu) \triangle I|} \right) \leq 3d_0 \ln \left(1 + \ln \left(\frac{tK}{d_0} \right) \right) + d_0 \mathcal{T} \left(\frac{\ln((|\mathcal{I}| - 1)/\delta)}{d_0} \right)$$

$$2 \sum_{a \in I^*(\mu) \triangle I} \ln(4 + \ln(\tilde{N}_{t,a})) + |I^*(\mu) \triangle I| \mathcal{C}^{gG} \left(\frac{\ln \left(\frac{|\mathcal{I}|-1}{\delta} \right)}{|I^*(\mu) \triangle I|} \right) \leq 2d_0 \ln \left(4 + \ln \left(\frac{tK}{d_0} \right) \right) + d_0 \mathcal{C}^{gG} \left(\frac{\ln((|\mathcal{I}| - 1)/\delta)}{d_0} \right)$$

Let $\beta(t, \delta)$ be the stopping threshold defined as:

$$\beta(t, \delta) := \begin{cases} 3d_0 \ln \left(1 + \ln \left(\frac{tK}{d_0} \right) \right) + d_0 \mathcal{T} \left(\frac{\ln \left(\frac{|\mathcal{I}|-1}{\delta} \right)}{d_0} \right) & \text{for (a)} \\ 2d_0 \ln \left(4 + \ln \left(\frac{tK}{d_0} \right) \right) + d_0 \mathcal{C}^{gG} \left(\frac{\ln \left(\frac{|\mathcal{I}|-1}{\delta} \right)}{d_0} \right) & \text{for (b)} \end{cases}$$

Since $x \leq y$ implies $\mathbb{P}[X > y] \leq \mathbb{P}[X > x]$, using Theorem 14 in Kaufmann and Koolen (2018) and Corollary 10 in Kaufmann and Koolen (2018) (Appendix D.1) yields the result:

$$\mathbb{P}_\nu \left[I_{\tau_\delta} \neq I^*(\mu) \right] \leq \sum_{I \neq I^*(\mu)} \frac{\delta}{|\mathcal{I}| - 1} = \delta$$

Therefore, we conclude that a strategy using the frequentist recommendation/stopping pair is δ-PAC. ■

D.1. Deviation Inequalities

The deviation inequality for sub-Gaussian bandit is rewritten in Appendix D.1.1, while the deviation inequality for Gaussian bandit is presented in Appendix D.1.2.

D.1.1. SUB-GAUSSIAN BANDIT

Theorem 14 in Kaufmann and Koolen (2018) holds for sub-Gaussian bandits.

Lemma *[Theorem 14 in Kaufmann and Koolen (2018)] Let $\delta > 0$, ν be independent one-parameter exponential families with mean μ and $S \subset [d]$. Then we have,*

$$\mathbb{P}_\nu\left[\exists t \in \mathbb{N} : \sum_{a\in S} \tilde{N}_{t,a} d_{KL}(\mu_{t,a}, \mu_a) \geq \sum_{a \in S} 3\ln(1+\ln(\tilde{N}_{t,a})) + |S|\mathcal{T}\left(\frac{\ln\left(\frac{1}{\delta}\right)}{|S|}\right)\right] \leq \delta$$

where $\mathcal{T} : \mathbb{R}^+ \to \mathbb{R}^+$ is such that $\mathcal{T}(x) = 2\tilde{h}_{3/2}\left(\frac{h^{-1}(1+x)+\ln\left(\frac{\pi^2}{3}\right)}{2}\right)$ with:

$$\forall u \geq 1, \quad h(u) = u - \ln(u)$$
$$\forall z \in [1,e], \forall x \geq 0, \quad \tilde{h}_z(x) = \begin{cases} \exp\left(\frac{1}{h^{-1}(x)}\right) h^{-1}(x) & \textit{if } x \geq h^{-1}\left(\frac{1}{\ln(z)}\right) \\ z(x - \ln(\ln(z))) & \textit{else} \end{cases}$$

D.1.2. GAUSSIAN BANDIT

Corollary 10 in Kaufmann and Koolen (2018) holds for Gaussian bandits. Lemma 6 gathers some properties of $\mathcal{C}^{g_G}$.

Lemma *[Corollary 10 in Kaufmann and Koolen (2018)] Let $\delta > 0$, ν be a family of independent Gaussian with mean μ and $S \subset [d]$. Then we have,*

$$\mathbb{P}_\nu\left[\exists t \in \mathbb{N} : \sum_{a\in S} \tilde{N}_{t,a} d_{KL}(\mu_{t,a}, \mu_a) \geq \sum_{a \in S} 2\ln(4+\ln(\tilde{N}_{t,a})) + |S|\mathcal{C}^{g_G}\left(\frac{\ln\left(\frac{1}{\delta}\right)}{|S|}\right)\right] \leq \delta$$

where $\mathcal{C}^{g_G}(x) = \min_{y \in]1/2,1]} \frac{g_G(y)+x}{y}$ with $g_G :]1/2, 1] \to \mathbb{R}$ such that $g_G(y) = 2y - 2y\ln(4y) + \ln(\zeta(2y)) - \frac{1}{2}\ln(1-y)$

Lemma 6 *The g_G is positive on $]1/2, 1[$ and satisfies $g_G(y) \to_{y \to \{1/2,1\}} +\infty$. The function $x \mapsto x C^{g_G}\left(\frac{c}{x}\right)$ is increasing.*

Proof Since $\zeta(1) = \lim_{n\to\infty} \sum_{s=1}^{n} \frac{1}{s} = +\infty$ and $g_G(y) \sim_{1/2} \ln(\zeta(2y))$, we have $g_G(y) \to_{y\to 1/2} +\infty$. Since $\zeta(2) = \frac{\pi^2}{6}$ and $g_G(y) \sim_1 -\frac{1}{2}\ln(1-y)$, we have $g_G(y) \to_{y \to 1} +\infty$.

Let $y \in]1/2, 1[$ and $h(y) = 2y - 2y\ln(4y) - \frac{1}{2}\ln(1-y)$. We have $h'(y) = \frac{1}{2}\frac{1}{1-y} - 2\ln(4y)$, hence $h'(y) \geq 0$ if and only if $1 \geq 4(1-y)\ln(4y)$. Numerically, this condition is always true, hence h is increasing. Since $h(y) \geq h(1/2) = 1$, we obtain $g_G(y) \geq 1 + \ln(\zeta(2y))$. Using that $\ln(\zeta(2)) = \ln\frac{\pi^2}{6} > 0$ and $x \mapsto \zeta(x)$ decreasing on $]1/2, 1]$, we obtain that $\ln(\zeta(2y)) \geq \ln(\zeta(2))$. Therefore we can conclude that $\forall y \in [1/2, 1[, g_G(y) \geq 1 \geq 0$.

Since $xC^{g_G}\left(\frac{c}{x}\right) = \min_{y \in]1/2,1]} \frac{xg_G(y)+c}{y}$ and g_G is positive on $]1/2, 1[$, we obtain directly that $x \mapsto xC^{g_G}\left(\frac{c}{x}\right)$ is increasing. ■

Appendix E. Optimistic Reward

In Appendix E.1, we prove an upper and lower bound on the optimistic reward (Lemma 7). The properties of r_t for Gaussian bandit are studied in Appendix E.2

E.1. Bounds on $\|r_t\|_\infty$

Due to the boundedness assumption, Lemma 7 below shows that the optimistic reward r_t is almost bounded. When an arm a is sampled less than a logarithmic number of times, $r_{t,a}$ becomes large enough to stir the sampling towards actions containing it.

Lemma 7 *Let $\mathcal{M}$ bounded. Under the event $\{\mu \in \mathcal{C}_t\}$, we have:*

$$\max\left\{\frac{f(t-1)}{\min_{a\in[d]} \tilde{N}_{t-1,a}}, \epsilon_\nu\right\} \leq \|r_t\|_\infty \leq \max\left\{\frac{f(t-1)}{\min_{a\in[d]} \tilde{N}_{t-1,a}}, D_{\mathcal{M}}\right\}$$

The upper bound is a consequence of the boundedness assumption, $D_{\mathcal{M}} := \sup_{(\phi,\lambda)\in\mathcal{M}^2} \|d_{\mathrm{KL}}(\phi,\lambda)\|_1$. Since μ has a unique correct answer, the lower bound stems from the Chernoff information lower bound ϵ_ν which holds for both (a) and (b): there exists $\epsilon_\nu > 0$ such that,

$$\forall \lambda \in \Theta^{\complement}_{I^*(\mu)}, \exists a \in [d], \quad \mathrm{ch}(\lambda_a, \mu_a) \geq \epsilon_\nu$$

where $\mathrm{ch}(x, y) := \inf_{u\in\Theta}\left(d_{\mathrm{KL}}(u, x) + d_{\mathrm{KL}}(u, y)\right)$.

Before proving Lemma 7, let's first prove that ϵ_ν exists. For (a) sub-Gaussian, we have $d_{\mathrm{KL}}(u, x) \geq \frac{(u-x)^2}{2\sigma_a^2}$, hence the chernoff information of the setting (a) is greater than the one for setting (b). For (b) Gaussian, we have: $\mathrm{ch}(x, y) = \frac{1}{2\sigma_a^2}\inf_{u\in\Theta}((u-x)^2 + (u-y)^2) = \frac{(x-y)^2}{8\sigma_a^2}$.

Let $\mu \in \mathcal{M}$ and $\lambda \in \Theta^{\complement}_{I^*(\mu)}$. Since $\mu \in \Theta_{I^*(\mu)}$, which is an open set, the euclidean distance to $\Theta^{\complement}_{I^*(\mu)}$ is strictly positive: there exists $a \in [d]$ such that $|\lambda_a - \mu_a| \geq \epsilon > 0$. Since $\mathrm{ch}(x, y) \geq \frac{(x-y)^2}{8\sigma_a^2}$, we can conclude for both (a) and (b) that there exists $\epsilon_\nu > 0$ as defined above.

Next, we prove the Lemma 7 itself.

Proof For all $a \in [d]$, let $\phi_{t,a} \in \operatorname{argmax}_{\phi\in\{\alpha_{t,a},\beta_{t,a}\}} d_{\mathrm{KL}}(\phi, \lambda_{t,a})$, the optimistic mean parameter. By convexity of $x \mapsto d_{\mathrm{KL}}(x, y)$, we have: $\max_{\phi\in[\alpha_{t,a},\beta_{t,a}]} d_{\mathrm{KL}}(\phi, \lambda_{t,a}) = d_{\mathrm{KL}}(\phi_{t,a}, \lambda_{t,a})$. The optimistic reward rewrites as: $r_{t,a} = \max\left\{\frac{f(t-1)}{\tilde{N}_{t-1,a}}, d_{\mathrm{KL}}(\phi_{t,a}, \lambda_{t,a})\right\}$ for all $a \in [d]$. Using that $\max_{a\in[d]}\max\{x_a, y_a\} = \max\{\max_{a\in[d]} x_a, \max_{a\in[d]} y_a\}$, we obtain:

$$\|r_t\|_\infty = \max\left\{\frac{f(t-1)}{\min_{a\in[d]} \tilde{N}_{t-1,a}}, \max_{a\in[d]} d_{\mathrm{KL}}(\phi_{t,a}, \lambda_{t,a})\right\}$$

Since $\mathcal{M}$ is bounded, $D_{\mathcal{M}} = \sup_{(\phi,\lambda)\in\mathcal{M}^2} \|d_{\mathrm{KL}}(\phi,\lambda)\|_1$ and $\|\cdot\|_\infty \leq \|\cdot\|_1$, we obtain the desired upper bound.

Due to concentration events, with high probability we have $\mu \in \mathcal{C}_t = \times_{a\in[d]}[\alpha_{t,a}, \beta_{t,a}]$. Assume $\{\mu \in \mathcal{C}_t\}$ holds. Combining the definition of $\phi_{t,a}$ and $\lambda_t \in \partial\Theta_{I_t}$, we obtain: $d_{\mathrm{KL}}(\phi_{t,a}, \lambda_{t,a}) \geq d_{\mathrm{KL}}(\mu_a, \lambda_{t,a}) \geq \min_{\lambda\in\partial\Theta_{I_t}} d_{\mathrm{KL}}(\mu_a, \lambda_a) \geq \min_{I\in\mathcal{I},\lambda\in\partial\Theta_I} d_{\mathrm{KL}}(\mu_a, \lambda_a)$. The function $y \mapsto d_{KL}(x,y)$ is not convex in general and is minimized in $x = y$. Hence, the geometry of the cells yields:

$$\min_{I\in\mathcal{I},\lambda\in\partial\Theta_I} d_{\mathrm{KL}}(\mu_a, \lambda_a) = \min_{\lambda\in\partial\Theta_{I^*(\mu)}} d_{\mathrm{KL}}(\mu_a, \lambda_a) \geq \min_{\lambda\in\Theta^{\complement}_{I^*(\mu)}} d_{\mathrm{KL}}(\mu_a, \lambda_a)$$

Since $\min\{d_{\mathrm{KL}}(y,x), d_{\mathrm{KL}}(x,y)\} \geq \mathrm{ch}(x,y)$, the previous inequalities yield: $d_{\mathrm{KL}}(\phi_{t,a}, \lambda_{t,a}) \geq \min_{\lambda\in\Theta^{\complement}_{I^*(\mu)}} \mathrm{ch}(\lambda_a, \mu_a)$ for all $a \in [d]$. Using the chernoff information lower bound and taking the maximum over $a \in [d]$, we conclude that: $\|r_t\|_\infty \geq \max\left\{\frac{f(t-1)}{\min_{a\in[d]} \tilde{N}_{t-1,a}}, \epsilon_\nu\right\}$. ■

E.2. Gaussian Bandit

For Gaussian bandit, we have $d_{\mathrm{KL}}(x,y) = \frac{(x-y)^2}{2\sigma_a^2}$. Direct computations yield: $\alpha_{t,a} = \mu_{t-1,a} - \sqrt{\frac{2f(t-1)\sigma_a^2}{\tilde{N}_{t-1,a}}}$ and $\beta_{t,a} = \mu_{t-1,a} + \sqrt{\frac{2f(t-1)\sigma_a^2}{\tilde{N}_{t-1,a}}}$. As a consequence of d_{KL} and $[\alpha_{t,a}, \beta_{t,a}]$ being symmetric, Lemma 8 shows that the clipping $\frac{f(t-1)}{\tilde{N}_{t-1,a}}$ is superfluous.

Lemma 8 *Let ν be independent Gaussian and $\lambda \in \mathcal{M}$. Then, for all $a \in [d]$,*

$$\phi_{t,a} = \alpha_{t,a}\mathbf{1}_{\lambda_a \geq \mu_{t-1,a}} + \beta_{t,a}(1 - \mathbf{1}_{\lambda_a \geq \mu_{t-1,a}})$$
$$d_{KL}(\phi_{t,a}, \lambda_a) = \frac{(\mu_{t-1,a} - \lambda_a)^2}{2\sigma_a^2} + \frac{f(t-1)}{\tilde{N}_{t-1,a}} + \sqrt{\frac{2f(t-1)}{\sigma_a^2 \tilde{N}_{t-1,a}}}|\mu_{t-1,a} - \lambda_a| \geq \frac{f(t-1)}{\tilde{N}_{t-1,a}}$$

where $\phi_{t,a} = \mathrm{argmax}_{\phi\in\{\alpha_{t,a},\beta_{t,a}\}} \frac{(\phi-\lambda_a)^2}{2\sigma_a^2}$.

Proof Let $\lambda \in \mathcal{M}$. Let $\phi_{t,a} = \mathrm{argmax}_{\phi\in\{\alpha_{t,a},\beta_{t,a}\}} \frac{(\phi-\lambda_a)^2}{2\sigma_a^2}$ for all $a \in [d]$. Assume $\lambda_a \geq \mu_{t-1,a}$. Since $\lambda_a \geq \mu_{t-1,a} \geq \alpha_{t,a}$ and $\mu_{t-1,a} \leq \beta_{t,a}$, we have $(\alpha_{t,a} - \lambda_a)^2 \geq (\beta_{t,a} - \lambda_a)^2$. Therefore $\phi_{t,a} = \alpha_{t,a}$. Assume $\lambda_a < \mu_{t-1,a}$. Since $\lambda_a < \mu_{t-1,a} \leq \beta_{t,a}$ and $\mu_{t-1,a} \geq \alpha_{t,a}$, we have $(\alpha_{t,a} - \lambda_a)^2 \leq (\beta_{t,a} - \lambda_a)^2$. This concludes the first statement.

Due to the explicit formulas for $\phi_{t,a}$, $\alpha_{t,a}$ and $\beta_{t,a}$, we have $(\phi_{t,a} - \mu_{t-1,a})(\mu_{t-1,a} - \lambda_a) = \sqrt{\frac{2f(t-1)\sigma_a^2}{\tilde{N}_{t-1,a}}}|\mu_{t-1,a} - \lambda_a| \geq 0$, hence we conclude:

$$\frac{(\phi_{t,a} - \lambda_a)^2}{2\sigma_a^2} = \frac{(\mu_{t-1,a} - \lambda_a)^2}{2\sigma_a^2} + \frac{f(t-1)}{\tilde{N}_{t-1,a}} + \sqrt{\frac{2f(t-1)}{\sigma_a^2 \tilde{N}_{t-1,a}}}|\mu_{t-1,a} - \lambda_{t,a}| \geq \frac{f(t-1)}{\tilde{N}_{t-1,a}}$$

■

Appendix F. Learner's Cumulative Regret

In the Appendix F, we show upper bounds on the cumulative regret for the different learners: Hedge in Lemma 9, AdaHedge in Lemma 10, OFW in Lemma 11 and LLOO in Lemma 12.

F.1. Learner on the Simplex

The extended optimistic reward is defined as: $U_{t,A} := \langle \mathbf{1}_A, r_t \rangle$ for all $A \in \mathcal{A}$. The cumulative regret rewrites as:

$$R_t^A = \sum_{s=1}^{t} b_s \langle w_s, l_s \rangle - \min_{A \in \mathcal{A}} \sum_{s=1}^{t} b_s l_{s,A} \leq \left(\sum_{s=1}^{t} \langle w_s, l_s \rangle - \min_{A \in \mathcal{A}} \sum_{s=1}^{t} l_{s,A} \right) \max_{s \leq t} b_s$$

where $l_{t,A} = \frac{\|U_t\|_\infty - U_{t,A}}{b_t} \in [0,1]$ and $b_t = \|U_t\|_\infty - \min_{A \in \mathcal{A}} U_{t,A}$ is the scale of the loss at time t. Since U_t is positive, $\tilde{N}_{t-1,a} \geq 1$ and $f(t) = \Omega(\ln(t))$, we obtain that: $b_t \leq \|U_t\|_\infty \leq d\|r_t\|_\infty$ and $\frac{f(t-1)}{\min_{a \in [d]} \tilde{N}_{t-1,a}} \leq f(t-1) = O(\ln(t))$. Therefore, Lemma 7 yields that: $\max_{s \leq t} b_s = O\left(\ln(t)\right)$.

Hedge Using Corollary 3 in Cesa-Bianchi et al. (2005), we obtain that Hedge has optimal cumulative regret (Lemma 9).

Lemma 9 *Hedge satisfies*

$$R_t^{Hedge} \leq \left(4\sqrt{\frac{L_t^*(t - L_t^*)}{t}} \ln(|\mathcal{A}|) + 39 \max\{1, \ln(|\mathcal{A}|)\} \right) \max_{s \leq t} b_s = O\left(\ln(t)\sqrt{t}\right)$$

where $L_t^* = \min_{A \in \mathcal{A}} \sum_{s=1}^{t} l_{s,A}$.

Proof For scaled losses in $[0,1]$, Corollary 3 in Cesa-Bianchi et al. (2005) yields that Hedge's cumulative regret R_t' satisfies: $R_t' \leq 4\sqrt{\frac{L_t^*(\sigma t - L_t^*)}{t}} \ln(|\mathcal{A}|) + 39\sigma \max\{1, \ln(|\mathcal{A}|)\}$ where σ is the range of observed loss. By definition of l_t, we have $\sigma = 1$. Factorizing the maximum of the scale of the loss $\max_{s \leq t} b_s$, we obtain the upper bound on R_t^{Hedge}. In the worst case this algorithm has a regret of order $O(\sqrt{t})$, but it performs much better when the loss of the best expert L_t^* is close to either 0 or t. Combined with $\max_{s \leq t} b_s = O\left(\ln(t)\right)$, this concludes the proof. ■

AdaHedge Using the results of Rooij et al. (2014), we obtain that AdaHedge has optimal cumulative regret (Lemma 10).

Lemma 10 *AdaHedge satisfies*

$$R_t^{AdaHedge} \leq \sqrt{\sum_{s \leq t} b_s^2 \ln(|\mathcal{A}|)} + \left(\frac{4}{3} \ln(|\mathcal{A}|) + 2 \right) \max_{s \leq t} b_s = O\left(\ln(t)\sqrt{t}\right)$$

Proof For scaled losses in $[0,1]$, Theorem 6 in Rooij et al. (2014) yields that AdaHedge's cumulative regret R_t' satisfies: $R_t' \leq 2\sqrt{V_t \ln(|\mathcal{A}|)} + \frac{4}{3} \ln(|\mathcal{A}|) + 2$ where $V_t = \sum_{s \in [t]} v_s$ with $v_s = \sum_{A \in \mathcal{A}} w_{s,A}(l_{s,A} - \langle w_s, l_s \rangle)^2$. We have $v_s \leq \|l_s - \langle w_s, l_s \rangle\|_\infty^2 = \frac{\|\langle w_s, U_s \rangle - U_s\|_\infty^2}{b_s^2} \leq \frac{b_s^2}{\sigma^2}$ where

$\sigma = \max_{s \leq t} b_s$. The upper bound on R'_t rewrites as: $R'_t \leq \frac{1}{\sigma}\sqrt{\sum_{s \leq t} b_s^2 \ln(|\mathcal{A}|)} + \frac{4}{3}\ln(|\mathcal{A}|) + 2$. Theorem 16 in Rooij et al. (2014) yields that $R_t^A = \sigma R'_t$. Therefore, we conclude that:

$$\begin{aligned} R_t^{AdaHedge} &\leq \sqrt{\sum_{s \leq t} b_s^2 \ln(|\mathcal{A}|)} + \left(\frac{4}{3}\ln(|\mathcal{A}|) + 2\right)\max_{s \leq t} b_s \\ &\leq \left(\sqrt{t \ln(|\mathcal{A}|)} + \left(\frac{4}{3}\ln(|\mathcal{A}|) + 2\right)\right)\max_{s \leq t} b_s \end{aligned}$$

Combined with $\max_{s \leq t} b_s = O\left(\ln(t)\right)$, this concludes the proof. ■

F.2. Learner on the Transformed Simplex

We recall the cumulative regret is defined as: $R_t^A = \max_{A \in \mathcal{A}} \sum_{s=1}^t \langle \mathbf{1}_A, r_s \rangle - \sum_{s=1}^t \langle \tilde{w}_s, r_s \rangle$. For the same reasons as in Appendix F.1 Lemma 7 yields: $\max_{s \leq t} \|r_s\|_2 = O(\ln(t))$.

OFW Slightly adapting the results of Hazan and Kale (2012b), we obtain that OFW has an upper bound on the cumulative regret in $O\left(\ln(t)^2 t^{3/4}\right)$ (Lemma 11). This is in general suboptimal for the online linear optimization setting.

Lemma 11 *OFW satisfies*

$$R_t^{OFW} \leq 18(2 + \max_{s \leq t} \|r_s\|_2)^2 \mathit{diam}(\mathcal{S}_{\mathcal{A}}) t^{3/4} + 3\mathit{diam}(\mathcal{S}_{\mathcal{A}}) t^{3/4} = O\left(\ln(t)^2 t^{3/4}\right)$$

Proof OFW described in Section 3.4 is exactly the algorithm used in the proof of Theorem 4.4 in Hazan and Kale (2012b), which is a result for adversarial cost functions. In their notations, the Lipschitz constant L satisfies: $L = \|r_t\|_2$. In order to conserve the anytime property of OFW, we use a different σ_t which is independent of L, $\sigma_t = \frac{1}{\text{diam}(\mathcal{S}_{\mathcal{A}})} t^{-1/4}$. The decrease in $t^{-1/4}$ is optimal. This modification doesn't change the idea of the proof and impact only the final bound by a multiplicative factor, $\max_{s \leq t} \|r_s\|_2$. A close examination of its proof shows that Theorem 3.1 in Hazan and Kale (2012b) still holds for time dependent Lipschitz constant L_t. Therefore, we follow the proof of Theorem 4.4 and apply Theorem 3.1 for $\hat{f}_t(x) = \langle l_t, x \rangle + \frac{1}{\text{diam}(\mathcal{S}_{\mathcal{A}})} t^{-1/4} \|x - x_1\|_2^2$. The exact same steps and using that $L_s \leq \max_{s \leq t} \|r_s\|_2$ for all $s \in [t]$ yield that:

$$\begin{aligned} R_t^{OFW} &\leq 18(\max_{s \leq t} \|r_s\|_2 + 2)\text{diam}(\mathcal{S}_{\mathcal{A}}) t^{3/4} \max_{s \leq t} \|r_s\|_2 + 3\text{diam}(\mathcal{S}_{\mathcal{A}}) t^{3/4} \\ &\leq 18(2 + \max_{s \leq t} \|r_s\|_2)^2 \text{diam}(\mathcal{S}_{\mathcal{A}}) t^{3/4} + 3\text{diam}(\mathcal{S}_{\mathcal{A}}) t^{3/4} \end{aligned}$$

Combined with $\max_{s \leq t} \|r_s\|_2 = O(\ln(t))$, this concludes the proof. ■

LLOO Using Theorem 3 in Garber and Hazan (2013), we obtain that OFW has optimal cumulative regret (Lemma 12).

Lemma 12 *Let T be the horizon and $\mu_{\mathcal{A}}$, defined in Garber and Hazan (2013). With $\gamma_{\mathcal{A}} = (3d\mu_{\mathcal{A}}^2)^{-1}$, $\eta_{\mathcal{A},T} = \frac{\mathit{diam}(\mathcal{S}_{\mathcal{A}})}{18\mu_{\mathcal{A}}\sqrt{dT}\max_{t \leq T}\|r_t\|_2}$ and $M_{\mathcal{A},T} = \min\left\{\mu_{\mathcal{A}}^2 \frac{d}{\sqrt{T}}\left(1 + \frac{1}{18d\mu_{\mathcal{A}}^2}\right), 1\right\}$, LLOO satisfies:*

$$R_t^{LLOO} = O\left(\mathit{diam}(\mathcal{S}_{\mathcal{A}})\mu_{\mathcal{A}}\sqrt{dt}\max_{s \leq t}\|r_s\|_2\right) = O\left(\ln(t)\sqrt{t}\right)$$

Proof Let T be the horizon and $\mu_{\mathcal{A}}$ as defined in Garber and Hazan (2013) (see Appendix I.1 for an explicit formula), which depends on $\mathcal{S}_{\mathcal{A}}$. For a non strongly convex function $\sigma = 0$, LLOO described in Section 3.4 is exactly the combination of Algorithm 5 and Algorithm 4 in Garber and Hazan (2013). The re-organization highlights the similarities with OFW. As parameters for the algorithm, we use the theoretically licensed: $\gamma_{\mathcal{A}} = (3d\mu_{\mathcal{A}}^2)^{-1}$, $\eta_{\mathcal{A},T} = \frac{\text{diam}(\mathcal{S}_{\mathcal{A}})}{18\mu_{\mathcal{A}}\sqrt{dT}\max_{t\leq T}\|r_t\|_2}$ and $M_{\mathcal{A},T} = \min\left\{\mu_{\mathcal{A}}^2\frac{d}{\sqrt{T}}\left(1+\frac{1}{18d\mu_{\mathcal{A}}^2}\right), 1\right\}$. Theorem 3 in Garber and Hazan (2013) yields that: $R_t^{LLOO} = O\left(\text{diam}(\mathcal{S}_{\mathcal{A}})\mu_{\mathcal{A}}\sqrt{dt}\max_{s\leq t}\|r_s\|_2\right) = O\left(\ln(t)\sqrt{t}\right)$. Combined with $\max_{s\leq t}\|r_s\|_2 = O(\ln(t))$, this concludes the proof. ■

Appendix G. Proof of Theorem 3

In Appendix G.1, we prove the preliminary Lemma 4. The saddle-point property of the algorithm $\mathcal{A}_I^A$ associated to I is proven in Appendix G.2. In Appendix G.3, we lower and upper bound the number of times when the candidate answer is correct. Combining them yields the definition of $T_0(\delta)$ and concludes the proof of Theorem 3. Technical arguments with respect to C-Tracking and concentration events are proven in Appendix G.5.

G.1. Proof of Lemma 4

Let $t_b := t^{1/(1+b)}$, with $b > 0$, and $(\mathcal{E}_t)_{t\geq 1}$ be a sequence of concentrations events for the exploration bonus f with parameters b and $c > 0$: for all $t \geq 1$,

$$\mathcal{E}_t := \left\{\forall s \leq t, \forall a \in [d], \quad \tilde{N}_{s,a} d_{\text{KL}}(\mu_{s,a}, \mu_a) \leq f\left(t_b\right)\right\} \tag{1}$$

where $f(t) = \overline{W}((1+c)(1+b)\ln(t))$ with $c > 0$, $b > 0$ and $\overline{W}(x) \approx x + \ln(x)$. More precisely, for $x \geq 1$, $\overline{W}(x) = -W_{-1}(-e^{-x})$ where W_{-1} denotes the negative branch of the Lambert W function. This sequence is theoretically validated due to Lemmas 5 and 6 in Degenne et al. (2019).

Lemma *[Lemmas 5 and 6 in Degenne et al. (2019)] Let $(Y_{s,a})_{s\in[t]}$ be i.i.d random variables in a canonical one-parameter exponential family with mean μ_a. Then, for $\alpha > 0$,*

$$\mathbb{P}_\nu\left[\exists s \leq t, d_{KL}\left(\frac{1}{s}\sum_{r=1}^{s} Y_{s,a}, \mu_a\right) \geq \frac{\alpha}{s}\right] \leq 2e\ln(t)e^{-(\alpha-\ln(\alpha))}$$

For independent ν and $(\mathcal{E}_t)_{t\geq 1}$ defined in Equation 1, we obtain:

$$\forall t \geq 3, \mathbb{P}_\nu\left[\mathcal{E}_t^{\complement}\right] \leq 2ed\frac{\ln(t)}{t^{1+c}} \quad \textit{and} \quad \sum_{t>T_0(\delta)} \mathbb{P}_\nu\left[\mathcal{E}_t^{\complement}\right] \leq \frac{2ed}{c^2}$$

Lemma *Let $(\mathcal{E}_t)_{t\geq 1}$ be a sequence of concentrations events for the exploration bonus f with parameters $c > 0$ and $b > 0$: for all $t \geq 1$,*

$$\mathcal{E}_t := \left\{\forall s \leq t, \forall a \in [d], \quad \tilde{N}_{s,a} d_{KL}(\mu_{s,a}, \mu_a) \leq f\left(t^{1/(1+b)}\right)\right\}$$

Suppose that there exists $T_0(\delta) \in \mathbb{N}$ such that for all $t > T_0(\delta)$, $\mathcal{E}_t \subset \{\tau_\delta \leq t\}$. Then

$$\mathbb{E}_\nu[\tau_\delta] \leq T_0(\delta) + \sum_{t > T_0(\delta)} \mathbb{P}_\nu\left[\mathcal{E}_t^{\complement}\right] \quad \textit{where} \quad \sum_{t > T_0(\delta)} \mathbb{P}_\nu\left[\mathcal{E}_t^{\complement}\right] \leq \frac{2ed}{c^2}$$

Proof First, let's prove the upper bound for an arbitrary sequence of concentrations events $(\mathcal{E}_t)_{t\geq 1}$ satisfying: there exists $T_0(\delta) \in \mathbb{N}$ such that for $t > T_0(\delta)$, $\mathcal{E}_t \subseteq \{\tau_\delta \leq t\}$. Since the stopping time is a positive random variable, we have: $\mathbb{E}_\nu[\tau_\delta] = \sum_{t=1}^{\infty} \mathbb{P}_\nu(\tau_\delta > t)$. For $t > T_0(\delta)$, $\{\tau_\delta > t\} \subseteq \mathcal{E}_t^{\complement}$, hence $\mathbb{P}_\nu(\tau_\delta > t) \leq \mathbb{P}_\nu\left[\mathcal{E}_t^{\complement}\right]$. For $t \leq T_0(\delta)$, we have $\mathbb{P}_\nu(\tau_\delta > t) \leq 1$. Combining those yields: $\mathbb{E}_\nu[\tau_\delta] \leq T_0(\delta) + \sum_{t>T_0(\delta)} \mathbb{P}_\nu\left[\mathcal{E}_t^{\complement}\right]$. Second, let's prove that $\sum_{t>T_0(\delta)} \mathbb{P}_\nu\left[\mathcal{E}_t^{\complement}\right] \leq \frac{2ed}{c^2}$ for $\mathcal{E}_t$ defined in Equation 1. Combining Lemma 5 and Lemma 6 from Degenne et al. (2019) yields the desired result. ■

G.2. Saddle-point Property

Let $T_{t,I} := \{s \in [t] : I_s = I\}$ for all $I \in \mathcal{I}$. Let $I \in \mathcal{I}$. Similarly to Degenne et al. (2019), we prove the saddle-point property of the algorithm $\mathcal{A}_I^A$ associated to I.

Definition 13 *An algorithm playing sequences $(\tilde{w}_s, \lambda_s)_{s \in T_{t,I}} \in \left(\mathcal{S}_\mathcal{A} \times \Theta_I^{\complement}\right)^{|T_{t,I}|}$ is an approximate optimistic saddle-point algorithm with slack x_t if:*

$$\inf_{\lambda \in \Theta_I^{\complement}} \sum_{s \in T_{t,I}} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda)\rangle \geq \max_{A \in \mathcal{A}} \sum_{s \in T_{t,I}} \langle \mathbf{1}_A, r_s \rangle - x_t$$

Using the standard result that $\min_x(f(x) + g(x)) \geq \min_x(f(x)) + \min_x(g(x))$ and the explicit definition of $\lambda_s \in \inf_{\lambda \in \Theta_I^{\complement}} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda)\rangle$, which is a best-response oracle without regret, we obtain that:

$$\inf_{\lambda \in \Theta_I^{\complement}} \sum_{s \in T_{t,I}} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda)\rangle \geq \sum_{s \in T_{t,I}} \inf_{\lambda \in \Theta_I^{\complement}} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda)\rangle = \sum_{s \in T_{t,I}} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda_s)\rangle$$

Let $C_{s,a} = r_{s,a} - d_{\mathrm{KL}}(\mu_{s-1,a}, \lambda_{s,a})$ and $C_s = (C_{s,a})_{a \in [d]}$ be the slack between the optimistic reward and the reward for the parameter μ_{s-1}. We have:

$$\sum_{s \in T_{t,I}} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda_s)\rangle \geq \sum_{s \in T_{t,I}} \langle \tilde{w}_s, r_s \rangle - \sum_{s \in T_{t,I}} \langle \tilde{w}_s, C_s \rangle$$

Since $\sum_{s \in T_{t,I}} \langle \tilde{w}_s, r_s \rangle = \sum_{s \in T_{t,I}} \langle w_s, U_s \rangle$ is The cumulative reward of the A-player with a learner on $\Delta_{|\mathcal{A}|}$ or a learner on $\mathcal{S}_\mathcal{A}$, introducing the cumulative regret R_t^A yields: $\sum_{s \in T_{t,I}} \langle \tilde{w}_s, r_s \rangle \geq \max_{A \in \mathcal{A}} \sum_{s \in T_{t,I}} \langle \mathbf{1}_A, r_s \rangle - R^A_{|T_{t,I}|}$. Combining these inequalities yield:

$$\inf_{\lambda \in \Theta_I^{\complement}} \sum_{s \in T_{t,I}} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda)\rangle \geq \max_{A \in \mathcal{A}} \sum_{s \in T_{t,I}} \langle \mathbf{1}_A, r_s \rangle - x_t$$

where $x_t = \sum_{s \in T_{t,I}} \langle \tilde{w}_s, C_s \rangle + R^A_{|T_{t,I}|}$ is the slack of the optimistic saddle-point algorithm.

G.3. Candidate Answer

The MLE μ_{t-1} summarizes the observations seen at the beginning of round t. Since $\mathcal{M}$ can have a peculiar geometry, $\mu_{t-1} \notin \mathcal{M}$ might happen. Due to concentration results, we have $\mu_{t-1} \in \mathcal{M}$ after a certain time. We consider $\tilde{\mu}_{t-1} \in \mathcal{M} \cap \mathcal{C}_t$, so that for all $a \in [d]$, $d_{\mathrm{KL}}(\mu_{t-1,a}, \mu^{\mathcal{M}}_{t-1,a}) \leq \frac{f(t-1)}{\tilde{N}_{t-1,a}}$. A more elaborate choice, but not necessary, would be: $\tilde{\mu}_{t-1} \in \operatorname{argmin}_{\lambda \in \mathcal{M} \cap \mathcal{C}_t} \langle \tilde{N}_{t-1}, d_{\mathrm{KL}}(\mu_{t-1}, \lambda) \rangle$. When $\mathcal{M} \cap \mathcal{C}_t = \emptyset$, $\tilde{\mu}_{t-1}$ is chosen randomly. The candidate answer is defined as: $I_t = I^*(\tilde{\mu}_{t-1})$.

In Appendix G.3.1, we show that I_t is not the correct answer for only $o(t)$ rounds. This provides a lower bound on the number of times the candidate answer is correct. An upper bound on the number of times the candidate answer is correct is proved in Appendix G.3.2.

G.3.1. INCORRECT ANSWER

Let $I^* = I^*(\mu)$, $t < \tau_\delta$, $T_{t,I} := \{s \in [t] : I_s = I\}$ for all $I \in \mathcal{I}$ and $t_b := t^{1/(1+b)}$. The number of time the recommended answer is not correct is $o(t)$ as a consequence of the following fact: when $I_t \neq I^*$ a quantity, denoted ϵ_t, is increasing linearly while being $O(\sqrt{t})$ due to concentration arguments. The proof of this fact uses a consequence of the chernoff information lower bound ϵ_ν (Appendix E.1): the Lemma 18 of Degenne et al. (2019).

Lemma *[Lemma 18 in Degenne et al. (2019)] For (a) sub-Gaussian or (b) Gaussian bandit, if $d_{KL}(\mu_{t-1,a}, \mu_a) \leq \frac{f(t-1)}{\tilde{N}_{t-1,a}}$ for all $a \in [d]$, then: $I_t \neq I^*(\mu)$ implies there exists $a_0 \in [d]$ such that $\frac{f(t-1)}{\tilde{N}_{t-1,a_0}} \geq \frac{\epsilon_\nu}{2}$.*

Let $s \in [t]$, such that $I_s \neq I^*$, and $\epsilon_t := \sum_{s \leq t, I_s \neq I^*} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \mu) \rangle$. Since $\mu \in \Theta^{\complement}_{I_s}$, we have: $\epsilon_t \geq \sum_{I \in \mathcal{I} \setminus \{I^*\}} \inf_{\lambda \in \Theta^{\complement}_{I_s}} \sum_{s \leq t, I_s = I} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda) \rangle$. For each $I \neq I^*$, the approximate optimistic saddle-point property of the learners with slack $x_t = R^{A}_{|T_{t,I}|} + \sum_{s \in T_{t,I}} \langle \tilde{w}_s, C_s \rangle$ (Appendix G.2) yields:

$$\epsilon_t \geq \sum_{I \in \mathcal{I} \setminus \{I^*\}} \max_{A \in \mathcal{A}} \sum_{s \in T_{t,I}} \langle \mathbf{1}_A, r_s \rangle - \sum_{I \in \mathcal{I} \setminus \{I^*\}} R^{A}_{|T_{t,I}|} - \sum_{s \leq t, I_s \neq I^*} \langle \tilde{w}_s, C_s \rangle$$

Since $f(s-1) \leq f(t-1)$, the condition of Lemma 18 in Degenne et al. (2019) is validated under the event $\mathcal{E}_t$. Hence, we obtain that: for all $s \in [t_b, t]$ such that $I_s \neq I^*$, there exists $a_0 \in [d]$ such that $\frac{f(s-1)}{\tilde{N}_{s-1,a_0}} \geq \frac{\epsilon_\nu}{2}$. Let $t' := \max\{s \in [t] : I_s \neq I^*\}$. We suppose $t' > t_b$, which is possible since $b > 0$. The higher b is, the weaker this assumption is. Let $a_0 \in [d]$ such that $\frac{f(t'-1)}{\tilde{N}_{t'-1,a_0}} \geq \frac{\epsilon_\nu}{2}$. Since f is increasing and, for $t > e$, $\frac{f(t_b)}{f(t)} \geq C_b = \frac{1}{3(1+b)}$, we have that: for all $s \in [t_b, t']$,

$$\frac{f(s-1)}{\tilde{N}_{s-1,a_0}} \geq \frac{f(s-1)}{\tilde{N}_{t'-1,a_0}} = \frac{f(s-1)}{f(t'-1)} \frac{f(t'-1)}{\tilde{N}_{t'-1,a_0}} \geq \frac{f(t_b)}{f(t)} \frac{\epsilon_\nu}{2} \geq C_b \frac{\epsilon_\nu}{2}$$

Let $A_0 \in \mathcal{A}$ such that $a_0 \in A_0$. By definition of r_s, we have $r_{s,a_0} \geq \frac{f(s-1)}{\tilde{N}_{s-1,a_0}}$ and $r_s \geq 0$. Combining these inequalities and dropping the time $s < t_b$ yield:

$$\max_{A \in \mathcal{A}} \sum_{s \in T_{t,I}} \langle \mathbf{1}_A, r_s \rangle \geq \sum_{s \in T_{t,I}} \langle \mathbf{1}_{A_0}, r_s \rangle \geq \sum_{s \in T_{t,I} : s \geq t_b} r_{s,a_0} \geq \sum_{s \in T_{t,I} : s \geq t_b} \frac{f(s-1)}{\tilde{N}_{s-1,a_0}} \geq C_b \frac{\epsilon_\nu}{2} \left(|T_{t,I}| - |T_{t_b,I}| \right)$$

Since $R^A_{|T_{t,I}|} \leq R^A_t$, we have $\sum_{I \in \mathcal{I} \setminus \{I^*\}} R^A_{|T_{t,I}|} \leq (|\mathcal{I}|-1) R^A_t$. For a concave cumulative regret such as $t \mapsto \sqrt{t}$, we would have $\sum_{I \in \mathcal{I} \setminus \{I^*\}} R^A_{|T_{t,I}|} \leq (|\mathcal{I}|-1) R^A_{\frac{t-|T_{t,I^*}|}{|\mathcal{I}|-1}}$. Since $\sum_{I \in \mathcal{I} \setminus \{I^*\}} (|T_{t,I}| - |T_{t_b,I}|) = t - t_b - |T_{t,I^*}|$, summing these inequalities yields:

$$\epsilon_t \geq \frac{C_b \epsilon_\nu}{2}(t - t_b - |T_{t,I^*}|) - (|\mathcal{I}|-1) R^A_t - \sum_{s \leq t, I_s \neq I^*} \langle \tilde{w}_s, C_s \rangle$$

Under event $\mathcal{E}_t$ we have: for all $s \leq t$ and all $a \in [d]$, $d_{\mathrm{KL}}(\mu_{s-1}, \mu) \leq \frac{f(t_b)}{\tilde{N}_{s-1,a}}$. Since $f(t_b) \leq f(t)$, by definition of ϵ_t and Lemma 15, we obtain:

$$\epsilon_t \leq f(t) \sum_{s \leq t} \sum_{a \in [d]} \frac{\tilde{w}_{s,a}}{\tilde{N}_{s-1,a}} \leq f(t) \left(d|\mathcal{A}|^2 + 2d \ln\left(\frac{tK}{d}\right)\right)$$

Combining these inequalities, we obtain a lower bound on $|T_{t,I^*}|$: for $t < \tau_\delta$, under $\mathcal{E}_t$,

$$|T_{t,I^*}| \geq t - t_b - \frac{2}{C_b \epsilon_\nu} \left((|\mathcal{I}|-1) R^A_t + \sum_{s \leq t, I_s \neq I^*} \langle \tilde{w}_s, C_s \rangle + f(t) \left(d|\mathcal{A}|^2 + 2d \ln\left(\frac{tK}{d}\right)\right)\right)$$

G.3.2. CORRECT ANSWER

Let $I^* = I^*(\mu)$, $t < \tau_\delta$ and $T_{t,I} := \{s \in [t] : I_s = I\}$ for all $I \in \mathcal{I}$. Let $t' := \max\{s \leq t : I_s = I^*\}$ be the last round in which we recommend the correct answer before the algorithm stops. Since $s \mapsto \beta(s,\delta)$ is increasing, we have $\beta(t,\delta) \geq \beta(t'-1,\delta)$. By definition of t', we have $T_{t,I^*} = T_{t',I^*}$. The non-satisfied stopping criterion rewrites as: $\beta(t'-1,\delta) \geq \inf_{\lambda \in \Theta^\complement_{I^*}} \langle \tilde{N}_{t'-1}, d_{\mathrm{KL}}(\mu_{t'-1}, \lambda)\rangle$. Lemma 16 and $t \mapsto tf(t)$ increasing yield that:

$$\begin{aligned}\inf_{\lambda \in \Theta^\complement_{I^*}} \langle \tilde{N}_{t'-1}, d_{\mathrm{KL}}(\mu_{t'-1}, \lambda)\rangle &\geq \inf_{\lambda \in \Theta^\complement_{I^*}} \langle \tilde{N}_{t'-1}, d_{\mathrm{KL}}(\mu, \lambda)\rangle - L_{\mathcal{M}} \sqrt{2(t'-1) f(t'-1) d \|\sigma^2\|_\infty K} \\ &\geq \inf_{\lambda \in \Theta^\complement_{I^*}} \langle \tilde{N}_{t'-1}, d_{\mathrm{KL}}(\mu, \lambda)\rangle - L_{\mathcal{M}} \sqrt{2t f(t) d \|\sigma^2\|_\infty K}\end{aligned}$$

Combining C-Tracking, Lemma 14 and $\langle \mathbf{1}, d_{\mathrm{KL}}(\mu, \lambda)\rangle = \|d_{\mathrm{KL}}(\mu,\lambda)\|_1 \leq D_{\mathcal{M}}$ ($\mathcal{M}$ bounded), we obtain:

$$\inf_{\lambda \in \Theta^\complement_{I^*}} \langle \tilde{N}_{t'-1}, d_{\mathrm{KL}}(\mu, \lambda)\rangle \geq \inf_{\lambda \in \Theta^\complement_{I^*}} \sum_{s=1}^{t'-1} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu, \lambda)\rangle - |\mathcal{A}|^2 D_{\mathcal{M}}$$

Lemma 14 in Degenne et al. (2019) (Appendix G.5.2) yields:

$$\inf_{\lambda \in \Theta^\complement_{I^*}} \sum_{s=1}^{t'-1} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu, \lambda)\rangle \geq \inf_{\lambda \in \Theta^\complement_{I^*}} \sum_{s=n_0+1}^{t'-1} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda)\rangle - L_{\mathcal{M}} \sqrt{2\|\sigma^2\|_\infty f(t)} \sum_{s=n_0+1}^{t'-1} \sum_{a \in [d]} \frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s-1,a}}}$$

Lemma 15 shows:

$$\sum_{a \in [d]} \sum_{s=n_0+1}^{t'-1} \frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s-1,a}}} \leq d|\mathcal{A}|^2 + 2\sqrt{2d(t'-1)K} \leq d|\mathcal{A}|^2 + 2\sqrt{2dtK}$$

Since $\langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda)\rangle \geq 0$, dropping all the rounds for which $I_s \neq I^*$ yields:

$$\inf_{\lambda \in \Theta^{\complement}_{I^*}} \sum_{s=n_0+1}^{t'-1} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda)\rangle \geq \inf_{\lambda \in \Theta^{\complement}_{I^*}} \sum_{n_0+1 \leq s \leq t'-1, I_s = I^*} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda)\rangle$$

Combining the saddle-point property of $\mathcal{A}^A_{I^*}$ and $R^A_{t'-1} \leq R^A_t$, we obtain:

$$\begin{aligned}\inf_{\lambda \in \Theta^{\complement}_{I^*}} \sum_{n_0+1 \leq s \leq t'-1, I_s = I^*} \langle \tilde{w}_s, d_{\mathrm{KL}}(\mu_{s-1}, \lambda)\rangle \geq \max_{A \in \mathcal{A}} &\sum_{n_0+1 \leq s \leq t'-1, I_s = I^*} \langle \mathbf{1}_A, r_s\rangle - R^A_t \\ &- \sum_{n_0+1 \leq s \leq t'-1, I_s = I^*} \langle \tilde{w}_s, C_s\rangle\end{aligned}$$

Under the concentration event $\mathcal{E}_t$, we have $d_{\mathrm{KL}}(\mu_{s,a}, \mu_a) \leq \frac{f(t_b)}{\tilde{N}_{s,a}} \leq \frac{f(s)}{\tilde{N}_{s,a}}$ for all $s \geq t_b$. Hence, we obtain that: $\mu \in [\alpha_{s,a}, \beta_{s,a}]$ for all $s \geq t_b$. Combined with the definition of r_s, this implies that: for all $a \in [d]$, $r_{s,a} \geq d_{\mathrm{KL}}(\mu_a, \lambda_{s,a})$. Dropping all the the rounds for which $s < t_b$ yields:

$$\max_{A \in \mathcal{A}} \sum_{n_0+1 \leq s \leq t'-1, I_s = I^*} \langle \mathbf{1}_A, r_s\rangle \geq \max_{A \in \mathcal{A}} \sum_{t_b \leq s \leq t'-1, I_s = I^*} \langle \mathbf{1}_A, d_{\mathrm{KL}}(\mu, \lambda_s)\rangle$$

Combining $\frac{1}{|T_{t'-1,I^*}| - |T_{t_b,I^*}|} \sum_{t_b \leq s \leq t'-1, I_s = I^*} \delta_{\lambda_s} \in \mathcal{P}\left(\Theta^{\complement}_{I^*}\right)$ (average of diracs in $(\lambda_s)_s$), the dual formulation of D_ν and the fact that $|T_{t'-1,I^*}| \geq |T_{t,I^*}| - 1$, we obtain that:

$$\begin{aligned}\max_{A \in \mathcal{A}} \sum_{t_b \leq s \leq t'-1, I_s = I^*} \langle \mathbf{1}_A, d_{\mathrm{KL}}(\mu, \lambda_s)\rangle &\geq (|T_{t'-1,I^*}| - |T_{t_b,I^*}|) \inf_{q \in \mathcal{P}\left(\Theta^{\complement}_{I^*}\right)} \max_{A \in \mathcal{A}} \mathbb{E}_{\lambda \sim q}\left[\langle \mathbf{1}_A, d_{\mathrm{KL}}(\mu, \lambda)]\right\rangle \\ &\geq (|T_{t,I^*}| - 1 - t_b) D_\nu\end{aligned}$$

Combining these inequalities, we obtain an upper bound on $|T_{t,I^*}|$: for $t < \tau_\delta$, under $\mathcal{E}_t$,

$$\begin{aligned}&\frac{\beta(t,\delta) + R^A_t + c_t}{D_\nu} \geq |T_{t,I^*}| - 1 - t_b \\ \text{where} \quad &c_t = L_{\mathcal{M}} \sqrt{2\|\sigma^2\|_\infty f(t)} \left(d|\mathcal{A}|^2 + 2\sqrt{2dtK}\right) + |\mathcal{A}|^2 D_{\mathcal{M}} \\ &\qquad + L_{\mathcal{M}} \sqrt{2tf(t)d\|\sigma^2\|_\infty K} + \sum_{n_0+1 \leq s \leq t'-1, I_s = I^*} \langle \tilde{w}_s, C_s\rangle\end{aligned}$$

G.4. Stopping Time Upper Bound

Combining the upper and lower bounds on $|T_{t,I^*}|$ (Appendices G.3.1 and G.3.2) yields: for $t < \tau_\delta$, under $\mathcal{E}_t$,

$$t \leq \frac{\beta(t,\delta)}{D_\nu} + C_\nu \left(R^A_t + h(t)\right) \tag{2}$$

where $C_\nu := \frac{1}{D_\nu} + \frac{2(|\mathcal{I}|-1)}{C_b \epsilon_\nu}$ and

$$h(t) := \frac{1}{C_\nu} \left(\frac{c_t}{D_\nu} + 2t_b + 1 + \frac{2}{C_b \epsilon} \left(\sum_{s \leq t, I_s \neq I^*} \langle \tilde{w}_s, C_s\rangle + f(t) \left(d|\mathcal{A}|^2 + 2d \ln\left(\frac{tK}{d}\right)\right)\right)\right)$$

Lemma 17 yields:

$$\sum_{s \geq n_0+1}^{t} \langle \tilde{w}_s, C_s \rangle \leq f(t) \left(d|\mathcal{A}|^2 + 2d \ln\left(\frac{tK}{d}\right)\right) \\ + 2L_{\mathcal{M}} \sqrt{2\|\sigma^2\|_\infty f(t)} \left(d|\mathcal{A}|^2 + 2\sqrt{2dtK} \right)$$

Let $b = 1$. Using that $f(t) = \Omega(\ln(t))$ and $t_b = \sqrt{t}$, we obtain that: $h(t) = O\left(\sqrt{t \ln(t)}\right)$. Let $T_0(\delta) := \max\left\{t \in \mathbb{N} : t \leq \frac{\beta(t,\delta)}{D_\nu} + C_\nu(R_t^A + h(t))\right\}$ be the upper bound on time such that the Equation 2 is satisfied. Since the set is non empty and bounded, $T_0(\delta) \in \mathbb{N}$. The set is bounded since the learner has sublinear cumulative regret, $R_t^A = o(t)$, and $\frac{\beta(t,\delta)}{D_\nu} + C_\nu(R_t^A + h(t)) = O\left(R_t^A + \sqrt{t\ln(t)}\right)$.

Theorem *Let $\mathcal{M}$ bounded. The sample complexity of the instantiated CombGame meta-algorithm on bandit $\mu \in \mathcal{M}$ satisfies:*

$$\mathbb{E}_\nu[\tau_\delta] \leq T_0(\delta) + \frac{2ed}{c^2} \quad \textit{with} \quad T_0(\delta) := \max\left\{t \in \mathbb{N} : t \leq \frac{\beta(t,\delta)}{D_\nu} + C_\nu(R_t^A + h(t))\right\}$$

where $c > 0$ is the parameter of the exploration bonus $f(t)$ when taking $b = 1$. The reminder terms are: the approximation error $h(t) = O\left(\sqrt{t\ln(t)}\right)$, the learner's cumulative regret R_t^A and a constant C_ν depending on the distribution.

The instantiated CombGame meta-algorithm is an asymptotically optimal algorithm.

Proof By the absurd, we assume there exists $t > T_0(\delta)$ such that $\mathcal{E}_t \cap \{t < \tau_\delta\} \neq \emptyset$. Under event $\mathcal{E}_t$ combining $t > T_0(\delta)$ and $t < \tau_\delta$ yields the following contradiction:

$$\frac{\beta(t,\delta)}{D_\nu} + C_\nu(R_t^A + h(t)) < t \leq \frac{\beta(t,\delta)}{D_\nu} + C_\nu(R_t^A + h(t))$$

Therefore, we have $\mathcal{E}_t \cap \{t < \tau_\delta\} = \emptyset$. Hence, for all $t > T_0(\delta)$, $\mathcal{E}_t \subset \{\tau_\delta \leq t\}$. Applying Lemma 4 concludes the proof of the finite-time upper bound. Taking the limit $\delta \to 0$ yields that the instantiated CombGame meta-algorithm is an asymptotically optimal algorithm. ■

G.5. Technical Arguments

In Appendix G.5, we prove technical arguments on C-Tracking (Appendix G.5.1) and on concentration events (Appendix G.5.2).

G.5.1. TRACKING ARGUMENTS

Let $\mathcal{A}_{|a} = \{A \in \mathcal{A} : a \in A\}$ be the set of actions containing the arm a and $B_t = \mathrm{supp}\left(\sum_{s=1}^t w_s\right)$. Sparse C-Tracking is defined as: $A_t \in \mathrm{argmin}_{A \in B_t} \frac{\tilde{N}_{t-1,a}}{\sum_{s=1}^t w_{s,A}}$ for all $t > n_0$. Lemma 14 controls the deviation between the empirical count of sampled actions, $N_{t,A}$, and the cumulative sum of pulling proportions, $\sum_{s=1}^t w_{s,A}$. This is an adaptation of Lemma 7 in Degenne et al. (2019).

Lemma 14 *Using sparse C-Tracking, we have: for all $t \geq n_0$ and for all $A \in \mathcal{A}$, and all $a \in [d]$,*

$$\sum_{s=1}^{t} w_{s,A} - (|\mathcal{A}| - 1) \leq N_{t,A} \leq 1 + \sum_{s=1}^{t} w_{s,A}$$

$$\sum_{s=1}^{t} \tilde{w}_{s,a} - (|\mathcal{A}| - 1)|\mathcal{A}_{|a}| \leq \tilde{N}_{t,a} \leq |\mathcal{A}_{|a}| + \sum_{s=1}^{t} \tilde{w}_{s,a}$$

Proof If $A \notin B_t$, we have $N_{t,A} = 0$ and $\sum_{s=1}^{t} w_{s,A} = 0$. Hence the first inequalities are immediate. Let $A \in B_t$ and $S_{t,A} = \sum_{s=1}^{t} w_{s,A}$. We will prove $N_{t,A} \leq 1 + S_{t,A}$ by induction. At $t = n_0$, the result is true based on the initialization: $S_{n_0,A} = 1$ if $A \in B_{n_0}$ and $S_{n_0,A} = 0$ else. Assume that $N_{s,A} \leq S_{s,A} + 1$ for all $A \in \mathcal{A}$ and all $s \leq t-1$. Let's prove that it holds at round t too. If $A \neq A_t$, the induction property yields: $N_{t,A} = N_{t-1,A} \leq S_{t-1,A} + 1 \leq S_{t,A} + 1$. Assume $A = A_t$, then:

$$\frac{N_{t,A_t}}{S_{t,A_t}} = \frac{N_{t-1,A_t}}{S_{t,A_t}} + \frac{1}{S_{t,A_t}} = \frac{1}{S_{t,A_t}} + \min_{A \in \mathcal{A}} \frac{N_{t-1,A}}{S_{t,A}} \leq \frac{1}{S_{t,A_t}} + 1$$

where the last inequality is shown by the absurd. If $\min_{A \in \mathcal{A}} \frac{N_{t-1,A}}{S_{t,A}} \leq 1$ doesn't hold, we have for all $A \in \mathcal{A}$, $N_{t-1,A} > S_{t,A}$. Summing these strict inequalities yields a contradiction: $t - 1 = \sum_{A \in \mathcal{A}} N_{t-1,A} > \sum_{A \in \mathcal{A}} S_{t,A} = \sum_{s=1}^{t} \sum_{A \in \mathcal{A}} w_{s,A} = t$. Therefore, we have $N_{t,A_t} \leq 1 + S_{t,A_t}$. This concludes the induction.

Combining the previous upper bound and $t = \sum_{A \in \mathcal{A}} N_{t,A} = \sum_{A \in \mathcal{A}} S_{t,A}$ yield the lower bound:

$$N_{t,A} = t - \sum_{A' \neq A} N_{t,A'} \geq t - \sum_{A' \neq A} (S_{t,A'} + 1) = S_{t,A} - (|\mathcal{A}| - 1)$$

Applying the linear map $W_{\mathcal{A}}$ on the previous inequalities yield the counterpart at the arms level:

$$\sum_{s=1}^{t} \tilde{w}_{s,a} - (|\mathcal{A}| - 1)|\mathcal{A}_{|a}| \leq \tilde{N}_{t,a} \leq |\mathcal{A}_{|a}| + \sum_{s=1}^{t} \tilde{w}_{s,a}$$

■

A better bound for C-Tracking was proven in Theorem 6 of Degenne et al. (2020b). They obtain that for all $t \in \mathbb{N}$ and $A \in \mathcal{A}$,

$$-\ln(|\mathcal{A}|) \leq N_{t,A} - \sum_{s=1}^{t} w_{s,A} \leq 1$$

Lemma 8 from Degenne et al. (2019) is a technical lemma on summations.

Lemma *[Lemma 8 in Degenne et al. (2019)] For $t \geq t_0 \geq 1$ and $(x_s)_{s \in [t]}$ non negative real numbers such that $\sum_{s=1}^{t_0-1} x_s > 0$,*

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

$$\sum_{s=t_0}^{t} \frac{x_s}{\sum_{r=1}^{s} x_r} \leq \ln\left(\sum_{s=1}^{t} x_s\right) - \ln\left(\sum_{s=1}^{t_0-1} x_s\right)$$

Lemma 15 controls the summation of ratios $\frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s,a}}}$ and $\frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s-1,a}}}$ over arms and time. This is an adaptation of Lemma 9 in Degenne et al. (2019) to our setting.

Lemma 15 *Let $(\tilde{w}_s)_{s\in\mathbb{N}} \in \mathcal{S}_{\mathcal{A}}^{\mathbb{N}}$ and $\tilde{N}_t$ obtained with sparse C-Tracking. Then,*

$$\sum_{a\in[d]} \sum_{s=n_0}^{t} \frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s,a}}} \leq d|\mathcal{A}|^2 + 2\sqrt{dtK} \quad \textit{and} \quad \sum_{a\in[d]} \sum_{s=n_0+1}^{t} \frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s-1,a}}} \leq d|\mathcal{A}|^2 + 2\sqrt{2dtK}$$

$$\sum_{a\in[d]} \sum_{s=n_0}^{t} \frac{\tilde{w}_{s,a}}{\tilde{N}_{s,a}} \leq d|\mathcal{A}|^2 + d\ln\left(\frac{tK}{d}\right) \quad \textit{and} \quad \sum_{a\in[d]} \sum_{s=n_0+1}^{t} \frac{\tilde{w}_{s,a}}{\tilde{N}_{s-1,a}} \leq d|\mathcal{A}|^2 + 2d\ln\left(\frac{tK}{d}\right)$$

Proof First, let's prove inequalities 1 and 3. Let $a \in [d]$ and $t_{0,a}$ be the first time such that: $\sum_{s=1}^{t_{0,a}-1} \tilde{w}_{s,a} > (|\mathcal{A}|-1)|\mathcal{A}_{|a}|+1$. Since $\tilde{w}_{t_{0,a}-1,a} \leq 1$, we have $\sum_{s=1}^{t_{0,a}-1} \tilde{w}_{s,a} \leq (|\mathcal{A}|-1)|\mathcal{A}_{|a}|+2$. Since $\tilde{N}_{s,a} \geq 1$ for $s \geq n_0$, we obtain:

$$\begin{aligned}
\sum_{s=n_0}^{t} \frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s,a}}} = \sum_{s=n_0}^{t_{0,a}-1} \frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s,a}}} + \sum_{s=t_{0,a}}^{t} \frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s,a}}} &\leq \sum_{s=n_0}^{t_{0,a}-1} \tilde{w}_{s,a} + \sum_{s=t_{0,a}}^{t} \frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s,a}}} \\
&\leq (|\mathcal{A}|-1)|\mathcal{A}_{|a}| + 2 + \sum_{s=t_{0,a}}^{t} \frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s,a}}} \\
\sum_{s=n_0}^{t} \frac{\tilde{w}_{s,a}}{\tilde{N}_{s,a}} = \sum_{s=n_0}^{t_{0,a}-1} \frac{\tilde{w}_{s,a}}{\tilde{N}_{s,a}} + \sum_{s=t_{0,a}}^{t} \frac{\tilde{w}_{s,a}}{\tilde{N}_{s,a}} &\leq \sum_{s=n_0}^{t_{0,a}-1} \tilde{w}_{s,a} + \sum_{s=t_{0,a}}^{t} \frac{\tilde{w}_{s,a}}{\tilde{N}_{s,a}} \\
&\leq (|\mathcal{A}|-1)|\mathcal{A}_{|a}| + 2 + \sum_{s=t_{0,a}}^{t} \frac{\tilde{w}_{s,a}}{\tilde{N}_{s,a}}
\end{aligned}$$

Combining Lemma 14 and Lemma 8 in Degenne et al. (2019) for $x_s = \tilde{w}_{s,a}$, we obtain:

$$\begin{aligned}
\sum_{s=t_{0,a}}^{t} \frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s,a}}} &\leq \sum_{s=t_{0,a}}^{t} \frac{\tilde{w}_{s,a}}{\sqrt{\sum_{r=1}^{s} \tilde{w}_{r,a} - (|\mathcal{A}|-1)|\mathcal{A}_{|a}|}} \\
&\leq 2\sqrt{\sum_{s=1}^{t} \tilde{w}_{s,a} - (|\mathcal{A}|-1)|\mathcal{A}_{|a}|} - 2\sqrt{\sum_{s=1}^{t_{0,a}-1} \tilde{w}_{s,a} - (|\mathcal{A}|-1)|\mathcal{A}_{|a}|} \\
\sum_{s=t_{0,a}}^{t} \frac{\tilde{w}_{s,a}}{\tilde{N}_{s,a}} &\leq \sum_{s=t_{0,a}}^{t} \frac{\tilde{w}_{s,a}}{\sum_{r=1}^{s} \tilde{w}_{r,a} - (|\mathcal{A}|-1)|\mathcal{A}_{|a}|} \\
&\leq \ln\left(\sum_{s=1}^{t} \tilde{w}_{s,a} - (|\mathcal{A}|-1)|\mathcal{A}_{|a}|\right) - \ln\left(\sum_{s=1}^{t_{0,a}-1} \tilde{w}_{s,a} - (|\mathcal{A}|-1)|\mathcal{A}_{|a}|\right)
\end{aligned}$$

Using that $\sum_{s=1}^{t_{0,a}-1} \tilde{w}_{s,a} - (|\mathcal{A}|-1)|\mathcal{A}_{|a}| > 1$, we obtain: $\sum_{s=t_{0,a}}^{t} \frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s,a}}} \leq 2\sqrt{\sum_{s=1}^{t} \tilde{w}_{s,a}}$ and $\sum_{s=t_{0,a}}^{t} \frac{\tilde{w}_{s,a}}{\tilde{N}_{s,a}} \leq \ln\left(\sum_{s=1}^{t} \tilde{w}_{s,a}\right)$. Combining the concavity of $x \mapsto \sqrt{x}$ and $x \mapsto \ln(x)$ and

$\sum_{a\in[d]}\sum_{s=1}^{t}\tilde{w}_{s,a} \leq tK$ yield by summation:

$$\sum_{a\in[d]}\sum_{s=t_{0,a}}^{t}\frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s,a}}} \leq 2\sum_{a\in[d]}\sqrt{\sum_{s=1}^{t}\tilde{w}_{s,a}} \leq 2\sqrt{dtK}$$

$$\sum_{a\in[d]}\sum_{s=t_{0,a}}^{t}\frac{\tilde{w}_{s,a}}{\tilde{N}_{s,a}} \leq \sum_{a\in[d]}\ln\left(\sum_{s=1}^{t}\tilde{w}_{s,a}\right) \leq d\ln\left(\frac{tK}{d}\right)$$

Therefore, we obtain: $\sum_{a\in[d]}\sum_{s=n_0}^{t}\frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s,a}}} \leq d|\mathcal{A}|^2 + 2\sqrt{dtK}$ and $\sum_{a\in[d]}\sum_{s=n_0}^{t}\frac{\tilde{w}_{s,a}}{\tilde{N}_{s,a}} \leq d|\mathcal{A}|^2 + d\ln\left(\frac{tK}{d}\right)$. For all $s \geq n_0$, we have $N_{s-1,a} \geq 1$, hence $N_{s-1,a} \geq \frac{1}{2}N_{s,a}$. Plugging this inequality in the sum starting from $t_{0,a}$ yields: $\sum_{a\in[d]}\sum_{s=n_0+1}^{t}\frac{\tilde{w}_{s,a}}{\sqrt{\tilde{N}_{s-1,a}}} \leq d|\mathcal{A}|^2 + 2\sqrt{2dtK}$ and $\sum_{a\in[d]}\sum_{s=n_0+1}^{t}\frac{\tilde{w}_{s,a}}{\tilde{N}_{s-1,a}} \leq d|\mathcal{A}|^2 + 2d\ln\left(\frac{tK}{d}\right)$. ■

G.5.2. CONCENTRATION ARGUMENTS

Let $t_b = t^{1/(1+b)} < t$, $L_{\mathcal{M}}$ the Lipschitz constant of $x \mapsto d_{KL}(x, y)$ ($\mathcal{M}$ bounded). The sequence of concentrations events $(\mathcal{E}_t)_{t\geq 1}$ for the exploration bonus f with parameters $c > 0$ and $b > 0$ was defined as:

$$\mathcal{E}_t := \left\{\forall s \leq t, \forall a \in [d], \quad \tilde{N}_{s,a}d_{\mathrm{KL}}(\mu_{s,a}, \mu_a) \leq f\left(t_b\right)\right\}$$

Lemma 14 in Degenne et al. (2019) controls the deviation $|d_{\mathrm{KL}}(\mu_{s-1,a}, \lambda_a) - d_{\mathrm{KL}}(\mu_a, \lambda_a)|$. Its proof is similar to the beginning of the proof of Lemma 16.

Lemma *[Lemma 14 in Degenne et al. (2019)] Let $\mathcal{M}$ bounded. Under $\mathcal{E}_t$, for all $s \in [t]$, $a \in [d]$ any $\lambda \in \mathcal{M}$,*

$$|d_{KL}(\mu_{s-1,a}, \lambda_a) - d_{KL}(\mu_a, \lambda_a)| \leq L_{\mathcal{M}}\sqrt{2\|\sigma^2\|_\infty\frac{f(t)}{\tilde{N}_{s-1,a}}}$$

Lemma 16 controls the weighted sum of deviations, $\langle\tilde{N}_t, d_{\mathrm{KL}}(\mu_t, \lambda) - d_{\mathrm{KL}}(\mu, \lambda)\rangle$. This is an adaptation of Lemma 17 in Degenne et al. (2019).

Lemma 16 *Let $\mathcal{M}$ be bounded. Under $\mathcal{E}_t$, for any $\lambda \in \mathcal{M}$,*

$$\langle\tilde{N}_t, d_{KL}(\mu_t, \lambda)\rangle \geq \langle\tilde{N}_t, d_{KL}(\mu, \lambda)\rangle - L_{\mathcal{M}}\sqrt{2tf(t)\|\sigma^2\|_\infty dK}$$

Proof Using the Lipschitz property of $x \mapsto d_{\mathrm{KL}}(x, y)$, we have $d_{\mathrm{KL}}(\mu_{t,a}, \lambda_a) - d_{\mathrm{KL}}(\mu_a, \lambda_a) \geq -L_{\mathcal{M}}|\mu_{t,a} - \mu_a|$. The sub-Gaussian property when (a) or the direct formula for Gaussian when (b), implies that $|\mu_{t,a} - \mu_a| \leq \sqrt{2\sigma_a^2 d_{\mathrm{KL}}(\mu_{t,a}, \mu_a)}$. Under $\mathcal{E}_t$, we have $d_{\mathrm{KL}}(\mu_{t,a}, \mu_a) \leq \frac{f(t_b)}{\tilde{N}_{t,a}}$. Combining these inequalities, f increasing and $\sigma_a^2 \leq \|\sigma^2\|_\infty$, we obtain: $d_{\mathrm{KL}}(\mu_{t,a}, \lambda_a) - d_{\mathrm{KL}}(\mu_a, \lambda_a) \geq -L_{\mathcal{M}}\sqrt{2\|\sigma^2\|_\infty\frac{f(t)}{\tilde{N}_{t,a}}}$ for all $a \in [d]$. Summing with weights $\tilde{N}_t$ yields:

$$\langle\tilde{N}_t, d_{\mathrm{KL}}(\mu_t, \lambda)\rangle - \langle\tilde{N}_t, d_{\mathrm{KL}}(\mu, \lambda)\rangle \geq -L_{\mathcal{M}}\sqrt{2f(t)\|\sigma^2\|_\infty}\sum_{a\in[d]}\sqrt{\tilde{N}_{t,a}}$$

Since $x \mapsto \sqrt{x}$ is concave, we have $\sum_{a\in[d]} \sqrt{\tilde{N}_{t,a}} \leq \sqrt{d \sum_{a\in[d]} \tilde{N}_{t,a}} \leq \sqrt{dtK}$. This concludes the proof: $\langle \tilde{N}_t, d_{\text{KL}}(\mu_t, \lambda)\rangle \geq \langle \tilde{N}_t, d_{\text{KL}}(\mu, \lambda)\rangle - L_{\mathcal{M}}\sqrt{2tf(t)\|\sigma^2\|_\infty dK}$. ■

Lemma 17 controls one term of the slack appearing in the saddle-point property, the one linked to r_s: $\sum_{s\geq n_0+1}^{t} \langle \tilde{w}_s, C_s\rangle$. This is an adaptation of Lemmas 15 and 16 in Degenne et al. (2019).

Lemma 17 *Let $\mathcal{M}$ be bounded and*

$$D_{s,a} = \max\left\{ 2L_{\mathcal{M}}\sqrt{2\sigma_a^2 \frac{f(\max\{s-1, t_b\})}{\tilde{N}_{s-1,a}}}, \frac{f(\max\{s-1,t_b\})}{\tilde{N}_{s-1,a}} \right\}$$

Under the event $\mathcal{E}_t$, for all $s \in [t]$: $\sup_{\phi\in[\alpha_{s,a},\beta_{s,a}]}(r_{s,a} - d_{KL}(\phi, \lambda_{s,a})) \leq D_{s,a}$. Let $C_{s,a} = r_{s,a} - d_{KL}(\mu_{s-1,a}, \lambda_{s,a})$, we obtain:

$$\sum_{s\geq n_0+1}^{t} \langle \tilde{w}_s, C_s\rangle \leq f(t)\left(d|\mathcal{A}|^2 + 2d\ln\left(\frac{tK}{d}\right)\right) + 2L_{\mathcal{M}}\sqrt{2\|\sigma^2\|_\infty f(t)}\left(d|\mathcal{A}|^2 + 2\sqrt{2dtK}\right)$$

Proof We recall that: $r_{s,a} = \max\left\{\frac{f(s-1)}{\tilde{N}_{s-1,a}}, \max_{\phi\in\{\alpha_{s,a},\beta_{s,a}\}} d_{\text{KL}}(\phi, \lambda_{s,a})\right\}$ for all $a \in [d]$. Assume $r_{s,a} = \frac{f(s-1)}{\tilde{N}_{s-1,a}}$. Since d_{KL} is positive, f is increasing and $s-1 \leq \max\{s-1, t_b\}$, we have:

$$\sup_{\phi\in[\alpha_{s,a},\beta_{s,a}]}(r_{s,a} - d_{\text{KL}}(\phi,\lambda_{s,a})) \leq \frac{f(s-1)}{\tilde{N}_{s-1,a}} \leq \frac{f(\max\{s-1,t_b\})}{\tilde{N}_{s-1,a}} \leq D_{s,a}$$

Assume $r_{s,a} = d_{\text{KL}}(\phi_{s,a}, \lambda_{s,a})$ where $\phi_{s,a} = \text{argmax}_{\phi\in\{\alpha_{s,a},\beta_{s,a}\}} d_{\text{KL}}(\phi, \lambda_{s,a})$. By convexity of $x \mapsto d_{\text{KL}}(x,y)$, we have $d_{\text{KL}}(\phi_{s,a}, \lambda_{s,a}) = \max_{\phi\in[\alpha_{s,a},\beta_{s,a}]} d_{\text{KL}}(\phi, \lambda_{s,a})$. Upper bounding yields: $\sup_{\phi\in[\alpha_{s,a},\beta_{s,a}]}(r_{s,a} - d_{\text{KL}}(\phi,\lambda_{s,a})) \leq \sup_{\phi,\eta\in[\alpha_{s,a},\beta_{s,a}]} |d_{\text{KL}}(\eta, \lambda_{s,a}) - d_{\text{KL}}(\phi,\lambda_{s,a})|$. The Lipschitz property of $x \mapsto d_{\text{KL}}(x,y)$ yields: $|d_{\text{KL}}(\eta,\lambda_{s,a}) - d_{\text{KL}}(\phi,\lambda_{s,a})| \leq L_{\mathcal{M}}|\eta - \phi|$. Under event $\mathcal{E}_t$, combining the sub-Gaussian property when (a) or the direct formula for Gaussian when (b) and f increasing, we obtain:

$$\sup_{\phi\in[\alpha_{s,a},\beta_{s,a}]}(r_{s,a} - d_{\text{KL}}(\phi,\lambda_{s,a})) \leq 2L_{\mathcal{M}}\sqrt{2\sigma_a^2\frac{f(\max\{s-1,t_b\})}{N_{s-1,a}}} \leq D_{s,a}$$

For the second part of the lemma, since $\mu_{s-1,a} \in [\alpha_{s,a}, \beta_{s,a}]$, we have $C_{s,a} \leq D_{s,a}$ and $\sum_s \langle \tilde{w}_s, C_s\rangle \leq \sum_s \langle \tilde{w}_s, D_s\rangle$. Applying Lemma 15 twice, we obtain:

$$\begin{aligned}
\sum_{s\geq n_0+1}^{t}\sum_{a\in[d]} \tilde{w}_{s,a} 2L\sqrt{2\sigma_a^2\frac{f(\max\{s-1,t_b\})}{N_{s-1,a}}} &\leq 2L_{\mathcal{M}}\sqrt{2\|\sigma^2\|_\infty f(t)}\sum_{s\geq n_0+1}^{t}\sum_{a\in[d]}\frac{\tilde{w}_{s,a}}{\sqrt{N_{s-1,a}}} \\
&\leq 2L_{\mathcal{M}}\sqrt{2\|\sigma^2\|_\infty f(t)}\left(d|\mathcal{A}|^2 + 2\sqrt{2dtK}\right) \\
\sum_{s\geq n_0+1}^{t}\sum_{a\in[d]} \tilde{w}_{s,a}\frac{f(\max\{s-1,t_b\})}{N_{s-1,a}} &\leq f(t)\sum_{s\geq n_0+1}^{t}\sum_{a\in[d]}\frac{\tilde{w}_{s,a}}{N_{s-1,a}} \\
&\leq f(t)\left(d|\mathcal{A}|^2 + 2d\ln\left(\frac{tK}{d}\right)\right)
\end{aligned}$$

Combining $\max(a+b) \leq a+b$ for $D_{s,a}$ and the previous inequalities concludes the proof. ■

Appendix H. Unbounded $\mathcal{M}$ for Gaussian Bandit

As already discussed in Appendix F of Degenne et al. (2019), the boundedness assumption of $\mathcal{M}$ can be weakened. In particular, for Gaussian bandit where $y \mapsto d_{\mathrm{KL}}(x, y) = \frac{(x-y)^2}{\sigma_a^2}$ is convex and symmetric, we can remove it completely using concentration events and explicit formulas.

We sketch the ideas of the required adaptations, the full proof is omitted for the sake of space. The concentration arguments of Appendix G.5.2 are replaced by weaker results: the deviation is controlled for a given λ, not for an arbitrary $\lambda \in \mathcal{M}$. Similarly, using explicit formulas, we can upper bound the optimistic reward and prove that $\tau_\delta < +\infty$. The adaptation is mainly technical and requires to be familiar with the detail of the proof of Theorem 3.

Bounded $\|\mu_t - \mu\|_\infty$ Under event $\mathcal{E}_t$, we have for all $s \leq t$ and all $a \in [d]$, $d_{\mathrm{KL}}(\mu_{s,a}, \mu_a) \leq \frac{f(t)}{\tilde{N}_{s,a}}$. The concentration event yields that $\|\mu_s - \mu\|_\infty \leq \sqrt{2\|\sigma^2\|_\infty \frac{f(t)}{\min_{a\in[d]} \tilde{N}_{s,a}}}$. Hence, μ_s belongs to a bounded set around μ. When all arms are sampled more than a logarithmic number of time, we have $\lim_{s\to\infty} \|\mu_s - \mu\|_\infty = 0$.

Explicit formula Let $I \in \mathcal{I}$, $\tilde{w} \in \mathcal{S}_{\mathcal{A}}$ and $\phi \in \Theta_I$. Let $\lambda(\phi, I, \tilde{w}) \in \operatorname{argmin}_{\lambda \in \Theta_I^\complement} \langle \tilde{w}, d_{\mathrm{KL}}(\phi, \lambda)\rangle$ and $\lambda(\phi, I, \tilde{w}, J) \in \operatorname{argmin}_{\lambda \in \bar{\Theta}_J^I} \langle \tilde{w}, d_{\mathrm{KL}}(\phi, \lambda)\rangle$. Using $\Theta_I^\complement = \bigcup_{J \neq I} \bar{\Theta}_J^I$, there exists $J(I) \in \mathcal{I}$ such that $\lambda(\phi, I, \tilde{w}) = \lambda(\phi, I, \tilde{w}, J(I))$. Lemma 18 proves an explicit formula for $\lambda(\phi, I, \tilde{w}, J)$, which implies an upper bound on $\|\phi - \lambda(\phi, I, \tilde{w})\|_\infty$. Let $\phi' \in \Theta_I$. When $\|\phi - \phi'\|$ is bounded, applying Lemma 18 twice allows to control $\|\lambda(\phi, I, \tilde{w}) - \lambda(\phi', I, \tilde{w})\|_\infty$.

Lemma 18 *Assume $\mathcal{M} = \mathbb{R}^d$. Let $(I, J) \in \mathcal{I}^2$, such that $I \neq J$, $\phi \in \mathbb{R}^d$ and $\tilde{w} \in \mathcal{S}_{\mathcal{A}}$. Let $\lambda(\phi, I, \tilde{w}, J) \in \operatorname{argmin}_{\lambda \in \bar{\Theta}_J^I} \langle \tilde{w}, \frac{(\phi-\lambda)^2}{\sigma^2}\rangle$. Then,*

$$\lambda(\phi, I, \tilde{w}, J) = \begin{cases} \phi & \textit{if } \phi \in \bar{\Theta}_J^I \\ \phi - (\mu_{a_0} - \alpha_{a_0})\delta_{a_0} & \textit{if } a_0 \in \textit{supp}(\tilde{w})^\complement \cap I \triangle J \neq \emptyset \\ \phi - \frac{\langle \mathbf{1}_J - \mathbf{1}_I, \phi\rangle}{\sum_{\tilde{a} \in I \triangle J} \frac{\sigma_{\tilde{a}}^2}{\tilde{w}_{\tilde{a}}}} \left(\frac{\sigma_a^2}{\tilde{w}_a} \left(\mathbf{1}_{a\in J} - \mathbf{1}_{a \in I}\right)\right)_{a\in[d]} & \textit{else} \end{cases}$$

where $\alpha_{a_0} = -\frac{(\mathbf{1}_{J\setminus\{a_0\}} - \mathbf{1}_{I\setminus\{a_0\}})^\intercal \phi}{\mathbf{1}_{a_0\in J} - \mathbf{1}_{a_0 \in I}}$.

Proof The proof uses the fact that $\bar{\Theta}_J^I = \{\lambda \in \mathbb{R}^d : \langle \mathbf{1}_J - \mathbf{1}_I, \lambda\rangle \geq 0\}$ and the KKT conditions. ■

Adapted Lemma 16 This lemma is used in Appendix G.3.2 when $I^*(\mu) = I_{t'}$. We apply Lemma 18 twice, for $\lambda(\mu_{t'-1}, I^*(\mu), \frac{\tilde{N}_{t'-1}}{t'-1})$ and $\lambda(\mu, I^*(\mu), \frac{\tilde{N}_{t'-1}}{t'-1})$ and we use that $\|\mu_{t'-1} - \mu\|_\infty \leq \sqrt{2\|\sigma^2\|_\infty f(t)}$ by concentration. Therefore, we can control $\|\lambda(\mu_{t'-1}, I^*(\mu), \frac{\tilde{N}_{t'-1}}{t'-1}) - \lambda(\mu, I^*(\mu), \frac{\tilde{N}_{t'-1}}{t'-1})\|_\infty$.

Adapted Lemma 17 This lemma is used in Appendix G.4. Using the closed-form formula for $r_{s,a}$ in Lemma 8, we obtain: for all $s \leq t$ and all $a \in [d]$, $C_{s,a} = \frac{f(s-1)}{\tilde{N}_{s-1,a}} + \sqrt{\frac{2f(s-1)}{\sigma_a^2 \tilde{N}_{s-1,a}}} |\mu_{s-1,a} - \lambda_{s,a}|$. Using Lemma 18 for $\lambda_s = \lambda(\mu_{s-1}, I_s, \tilde{w}_s)$, we can control $\|\mu_{s-1} - \lambda_s\|_\infty$.

Adapted Lemma 14 in Degenne et al. (2019) This lemma is used in Appendix G.3.2. Applying Lemma 18 for $\lambda(\mu, I^*, \sum_{s=1}^{t'-1} \tilde{w}_s)$ and using that $\|\mu_s - \mu\|_\infty \leq \sqrt{2\|\sigma^2\|_\infty f(t)}$ for all $s \leq t$, by concentration, we can control $\|d_{\mathrm{KL}}(\mu, \lambda(\mu, I^*, \sum_{s=1}^{t'-1} \tilde{w}_s)) - d_{\mathrm{KL}}(\mu_{s-1}, \lambda(\mu, I^*, \sum_{s=1}^{t'-1} \tilde{w}_s))\|_\infty$.

Adapted Lemma 7 This Lemma is used in Appendix F. The closed-form formula for $r_{t,a}$ in Lemma 8 is: $r_{t,a} = \frac{(\mu_{t-1,a}-\lambda_{t,a})^2}{2\sigma_a^2} + \frac{f(t-1)}{\tilde{N}_{t-1,a}} + \sqrt{\frac{2f(t-1)}{\sigma_a^2 \tilde{N}_{t-1,a}}}|\mu_{t-1,a} - \lambda_{t,a}|$ for all $a \in [d]$. Using Lemma 18 for $\lambda_s = \lambda(\mu_{s-1}, I_s, \tilde{w}_s)$, we obtain an upper bound on $\|r_t\|_\infty$, which will be used to bound R_t^A.

Adapted proof of $\tau_\delta < +\infty$ We use this result in the proof of Theorem 2 (Appendix D). Applying Lemma 18 for $\lambda(\mu_{t-1}, I^*(\mu), \frac{\tilde{N}_{t-1}}{t-1})$ and using $\lim_\infty \langle \frac{\tilde{N}_{t-1}}{t-1}, d_{\mathrm{KL}}(\mu_{t-1}, \lambda(\mu_{t-1}, I^*(\mu), \frac{\tilde{N}_{t-1}}{t-1})) \rangle$, we can conclude similarly.

Appendix I. Implementation Details

D-Tracking D-Tracking tracks w_t instead of $\sum_{s=1}^t w_s$ (Garivier and Kaufmann, 2016). It can be used instead of C-Tracking. Sparse D-Tracking is defined as: $A_t \in \operatorname{argmin}_{A \in B_t} \frac{N_{t-1,A}}{w_{t,A}}$ where $B_t = \mathrm{supp}(w_t)$. D-Tracking has been shown to empirically outperform C-Tracking (Degenne et al., 2019; Garivier and Kaufmann, 2016). In our experiments, C-Tracking and D-Tracking have similar results, up to a few percent. Therefore, we omit C-Tracking from the graphs.

In Appendix C of Degenne and Koolen (2019), the reason why D-Tracking might fail to converge is discussed. It stems from the fact that D-Tracking does not in general converge to the convex hull of the points it tracks. Due to the non-uniqueness of the optimal allocations, D-Tracking might also fail in our setting. For linear bandits Degenne et al. (2020a) showed that D-Tracking is licensed theoretically in order to obtain asymptotically optimal algorithms. In lights of those facts, whether D-Tracking is theoretically validated in our setting remains open.

One learner As in Degenne et al. (2019), we consider only one learner $\mathcal{A}^A$ instead of partitioning the rounds according to the candidate answer I_t. Experimentally, the results when considering $|\mathcal{I}|$ learners are always within a few percent of the one learner implementation. Therefore, we omit them from the graphs.

When considering $|\mathcal{I}|$ learners, one might ask what is the number of called learners before stopping. Since a learner is not used until its corresponding answer is the candidate answer, we expect this number to be small in comparison to $|\mathcal{I}|$. Our experiments validate this intuition: the used learners are the one for I^* and the ones for the most confusing alternatives. Considering a similar game-inspired algorithm, Tirinzoni et al. (2020) present a rigorous reason for using only one learner instead of $|\mathcal{I}|$ different ones.

Stylized stopping threshold and exploration bonus As in Degenne et al. (2019), we use stylized stopping threshold $\beta(t, \delta) = \ln\left(\frac{1+\ln(t)}{\delta}\right)$ and exploration bonus $f(t) = \ln(t)$ instead of the ones licensed by the theory. Despite being unlicensed yet, they are both empirically conservative since the empirical error rate is order of magnitude lower than the theoretical confidence error δ.

Sparsification As shown in Table 1, the computational complexity of both OFW and LLOO can become a hurdle when $|B_t| \gg d$. This problem was mentioned and tackled in Garber and Hazan

(2013). To circumvent it, we use an offline sparsification procedure to obtain an approximation $\tilde{w}_{t,0}$ of $\tilde{w}_t$ with sparse support.

Let $\tilde{w}_{t,0}$ be the approximate solution to the optimization problem $\min_{y \in \text{Im}(W_{\mathcal{A}})} \|y - \tilde{w}_t\|_2^2$ obtained thanks to Algorithm 2 in Garber and Hazan (2013), up to precision r^2. By Theorem 2 in Garber and Hazan (2013), this offline smooth and strongly convex optimization algorithm satisfies: $\|x_{s+1} - \tilde{w}_t\|_2^2 \leq C \exp\left(-\frac{1}{4\rho^2} s\right)$. The algorithm maintains a representation $w_{t,0} \in \Delta_{|\mathcal{A}|}$.

When $|B_t| \gg d$, we solve this optimization and use $(\tilde{w}_{t,0}, w_{t,0})$ instead of $(\tilde{w}_t, w_t)$. The parameters of LLOO are modified accordingly to Lemma 10 in Garber and Hazan (2013).

Doubling trick The horizon T corresponds to the stopping time τ_δ which is unknown. Therefore, we need to convert the non-anytime learners, Hedge and LLOO, into anytime learners. The geometric doubling trick (Besson and Kaufmann, 2018) can be used for that purpose. It preserves the minimax bounds in $R_t = O(\sqrt{t})$. In our experiments, we use the geometric doubling trick sequence $\left(\lfloor T_0 b^i \rfloor\right)_{i \in \mathbb{N}}$ where $T_0 = 200$ and $b = \frac{3+\sqrt{5}}{2}$ as advocated in Besson and Kaufmann (2018).

Covering initialization When considering a covering initialization, the sole requirement is to observe each arm at least once. Due to the combinatorial nature of the problem, numerous combinations of actions are valid initialization. Since our algorithms on the transformed simplex have a computational cost which is sensitive to $|B_{n_t}|$, we will consider an initialization such that the number of actions n_0 required to observe all arms is the smallest. When numerous choices achieve lowest n_0, we choose one arbitrarily. Alternatively one could sample randomly the actions without replacement till observing each arm at least once. This random covering initialization often damages simultaneously the sample complexity and the computational cost.

LLOO's parameters We recall here the definitions of the geometric parameters for the polytope $\mathcal{S}_{\mathcal{A}}$ used in Garber and Hazan (2013). The diameter of $\mathcal{S}_{\mathcal{A}}$ is $\text{diam}(\mathcal{S}_{\mathcal{A}}) := \max_{x,y \in \mathcal{S}_{\mathcal{A}}} \|x - y\|_2$. The parameter $\mu_{\mathcal{A}}$ is defined as $\mu_{\mathcal{A}} := \frac{\psi_{\mathcal{A}} \text{diam}(\mathcal{S}_{\mathcal{A}})}{\phi_{\mathcal{A}}}$ where $\psi_{\mathcal{A}}$ and $\phi_{\mathcal{A}}$ are also geometric parameters. A convex polytope admits a description with linear inequalities, $\mathcal{S}_{\mathcal{A}} = \{x \in \mathbb{R}^d : A_1 x = b_1 \wedge A_2 x \leq b_2\}$. $\phi_{\mathcal{A}}$ is defined as $\phi_{\mathcal{A}} := \min_{A \in \mathcal{A}} \{\min\{b_2(j) - \langle A_2(j), \mathbf{1}_A \rangle : j \in [m], b_2(j) > \langle A_2(j), \mathbf{1}_A \rangle\}\}$. It measures the deviation from equality constraints. $\psi_{\mathcal{A}}$ is defined as $\psi_{\mathcal{A}} := \max_{M \in \mathbb{A}_{\mathcal{A}}} \|M\|$, where $\|.\|$ is the spectral norm, $r(A_2)$ is the row rank of A_2 and $\mathbb{A}_{\mathcal{A}}$ is the set of $r(A_2) \times d$ matrices whose rows are linearly independent vectors chosen from the rows of A_2. Computing $\psi_{\mathcal{A}}$ is computationally expensive for high dimensional polytope. In such case we use an approximate $\psi_{\mathcal{A}}$, computed with a greedy algorithm. The parameter $\mu_{\mathcal{A}}$ is invariant to translation, rotation and scaling.

GCB-PE In the concurrent work of Chen et al. (2020), GCB-PE aims at solving the best-action problem for partial linear feedback. Chen et al. (2020) use a different notion of sample complexity, which is defined as a time T such that with probability $1 - \delta$, the algorithm returns the correct answer before time T. In our work, the sample complexity is the expected stopping time of the algorithm, which is required to be correct with probability $1 - \delta$.

The correspondence between our notations and theirs is: $M_x = S_A := \left(\mathbf{1}_{(\tilde{a}=a)}\right)_{\tilde{a} \in A, a \in [d]}$, $\bar{r}(I, \theta) = \langle \mathbf{1}_I, \theta \rangle$, $L_p = \sqrt{\max_{I \in \mathcal{I}} |I|}$. Since our experiments consider BAI with semi-bandit feedback, we need to adapt the Algorithm 1 of Chen et al. (2020). The sole modification is to consider $\hat{I} = \text{argmax}_{I \in \mathcal{I}} \bar{r}(I, \hat{\theta}(n))$ and $\hat{I}^- = \text{argmax}_{I \in \mathcal{I} \setminus \{\hat{I}\}} \bar{r}(I, \hat{\theta}(n))$ instead of $\hat{A}$ and $\hat{A}^-$.

The computational complexity of GCB-PE is sensitive to the choice of the global observer set. This choice corresponds to the random covering initialization in our setting, $\sigma = B_{n_0}$. Based on σ,

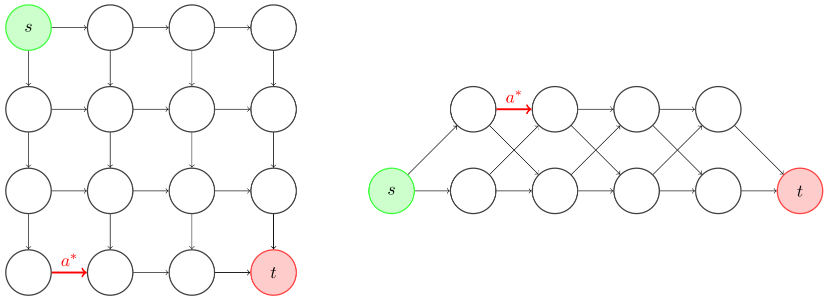

Figure 2: Paths examples: (a) grid network with $n_s = 6$ and (b) line network with $(n_n, n_l) = (2, 4)$

they define a constant β_σ which is used for the stopping rule. Unfortunately, β_σ is the solution of the following NP-hard binary quadratic program:

$$\beta_\sigma^2 = \max_{(\eta_i)_{i\in[|\sigma|]}\in[-1,1]^{m_\sigma}} \left\| \frac{1}{N_{n_0}} \odot \sum_{i=1}^{|\sigma|} S_{A_i}^\intercal \eta_i \right\|_2^2 = \max_{\eta\in\{-1,1\}^{m_\sigma}} \eta^\intercal P_\sigma \eta$$

where $m_\sigma = \sum_{i=1}^{|\sigma|} |A_i| \gg d$, $P_\sigma = M_\sigma \mathrm{diag}\left(\frac{1}{N_{n_0}^2}\right) M_\sigma^\intercal$, $M_\sigma^\intercal = \left[S_{A_1}^\intercal \cdots S_{A_{|\sigma|}}^\intercal\right]$ and $\odot$ denotes the component-wise multiplication. To our knowledge, there is no efficient solver for this optimization.

In our experiments on GCB-PE we will compute β_σ by testing the 2^{m_σ} possibilities. This restricts our results to small examples since the computational cost is increasing exponentially.

I.1. Experimental Results

As illustrative examples we use the best-arm identification by sampling actions. The bandit is Gaussian, $\nu = \mathcal{N}(\mu, \sigma^2 I_d)$. As regards the action set, we will consider:

- uniform matroid, $\mathcal{A} = \{A \subset [d] : |A| = k\}$, where the agent samples batches of size k. The batch setting is useful for real-world applications and admits an efficient oracle, the greedy algorithm.

- paths, $\mathcal{A} = \{A \subset [d] : A \in \mathrm{path}(s, t, \mathcal{G})\}$, where the agent samples paths connecting (s, t) in the graph $\mathcal{G}$. The path setting is omnipresent for network applications and admits efficient oracles, such as Dijkstra's algorithm. As a first illustrative example, we will consider a grid network with n_s stages, also known as binomial bridges. A grid network with $n_s = 6$ is represented in Figure 2(a). Grid networks appear in real-world applications. They were also studied in Kveton et al. (2015). As a second illustrative example we will consider a line network with n_l layers and redundancy n_n (number of nodes per layer). A line network with $(n_n, n_l) = (2, 4)$ is represented in Figure 2(b). Line networks appear in real-world applications. The redundancy ensures the system to be robust against failures.

- almost all sets, $\mathcal{A} = \left(\{I^*\} \cup \{A \in 2^d : I^* \not\subset A\}\right) \setminus \{\emptyset\}$, where the agent samples a set. This example is purely artificial. There is no efficient oracle. We designed it as an extreme needle-in-haystack problem where there is only one informative action among an exponential number of actions.

	d	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{A}^*\rvert$	$\frac{\lvert\mathcal{A}^*\rvert}{\lvert\mathcal{A}\rvert}$	n_0
Uniform matroid	d	$\binom{d}{k}$	$\binom{d-1}{k-1}$	$\frac{k}{d}$	$\lceil \frac{d}{k} \rceil$
Grid network	$n_s(\frac{n_s}{2}+1)$	$\binom{n_s}{n_s/2}$	1	$1/\binom{n_s}{n_s/2}$	n_s
Line network	$2n_n + (n_l - 1)n_n^2$	$n_n^{n_l}$	$n_n^{n_l-2}$	$1/n_n^2$	n_n^2
Almost all sets	d	2^{d-1}	1	$1/2^{d-1}$	2

Table 3: Central quantities

The central quantities of interest are summarized in Table 3: the dimension d, the size of the action sets $|\mathcal{A}|$, the size of the informative action set (actions containing the best arm) $|\mathcal{A}^*|$ where $\mathcal{A}^* = \{A \subset [d] : I^* \subset A\}$, the ratio of informative actions $\frac{|\mathcal{A}^*|}{|\mathcal{A}|}$ and the minimal number of actions to perform a covering initialization n_0. Intuitively, the lower the ratio of informative actions is, the harder the problem is for naive algorithms. For example, uniform sampling fails drastically when $|\mathcal{A}^*|$ is low and σ is high. When comparing learners on the simplex and the ones on the transformed simplex, the difference between the sizes of the respective initialization can have an important role, $|\mathcal{A}| - n_0$. The learners on $\mathcal{S}_{\mathcal{A}}$ spend this additional budget on exploring relevant actions instead of merely sampling them all. The lower the noise, the more significant this difference is. In the no-noise setting, at most n_0 samples are necessary for the learners on the transformed simplex, while at most $|\mathcal{A}|$ samples are necessary for the ones on the simplex. The exact sample complexity depends on $|\mathcal{A}^*|$ and on the random draw of actions.

In the additional experiments, we will only compare AdaHedge and LLOO since they are the best instance in their family of learner (Figure 1).

I.1.1. UNIFORM MATROID

By increasing the dimension d, we observe the effect of an exponential increase of $|\mathcal{A}| = \binom{d}{k}$ while the ratio of informative actions $\frac{|\mathcal{A}^*|}{|\mathcal{A}|} = \frac{k}{d}$ is decreasing harmonically. Since $|\mathcal{A}^*| = \binom{d-1}{k-1}$ is also increasing with d and k, we need to consider higher noise for $k = 3$ than for $k = 2$. Otherwise, the sampling rules using a full initialization will satisfy the stopping criterion before the end of the initialization.

For experiments on uniform matroids in Figures 1 and 3, we consider $\mu^{(d)} \in \mathbb{R}^d$ for all $d \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$, such that $\mu_1^{(d)} = 0.3$, $\mu_2^{(d)} = 0.29$, $\mu_3^{(d)} = 0.28$ and $\mu_i^{(d)} < 0.24$ for $i > 3$. Those values ensure that the best arm is always $I^*(\mu) = \{1\}$, while having two serious contenders $J \in \{\{2\}, \{3\}\}$. The rest of the arms are chosen ordered such as they are clearly suboptimal: $\mu_{4:5}^{(5)} = \{0.23, 0.2\}$, $\mu_{4:10}^{(10)} = \{0.232, 0.224, 0.207, 0.200, 0.192, 0.182, 0.176\}$, $\mu_{4:15}^{(15)} = \mu_{4:10}^{(10)} \cup \{0.214, 0.199, 0.195, 0.190, 0.164\}$, $\mu_{4:20}^{(20)} = \mu_{4:15}^{(15)} \cup \{0.185, 0.19, 0.195, 0.199, 0.214\}$, $\mu_{4:25}^{(25)} = \mu_{4:20}^{(20)} \cup \{0.158, 0.172, 0.211, 0.228, 0.244\}$, $\mu_{4:30}^{(30)} = \mu_{4:25}^{(25)} \cup \{0.174, 0.18, 0.194, 0.202, 0.23, 0.242\}$, $\mu_{4:35}^{(35)} = \mu_{4:30}^{(30)} \cup \{0.17, 0.178, 0.219, 0.222, 0.226\}$, $\mu_{4:40}^{(40)} = \mu_{4:35}^{(35)} \cup \{0.197, 0.198, 0.201, 0.203, 0.205\}$, $\mu_{4:45}^{(45)} = \mu_{4:40}^{(40)} \cup \{0.193, 0.206, 0.208, 0.21\}$ and $\mu_{4:50}^{(50)} = \mu_{4:45}^{(45)} \cup \{0.188, 0.189, 0.191, 0.212, 0.213\}$.

In Figures 3(a) and 3(b), we observe an identical behavior as in Figures 1(a) and 1(b). LLOO has competitive sample complexity for a low and almost constant computational cost compared to AdaHedge.

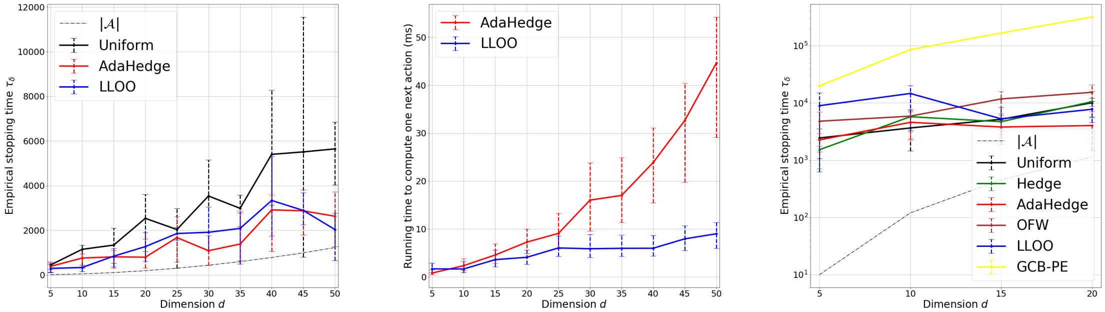

Figure 3: Uniform matroid, $k = 2$ (resp. $k = 3$ for plot (c)), Gaussian bandit, $\nu = \mathcal{N}(\mu, \sigma^2 I_d)$ with $\sigma = 0.035$ (resp. $\sigma = 0.1$). Influence of the dimension d on: (a) (resp. (c)) the empirical stopping time τ_δ and (b) the average running time to compute the next action.

I.1.2. Grid Network

By increasing the number of stages n_s, we observe the effect of an exponential increase of $|\mathcal{A}| = \binom{n_s}{n_s/2}$ and an exponential decrease of $\frac{|\mathcal{A}^*|}{|\mathcal{A}|} = \frac{1}{\binom{n_s}{n_s/2}}$ since $|A^*| = 1$.

For experiments on the grid networks in Figure 4, we consider $\mu^{(n_l)} \in \mathbb{R}^d$ for all $n_s \in \{6, 8, 10, 12, 14, 16\}$. The values for the parameters $\mu^{(n_l)}$ were obtained by random sampling with a Gaussian of mean 0.2 and standard deviation 0.025. After sorting, we increment $\mu_1^{(n_l)}$ by 0.025 to ensure that $I^*(\mu) = \{1\}$ with a statistically significant gap.

The Figure 4(a) highlights two important intuitive facts. First, the uniform sampling is highly inefficient in terms of samples when few informative actions are available, here $|A^*| = 1$. Second, the empirical performance of a learner on the simplex is limited by the initialization of size $n_0 = |\mathcal{A}|$. The Figure 4(b) highlights the lower computational cost of LLOO compared to AdaHedge. The slightly higher cost stems from the more expensive efficient oracle to solve the shortest path offline problem.

I.1.3. Line Network

By increasing the number of layers n_l, we observe the effect of an exponential increase of $|\mathcal{A}| = n_n^{n_l}$ while the ratio of informative actions is decreasing as $\frac{|\mathcal{A}^*|}{|\mathcal{A}|} = \frac{1}{n_n^2}$. The number of informative actions $|\mathcal{A}^*| = n_n^{n_l - 2}$ is also increasing with n_l, slowly for low n_n.

For experiments on the line networks in Figure 5, we consider $\mu^{(n_l)} \in \mathbb{R}^d$ for all $n_l \in \{5, \cdots, 12\}$ when $n_n = 2$ and $n_l \in \{4, \cdots, 8\}$ when $n_n = 3$. The values for the parameters $\mu^{(n_l)}$ were obtained by sampling randomly from a Gaussian with mean 0.2 and standard deviation 0.025. After sorting, we increment $\mu_1^{(n_l)}$ by 0.025 to ensure that $I^*(\mu) = \{1\}$ with a statistically significant gap.

In Figure 5, the take-away message is similar as for uniform matroids. LLOO has competitive sample complexity for a low computational cost compared to AdaHedge.

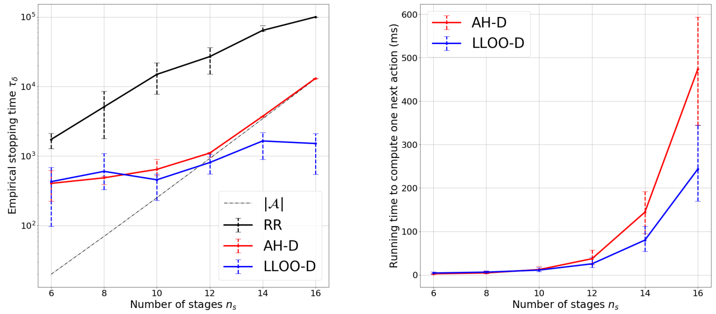

Figure 4: Grid network, Gaussian bandit, $\nu = \mathcal{N}(\mu, \sigma^2 I_d)$ with $\sigma = 0.075$. Influence of the number of stages n_s on: (a) the empirical stopping time τ_δ and (b) the average running time to compute the next action.

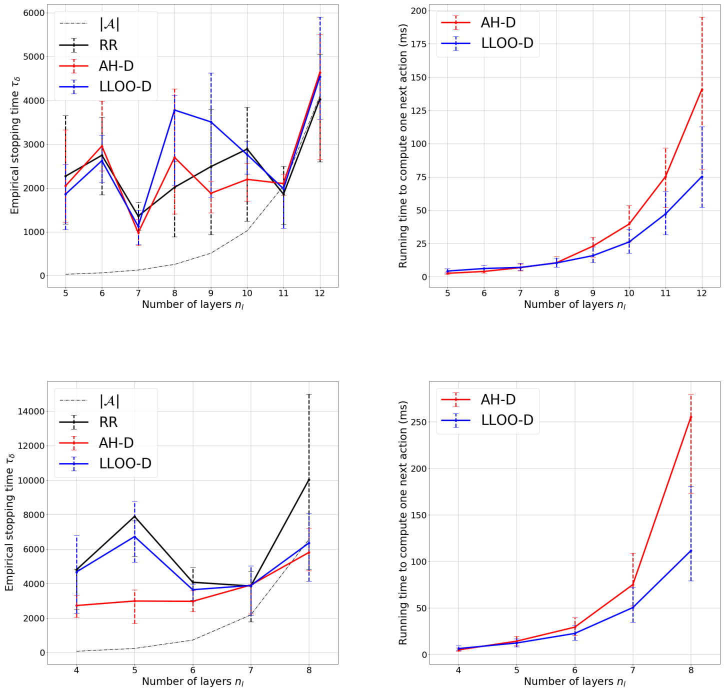

Figure 5: Line network, $n_n = 2$ (resp. $n_n = 3$ for the bottom plots), Gaussian bandit, $\nu = \mathcal{N}(\mu, \sigma^2 I_d)$ with $\sigma = 0.2$. Influence of the number of layers n_l on: (a) (resp. (c)) the empirical stopping time τ_δ and (b) (resp. (d)) the average running time to compute the next action.

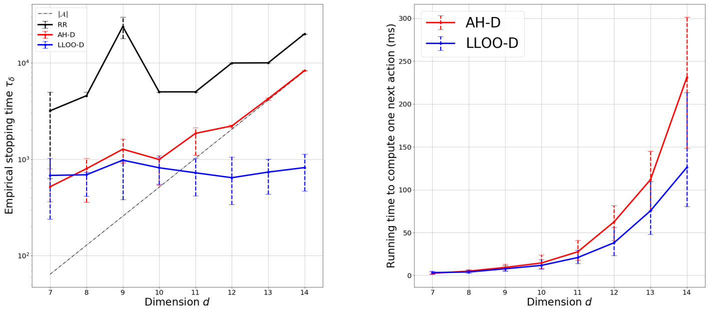

Figure 6: Almost all sets, Gaussian bandit, $\nu = \mathcal{N}(\mu, \sigma^2 I_d)$ with $\sigma = 0.25$. Influence of the dimension d on: (a) the empirical stopping time τ_δ and (b) the average running time to compute the next action.

I.1.4. ALMOST ALL SETS

By increasing the dimension d, we observe the effect of an exponential increase of $|\mathcal{A}| = 2^{d-1}$ and an exponential decrease of $\frac{|\mathcal{A}^*|}{|\mathcal{A}|} = \frac{1}{2^{d-1}}$ since $|A^*| = 1$.

For experiments on almost all sets in Figure 6, we consider $\mu^{(d)} \in \mathbb{R}^d$ for all $d \in \{7, \cdots, 14\}$, such that $\mu_1^{(d)} = 0.3$ and $\mu_i^{(d)} \leq 0.24$ for $i > 1$. Those values ensure that the best arm is always $I^*(\mu) = \{1\}$. The rest of the arms are chosen ordered such as they are clearly suboptimal: $\mu_{2:7}^{(7)} = \{0.24, 0.23, 0.22, 0.21, 0.2, 0.19\}$, $\mu_{2:8}^{(8)} = \mu_{2:7}^{(7)} \cup \{0.18\}$, $\mu_{2:9}^{(9)} = \mu_{2:8}^{(8)} \cup \{0.17\}$, $\mu_{2:10}^{(10)} = \mu_{2:9}^{(9)} \cup \{0.16\}$, $\mu_{2:11}^{(11)} = \mu_{2:10}^{(10)} \cup \{0.215\}$, $\mu_{2:12}^{(12)} = \mu_{2:11}^{(11)} \cup \{0.195\}$, $\mu_{2:13}^{(13)} = \mu_{2:12}^{(12)} \cup \{0.205\}$ and $\mu_{2:14}^{(14)} = \mu_{2:13}^{(13)} \cup \{0.185\}$.

In Figure 6(a), the take-away message is similar as for the grid networks. Even though no efficient oracle exists, the computational cost of LLOO is still lower than the one of AdaHedge, see Figure 6(b).