Dealing With Misspecification In Fixed-Confidence Linear Top-m Identification

Clémence Réda, Andrea Tirinzoni, Rémy Degenne

Dealing With Misspecification In Fixed-Confidence Linear Top-m Identification

Clémence Réda
Université de Paris, NeuroDiderot, Inserm, F-75019 Paris, France
clemence.reda@inria.fr

Andrea Tirinzoni
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
andrea.tirinzoni@inria.fr

Rémy Degenne
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
remy.degenne@inria.fr

Abstract

We study the problem of the identification of m arms with largest means under a fixed error rate δ (fixed-confidence Top-m identification), for misspecified linear bandit models. This problem is motivated by practical applications, especially in medicine and recommendation systems, where linear models are popular due to their simplicity and the existence of efficient algorithms, but in which data inevitably deviates from linearity. In this work, we first derive a tractable lower bound on the sample complexity of any δ-correct algorithm for the general Top-m identification problem. We show that knowing the scale of the deviation from linearity is necessary to exploit the structure of the problem. We then describe the first algorithm for this setting, which is both practical and adapts to the amount of misspecification. We derive an upper bound to its sample complexity which confirms this adaptivity and that matches the lower bound when $\delta \to 0$. Finally, we evaluate our algorithm on both synthetic and real-world data, showing competitive performance with respect to existing baselines.

1 Introduction

The multi-armed bandit (MAB) is a popular framework to model sequential decision making problems. At each round $t > 0$, a learner chooses an *arm* k_t among a finite set of $K \in \mathbb{N}$ possible options, and it receives a random reward $X_t^{k_t} \in \mathbb{R}$ drawn from a distribution ν^{k_t} with unknown mean μ^{k_t}. Among the many problem settings studied in this context, we focus on *pure exploration*, where the learner aims at maximizing the information gain for answering a given query about the arms [5]. In particular, we are interested in finding a subset of $m \geq 1$ arms with largest expected reward, which is known as the *Top-m identification* problem [22]. This generalizes the widely-studied best-arm (i.e., Top-1) identification problem [16]. This problem has several important applications, including online recommendation and drug repurposing [31, 35]. Two objectives are typically studied. On the one hand, in the *fixed-budget* setting [2], the learner is given a finite amount of samples and must return a subset of m best arms while minimizing the probability of error in identification. On the other hand, in the *fixed-confidence* setting [16], the learner aims at minimizing the *sample complexity* for returning a subset of m best arms with a fixed maximum error rate $\delta \in (0, 1)$, defined as the number of samples collected before the algorithm stops. This paper focuses on the latter.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

In practice, information about the arms is typically available (e.g., the characteristics of an item in a recommendation system, or the influence of a drug on protein production in a clinical application). This side information influence the expected rewards of the arms, thus adding *structure* (*i.e.*, prior knowledge) to the problem. This is in contrast to the classic *unstructured* MAB setting, where the learner has no prior knowledge about the arms. Due to their simplicity and flexibility, linear models have become the most popular to represent this structure. Formally, in the *linear bandit* setting [3], the mean reward μ^k of each arm $k \in [K] := \{1, 2, \dots, K\}$ is assumed to be an inner product between known d-dimensional arm features $\phi_k \in \mathbb{R}^d$ and an unknown parameter $\theta \in \mathbb{R}^d$. This model has led to many provably-efficient algorithms for both best-arm [38, 42, 17, 43, 13] and Top-m identification [24, 35]. Unfortunately, the strong guarantees provided by these algorithms hold only when the expected rewards are perfectly linear in the given features, a property that is often violated in real-world applications. In fact, when using linear models with real data, one inevitably faces the problem of *misspecification*, i.e., the situation in which the data deviates from linearity.

A *misspecified linear bandit* model is often described as a linear bandit model with an additive term to encode deviation from linearity. Formally, the expected reward $\mu^k = \phi_k^\top \theta + \eta^k$ of each arm $k \in [K]$ can be decomposed into its linear part $\phi_k^\top \theta$ and its misspecification $\eta^k \in \mathbb{R}$. Note the flexibility of this model: for $\|\eta\| = 0$, where $\eta = [\eta^1, \eta^2, \dots, \eta^K]^\top$, the problem is perfectly linear and thus highly structured, as the mean rewards of different arms are related through the common parameter θ; whereas when the misspecification vector η is large in all components, the problem reduces to an unstructured one, since knowing the linear part alone provides almost no information about the expected rewards. Learning in this setting thus requires adapting to the scale of misspecification, typically under the assumption that some information about the latter is known (e.g., an upper bound ε to $\|\eta\|$). Due to its importance, this problem has recently gained increasing attention in the bandit community for regret minimization [20, 29, 18, 33, 39]. However it has not been addressed in the context of pure exploration. In this paper, we take a step towards bridging this gap by studying fixed-confidence Top-m identification in the context of misspecified linear bandits. Our detailed contributions are as follows.

Contributions. (1) We derive a tractable lower bound on the sample complexity of any δ-correct algorithm for the general Top-m identification problem. (2) Leveraging this lower bound, we show that knowing an upper bound ε to $\|\eta\|$ is necessary for adapting to the scale of misspecification, in the sense that any δ-correct algorithm without such information cannot achieve a better sample complexity than that obtainable when no structure is available. (3) We design the first algorithm for Top-m identification in misspecified linear bandits. We derive an upper bound to its sample complexity that holds for any $\delta \in (0, 1)$ and that matches our lower bound for $\delta \to 0$. Notably, our analysis reveals a nice adaptation to the value of ε, recovering state-of-the-art dependences in the linear case ($\varepsilon = 0$), where the sample complexity scales polynomially in d and not in K, and in the unstructured case (ε large), where only polynomial terms in K appear. (4) We evaluate our algorithm on synthetic problems and real datasets from drug repurposing and recommendation system applications, while showing competitive performance with state-of-the-art methods.

Related work. While model misspecification has not been addressed in the pure exploration literature, several attempts to tackle this problem in the context of regret minimization exist. In [20], the authors show that, if T is the learning horizon, for any bandit algorithm which enjoys $\mathcal{O}(d\sqrt{T})$ regret scaling on linear models, there exists a misspecified instance where the regret is necessarily linear. As a workaround, the authors design a statistical test based on sampling a subset of arms prior to learning to decide whether a linear or an unstructured bandit algorithm should be run on the data. Similar ideas are presented in [8], where the authors design a sequential test to switch online between linear and unstructured models. More recently, elimination-based algorithms [29, 39] and model selection methods [33, 18] have attracted increasing attention. Notably, these algorithms adapt to the amount of misspecification ε *without* knowing it beforehand, at the cost of an additive linear term that scales with ε. Moreover, while best-arm identification has been the focus of many prior works in the realizable linear setting, some suggesting asymptotically-optimal algorithms [13, 21], Top-m identification has been seldom studied in terms of problem-dependent lower bounds. Lower bounds for the unstructured Top-m problem have been derived previously, focusing on explicit bounds [26], on getting the correct dependence in the problem parameters for any confidence δ [9, 37], or on asymptotic optimality (as $\delta \to 0$) [19]. Because of the combinatorial nature of the Top-m identification problem, obtaining a tractable, tight, problem-dependent lower bound is not straightforward.

2 Setting

At successive stages $t \in \mathbb{N}$, the learner samples an arm $k_t \in [K]$ based on previous observations and internal randomization (a random variable $U_t \in [0,1]$) and observes a reward $X_t^{k_t}$. Let $\mathcal{F}_t := \sigma(\{U_1, k_1, X_1^{k_1}, \dots, U_t, k_t, X_t^{k_t}, U_{t+1}\})$ be the σ-algebra associated with past sampled arms and rewards until time t. Then k_t is a $\mathcal{F}_{t-1}$-measurable random variable. The reward $X_t^{k_t}$ is sampled from ν^{k_t} and is independent of all past observations, conditionally on k_t. We suppose that the noise is Gaussian with variance 1, such that the observation when pulling arm k_t at time t is $X_t^{k_t} \sim \mathcal{N}(\mu^{k_t}, 1)$. The mean vector $\mu = (\mu^k)_{k\in[K]} \in \mathbb{R}^K$ then fully describes the reward distributions.

In a misspecified linear bandit, each arm $k \in [K]$ is described by a feature vector $\phi_k \in \mathbb{R}^d$. The corresponding feature matrix is denoted by $A := [\phi_1, \phi_2, \dots, \phi_K]^\top \in \mathbb{R}^{K\times d}$ and the maximum ℓ_2 -norm of these vectors is $L := \max_{k\in[K]} \|\phi_k\|_2$. We assume that the feature vectors span $\mathbb{R}^d$ (otherwise we could rewrite those vectors in a subspace of smaller dimension). We assume that the learner is provided with a set of realizable models

$$\mathcal{M} := \left\{\mu \in \mathbb{R}^K \mid \exists \theta \in \mathbb{R}^d \, \exists \eta \in \mathbb{R}^K, \; \mu = A\theta + \eta \;\wedge\; \|\mu\|_\infty \le M \;\wedge\; \|\eta\|_\infty \le \varepsilon\right\}, \qquad (1)$$

where $M, \varepsilon \in \mathbb{R}$ are known upper bounds on the ℓ^∞-norm of the mean[1] and misspecification vectors, respectively. Intuitively, $\mathcal{M}$ represents the set of bandit models whose mean vector μ is linear in the features A only up to some misspecification η.

We consider Top-m identification in the fixed-confidence setting. Given a confidence parameter $\delta \in (0,1)$, the learner is required to output the $m \in [K]$ arms of the unknown bandit model $\mu \in \mathcal{M}$ with highest means with probability at least $1-\delta$. The strategy of a bandit algorithm designed for Top-m identification can be decomposed into three rules: a *sampling* rule, which selects the arm k_t to sample at a given learning round t according to past observations; a *stopping* rule, which determines the end of the learning phase, and is a stopping time with respect to the filtration $(\mathcal{F}_t)_{t>0}$, denoted by τ_δ; finally, a *decision* rule, which returns a $\mathcal{F}_{\tau_\delta}$-measurable answer to the pure exploration problem. An answer is a set $\hat{S}_m \subseteq [K]$ with exactly m arms: $|\hat{S}_m| = m$. In our context, the "m best arms of μ" might not be well defined since the set $S^\star(\mu) := \{k \in [K] \mid \mu^k \ge \max^m_{i\in[K]} \mu^i\}$[2] might contain more than m elements if some arms have the same mean. Thus, let $\mathcal{S}_m(\mu) = \{S \subseteq S^\star(\mu) \mid |S| = m\}$ be the set containing all subsets of m elements of $S^\star(\mu)$.

Definition 1 (δ-correctness)**.** For $\delta \in (0,1)$, we say that an algorithm $\mathfrak{A}$ is δ-correct on $\mathcal{M}$ if, for all $\mu \in \mathcal{M}$, $\tau_\delta < +\infty$ almost surely and $\mathbb{P}^{\mathfrak{A}}_\mu\left(\hat{S}_m \notin \mathcal{S}_m(\mu)\right) \le \delta$.

3 Tractable lower bound for the general Top-m identification problem

Let N_t^k denote the number of times arm k has been sampled until time t included. Suppose that the true model μ has exactly m arms that are among the top-m, *i.e.*, that $|S^\star(\mu)| = m$ and $\mathcal{S}_m(\mu) = \{S^\star(\mu)\}$. Consider the following set of alternatives to μ,

$$\Lambda_m(\mu) := \{\lambda \in \mathcal{M} \mid \mathcal{S}_m(\lambda) \cap \mathcal{S}_m(\mu) = \emptyset\} \;,$$

that is, the set of all bandit models λ in $\mathcal{M}$ where the top-m arms of μ are not among the top-m arms of λ. Note that, while we assumed that the set of top-m arms in μ is unique, this might not be the case for λ. Define the event $E_{\tau_\delta} := \{\hat{S}_m \in \mathcal{S}_m(\mu)\}$ that the answer returned by the algorithm at τ_δ is correct for μ and consider any δ-correct algorithm $\mathfrak{A}$. Let us call KL the Kullback-Leibler divergence[3] and kl the binary relative entropy. Then, using the change-of-measure argument proposed in [19, Theorem 1], for any $\lambda \in \Lambda_m(\mu)$ and $\delta \le 1/2$,

$$\sum_{k\in[K]} \mathbb{E}^{\mathfrak{A}}_\mu[N^k_\tau]\,\mathrm{KL}\left(\mu^k, \lambda^k\right) \ge \mathrm{kl}\left(\mathbb{P}^{\mathfrak{A}}_\mu(E_{\tau_\delta}), \mathbb{P}^{\mathfrak{A}}_\lambda(E_{\tau_\delta})\right) \ge \mathrm{kl}(1-\delta, \delta) \ge \log\left(\frac{1}{2.4\delta}\right),$$

[1] The restriction to $\|\mu\|_\infty \le M$ is required only for our analysis, while it can be safely dropped in practice.
[2] The expression $\max^m_{i\in S} f(i)$ denotes the m^{th} maximal value in $\{f(i) \mid i \in S\}$.
[3] We abuse notation by denoting distributions in the same one-dimensional exponential family by their means.

where the second-last inequality follows from the δ-correctness of the algorithm and the monotonicity of the function kl. This holds for any $\lambda \in \Lambda_m(\mu)$, so we have that

$$\mathbb{E}_\mu^{\mathfrak{A}}[\tau] \geq \left(\sup_{\omega \in \Delta_K} \inf_{\lambda \in \Lambda_m(\mu)} \sum_{k \in [K]} \omega^k \mathrm{KL}\left(\mu^k, \lambda^k\right) \right)^{-1} \log\left(\frac{1}{2.4\delta}\right), \tag{2}$$

with $\Delta_K := \{p \in [0,1]^K \mid \sum_{k=1}^K p_k = 1\}$ the simplex on $[K]$. We define the inverse complexity $H_\mu := \sup_{\omega \in \Delta_K} \inf_{\lambda \in \Lambda_m(\mu)} \sum_{k \in [K]} \omega^k \mathrm{KL}\left(\mu^k, \lambda^k\right)$. Computing that lower bound might be difficult: while the Kullback-Leibler is convex for Gaussians, the set $\Lambda_m(\mu)$ over which it is minimized is non-convex. Its description using $\mathcal{S}_m(\lambda)$ is combinatorial: we can write $\Lambda_m(\mu)$ as a union of convex sets, one for each subset of top-m arms of λ, but this implies minimizing over $\binom{K}{m}$ sets, which is not practical. In order to rewrite this lower bound, we prove the following lemma in Appendix C.

Lemma 1. $\forall \mu, \lambda \in \mathbb{R}^K$ *s.t.* $|S^\star(\mu)| = m$, $\mathcal{S}_m(\lambda) \cap \mathcal{S}_m(\mu) = \emptyset \Leftrightarrow \exists i \notin S^\star(\mu)\ \exists j \in S^\star(\mu), \lambda^i > \lambda^j$.

Lemma 1 allows us to go from an exponentially costly optimization problem, which implied minimizing over $\binom{K}{m}$ sets, to optimizing across $m(K-m)$ halfspaces. Therefore, by replacing the set of alternative models as derived in Lemma 1, the lower bound in Equation 2 can be rewritten in the following more convenient form :

Theorem 1. *For any $\delta \leq 1/2$, for any δ-correct algorithm $\mathfrak{A}$ on $\mathcal{M}$, for any bandit instance $\mu \in \mathbb{R}^K$ such that $|S^\star(\mu)| = m$, the following lower bound holds on the stopping time τ_δ of $\mathfrak{A}$ on instance μ:*

$$\mathbb{E}_\mu^{\mathfrak{A}}[\tau_\delta] \geq \left(\sup_{\omega \in \Delta_K} \min_{i \notin S^\star(\mu)} \min_{j \in S^\star(\mu)} \inf_{\lambda \in \mathcal{M}: \lambda^i > \lambda^j} \sum_{k \in [K]} \omega^k \mathrm{KL}\left(\mu^k, \lambda^k\right) \right)^{-1} \log\left(\frac{1}{2.4\delta}\right).$$

Computing the lower bound now requires performing one maximization over the simplex (which can be still hard), and $m(K-m)$ minimizations over half-spaces $\{\lambda \in \mathcal{M} : \lambda^i > \lambda^j\}$, where $(i,j) \in (\mathcal{S}^\star(\mu))^c \times \mathcal{S}^\star(\mu)$. The minimizations are convex optimization problems and can be solved efficiently. Our algorithm inspired from that bound will need to perform only those minimizations.

Note that a lower bound for Top-m identification using the cited change-of-measure argument has been obtained in [26]. Aiming to be more explicit, it relies on alternative models where one of the best arms is switched with the $(m+1)^{th}$ best one (or one of the $K-m$ worst ones with the m^{th} best one). These models are a strict subset of $\Lambda_m(\mu)$. Hence this bound is not as tight as the one in Theorem 1, which is why the algorithm we detail in the next sections will rely on the latter instead.

Note that with $\varepsilon = 0$ and $m = 1$, this lower bound is exactly the one for best arm identification in perfectly linear models [17]. As the misspecification ε grows, the set $\mathcal{M}$ becomes larger and so does the set of alternative models $\Lambda_m(\mu)$, thus the lower bound grows. In the limit $\varepsilon \to +\infty$, the model becomes the same as the unstructured model. We show that in fact the lower bound becomes exactly equal to the unstructured lower bound as soon as $\varepsilon > \varepsilon_\mu$, a finite value.

Lemma 2. *There exists $\varepsilon_\mu \in \mathbb{R}$ with $\varepsilon_\mu \leq \max_k \mu^k - \min_k \mu^k$ such that if $\varepsilon > \varepsilon_\mu$, then the lower bound of Theorem 1 is equal to the unstructured top-m lower bound.*

The proof is in Appendix C. It considers finitely supported distributions over $\Lambda_m(\mu)$ that realize the equilibrium in the max-min game of the lower bound. As soon as one of these equilibrium distributions for the unstructured problem has its whole support in the misspecified model, the two complexities are equal.

3.1 Adaptation to unknown misspecification is impossible

We now make an important observation: knowing that a problem is misspecified without knowing an upper bound ε on $\|\eta\|_\infty$ is the same as not knowing anything about the structure of that problem.

The lower bound of Equation (2) is a function of the set $\mathcal{M}$ of realizable models μ. Let $B(\mu, \delta, \mathcal{M})$ be the right-hand side of that equation, such that $\mathbb{E}_\mu^{\mathfrak{A}}[\tau_\delta] \geq B(\mu, \delta, \mathcal{M})$ for any algorithm $\mathfrak{A}$ which is δ-correct on $\mathcal{M}$. Suppose that we have $\mathcal{M}_1 \subseteq \mathcal{M}$, a subset of the model, for which we would like to have lower sample complexity (possibly at the cost of a higher sample complexity on $\mathcal{M} \setminus \mathcal{M}_1$). If

Algorithm 1 MISLID

Require: Set of models $\mathcal{M}$, online learner $\mathcal{L}$, stopping thresholds $\{\beta_{t,\delta}\}_{t\geq 1}$

Compute a sequence of arms $k_1, \dots, k_{t_0}$ such that $\sum_{t=1}^{t_0} \phi_{k_t} \phi_{k_t}^T \succeq 2L^2 I_d$ // INITIALIZATION
for $t = 1, \dots, t_0$ **do**
  Pull k_t, receive $X_t^{k_t}$, and set $\omega_t \leftarrow e_{k_t}$ // PULL SPANNER
end for
Compute empirical mean $\widehat{\mu}_{t_0}$ and its projection $\tilde{\mu}_{t_0} \leftarrow \arg\min_{\lambda \in \mathcal{M}} \|\lambda - \widehat{\mu}_{t_0}\|^2_{D_{N_{t_0}}}$
for $t = t_0 + 1, t_0 + 2, \dots,$ **do**
  if $\inf_{\lambda \in \Lambda_m(\tilde{\mu}_{t-1})} \|\tilde{\mu}_{t-1} - \lambda\|^2_{D_{N_{t-1}}} > 2\beta_{t-1,\delta}$ **then** // STOPPING RULE
    Stop and return $\mathcal{S}_m^\star(\tilde{\mu}_{t-1})$
  end if
  Obtain ω_t from $\mathcal{L}$
  Compute closest alternative: $\lambda_t \leftarrow \arg\min_{\lambda \in \Lambda_m(\tilde{\mu}_{t-1})} \|\tilde{\mu}_{t-1} - \lambda\|^2_{D_{\omega_t}}$
  Update $\mathcal{L}$ with gain $g_t : \omega \mapsto \sum_{k \in [K]} \omega^k \left(|\tilde{\mu}_{t-1}^k - \lambda_t^k| + \sqrt{c_{t-1}^k} \right)^2$ // UPDATE LEARNER
  Pull $k_t \sim \omega_t$ and receive reward $X_t^{k_t}$ // ACTION SAMPLING
  Update $\widehat{\mu}_t$ and compute projection $\tilde{\mu}_t \leftarrow \arg\min_{\lambda \in \mathcal{M}} \|\lambda - \widehat{\mu}_t\|^2_{D_{N_t}}$ // ESTIMATION
end for

$\mathcal{M}$ is the misspecified linear model with deviation ε, let us say that $\mathcal{M}_1$ is the set of problems with deviation lower than $\varepsilon_1 < \varepsilon$; that is, we want the algorithm to be faster on more linear models. This is not achievable. The lower bound states that it is *not* possible for an algorithm to have lower sample complexity on $\mathcal{M}_1$ while being δ-correct on $\mathcal{M}$. On every $\mu \in \mathcal{M}$, the lower bound is $B(\mu, \delta, \mathcal{M})$.

An algorithm cannot adapt to the deviation to linearity: it has to use a parameter ε set in advance, and its sample complexity will depend on that ε, not on the actual deviation of the problem. Note that this observation does not contradict recent results for regret minimization [e.g., 29, 39], which show that adapting to an unknown scale of misspecification is possible. In fact, such results involve a "weak" form of adaptivity, where the algorithms provably leverage the linear structure at the price of suffering an additive *linear* regret term of order $\mathcal{O}(\varepsilon\sqrt{d}T)$, where T is the learning horizon. Since the counterpart of δ-correctness for regret minimization is "the algorithm suffers sub-linear regret in T for all instances of the given family", this implies that algorithms with such "weak" adaptivity loose this important property of consistency.

4 The MISLID algorithm

We introduce MISLID (Misspecified Linear Identification), an algorithm to tackle misspecification in linear bandit models for fixed-confidence Top-m identification. We describe the algorithm in Section 4.1, while in Section 4.2 we report its sample complexity analysis.

4.1 Algorithm

The pseudocode of MISLID is outlined in Algorithm 1. On the one hand, the design of MISLID builds on top of recent approaches for constructing pure exploration algorithms from lower bounds [12, 13, 43, 21]. On the other hand, its main components and their analysis introduce several technical novelties to deal with misspecified Top-m identification, that might be of independent interest for other settings. We describe these components below. Let us define $D_v := \mathrm{diag}(v^1, v^2, \dots, v^K)$ for any vector $v \in \mathbb{R}^K$, and $V_t := \sum_{s=1}^t \phi_{k_s} \phi_{k_s}^\top$.

Initialization phase. MISLID starts by pulling a deterministic sequence of t_0 arms that make the minimum eigenvalue of the resulting design matrix V_{t_0} larger than $2L^2$. Since the rows of A span $\mathbb{R}^d$, such sequence can be easily found by taking any subset of d arms that span the whole space (e.g., by computing a barycentric spanner [4]) and pulling them in a round robin fashion until the desired condition is met. This is required to make the design matrix invertible. While the literature typically avoid this step by regularizing (*e.g.*, [1]), in our misspecified setting it is crucial not to do

so to obtain tight concentration results for the estimator of μ, as explained in the next paragraph. See Appendix D.1 for a discussion of the length t_0 of that initialization phase.

Estimation. At each time step $t \geq t_0$, MISLID maintains an estimator $\tilde{\mu}_t$ of the true bandit model μ. This is obtained by first computing the empirical mean $\widehat{\mu}_t$, such that $\widehat{\mu}_t^k = \frac{1}{N_t^k} \sum_{s=1}^t \mathbb{1}\{k_s = k\} X_s^{k_s}$, and then projecting it onto the family of realizable models $\mathcal{M}$ according to the D_{N_t}-weighted norm, i.e., $\tilde{\mu}_t := \arg\min_{\lambda \in \mathcal{M}} \|\lambda - \widehat{\mu}_t\|_{D_{N_t}}^2$. Since each $\lambda \in \mathcal{M}$ can be decomposed into $\lambda = A\theta' + \eta'$ for some $\theta' \in \mathbb{R}^d$ and $\eta' \in \mathbb{R}^K$, this can be solved efficiently as the minimization of a quadratic objective in $K + d$ variables subject to the linear constraints $\|\eta'\|_\infty \leq \varepsilon$ and $\|A\theta' + \eta'\|_\infty \leq M$. The second constraint is only required for the analysis, while it often has a negligible impact in practice. Thus, we shall drop it in our implementation, which yields two independent optimization problems for the projection $\tilde{\mu}_t = A\tilde{\theta}_t + \tilde{\eta}_t$: one for $\tilde{\theta}_t$, whose solution is available in closed form as the standard least-squares estimator $\tilde{\theta}_t = \hat{\theta}_t := V_t^{-1} \sum_{s=1}^t X_s^{k_s} \phi_{k_s}$, and one for $\tilde{\eta}_t$, which is another quadratic program with K variables (see Appendix D).

A crucial component in the concentration of these estimators, and a key novelty of our work, is the adoption of an orthogonal parametrization of mean vectors. In particular, we leverage the following observation: any mean vector $\mu = A\theta + \eta$ can be equivalently represented, at any time t, as $\mu = A\theta_t + \eta_t$, where $\theta_t = V_t^{-1} \sum_{s=1}^t \mu^{k_s} \phi_{k_s}$ is the orthogonal projection (according to the design matrix V_t) of μ onto the feature space and $\eta_t = \mu - A\theta_t$ is the residual. Then, it is possible to show that $\|\hat{\theta}_t - \theta_t\|_{V_t}^2$ is *exactly* the self-normalized martingale considered in [1] and, thus, it enjoys the *same* bound we have in linear bandits with no misspecification (refer to Appendix B). This is an important advantage over prior works [29, 44] that, in order to concentrate $\hat{\theta}_t$ to θ, need to inflate the concentration rate by a factor $\varepsilon^2 t$, which often makes the bound too large to be practical for misspecified models with $\varepsilon \gg 0$. It allows us to also avoid superlinear terms of the form $\epsilon^2 t \log(t)$ which are present in related works and which would prevent us from deriving good problem-dependent guarantees.

Stopping rule. MISLID uses the standard stopping rule adopted in most existing algorithms for pure exploration [19, 12, 36]. What makes it peculiar is the definition of the thresholds $\beta_{t,\delta}$. MISLID requires a careful combination of concentration inequalities for (1) linear bandits, to make the algorithm adapt well to linear models with low ε, and (2) unstructured bandits, to guarantee asymptotic optimality. The precise definition of $\beta_{t,\delta}$ is shown in the following result.

Lemma 3 (MISLID is δ-correct)**.** *Let W_{-1} be the negative branch of the Lambert W function and let $\overline{W}(x) = -W_{-1}(-e^{-x}) \approx x + \log x$. For $\delta \in (0,1)$, define*

$$\beta_{t,\delta}^{\text{uns}} := 2K\overline{W}\left(\frac{1}{2K}\log\frac{2e}{\delta} + \frac{1}{2}\log(8eK\log t)\right), \tag{3}$$

$$\beta_{t,\delta}^{\text{lin}} := \frac{1}{2}\left(4\sqrt{t}\varepsilon + \sqrt{2}\sqrt{1 + \log\frac{1}{\delta} + \left(1 + \frac{1}{\log(1/\delta)}\right)\frac{d}{2}\log\left(1 + \frac{t}{2d}\log\frac{1}{\delta}\right)}\right)^2. \tag{4}$$

Then, for the choice $\beta_{t,\delta} := \min\{\beta_{t,\delta}^{\text{uns}}, \beta_{t,\delta}^{\text{lin}}\}$, MISLID is δ-correct.

This result is a simple consequence of two (linear and unstructured) concentration inequalities. See Appendix F.

Sampling strategy and online learners. The sampling strategy of MISLID aims at achieving the optimal sample complexity from the lower bound in Theorem 1. As popularized by recent works [12, 13, 43], instead of relying on inefficient max-min oracles to repeatedly solve the optimization problem of Theorem 1 [17, 21], we compute it incrementally by employing no-regret online learners. At each step t, the learner $\mathcal{L}$ plays a distribution over arms $\omega_t \in \Delta_K$ and it is updated with a gain function g_t whose precise definition will be specified shortly. Then, MISLID directly samples the next arm to pull from the distribution ω_t, instead of using tracking as in the majority of previously mentioned works. Similarly to what was recently shown by [40] for regret minimization in linear bandits, sampling will be crucial in our analysis to reduce dependencies on K and, in particular, to obtain only logarithmic dependencies in the realizable linear case.

Regarding the choice of $\mathcal{L}$, two important properties are worth mentioning. First, MISLID requires only a *single* learner, while existing asymptotically optimal algorithms for pure exploration [12, 13]

need to allocate one learner for each possible answer. Since the number of answers is $\binom{K}{m}$, a direct extension of these algorithms to the Top-m setting would yield an impractical method with exponential (in K) number of learners, hence space complexity, and possibly sample complexity.[4] Second, the choice of $\mathcal{L}$ is highly flexible since any learner that satisfies the following property suffices.

Definition 2 (No-regret learner)**.** A learner $\mathcal{L}$ over Δ_K is said to be no-regret if, for any $t \geq 1$ and any sequence of gains $\{g_s(\omega)\}_{s\leq t}$ bounded in absolute value by $B \in \mathbb{R}^+$, there exists a positive constant $C_{\mathcal{L}}(K, B)$ such that $\max_{w\in\Delta_K} \sum_{s=1}^{t} \big(g_s(w) - g_s(w_s)\big) \leq C_{\mathcal{L}}(K, B)\sqrt{t}$.

Examples of algorithms in this class are Exponential Weights [7] and AdaHedge [15]. The latter shall be our choice for the implementation since it does not use B as a parameter but adapts to it, and thus does not suffer from a possibly loose bound on B.

Optimistic gains. Finally, we need to specify how the gains g_t are computed. Clearly, if μ were known, one would directly use $g_t : \omega \mapsto \inf_{\lambda\in\Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_\omega}$. Since μ is unknown and must be estimated, we set $g_t(\omega)$ to an *optimistic* proxy for that quantity. In particular, we choose a sequence of bonuses $\{c_t^k\}_{t\geq t_0, k\in[K]}$ such that, with high probability, $g_t(\omega_t) := \sum_{k\in[K]} \omega_t^k \left(|\tilde{\mu}_{t-1}^k - \lambda_t^k| + \sqrt{c_{t-1}^k}\right)^2 \geq \inf_{\Lambda\in\lambda_m(\mu)} \|\mu - \lambda\|^2_{D_{\omega_t}}$, for $\lambda_t := \arg\min_{\lambda\in\Lambda_m(\tilde{\mu}_{t-1})} \|\tilde{\mu}_{t-1} - \lambda\|^2_{D_{\omega_t}}$. As for the stopping thresholds, we construct c_t^k by a careful combination of structured and unstructured concentration bounds:

$$c_t^k := \min\left\{8(LK+1)^2\varepsilon^2 + 4\alpha^{\text{lin}}_{t^2}\|\phi_k\|^2_{V_t^{-1}}, \frac{2\alpha^{\text{uns}}_{t^2}}{N_t^k}, 4M^2\right\},$$

where $\alpha_t^{\text{uns}} := \beta^{\text{uns}}_{t,1/(5t^3)}$ and $\alpha_t^{\text{lin}} := \log(5t^2) + d\log(1 + t/(2d))$. We show in Appendix F that this choice of c_t^k suffices to guarantee optimism with high probability.

4.2 Sample complexity

Theorem 2. MISLID *has expected sample complexity* $\mathbb{E}_\mu[\tau_\delta] \leq T_0(\delta) + 2$*, where* $T_0(\delta)$ *is the solution to the equation in* t

$$\beta_{t,\delta} \geq tH_\mu + \widehat{\mathcal{O}}\left(\min\{tK^2\varepsilon^2 + d\sqrt{t}\ell_t, \sqrt{Kt}\ell_t\}; \log K\sqrt{t}; \sqrt{\min\{tK^2\varepsilon^2 + d\ell_t, K\ell_t\}\log(1/\delta)}\right), \tag{5}$$

where $\ell_t := \log t$*,* H_μ *is the inverse complexity appearing in the lower bound (see Equation 2), and* $\widehat{\mathcal{O}}(a; b; c)$ *represent a sum of terms, each of which is* $\mathcal{O}$ *of one of the expressions shown.*

See Appendix F for the proof. Since $\beta^{\text{uns}}_{t,\delta} \approx \log(1/\delta)$ for small δ, $T_0(\delta) = H_\mu^{-1}\log(1/\delta) + C_\mu o(\log(1/\delta))$, where C_μ is a problem-dependent constant. Then $\liminf_{\delta\to 0} \mathbb{E}_\mu[\tau_\delta]/\log(1/\delta) = \liminf_{\delta\to 0} T_0(\delta)/\log(1/\delta) = H_\mu^{-1}$ and thus the upper bound matches the lower bound in that limit: MISLID is asymptotically optimal. The only polynomial factors in K are in a minimum with a term that depends on ε. In the linear setting, when $\varepsilon = 0$, we have only logarithmic (and no polynomial) dependence on the number of arms, which is on par with the state of the art [40, 21, 27]. Moreover, the bound exhibits an adaptation to the value of ε. If ε is small, then the minimums in $\beta_{t,\delta}$ and in the inequality (5) are equal to the "linear" values which involve $K\varepsilon$ and d instead of K. As ε grows, the upper bound transitions to terms matching the optimal unstructured bound.

Decoupling the stopping and sampling analyses. Our analysis decomposes into two parts: first, a result on the stopping rule, then, a discussion of the sampling rule. The algorithm is shown to verify that, under a favorable event, if it does not stop at time t,

$$2\beta_{t,\delta} \geq \inf_{\lambda\in\Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_{N_t}} - \mathcal{O}(\sqrt{t}) \geq 2tH_\mu - \mathcal{O}(\sqrt{t}) .$$

The sample complexity result is a consequence of that bound on t. The first inequality is due solely to the stopping rule, and the second one only to the sampling mechanism. The expression

[4] The fact that the optimization problem of the lower bound decomposes into $m(K-m)$ minimizations does not reduce the number of possible answers, which is still combinatorial in K.

$\inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_{N_t}}$ does not feature any variable specific to the algorithm: we can combine any stopping rule and any sampling rule, as long as they each verify the corresponding inequality.

A more aggressive optimism. The optimistic gains that we have chosen, $g_t(\omega) = \sum_{k \in [K]} \omega^k (|\tilde{\mu}^k_{t-1} - \lambda^k_t| + \sqrt{c^k_{t-1}})^2$, are tuned to ensure asymptotic optimality (with a factor 1 in the leading term). If we instead accept to be asymptotically optimal up to a factor 2, we can use the gains $g_t(\omega) = \sum_{k \in [K]} \omega^k \left((\tilde{\mu}^k_{t-1} - \lambda^k_t)^2 + c^k_{t-1} \right)$. When using those, the learner takes decisions which are much closer to those it would take if using the empirical gains $\sum_{k \in [K]} \omega^k (\tilde{\mu}^k_{t-1} - \lambda^k_t)^2$ and the theoretical bound, while worse in the leading factor, has better lower order terms. The aggressive optimism sometimes has significantly better practical performance (see Experiment (C) in Figure 1).

5 Experimental evaluation

Since our algorithm is the first to apply to Top-m identification in misspecified linear models, we compare it against an efficient linear algorithm, LinGapE [42] (that is, its extension to Top-m as described in [35], which coincides with LinGapE for $m = 1$), and an unstructured one, LUCB [23]. In all experiments, we consider $\delta = 5\%$. [5] For each algorithm, we show boxplots reporting the average sample complexity on the y-axis, and the error frequency $\hat{\delta}$ across 500 (resp. 100) repetitions for simulated (resp. real-life) instances rounding up to the 5^{th} decimal place. Individual outcomes are shown as gray dots. It has frequently been noted in the fixed-confidence literature that stopping thresholds which guarantee δ-correctness tend to be too conservative and to yield empirical error frequencies that are actually much lower than δ. Moreover, these thresholds are different from linear to unstructured models. In order to ensure a good trade-off between performance and computing speed, and fairness between tested algorithms, we use a heuristic value for the stopping rule $\beta_{t,\delta} := \ln((1 + \ln(t+1))/\delta)$ unless otherwise specified. For each experiment, we report the number of arms (K), the dimension of features (d), the size of the answer (m), the misspecification (ε) and the gap between the m^{th} and $(m+1)^{th}$ best arms ($\Delta := \max^m_{a \in [K]} \mu^a - \max^{m+1}_{b \in [K]} \mu^b$). The computational resources used, data licenses and further experimental details can be found in Appendix G.

(A) Simulated misspecified instances. ($K = 10$, $d = 5$, $m = 3$, $\varepsilon \in \{0, 5\}$, $\Delta \approx 0.28$) First, we fix a linear instance $\mu := A\theta$ by randomly sampling the values of $\theta \in \mathbb{R}^d$ and $A \in \mathbb{R}^{K \times d}$ from a zero-mean Gaussian distribution, and renormalizing them by their respective ℓ_∞ norm. Then, for $\varepsilon \in \{0, 5\}$, we build a misspecified linear instance $\mu_\varepsilon = A\theta + \eta_\varepsilon$, such that, if (4) is the index of the fourth best arm, $\forall k \neq (4), \eta^k_\varepsilon = 0$, and $\eta^{(4)}_\varepsilon = \varepsilon$. Note that any value of $\varepsilon < \Delta$ does not switch the third and fourth arms in the set of best arms of μ_ε, contrary to values greater than Δ. The greater ε is, the more different the answers from the linear and misspecified models are. This experiment was inspired by [20], where a similar model is used to prove a lower bound in the setting of regret minimization. See the leftmost two plots on Figure 1. As expected, LUCB is always δ-correct, but suffers from a significantly larger sample complexity than its structured counterparts. Moreover, LinGapE does not preserve the δ-correctness under large misspecification level $\varepsilon = 5$ (with error rate $\hat{\delta} = 0.96$), which illustrates the effect of ε on the answer set. Note that it is not due to the choice of stopping threshold, as running it with the theoretically-supported threshold derived in [1] also yields an empirical error rate $\hat{\delta} = 1$. MISLID proves to be competitive against LinGapE. Note that the case $\varepsilon = 0$ is a perfectly linear instance. See Table 2 in Appendix for numerical results for algorithms LinGapE and MISLID.

(B) Discrepancy between user-selected ε and true $\|\eta\|$. ($K = 15$, $d = 8$, $m = 3$, $\varepsilon \in \{0.5, 1, 2\}$, $\Delta \approx 0.4$) MISLID crucially relies on a user-provided upper bound on the scale of deviation from linearity. We test its robustness against perturbations to the input value ε compared to the value $\varepsilon^\star := \|\eta\|_\infty$ in the misspecified model $\mu := A\theta + \eta$. Values are sampled randomly for θ, A, η, and the associated vectors are normalized by their ℓ^∞ norm (for η, by $\|\eta\|_\infty / \varepsilon^\star$), where $\varepsilon^\star = 1 > \Delta$ is the true deviation to linearity. The results, shown in the third plot of Figure 1, display the behavior predicted by Lemma 2. Indeed, as the user-provided value ε increases, the associated sample complexity increases as well. The plateau in sample complexity when ε is large enough is noticeable. Cases $\varepsilon \in \{1, 2\}$ display a sample complexity close to that of unstructured bandits.

[5] All the code and scripts are available at https://github.com/clreda/misspecified-top-m.

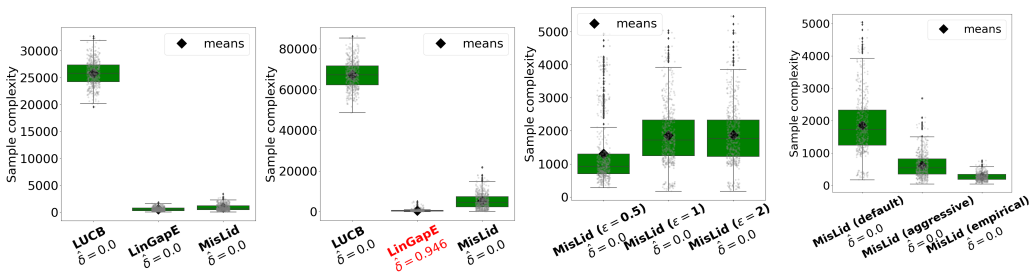

Figure 1: Experiment (A) for $\varepsilon \in \{0, 5\}$ (*first two from the left*). Experiment (B) with $\varepsilon \in \{0.5, 1, 2\}$. Experiment (C) to compare different optimistic gains (*right*).

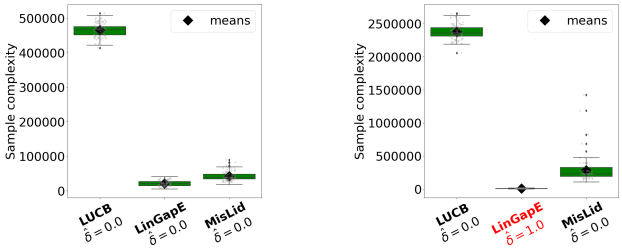

Figure 2: Experiment (D) for drug repurposing in epilepsy (*left*). Experiment (E) for online recommendation.

(C) Comparing different optimisms. ($K = 15$, $d = 8$, $m = 3$, $\varepsilon = 1$, $\Delta \approx 0.4$) We use the same bandit model as in Experiment (B), and use $\varepsilon = \varepsilon^\star = 1$. We compare the aggressive optimism described in Section 4.2, no optimism (that is, $\forall k \in [K], \forall t > 0, c_t^k = 0$), and the default optimistic gains given in Section 4.1. See the rightmost plot in Figure 1. The algorithm with no optimism is denoted "empirical", and is significantly faster than the optimistic variants.

(D) Application to drug repurposing. ($K = 10$, $d = 5$, $m = 5$, $\hat{\varepsilon} \approx 0.02$, $\Delta \approx 0.062$) We use the drug repurposing problem for epilepsy proposed by [35] to investigate the practicality of our method. In order to speed up LUCB, we consider the PAC version of Top-m identification, choosing as stopping threshold $0.06 \approx \Delta$, such that the algorithm stops earlier while returning the exact set of m best arms. Following [34, Appendix F.4], we extract a linear model from the data by fitting a neural network and taking the features learned in the last layer. We compute ε as the ℓ^∞ norm of the difference between the predictions of this linear model and the average rewards from the data, which yields $\hat{\varepsilon} = 0.02$. Since the misspecification is way below the minimum gap, and the linear model thus accurately fits the data, the results (leftmost plot in Figure 2) show that MisLid and LinGapE perform comparably on this instance. Moreover, both are an order of magnitude better than an unstructured bandit algorithm sample complexity-wise. Please refer to Table 3 in Appendix for numerical results for LinGapE and MISLID.

(E) Application to a larger instance of online recommendation. ($K = 103$, $d = 8$, $m = 4$, $\hat{\varepsilon} \approx 0.206$, $\Delta \approx 0.022$) As in Experiment (D), a linear representation is extracted for an instance of online recommendation of music artists to users (Last.fm dataset [6]). We compute a proxy for ε and feed the value Δ to the stopping threshold in LUCB in a similar fashion. Differently from Experiment (D), this yields a misspecification that is much larger than the minimum gap. To improve performance on these instances, we modified MISLID. To reduce the sample complexity, we use empirical gains instead of optimism. To reduce the computational complexity, we check the stopping rule infrequently (on a geometric grid) and use only a random subset of arms in each round to compute the sampling rule (see Appendix G for details and an empirical comparison to the theoretically supported MISLID). See the rightmost plot in Figure 2. This plot particularly illustrates our introductory claim: an unstructured bandit algorithm is δ-correct, but too slow in practice for misspecified instances, whereas the guarantee on correctness for a linear bandit does not hold anymore on these models with large misspecification. Numerical results for LinGapE and MISLID are listed in Table 3 in Appendix.

6 Discussion

We have designed the first algorithm to tackle misspecification in fixed-confidence Top-m identification, which has applications in online recommendation. However, the algorithm relies exclusively on the features provided in the input data, and as such might be subjected to bias and lack of fairness in its recommendation, depending on the dataset. The proposed algorithm can be applied to misspecified models which deviate from linearity (*i.e.*, $\varepsilon > 0$), encompassing unstructured settings (for large values of ε) and linear models (*i.e.*, $\varepsilon = 0$).

Our tests on variants of our algorithm suggest that the optimistic estimates have a big influence on the sample complexity. Removing the optimism completely and using the empirical gains leads to the best performance. We conjecture that other components of the algorithm like the learner are conservative enough for the optimism to be superfluous. The main limitation of our method is its computational complexity: at each round, $\mathcal{O}(Km)$ convex optimization problems need to be solved for both the sampling and stopping rules, which can be expensive if the number of arms is large. However, the "interesting" arms are much less numerous and we observed empirically that the sample complexity is not increased significantly if we consider only a few arms. In general, theoretically supported methods to replace the alternative set by computationally simpler approximations would greatly help in reducing the computational cost of our algorithm.

Since the sampling of our algorithm is designed to minimize a lower bound, we can expect it to suffer from the same shortcomings as that bound. It is known that the bound in question does not capture some lower order (in $1/\delta$) effects, in particular those due to the multiple-hypothesis nature of the test we perform, which can be very large for small times. Work to take these effects into account to design algorithms has started recently [24, 25, 41] and we believe that it is an essential avenue for further improvements in misspecified linear identification.

Acknowledgments and Disclosure of Funding

Clémence Réda was supported by the "Digital health challenge" Inserm-CNRS joint program, the French Ministry of Higher Education and Research [ENS.X19RDTME-SACLAY19-22], and the French National Research Agency [ANR-19-CE23-0026-04] (BOLD project).

References

[1] Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

[2] Audibert, J.-Y., Bubeck, S., and Munos, R. (2010). Best arm identification in multi-armed bandits. In *COLT*, pages 41–53.

[3] Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

[4] Awerbuch, B. and Kleinberg, R. (2008). Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114.

[5] Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer.

[6] Cantador, I., Brusilovsky, P., and Kuflik, T. (2011). Second workshop on information heterogeneity and fusion in recommender systems (hetrec2011). pages 387–388.

[7] Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

[8] Chatterji, N. S., Muthukumar, V., and Bartlett, P. L. (2020). OSOM: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 1844–1854. PMLR.

[9] Chen, L., Li, J., and Qiao, M. (2017). Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics*, pages 101–110. PMLR.

[10] De Rooij, S., Van Erven, T., Grünwald, P. D., and Koolen, W. M. (2014). Follow the leader if you can, hedge if you must. *The Journal of Machine Learning Research*, 15(1):1281–1316.

[11] Degenne, R. and Koolen, W. M. (2019). Pure exploration with multiple correct answers. *arXiv preprint arXiv:1902.03475*.

[12] Degenne, R., Koolen, W. M., and Ménard, P. (2019). Non-asymptotic pure exploration by solving games. In *NeurIPS*, pages 14465–14474.

[13] Degenne, R., Ménard, P., Shang, X., and Valko, M. (2020a). Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR.

[14] Degenne, R., Shao, H., and Koolen, W. (2020b). Structure adaptive algorithms for stochastic bandits. In *International Conference on Machine Learning*, pages 2443–2452. PMLR.

[15] Erven, T., Koolen, W. M., Rooij, S., and Grünwald, P. (2011). Adaptive hedge. *Advances in Neural Information Processing Systems*, 24:1656–1664.

[16] Even-Dar, E., Mannor, S., and Mansour, Y. (2003). Action elimination and stopping conditions for reinforcement learning. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 162–169.

[17] Fiez, T., Jain, L., Jamieson, K. G., and Ratliff, L. (2019). Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, volume 32.

[18] Foster, D. J., Gentile, C., Mohri, M., and Zimmert, J. (2020). Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33.

[19] Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR.

[20] Ghosh, A., Chowdhury, S. R., and Gopalan, A. (2017). Misspecified linear bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

[21] Jedra, Y. and Proutiere, A. (2020). Optimal best-arm identification in linear bandits. *arXiv preprint arXiv:2006.16073*.

[22] Kalyanakrishnan, S. and Stone, P. (2010). Efficient selection of multiple bandit arms: Theory and practice. In *ICML*.

[23] Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662.

[24] Katz-Samuels, J., Jain, L., Karnin, Z., and Jamieson, K. G. (2020). An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pages 10371–10382.

[25] Katz-Samuels, J. and Jamieson, K. (2020). The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, pages 1781–1791. PMLR.

[26] Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42.

[27] Kirschner, J., Lattimore, T., Vernade, C., and Szepesvári, C. (2020). Asymptotically optimal information-directed sampling. *arXiv preprint arXiv:2011.05944*.

[28] Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

[29] Lattimore, T., Szepesvari, C., and Weisz, G. (2020). Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR.

[30] Magureanu, S., Combes, R., and Proutiere, A. (2014). Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pages 975–999. PMLR.

[31] Mason, B., Jain, L., Tripathy, A., and Nowak, R. (2020). Finding all ε-good arms in stochastic bandits. *Advances in Neural Information Processing Systems*, 33.

[32] Orabona, F. (2019). A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*.

[33] Pacchiano, A., Phan, M., Abbasi Yadkori, Y., Rao, A., Zimmert, J., Lattimore, T., and Szepesvari, C. (2020). Model selection in contextual stochastic bandit problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 10328–10337.

[34] Papini, M., Tirinzoni, A., Restelli, M., Lazaric, A., and Pirotta, M. (2021). Leveraging good representations in linear contextual bandits. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8371–8380. PMLR.

[35] Réda, C., Kaufmann, E., and Delahaye-Duriez, A. (2021). Top-m identification for linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1108–1116. PMLR.

[36] Shang, X., Heide, R., Menard, P., Kaufmann, E., and Valko, M. (2020). Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, pages 1823–1832. PMLR.

[37] Simchowitz, M., Jamieson, K., and Recht, B. (2017). The simulator: Understanding adaptive sampling in the moderate-confidence regime. In *Conference on Learning Theory*, pages 1794–1834. PMLR.

[38] Soare, M., Lazaric, A., and Munos, R. (2014). Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, volume 27.

[39] Takemura, K., Ito, S., Hatano, D., Sumita, H., Fukunaga, T., Kakimura, N., and Kawarabayashi, K.-i. (2021). A parameter-free algorithm for misspecified linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3367–3375. PMLR.

[40] Tirinzoni, A., Pirotta, M., Restelli, M., and Lazaric, A. (2020). An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Advances in Neural Information Processing Systems*, 33.

[41] Wagenmaker, A., Katz-Samuels, J., and Jamieson, K. (2021). Experimental design for regret minimization in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3088–3096. PMLR.

[42] Xu, L., Honda, J., and Sugiyama, M. (2018). A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 843–851. PMLR.

[43] Zaki, M., Mohan, A., and Gopalan, A. (2020). Explicit best arm identification in linear bandits using no-regret learners. *arXiv preprint arXiv:2006.07562*.

[44] Zanette, A., Lazaric, A., Kochenderfer, M. J., and Brunskill, E. (2020). Learning near optimal policies with low inherent bellman error. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 10978–10989. PMLR.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 3 (lower bound for the Top-m identification problem), Section 3.1 (adaptivity to the misspecification), Section 4 (algorithm) and Section 5 (experiments).
   (b) Did you describe the limitations of your work? [Yes] See paragraph 2 in Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See paragraph 1 in Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendices.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Refer to the following GitHub repository: `https://github.com/clreda/misspecified-top-m`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See introductory and experiment-specific paragraphs in Section 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See boxplots in Section 5.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix G.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See paragraphs D and E in Section 5. Both real-life datasets are publicly available online.
   (b) Did you mention the license of the assets? [Yes] See Appendix G.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Table of Contents

A Notation

Table 1: Notation table

Name	Description
$d \in \mathbb{N}^*$	Dimension of the feature vectors
$K \in \mathbb{N}^*$	Number of arms
$[K] := \{1, 2, \dots, K\}$	Enumeration
$m \in [K-1]$	Number of best arms to return
$\mathbb{1}\{c\}$	Kronecker's symbol, equal to 1 iff. claim c is true
$\varepsilon \in \mathbb{R}^{*+}$	Upper bound on the ℓ_∞ norm of the deviation to linearity
$M \in \mathbb{R}^{*+}$	Upper bound on the ℓ_∞ norm on the mean vector
$L \in \mathbb{R}^{*+}$	Upper bound on the ℓ_2 norm on the arm feature vectors
$\delta \in (0,1)$	Upper bound for the probability of error in identification
$e_k \in \mathbb{R}^k, k \in \mathbb{N}$	k^{th} vector of the canonical basis of $\mathbb{R}^k$
$\Delta_K = \{p \in [0,1]^K \mid \sum_{k=1}^K p^k = 1\}$	Set of probability distributions over finite set of size K
$\phi_k \in \mathbb{R}^d, k \in [K]$	Feature vector for arm k
$A = [\phi_1, \phi_2, \dots, \phi_K]^\top \in \mathbb{R}^{K \times d}$	Feature matrix of arm contexts
$\Delta_K = \{p \in [0,1]^K \mid \sum_{k=1}^K p^k = 1\}$	Set of probability distributions on finite set of size K
$V_\omega := \sum_{k \le K} \omega_k \phi_k \phi_k^\top, \omega \in \Delta_K$	Design matrix associated with ω
$V_t := \sum_{s \le t} \phi_{k_s} \phi_{k_s}^\top, t > 0$	Design matrix at time t
$\mathcal{M} \subset \mathbb{R}^K$	Set of realizable models: $\{\mu \in \mathbb{R}^K \mid \exists \theta \in \mathbb{R}^d, \eta \in \mathbb{R}^K : \mu = A\theta + \eta, \|\mu\|_\infty \le M, \|\eta\|_\infty \le \varepsilon\}$
$\mu \in \mathcal{M}$	True mean vector: $\mu = A\theta + \eta$
$N_t^k \in \mathbb{N}, k \in [K], t > 0$	Number of times arm k has been sampled until time t included
$N_t = [N_t^1, N_t^2, \dots, N_t^K]^\top \in \mathbb{N}^K$	Vector of numbers of samplings for each arm at time t included
$D_N \in \mathbb{R}^{K \times K}, N \in \mathbb{R}^K$	Diagonal matrix with coefficients $N^1, N^2, \dots, N^K$
$k_s, s > 0$	Arm sampled at time s
$X_s^k, s > 0, k \in [K]$	Reward observed at time s from arm k
$\tau_\delta, \delta \in (0,1)$	Stopping time under δ-correctness
E_{τ_δ}	Event on δ-correctness: $E_{\tau_\delta} := \left\{\hat{S}_m \in \mathcal{S}_m(\mu)\right\}$
$\widehat{\mu}_t \in \mathbb{R}^K, t > 0$	Empirical mean vector at time t: $\widehat{\mu}_t^a := \frac{1}{N_t^a} \sum_{s \le t} X_s^a \mathbb{1}\{k_s = a\}$
$\tilde{\mu}_t \in \mathbb{R}^K, t > 0$	Projection of $\widehat{\mu}_t$ onto set $\mathcal{M}$ at time t
$\hat{S}_m \subseteq [K], m \in [K-1]$	Answer to Top-m identification as returned by the algorithm
$S^\star(\mu) \subseteq [K], \mu \in \mathbb{R}^K$	Set of best arms compared to the m^{th} greatest mean: $S^\star(\mu) := \left\{k \in [K] \mid \mu^k \ge \max_{i \in [K]}^m \mu^i\right\}$
$\mathcal{S}_m(\mu), \mu \in \mathcal{M}, m \in [K-1]$	Set of all subsets of size m in $S^\star(\mu)$: $\mathcal{S}_m(\mu) := \{S \subseteq S^\star(\mu) \mid \lvert S\rvert = m\}$
$\Lambda_m(\mu), \mu \in \mathcal{M}$	Set of alternative models to model μ: $\Lambda_m(\mu) := \{\lambda \in \mathcal{M} \mid \mathcal{S}_m(\lambda) \cap \mathcal{S}_m(\mu) = \emptyset\}$
$H_\mu, \mu \in \mathcal{M}$	Inverse complexity constant: $H_\mu := \sup_{\omega \in \Delta_K} \inf_{\lambda \in \Lambda_m(\mu)} \sum_{k \in [K]} \omega^k \operatorname{KL}(\mu^k, \lambda^k)$
KL	Kullback-Leibler divergence
kl	Binary relative enthropy
W_{-1}	Negative branch of the Lambert W function
$\overline{W} : x \mapsto -W_{-1}(-e^{-x})$	
$\mathcal{L}$	Learner algorithm
$g_t(\omega), \omega \in \mathbb{R}^K, t > 0$	Gains fed to the learner at time t
$c_t^k, k \in [K], t > 0$	Optimistic bonus, such that $(\tilde{\mu}_t^k - \mu^k)^2 \le c_t^k$ for any $k \in [K]$ and large enough $t > 0$, with high probability

Please refer to Table 1. Moreover, if $\omega \in \mathbb{R}^K$, at $t > 0$, we also introduce the following notation related to orthogonal parameterizations (see Appendix B):

- $A_\omega := D_\omega^{1/2} A \in \mathbb{R}^{K \times d}$.
- $P_\omega := A_\omega (A_\omega^\top A_\omega)^\dagger A_\omega^\top \in \mathbb{R}^{K \times K}$.
- $R_\omega := I_K - P_\omega \in \mathbb{R}^{K \times K}$, where I_K is the identity matrix of dimension K.
- $V_t = A_{N_t}^\top A_{N_t} = A^\top D_{N_t} A = \sum_{k \in [K]} N_t^k \phi_k \phi_k^\top = \sum_{s \leq t} \phi_{k_s} \phi_{k_s}^\top$.
- $\widehat{\theta}_t := (A_{N_t}^\top A_{N_t})^\dagger A_{N_t}^\top D_{N_t}^{1/2} \hat{\mu}_t$, which is the standard least-squares estimator, where † denotes the matrix pseudo-inverse.
- $\tilde{\theta}_t$ and $\tilde{\eta}_t$, parameters for the linear and misspecification parts of the projection $\tilde{\mu}_t$ of empirical mean $\widehat{\mu}_t$ onto set $\mathcal{M}$, such that $\tilde{\mu}_t = A\tilde{\theta}_t + \tilde{\eta}_t$.
- $\theta_t := (A_{N_t}^\top A_{N_t})^\dagger A_{N_t}^\top D_{N_t}^{1/2} \mu$, such that $A\theta_t = D_{N_t}^{-1/2} P_{N_t} D_{N_t}^{1/2} \mu$ if D_{N_t} is invertible. θ_t is the linear part of the orthogonal parametrization of μ at time t (see paragraph "Estimation" in Section 4.1 in the main paper).
- $\eta_t := \mu - A\theta_t$, equal to $D_{N_t}^{-1/2} R_{N_t} D_{N_t}^{1/2} \mu$ if D_{N_t} is invertible, is the misspecification part of the orthogonal parametrization of model μ at time t.
- $S_t := D_{N_t}(\widehat{\mu}_t - \mu) \in \mathbb{R}^K$.

B The orthogonal parameterization and its properties

Throughout the appendix, we shall adopt an orthogonal parametrization for mean vectors in the model $\mathcal{M}$. In particular, we leverage the following observation: any mean vector $\mu = A\theta + \eta$ can be equivalently represented, at any time t, as $\mu = A\theta_t + \eta_t$, where

$$\theta_t := (A_{N_t}^\top A_{N_t})^\dagger A_{N_t}^\top D_{N_t}^{1/2} \mu = V_t^{-1} \sum_{s=1}^{t} \mu^{k_s} \phi_{k_s}$$

is the orthogonal projection (according to the design matrix V_t) of μ onto the feature space and $\eta_t = \mu - A\theta_t$ is the residual. We now introduce some important properties of this parameterization.

Projecting the empirical mean When we use the orthogonal projection described above on the empirical mean $\widehat{\mu}_t$, the resulting linear part is exactly the standard least squares estimator. That is,

$$\widehat{\theta}_t := (A_{N_t}^\top A_{N_t})^\dagger A_{N_t}^\top D_{N_t}^{1/2} \hat{\mu}_t$$

Projection matrices For $\omega \in \mathbb{R}^K_{\geq 0}$, let us define the projection matrix $P_\omega := A_\omega (A_\omega^\top A_\omega)^\dagger A_\omega^\top \in \mathbb{R}^{K \times K}$ and the residual matrix $R_\omega := I_K - P_\omega \in \mathbb{R}^{K \times K}$. It is easy to check that both are orthogonal projection matrices, i.e., they are symmetric and idempotent ($P_\omega^2 = P_\omega$ and $R_\omega^2 = R_\omega$). Moreover, $P_\omega R_\omega = R_\omega P_\omega = 0$. Equipped with these matrices, we have the following useful identities:

$$A_{N_t} \theta_t = P_{N_t} D_{N_t}^{1/2} \mu = P_{N_t} D_{N_t}^{1/2} A\theta_t,$$

$$D_{N_t}^{1/2} \eta_t = R_{N_t} D_{N_t}^{1/2} \mu = R_{N_t} D_{N_t}^{1/2} \eta_t.$$

Distances between mean vectors in the model Often we will need to compute quantities of the form $\|\lambda - \mu\|_{D_{N_t}}^2$ for different mean vectors in the model. The following lemma shows how to leverage their orthogonal decomposition to split the norm into a distance between their linear parts and a distance between their deviation from linearity.

Lemma 4 (Linear/non-linear decomposition)**.** *For any $\lambda \in \mathcal{M}$ and $t \geq 1$, there exist $\theta'_t \in \mathbb{R}^d$ and $\eta'_t \in \mathbb{R}^K$ such that $\lambda = A\theta'_t + \eta'_t$ and*

$$\|\lambda - \mu\|_{D_{N_t}}^2 = \|\theta'_t - \theta_t\|_{V_t}^2 + \left\| R_{N_t} D_{N_t}^{1/2} \eta'_t - R_{N_t} D_{N_t}^{1/2} \eta_t \right\|_2^2,$$

$$\|\lambda - \widehat{\mu}_t\|_{D_{N_t}}^2 = \left\| \theta'_t - \widehat{\theta}_t \right\|_{V_t}^2 + \left\| R_{N_t} D_{N_t}^{1/2} \eta'_t - R_{N_t} D_{N_t}^{1/2} \widehat{\mu}_t \right\|_2^2.$$

Proof. By leveraging the properties of the orthogonal decomposition and of the matrices P_{N_t}, R_{N_t} (in particular, $P_{N_t}R_{N_t} = 0$ and $P_{N_t} + R_{N_t} = I_K$),

$$\begin{aligned}
\|\lambda - \mu\|^2_{D_{N_t}} &= \|P_{N_t}D_{N_t}(\lambda - \mu) + R_{N_t}D_{N_t}(\lambda - \mu)\|_2^2 \\
&= \left\|P_{N_t}D_{N_t}^{1/2}\lambda - P_{N_t}D_{N_t}^{1/2}\mu\right\|_2^2 + \left\|R_{N_t}D_{N_t}^{1/2}\lambda - R_{N_t}D_{N_t}^{1/2}\mu\right\|_2^2 \\
&= \|P_{N_t}A_{N_t}\theta'_t - P_{N_t}A_{N_t}\theta_t\|_2^2 + \left\|R_{N_t}D_{N_t}^{1/2}\eta'_t - R_{N_t}D_{N_t}^{1/2}\eta_t\right\|_2^2 \\
&= \|\theta'_t - \theta_t\|^2_{V_t} + \left\|R_{N_t}D_{N_t}^{1/2}\eta'_t - R_{N_t}D_{N_t}^{1/2}\eta_t\right\|_2^2 .
\end{aligned}$$

The second result can be shown analogously by noting that the projection of $\widehat{\mu}_t$ onto the linear space spanned by A is exactly the least-squares estimator $\widehat{\theta}_t$. □

The non-linear part of orthogonal parameterizations When applying the orthogonal parameterization to a mean vector $\mu = A\theta + \eta$ with $\|\eta\|_\infty \leq \varepsilon$, while we get some crucial properties for the linear part θ_t (like concentration, see Appendix E), it may be that the resulting non-linear part η_t is such that $\|\eta_t\|_\infty > \epsilon$. However, the following result shows that η_t cannot be too distant from η and, in particular, that $\|\eta_t\|_\infty$ still decreases with ϵ.

Lemma 5 (Maximum deviation)**.** *Let t any time step such that V_t is invertible. Consider the orthogonal parameterization (θ_t, η_t) for $\mu = A\theta + \eta$ with $\|\eta\|_\infty \leq \varepsilon$. Then,*

$$\|\eta_t\|_\infty \leq (LK + 1)\varepsilon.$$

Proof. By definition of the orthogonal parameterization, it is easy to see that $\eta_t - \eta = A(\theta - \theta_t)$. Moreover,

$$\begin{aligned}
\theta_t &:= (A_{N_t}^\top A_{N_t})^\dagger A_{N_t}^\top D_{N_t}^{1/2}\mu = (A_{N_t}^\top A_{N_t})^\dagger A_{N_t}^\top D_{N_t}^{1/2}(A\theta + \eta) \\
&= \theta + (A_{N_t}^\top A_{N_t})^\dagger A_{N_t}^\top D_{N_t}^{1/2}\eta = \theta + V_t^{-1}A^\top D_{N_t}\eta = \theta + V_t^{-1}\sum_{k\in[K]} N_t^k\phi_k\eta^k.
\end{aligned}$$

Therefore, for any arm $k \in [K]$:

$$\begin{aligned}
\left|\eta_t^k - \eta^k\right| &= \left|\phi_k^\top V_t^{-1}\sum_{j\in[K]} N_t^j\phi_j\eta^j\right| \overset{(a)}{\leq} \|\phi_k\|_2\left\|V_t^{-1}\sum_{j\in[K]} N_t^j\phi_j\eta^j\right\|_2 \\
&= \|\phi_k\|_2\left\|\sum_{j\in[K]} N_t^j\phi_j\eta^j\right\|_{V_t^{-2}} \overset{(b)}{\leq} \|\phi_k\|_2\epsilon\sum_{j\in[K]} N_t^j\left\|\phi_j\right\|_{V_t^{-2}} \overset{(c)}{\leq} \|\phi_k\|_2\epsilon K,
\end{aligned}$$

where (a) is from Cauchy-Schwartz inequality, (b) uses the sub-additivity of the norm, and (c) uses that, for each $j \in [K]$, $V_t = \sum_{q\in[K]} N_t^q\phi_q\phi_q^\top \succeq N_t^j\phi_j\phi_j^\top$ (in the sense of the partial order on positive definite matrices). Using that features are bounded by L in ℓ_2-norm,

$$\|\eta_t - \eta\|_\infty \leq LK\epsilon,$$

from which the result easily follows. □

The linear parts of different parametrizations We consider mainly two parametrizations of μ: the orthogonal parametrization with respect to N_t and another (θ, η) for which $\|\eta\|_\infty \leq \varepsilon$. We will now relate the linear parts of these two parametrizations.

Lemma 6. *Let t any time step such that V_t is invertible. Consider the orthogonal parameterization (θ_t, η_t) for $\mu = A\theta + \eta$ with $\|\eta\|_\infty \leq \varepsilon$. Then*

$$\|\theta_t - \theta\|_{V_t} \leq \sqrt{t}\varepsilon\,.$$

Proof. We use the expression $\theta_t = \theta + V_t^{-1} A^\top D_{N_t} \eta$ derived in the last paragraph, the fact that P_{N_t} is a projection and lastly $\|\eta\|_\infty \leq \varepsilon$:

$$\begin{aligned}
\|\theta_t - \theta\|_{V_t} &= \left\|V_t^{-1} A^\top D_{N_t} \eta\right\|_{V_t} \\
&= \sqrt{\eta^\top D_{N_t} A V_t^{-1} A^\top D_{N_t} \eta} \\
&= \left\|D_{N_t}^{1/2} \eta\right\|_{P_{N_t}} \leq \left\|D_{N_t}^{1/2} \eta\right\| = \|\eta\|_{D_{N_t}} \leq \sqrt{t}\varepsilon \,.
\end{aligned}$$

□

C Tractable lower bound for the general Top-m identification problem

We present here the proofs for the claims made in the main paper in Section 3.

C.1 Proof of Lemma 1 and Theorem 1

Lemma. (Lemma 1 in the main paper) *$\forall \mu, \lambda \in \mathbb{R}^K$ s.t. $|S^\star(\mu)| = m$,*

$$\mathcal{S}_m(\lambda) \cap \mathcal{S}_m(\mu) = \emptyset \quad \Leftrightarrow \quad \exists i \notin S^\star(\mu)\ \exists j \in S^\star(\mu), \lambda^i > \lambda^j \,.$$

Proof. To see this, first suppose that the condition on the right holds. That is, there exist $(i, j) \in (S^\star(\mu))^c \times S^\star(\mu)$, where $|S^\star(\mu)| = m$, such that $\lambda^i > \lambda^j$. Then, we have two cases. If j does not belong to any of the top-m sets of λ, that is, $j \notin S^\star(\lambda)$, the result follows trivially since j belongs to the top-m set of μ $S^\star(\mu)$ and $\mathcal{S}_m(\mu) = \{S^\star(\mu)\}$. If, on the other hand, j belongs to at least one top-m set of λ, that is, $j \in S^\star(\lambda)$, then $i \in S^\star(\lambda)$ as well since $\lambda^i > \lambda^j$. But $i \notin S^\star(\mu)$, which proves that $\mathcal{S}_m(\lambda) \cap \mathcal{S}_m(\mu) = \emptyset$. Suppose now that $\mathcal{S}_m(\lambda) \cap \mathcal{S}_m(\mu) = \emptyset$ holds and, by contradiction, that $\forall i \notin S^\star(\mu)\ \forall j \in S^\star(\mu), \lambda^i \leq \lambda^j$. This trivially implies that $S^\star(\mu)$ is a valid top-m set of λ. That is, $\mathcal{S}_m(\lambda) \cap \mathcal{S}_m(\mu) \neq \emptyset$ and we have our desired contradiction. □

Theorem. (Theorem 1 in the main paper) *For any $\delta \leq 1/2$, for any δ-correct algorithm $\mathfrak{A}$ on $\mathcal{M}$, for any bandit instance $\mu \in \mathcal{M}$ such that $|S^\star(\mu)| = m$, the following lower bound holds on the stopping time τ_δ of $\mathfrak{A}$ on instance μ:*

$$\mathbb{E}_\mu^{\mathfrak{A}}[\tau_\delta] \geq \left(\sup_{\omega \in \Delta_K} \min_{i \notin S^\star(\mu)} \min_{j \in S^\star(\mu)} \inf_{\lambda \in \mathcal{M}: \lambda^i > \lambda^j} \sum_{k \in [K]} \omega^k \mathrm{KL}(\mu^k, \lambda^k) \right)^{-1} \log\left(\frac{1}{2.4\delta}\right) .$$

Proof. We start from Equation 2 (main paper), and using Lemma 1, we can rewrite the $\inf$ operator. That yields the desired expression. □

C.2 Proof of Lemma 2

Let $\Lambda_m(\mu, \mathcal{M}') \subseteq \mathcal{M}'$ denote the set of alternative models to $\mu \in \mathbb{R}^K$ in the model $\mathcal{M}'$ (which might be different from $\mathcal{M}$). Consider the lower bound problem

$$H_\mu(\mathcal{M}') := \sup_{\omega \in \Delta_K} \inf_{\lambda \in \Lambda_m(\mu, \mathcal{M}')} \sum_{k \in [K]} \omega^k \,\mathrm{KL}(\mu^k, \lambda^k) \,.$$

A pair of equilibrium strategies for that problem is composed of $\omega \in \Delta_K$ and $q \in \mathcal{P}(\Lambda_m(\mu, \mathcal{M}'))$ (which is the set of probability distributions on $\Lambda_m(\mu, \mathcal{M}')$). Let $Q_{\mathcal{M}'}$ be the set of equilibrium distributions. For $q \in Q_{\mathcal{M}'}$, let $\Lambda_q \subseteq \Lambda_m(\mu, \mathcal{M}')$ be its support.

Lemma 7. *Let $\mathcal{M}_1, \mathcal{M}_2$ be models such that $\mathcal{M}_1 \subseteq \mathcal{M}_2$. For any $q \in Q_{\mathcal{M}_2}$, if $\Lambda_q \subseteq \mathcal{M}_1$, then $H_\mu(\mathcal{M}_1) = H_\mu(\mathcal{M}_2)$.*

Proof. First, we have $H_\mu(\mathcal{M}_1) \geq H_\mu(\mathcal{M}_2)$ since $\mathcal{M}_1 \subseteq \mathcal{M}_2$. If $\Lambda_q \subseteq \mathcal{M}_1$, then using successively $q \in \mathcal{P}(\Lambda(\mu, \mathcal{M}_1))$ and $q \in Q_{\mathcal{M}_2}$,

$$\begin{aligned} H_\mu(\mathcal{M}_1) &= \sup_{\omega \in \Delta_K} \inf_{\lambda \in \Lambda_m(\mu, \mathcal{M}_1)} \sum_{k \in [K]} \omega^k \operatorname{KL}(\mu^k, \lambda^k) \\ &\leq \sup_{\omega \in \Delta_K} \mathbb{E}_{\lambda \sim q} \sum_{k \in [K]} \omega^k \operatorname{KL}(\mu^k, \lambda^k) = H_\mu(\mathcal{M}_2) \,. \end{aligned}$$

□

For $\lambda \in \mathbb{R}^K$, let $|\lambda|_\varepsilon = \inf\{\|\eta\|_\infty \mid \exists \theta \in \mathbb{R}^d,\ \lambda = A\theta + \eta\}$. Let us now consider $\mathcal{M}$ as defined in Equation 1 in the main paper, with misspecification upper bound $\varepsilon \geq 0$.

Lemma 8. *Let $\mathcal{M}' \subseteq \{\lambda \in \mathbb{R}^K \mid \|\lambda\|_\infty \leq M\}$ be a set of models such that $\mathcal{M} \subseteq \mathcal{M}'$ and $\varepsilon > \varepsilon_\mu(\mathcal{M}') := \inf_{q \in Q_{\mathcal{M}'}} \sup_{\lambda \in \Lambda_q} |\lambda|_\varepsilon$.*[6] *Then $H_\mu(\mathcal{M}) = H_\mu(\mathcal{M}')$.*

Proof. If $\varepsilon > \inf_{q \in Q_{\mathcal{M}'}} \sup_{\lambda \in \Lambda_q} |\lambda|_\varepsilon$, then there exists $q \in Q_{\mathcal{M}'}$ such that for all $\lambda \in \Lambda_q$, $|\lambda|_\varepsilon \leq \varepsilon$. Hence $\Lambda_q \subseteq \mathcal{M}$ and we apply Lemma 7. □

For any model $\mathcal{M}'$, there exist equilibrium strategies for which q is supported on K points [11]. Hence $\varepsilon_\mu(\mathcal{M}')$ is always finite.

Let $\mathcal{M}_u := \mathbb{R}^K$ be the set of unstructured models, and for $a, b \in \mathbb{R}$, $\mathcal{M}_{[a,b]} := \{\lambda \in \mathbb{R}^K \mid \forall k \in [K], \lambda^k \in [a, b]\}$ be the set of models that verify a boundedness assumption.

Lemma 9. *Let $\mu^{(K)} := \min_j \mu^j$ and $\mu^{(1)} := \max_j \mu^j$. For all $\mu \in \mathbb{R}^K$, $H_\mu(\mathcal{M}_u) = H_\mu(\mathcal{M}_{[\mu^{(K)}, \mu^{(1)}]})$.*

Proof. Let us consider any $\lambda \in \Lambda_m(\mu, \mathcal{M}_u)$, such that there exists $k \in [K]$ with $\lambda^k \notin [\mu^{(K)}, \mu^{(1)}]$. Let us define $\tilde{\lambda}$ as the projection of λ onto $[\mu^{(K)}, \mu^{(1)}]^K$. Then $\tilde{\lambda}$ satisfies $\tilde{\lambda} \in \Lambda_m(\mu, \mathcal{M}_{[\mu^{(K)}, \mu^{(1)}]}) \subseteq \Lambda_m(\mu, \mathcal{M}_u)$, and by monotonicity of the Kullback-Leibler divergence in one-parameter exponential families, for all $k \in [K]$, $\operatorname{KL}(\mu^k, \tilde{\lambda}^k) \leq \operatorname{KL}(\mu^k, \lambda^k)$. Thus for all $\omega \in \Delta_K$

$$\sum_{k \in [K]} \omega^k \operatorname{KL}(\mu^k, \tilde{\lambda}^k) \leq \sum_{k \in [K]} \omega^k \operatorname{KL}(\mu^k, \lambda^k) \,.$$

For $q \in Q_{\mathcal{M}_u}$, let $\tilde{q}$ be the distribution in which every support point λ of q is transported onto its projection $\tilde{\lambda}$. Then for all $\omega \in \triangle_K$,

$$\mathbb{E}_{\lambda \sim \tilde{q}} \sum_{k \in [K]} \omega^k \operatorname{KL}(\mu^k, \lambda^k) \leq \mathbb{E}_{\lambda \sim q} \sum_{k \in [K]} \omega^k \operatorname{KL}(\mu^k, \lambda^k) \,,$$

from which we obtain that $\tilde{q}$ has lower objective value than q. Since $q \in Q_{\mathcal{M}_u}$, then $\tilde{q} \in Q_{\mathcal{M}_u}$ as well. By construction, its support verifies $\Lambda_{\tilde{q}} \subseteq \mathcal{M}_{[\mu^{(K)}, \mu^{(1)}]}$. We conclude with Lemma 7. □

Applying Lemma 8 to $\mathcal{M}_{[\mu^{(K)}, \mu^{(1)}]}$, together with Lemma 9, we finally obtain Lemma 2 from the main paper, restated here using the notations we introduced:

Lemma. *If $\varepsilon > \mu^{(1)} - \mu^{(K)}$ then $H_\mu(\mathcal{M}) = H_\mu(\mathcal{M}_{[\mu^{(K)}, \mu^{(1)}]}) = H_\mu(\mathcal{M}_u)$.*

C.3 Computing the closest alternative

In order to compute the closest alternative to $\mu \in \mathcal{M}$ in the half-space $\{\lambda \in \mathcal{M} \mid \lambda^k \geq \lambda^j\}$, the optimization problem we need to solve is

$$\begin{aligned} \inf_{\theta, \eta} \ & \frac{1}{2} \|A\theta + \eta - \mu\|^2_{D_{N_t}} \\ \text{s.t } & (e_k - e_j)^\top (A\theta + \eta) \geq 0 \\ & \|A\theta + \eta\|_\infty \leq M \\ & \|\eta\|_\infty \leq \varepsilon \,. \end{aligned}$$

[6] Note that indeed quantity $\varepsilon_\mu(\mathcal{M}')$ depends on μ, since $Q_{\mathcal{M}'}$ is defined with respect to μ.

In our implementation, and thus in the remainder of this section, we shall drop the boundedness constraint $\|A\theta + \eta\|_\infty \leq M$ which has typically a negligible effect on the algorithm's behavior.

Quadratic problem We express the problem as function of the variable $(\theta^\top, \eta^\top)^\top$. Up to the constant term, this problem is equivalent to

$$\begin{aligned}
&\inf_{\theta,\eta} \begin{pmatrix} \theta \\ \eta \end{pmatrix}^\top \begin{pmatrix} A^\top D_N A & A^\top D_N \\ D_N A & D_N \end{pmatrix} \begin{pmatrix} \theta \\ \eta \end{pmatrix} - \begin{pmatrix} \theta \\ \eta \end{pmatrix}^\top \begin{pmatrix} A^\top D_N \mu \\ D_N \mu \end{pmatrix} \\
&\text{s.t.} \begin{pmatrix} A^\top (e_j - e_k) \\ e_j - e_k \end{pmatrix}^\top \begin{pmatrix} \theta \\ \eta \end{pmatrix} \leq 0 \\
&\quad \|\eta\|_\infty \leq \varepsilon \,.
\end{aligned}$$

In the code, we directly solve the problem under this form using a quadratic problem solver.

Computing the closest alternative We now detail the form of the solutions analytically (as much as possible). Let $j, k \in [K]$, $j \neq k$. We want to compute the closest alternative in the half-space $\{\lambda \in \mathcal{M} \mid \lambda^k \geq \lambda^j\}$ to $\mu \in \mathbb{R}^K$. That is, we compute the solution to

$$\begin{aligned}
&\inf_{\theta,\eta} \frac{1}{2} \|A\theta + \eta - \mu\|^2_{D_{N_t}} \\
&\text{s.t } (e_k - e_j)^\top (A\theta + \eta) \geq 0 \\
&\quad \eta \in \mathcal{C}
\end{aligned}$$

Here, to highlight the generality of the following derivation, we replace the ℓ_∞ norm constraint on η with any convex set $\mathcal{C}$. To simplify the notation, we denote by D_N the diagonal matrix with N_t on the diagonal and $u := e_j - e_k$. The problem above is then written as

$$\begin{aligned}
&\inf_{\theta,\eta} \frac{1}{2} \left\| D_N^{1/2} A\theta + D_N^{1/2} \eta - D_N^{1/2} \mu \right\|_2^2 \\
&\text{s.t } u^\top (A\theta + \eta) \leq 0 \\
&\quad \eta \in \mathcal{C}
\end{aligned}$$

Assumption 1. At t_0, $A^\top D_{N_{t_0}} A = V_{t_0}$ is invertible.

See paragraph "Initialization phase" in Subsection 4.1 to see how that assumption is ensured in practice. We now suppose that $t \geq t_0$. Minimizing first in θ at fixed η, we solve the problem

$$\begin{aligned}
&\inf_{\theta} \frac{1}{2} \left\| D_N^{1/2} A\theta + D_N^{1/2} \eta - D_N^{1/2} \mu \right\|_2^2 \\
&\text{s.t } u^\top (A\theta + \eta) \leq 0
\end{aligned}$$

The Lagrangian is $L(\theta, \alpha) = \frac{1}{2} \left\| D_N^{1/2} A\theta + D_N^{1/2} \eta - D_N^{1/2} \mu \right\|_2^2 + \alpha u^\top (\eta + A\theta)$ with $\alpha \geq 0$. We get that at the optimal θ,

$$A^\top D_N (A\theta + \eta - \mu) = -\alpha A^\top u \quad \Longrightarrow \quad \theta = (A^\top D_N A)^{-1} A^\top (-\alpha u + D_N \mu - D_N \eta) \,.$$

At the optimum, from the KKT conditions, either $\alpha = 0$ and $u^\top (A\theta + \eta) \leq 0$, or $\alpha > 0$ and $u^\top A\theta = -u^\top \eta$.

Case $\alpha = 0$. If $\alpha = 0$, then $\theta = (A^\top D_N A)^{-1} A^\top D_N (\mu - \eta)$, $D_N^{1/2}(A\theta + \eta - \mu) = (D_N^{1/2} A (A^\top D_N A)^{-1} A^\top D_N^{1/2} - I) D_N^{1/2} (\mu - \eta)$ and the value of the optimization problem is the norm of this quantity.

Let $P_N = D_N^{1/2} A (A^\top D_N A)^{-1} A^\top D_N^{1/2}$. Note: it is symmetric and idempotent ($P_N^2 = P_N$), meaning that it is an orthogonal projection. Let $R_N = I - P_N$ be the residual matrix. We also have $R_N^2 = R_N$. Furthermore, $P_N R_N = R_N P_N = 0$.

With these notations, $D_N^{1/2} A\theta = P_N D_N^{1/2} (\mu - \eta)$, $D_N^{1/2}(A\theta + \eta - \mu) = -R_N D_N^{1/2} (\mu - \eta)$ and the value of the optimization problem is $\frac{1}{2} \|R_N D_N^{1/2} \eta - R_N D_N^{1/2} \mu\|^2$. The case $\alpha = 0$ is

possible only if the constraint is then satisfied, that is if $u^\top(A\theta+\eta)\leq 0$ at the optimum, i.e. if $u^\top(A^\top(A^\top D_N A)^{-1}A^\top D_N\mu+(I-A^\top(A^\top D_N A)^{-1}A^\top D_N)\eta)\leq 0$. The problem we need to solve in that case is

$$\begin{aligned}\min_{\eta_N}\ & \frac{1}{2}\left\|R_N D_N^{1/2}\eta - R_N D_N^{1/2}\mu\right\|_2^2\\ \text{s.t. } & u^\top(I-A^\top(A^\top D_N A)^{-1}A^\top D_N)\eta \leq -u^\top A^\top(A^\top D_N A)^{-1}A^\top D_N\mu\\ & \eta\in\mathcal{C}\end{aligned}$$

If $\mathcal{C}$ is convex this is a convex optimization problem. It can happen that there is no feasible point, which simply means that there is no solution with $\alpha=0$.

Case $\alpha\neq 0$. Consider now the case $\alpha>0$. We get

$$\begin{aligned} & u^\top A\theta = -u^\top\eta\\ \implies\quad & u^\top A(A^\top D_N A)^{-1}A^\top(-\alpha u+D_N\mu-D_N\eta)=-u^\top\eta\\ \implies\quad & \alpha u^\top A(A^\top D_N A)^{-1}A^\top u = u^\top A(A^\top D_N A)^{-1}A^\top D_N(\mu-\eta)+u^\top\eta\end{aligned}$$

Then

$$\begin{aligned} D_N^{1/2}A\theta &= D_N^{1/2}A(A^\top D_N A)^{-1}A^\top(-\alpha u+D_N\mu-D_N\eta)\\ &= -\alpha D_N^{1/2}A(A^\top D_N A)^{-1}A^\top u + P_N D_N^{1/2}(\mu-\eta)\\ D_N^{1/2}(A\theta+\eta-\mu) &= -\alpha D_N^{1/2}A(A^\top D_N A)^{-1}A^\top u - R_N D_N^{1/2}(\mu-\eta)\\ &= -\frac{u^\top A(A^\top D_N A)^{-1}A^\top D_N(\mu-\eta)+u^\top\eta}{u^\top A(A^\top D_N A)^{-1}A^\top u} D_N^{1/2}A(A^\top D_N A)^{-1}A^\top u\\ &\quad - R_N D_N^{1/2}(\mu-\eta)\end{aligned}$$

We can now see that $D_N^{1/2}(A\theta+\eta-\mu)$ is linear in η and the objective value $\frac{1}{2}\left\|D_N^{1/2}(A\theta+\eta-\mu)\right\|^2$ is quadratic in η. We need to solve a quadratic optimization problem under the constraint $\eta\in\mathcal{C}$. Let's now simplify that optimization problem. We first show that the cross term in $\frac{1}{2}\left\|D_N^{1/2}(A\theta+\eta-\mu)\right\|_2^2 = \frac{1}{2}\left\|-\alpha D_N^{1/2}A(A^\top D_N A)^{-1}A^\top u - R_N D_N^{1/2}(\mu-\eta)\right\|_2^2$ is zero. Note: if D_N is invertible, then $\frac{1}{2}\left\|-\alpha D_N^{1/2}A(A^\top D_N A)^{-1}A^\top u - R_N D_N^{1/2}(\mu-\eta)\right\|_2^2 = \frac{1}{2}\left\|-\alpha P_N D_N^{-1/2}u - R_N D_N^{1/2}(\mu-\eta)\right\|_2^2$ and the fact that the cross term is 0 is a simple consequence of $P_N R_N = R_N P_N = 0$.

$$\begin{aligned} & (R_N D_N^{1/2}(\mu-\eta))^\top D_N^{1/2}A(A^\top D_N A)^{-1}A^\top u\\ &= ((I-P_N)D_N^{1/2}(\mu-\eta))^\top D_N^{1/2}A(A^\top D_N A)^{-1}A^\top u\\ &= (\mu-\eta)^\top D_N A(A^\top D_N A)^{-1}A^\top u - (\mu-\eta)^\top D_N^{1/2}P_N D_N^{1/2}A(A^\top D_N A)^{-1}A^\top u\\ &= (\mu-\eta)^\top D_N A(A^\top D_N A)^{-1}A^\top u - (\mu-\eta)^\top D_N A(A^\top D_N A)^{-1}A^\top D_N A(A^\top D_N A)^{-1}A^\top u\\ &= 0\,.\end{aligned}$$

Now that we established that the cross term is zero, the objective value is simply the sum of two square terms,

$$\begin{aligned}\frac{1}{2}\left\|D_N^{1/2}(A\theta+\eta-\mu)\right\|_2^2 &= \frac{1}{2}\alpha^2 u^\top A(A^\top D_N A)^{-1}A^\top u + \frac{1}{2}(\mu-\eta)^\top D_N^{1/2}R_N D_N^{1/2}(\mu-\eta)\\ &= \frac{1}{2}\frac{\left(u^\top A(A^\top D_N A)^{-1}A^\top D_N(\mu-\eta)+u^\top\eta\right)^2}{u^\top A(A^\top D_N A)^{-1}A^\top u}\\ &\quad + \frac{1}{2}(\mu-\eta)^\top D_N^{1/2}R_N D_N^{1/2}(\mu-\eta)\\ &= \frac{1}{2}\eta^\top Q\eta + q^\top\eta + C\end{aligned}$$

where C doesn't depend on η and

$$Q = D_N^{1/2} R_N D_N^{1/2} + \frac{1}{u^\top A(A^\top D_N A)^{-1} A^\top u} \left((I - D_N A(A^\top D_N A)^{-1} A^\top)u\right) \left((I - D_N A(A^\top D_N A)^{-1} A^\top)u\right)^\top$$
$$q = \frac{u^\top A(A^\top D_N A)^{-1} A^\top D_N \mu}{u^\top A(A^\top D_N A)^{-1} A^\top u}(I - D_N A(A^\top D_N A)^{-1} A^\top)u - D_N^{1/2} R_N D_N^{1/2} \mu \,.$$

Again if D_N is invertible these have simpler expressions:

$$Q = D_N^{1/2} \left(R_N + \frac{1}{u^\top D_N^{-1/2} P_N D_N^{-1/2} u} (R_N D_N^{-1/2} u)(R_N D_N^{-1/2} u)^\top \right) D_N^{1/2}$$
$$q = D_N^{1/2} R_N \left(\frac{u^\top D_N^{-1/2} P_N D_N^{1/2} \mu}{u^\top D_N^{-1/2} P_N D_N^{-1/2} u} D_N^{-1/2} u - D_N^{1/2} \mu \right) \,.$$

We are looking for a solution to

$$\arg\min_{\eta \in \mathcal{C}} \ \frac{1}{2} \eta^\top Q \eta + q^\top \eta \,.$$

This is a quadratic objective. The difficulty of finding the minimum depends on $\mathcal{C}$.

Summary. To compute the closest alternative in a half-space, we compute the solution to two quadratic problems corresponding to the possibilities that Lagrangian multiplier α satisfies either $\alpha = 0$ or $\alpha > 0$. Then we retain the solution with the minimal objective value.

D The MISLID algorithm

D.1 Initialization

MISLID starts by pulling a deterministic sequence of t_0 arms that make the minimum eigenvalue of the resulting design matrix V_{t_0} larger than $2L^2$. Since the rows of A span $\mathbb{R}^d$, such sequence can be found by taking any subset of d arms that span the whole space (e.g., by computing a barycentric spanner [4]) and pulling them in a round robin fashion until the desired condition is met.

In order to get an approximation of the length t_0 of the initialization phase, let us denote $\sigma_{\min}(M)$ the minimal singular value of a matrix M. Let us consider $\mathcal{B} = \{b_1, b_2, \ldots, b_d\} \subseteq [K]$, $|\mathcal{B}| = d$, the barycentric spanner of size d computed on matrix A. Then, if we stopped the round-robin sampling such that each arm in the barycentric spanner is sampled exactly u_0 times, $V_{t_0} = u_0 \sum_{k \in \mathcal{B}} \phi_k \phi_k^\top$. To ensure that $V_t \succeq 2L^2 I_d$, we need $u_0 \sigma_{\min}\left(\sum_{k \in \mathcal{B}} \phi_k \phi_k^\top\right) \geq 2L^2$. Let $\Gamma'(A) := \min\left\{\sigma_{\min}\left(\sum_{k \in \mathcal{B}} \phi_k \phi_k^\top\right) \mid \mathcal{B}\ d\text{-sized spanner of } A\right\}$. Then $u_0 = \left\lceil \frac{2L^2}{\Gamma'(A)} \right\rceil$ is large enough.

We obtain the bound $t_0 \leq d \left\lceil \frac{2L^2}{\Gamma'(A)} \right\rceil$.

D.2 Projection of the empirical mean onto the set of realizable models $\mathcal{M}$

As done in Equation 1 in the main paper, we define the set of realizable models as

$$\mathcal{M} := \left\{\mu = A\theta + \eta \in \mathbb{R}^K \mid \exists \theta \in \mathbb{R}^d \exists \eta \in \mathbb{R}^K, \|\eta\|_\infty \leq \varepsilon \wedge \|A\theta + \eta\|_\infty \leq M\right\} \,.$$

We require our estimates of μ to be in this set, but the estimate at time t $\widehat{\mu}_t$ might not satisfy the constraint on its ℓ_∞ norm (*i.e.*, $\|\widehat{\mu}_t\|_\infty > M$). We then directly project the empirical mean vector onto $\mathcal{M}$. Define

$$(\tilde{\theta}_t, \tilde{\eta}_t) := \arg\min_{\theta', \eta' : A\theta' + \eta' \in \mathcal{M}} \ \|A\theta' + \eta' - \widehat{\mu}_t\|^2_{D_{N_t}} \,. \tag{6}$$

Lemma 10. *Let $\tilde{\mu}_t = A\tilde{\theta}_t + \tilde{\eta}_t$,[7] where $(\tilde{\theta}_t, \tilde{\eta}_t)$ are the solution of (6). Then, all the following hold:*

$$\begin{aligned}
\|\tilde{\mu}_t - \mu\|^2_{D_{N_t}} &\leq \|\mu - \widehat{\mu}_t\|^2_{D_{N_t}} \ , \\
\left\|\tilde{\theta}_t - \theta_t\right\|^2_{V_t} &\leq \left\|\widehat{\theta}_t - \theta_t\right\|^2_{V_t} \ , \\
\left\|R_{N_t} D_{N_t}^{1/2}\tilde{\eta}_t - R_{N_t} D_{N_t}^{1/2}\eta_t\right\|_2^2 &\leq \left\|R_{N_t} D_{N_t}^{1/2}\widehat{\mu}_t - R_{N_t} D_{N_t}^{1/2}\eta_t\right\|_2^2 \ , \\
\left\|\tilde{\theta}_t - \theta\right\|^2_{V_t} &\leq \left\|\widehat{\theta}_t - \theta\right\|^2_{V_t} \ , \\
\left\|R_{N_t} D_{N_t}^{1/2}\tilde{\eta}_t - R_{N_t} D_{N_t}^{1/2}\eta\right\|_2^2 &\leq \left\|R_{N_t} D_{N_t}^{1/2}\widehat{\mu}_t - R_{N_t} D_{N_t}^{1/2}\eta\right\|_2^2
\end{aligned}$$

Proof. The first inequality is easy to check by using $\mu \in \mathcal{M}$ together with the non-expansion of the projection in the optimized norm.

The proof of the other inequalities extends Lemma 9 in [40]. Note that, using Lemma 4, an equivalent formulation of (6) is

$$\begin{aligned}
(\tilde{\theta}_t, \tilde{\eta}_t) &:= \underset{\theta', \eta' : A\theta' + \eta' \in \mathcal{M}}{\arg\min} \left\{ \left\|P_{N_t} A_{N_t}\theta' - P_{N_t} D_{N_t}^{1/2}\widehat{\mu}_t\right\|_2^2 + \left\|R_{N_t} D_{N_t}^{1/2}\eta' - R_{N_t} D_{N_t}^{1/2}\widehat{\mu}_t\right\|_2^2 \right\} \\
&= \underset{\theta', \eta' : A\theta' + \eta' \in \mathcal{M}}{\arg\min} \left\{ \left\|\theta' - \widehat{\theta}_t\right\|^2_{V_t} + \|\eta' - \widehat{\mu}_t\|^2_{D_{N_t}^{1/2} R_{N_t} D_{N_t}^{1/2}} \right\}
\end{aligned}$$

This is the minimization of a convex function over a convex set. For any $\theta' \in \mathbb{R}^d$, $\eta' \in \mathbb{R}^K$, let $f(\theta') = \left\|\theta' - \widehat{\theta}_t\right\|^2_{V_t}$ and $g(\eta') = \|\eta' - \widehat{\mu}_t\|^2_{D_{N_t}^{1/2} R_{N_t} D_{N_t}^{1/2}}$. Therefore, using the first-order optimality conditions for convex functions (see, e.g., Theorem 2.8 in [32]), $(\tilde{\theta}_t, \tilde{\eta}_t)$ are minimizers if and only if for each $\theta', \eta' : A\theta' + \eta' \in \mathcal{M}$,

$$\begin{aligned}
\langle \nabla_\theta f(\tilde{\theta}_t), \theta' - \tilde{\theta}_t \rangle \geq 0 \quad &\Longrightarrow \quad (\tilde{\theta}_t - \widehat{\theta}_t)^T V_t (\theta' - \tilde{\theta}_t) \geq 0 \\
\langle \nabla_\eta g(\tilde{\eta}_t), \eta' - \tilde{\eta}_t \rangle \geq 0 \quad &\Longrightarrow \quad (\tilde{\eta}_t - \widehat{\mu}_t)^T D_{N_t}^{1/2} R_{N_t} D_{N_t}^{1/2} (\eta' - \tilde{\eta}_t) \geq 0
\end{aligned}$$

Note that $\mu = A\theta_t + \eta_t$, thus the orthogonal parametrization (θ_t, η_t) is such that $A\theta_t + \eta_t \in \mathcal{M}$. Thus, (θ_t, η_t) are feasible solutions. This implies

$$\left\|\widehat{\theta}_t - \theta_t\right\|^2_{V_t} = \left\|\widehat{\theta}_t - \tilde{\theta}_t\right\|^2_{V_t} + \left\|\tilde{\theta}_t - \theta_t\right\|^2_{V_t} + 2(\widehat{\theta}_t - \tilde{\theta}_t)^T V_t (\tilde{\theta}_t - \theta_t) \geq \left\|\tilde{\theta}_t - \theta_t\right\|^2_{V_t}$$

and

$$\begin{aligned}
\|\widehat{\mu}_t - \eta_t\|^2_{D_{N_t}^{1/2} R_{N_t} D_{N_t}^{1/2}} &= \|\widehat{\mu}_t - \tilde{\eta}_t\|^2_{D_{N_t}^{1/2} R_{N_t} D_{N_t}^{1/2}} + \|\tilde{\eta}_t - \eta_t\|^2_{D_{N_t}^{1/2} R_{N_t} D_{N_t}^{1/2}} \\
&\quad + 2(\widehat{\mu}_t - \tilde{\eta}_t)^T D_{N_t}^{1/2} R_{N_t} D_{N_t}^{1/2} (\tilde{\eta}_t - \eta_t) \\
&\geq \|\tilde{\eta}_t - \eta_t\|^2_{D_{N_t}^{1/2} R_{N_t} D_{N_t}^{1/2}} \, .
\end{aligned}$$

Rearranging concludes the proof of the second and the third inequalities. To show the last two inequalities, simply use the same argument by noting that (θ, η) is also a feasible solution (since $\mu = A\theta + \eta \in \mathcal{M}$). □

E Concentration results

E.1 Concentration of the linear part

In this section we derive concentration results for

[7] Note that the equation $\widehat{\theta}_t = \tilde{\theta}_t$ mentioned in Section 4.1 in the main paper no longer holds, because we consider the boundedness assumption $\mu \in \mathcal{M} \implies \|\mu\|_\infty \leq M$

$$\left\|\widehat{\theta}_t - \theta_t\right\|_{V_t}^2 = \left\|V_t^{-1} A^\top S_t\right\|_{V_t}^2 = \left\|A^\top S_t\right\|_{V_t^{-1}}^2 .$$

We rewrite the quantities involved to make obvious that this is the usual self-normalized quantity from the linear bandit literature [1]:

$$A^\top S_t = \sum_{s=1}^{t} \left(X_s^{k_s} - \mu^{k_s}\right) A^\top e_{k_s} = \sum_{s=1}^{t} \left(X_s^{k_s} - \mu^{k_s}\right) \phi_{k_s} \quad \text{and} \quad V_t = \sum_{s=1}^{t} \phi_{k_s} \phi_{k_s}^\top .$$

We restate here Theorem 20.4 (in combination with the Equation 20.9) of [28], which states a result due to [1].

Theorem 3. *Suppose that for all $k \in [K]$, $\|\phi_k\|_2 \leq L$. For all $x > 0$ and $\delta \in (0, 1]$,*

$$\mathbb{P}\left(\exists t \in \mathbb{N},\ \frac{1}{2}\left\|A^\top S_t\right\|_{(V_t + x I_d)^{-1}}^2 \geq \log\frac{1}{\delta} + \frac{d}{2}\log\left(1 + \frac{tL^2}{xd}\right)\right) \leq \delta .$$

Corollary 1. *If we ensure that $V_{t_0} \succeq x I_d$ (in the sense of positive definite matrices), then*

$$\mathbb{P}\left(\exists t > t_0,\ \frac{1}{2}\left\|\widehat{\theta}_t - \theta_t\right\|_{V_t}^2 \geq 2\log\frac{1}{\delta} + d\log\left(1 + \frac{tL^2}{xd}\right)\right) \leq \delta .$$

Proof. If $V_t \succeq x I_d$ then $2V_t \succeq V_t + x I_d$ and

$$\left\|\widehat{\theta}_t - \theta_t\right\|_{V_t}^2 = \left\|A^\top S_t\right\|_{V_t^{-1}}^2 \leq 2\left\|A^\top S_t\right\|_{(V_t + x I_d)^{-1}}^2 .$$

□

The $2\log(1/\delta)$ term is fine for some steps of the analysis but not for the stopping rule. For the stopping rule concentration inequality, we need $\log(1/\delta)$.

Corollary 2. *Suppose that $V_{t_0} \succeq x I_d$. Then*

$$\mathbb{P}\left(\exists t \geq t_0,\ \frac{1}{2}\left\|\widehat{\theta}_t - \theta_t\right\|_{V_t}^2 \geq 1 + \log\frac{1}{\delta} + \left(1 + \frac{1}{\log(1/\delta)}\right)\frac{d}{2}\log\left(1 + \frac{tL^2}{xd}\log\frac{1}{\delta}\right)\right) \leq \delta .$$

Proof. Suppose that $V_{t_0} \succeq x I_d$ and let $\gamma(\delta) := \log(1/\delta)^{-1}$. For any $t \geq t_0$,

$$\left\|\widehat{\theta}_t - \theta_t\right\|_{V_t}^2 = \left\|A^\top S_t\right\|_{V_t^{-1}}^2 \leq (1 + \gamma(\delta))\left\|A^\top S_t\right\|_{(V_t + x\gamma(\delta) I_d)^{-1}}^2 .$$

Then we conclude by applying Theorem 3. □

E.2 Unstructured concentration

Let W_{-1} be the negative branch of the Lambert W function and let $\overline{W}(x) = -W_{-1}(-e^{-x})$. Note that for $x \geq 1$, $x + \log x \leq \overline{W}(x) \leq x + \log x + \min\{\frac{1}{2}, \frac{1}{\sqrt{x}}\}$.

Lemma 11. *For $t > 1$, with probability $1 - \delta$,*

$$\frac{1}{2}\left\|\widehat{\mu}_t - \mu\right\|_{D_{N_t}}^2 \leq 2K\overline{W}\left(\frac{1}{2K}\log\frac{e}{\delta} + \frac{1}{2}\log(8eK\log t)\right) .$$

Proof. See [14, Appendix A, Theorem 4] for that form of the lemma, which is a small reformulation of a result due to [30]. □

The concentration inequality of Lemma 11 is also valid for $\left\|\tilde{\mu}_t - \mu\right\|_{D_{N_t}}^2$ since the first inequality of Lemma 10 states that $\left\|\tilde{\mu}_t - \mu\right\|_{D_{N_t}}^2 \leq \left\|\widehat{\mu}_t - \mu\right\|_{D_{N_t}}^2$.

E.3 Elliptic potential lemmas

All lemmas in this section are derived under the following assumption.

Assumption 2. For $t \geq t_0$, $V_t \succeq 2L^2 I_d$.

In the remainder of the section, we consider $\omega_t \in \Delta^K$, for any time $t > 0$.

Lemma 12. *Under Assumption 2, with probability* $1-\delta$,

$$\sum_{s=t_0+1}^{t} \sum_{k=1}^{K} \omega_s^k \left\|\phi_k\right\|^2_{V_{s-1}^{-1}} \leq \sqrt{2t \log \frac{1}{\delta}} + d \log\left(1 + \frac{t}{d}\right) .$$

Proof.

$$\sum_{s=t_0+1}^{t} \sum_{k=1}^{K} \omega_s^k \left\|\phi_k\right\|^2_{V_{s-1}^{-1}} = \sum_{s=t_0+1}^{t} \left(\sum_{k=1}^{K} \omega_s^k \left\|\phi_k\right\|^2_{V_{s-1}^{-1}} - \left\|\phi_{k_s}\right\|^2_{V_{s-1}^{-1}} \right) + \sum_{s=t_0+1}^{t} \left\|\phi_{k_s}\right\|^2_{V_{s-1}^{-1}} .$$

The first term is the sum of a martingale difference sequence with bounded increments

$$\mathbb{E}\left[\sum_{k=1}^{K} \omega_s^k \left\|\phi_k\right\|^2_{V_{s-1}^{-1}} - \left\|\phi_{k_s}\right\|^2_{V_{s-1}^{-1}} \mid \mathcal{F}_{s-1} \right] = 0 ,$$

$$\left| \sum_{k=1}^{K} \omega_s^k \left\|\phi_k\right\|^2_{V_{s-1}^{-1}} - \left\|\phi_{k_s}\right\|^2_{V_{s-1}^{-1}} \right| \leq 1 .$$

since $V_{s-1} \succeq 2L^2 I_d$ and $\|\phi_k\| \leq L$. By the Azuma-Hoeffding inequality, with probability $1-\delta$,

$$\sum_{s=t_0+1}^{t} \left(\sum_{k=1}^{K} \omega_s^k \left\|\phi_k\right\|^2_{V_{s-1}^{-1}} - \left\|\phi_{k_s}\right\|^2_{V_{s-1}^{-1}} \right) \leq \sqrt{2t \log \frac{1}{\delta}} .$$

The second term is an elliptic potential, bounded in Lemma 13 below. □

Lemma 13. *Under Assumption 2, for* $t > t_0$,

$$\sum_{s=t_0+1}^{t} \left\|\phi_{k_s}\right\|^2_{V_{s-1}^{-1}} \leq d \log\left(1 + \frac{t}{d}\right) .$$

Proof. Let $V_{s:t}$ denote the design matrix using only rounds from s to t. We use Lemma 14,

$$\sum_{s=t_0+1}^{t} \left\|\phi_{k_s}\right\|^2_{V_{s-1}^{-1}} \leq \sum_{s=t_0+1}^{t} \left\|\phi_{k_s}\right\|^2_{(2L^2 I_d + V_{t_0+1:s-1})^{-1}} \leq d \log\left(1 + \frac{t}{d}\right) .$$

□

Lemma 14. *Under Assumption 2, for* $t > t_0$,

$$\sum_{s=t_0+1}^{t} \left\|\phi_{k_s}\right\|^2_{(V_{t_0+1:s-1} + 2L^2 I_d)^{-1}} \leq d \log\left(1 + \frac{t}{d}\right) .$$

Proof. By definition of L, for all $k \in [K]$, $\phi_k \phi_k^\top \preceq L^2 I_d$. From Lemma 15 below, we have

$$\begin{aligned}
\sum_{s=t_0+1}^{t} \left\|\phi_{k_s}\right\|^2_{(V_{t_0+1:s-1} + 2L^2 I_d)^{-1}} &= \sum_{s=t_0+1}^{t} \left\|\phi_{k_s}\right\|^2_{(V_{t_0+1:s-1} + L^2 I_d + L^2 I_d)^{-1}} \\
&\leq \sum_{s=t_0+1}^{t} \left\|\phi_{k_s}\right\|^2_{(V_{t_0+1:s} + L^2 I_d)^{-1}} \\
&\leq d \log\left(1 + \frac{2t}{d}\right) .
\end{aligned}$$

□

A general statement (extracted from [13] but widely known, see for example [28]) is

Lemma 15. *Let $(\omega_t)_{t\geq 1}$ be a sequence in the simplex Δ_K and $x > 0$. Let $W_t := \sum_{s=1}^t \omega_s$ and $V_{W_t} := \sum_{s=1}^t \sum_{k=1}^K \omega_s^k \phi_k \phi_k^\top$. Then*

$$\sum_{s=1}^t \sum_{k=1}^K \omega_s^k \left\|\phi_k\right\|^2_{(V_{W_s}+xI_d)^{-1}} \leq d \log\left(1 + \frac{tL^2}{d\eta}\right) .$$

Proof. Define the function $f(W) = \log\det(V_W + xI_d)$ for any $W \in (\mathbb{R}^+)^K$. It is a concave function since the function $V \mapsto \log\det(V)$ is a concave function over the set of positive definite matrices (see Exercise 21.2 of [28]). Its partial derivative with respect to the coordinate k at W is

$$\nabla_k f(W) = \left\|\phi_k\right\|^2_{(V_W+xI_d)^{-1}} .$$

Hence using the concavity of f we have

$$\sum_{k=1}^K \omega_s^k \left\|\phi_k\right\|^2_{(V_{W_s}+xI_d)^{-1}} = (W_s - W_{s-1})^\top \nabla_a f(W_s) \leq f(W_s) - f(W_{s-1}) ,$$

which implies that

$$\sum_{s=1}^t \sum_{k=1}^K \omega_s^k \left\|\phi_k\right\|^2_{V_{W_s}+xI_d} \leq f(W_t) - f(W_0) = \log\left(\frac{\det(V_{W_t} + xI_d)}{\det(xI_d)}\right) \leq d\log\left(1 + \frac{tL^2}{dx}\right) ,$$

where for the last inequality we use the inequality of arithmetic and geometric means in combination with $\mathrm{Tr}(V_{W_t}) \leq tL^2$. □

Lemma 16. *Let $C > 0$ be a constant. With probability $1 - \delta$,*

$$\sum_{s=t_0+1}^t \sum_{k=1}^K \omega_{s-1}^k \min\left\{C, \frac{1}{N_{s-1}^k}\right\} \leq C\sqrt{2t\log\frac{1}{\delta}} + K(C + 1 + \log t)$$

Proof. The first term is due to a martingale argument to bound $\sum_s \left(\sum_k \omega_{s-1}^k \min\left\{C, \frac{1}{N_{s-1}^k}\right\} - \min\left\{C, \frac{1}{N_{s-1}^{k_s}}\right\}\right)$. Then

$$\sum_{s=t_0+1}^t \min\left\{C, \frac{1}{N_{s-1}^{k_s}}\right\} \leq CK + \sum_{k=1}^K \mathbb{I}\left\{N_{t-1}^k > 0\right\} \sum_{j=1}^{N_{t-1}^k} \frac{1}{j} \leq K(C + 1 + \log t) .$$

□

E.4 Martingale concentration

Lemma 17. *Let $\mu \in \mathcal{M}$ (with upper bounds M and ε) and $Z_s(\lambda) := (\mu^{k_s} - \lambda^{k_s})^2 - \mathbb{E}_{k\sim\omega_s}[(\mu^k - \lambda^k)^2]$. For any $\delta' \in (0, 1)$,*

$$\mathbb{P}\left\{\exists t \geq 1 : \sup_{\lambda\in\mathcal{M}} \left|\sum_{s=1}^t Z_s(\lambda)\right| > r(t, \delta')\right\} \leq \delta',$$

where

$$r(t, \delta') := 2M^2\sqrt{\frac{t}{2}\left(\log\frac{4t^2}{\delta'} + d\log\frac{6(M+\varepsilon)Lt}{\sqrt{\Gamma(A)}} + K\log\max\{4\varepsilon t, 1\}\right)} + 2 + 8M ,$$

$\Gamma(A) := \max_{\omega\in\Delta_K} \sigma_{\min}\left(\sum_{k=1}^K \omega^k \phi_k \phi_k^\top\right)$, *and $\sigma_{\min}(M)$ is the minimal eigenvalue of matrix M.*

Proof. First note that ω_s is $\mathcal{F}_{s-1}$-measurable. Thus, for any fixed λ,

$$\mathbb{E}[Z_s(\lambda)|\mathcal{F}_{s-1}] = \mathbb{E}[(\mu^{k_s} - \lambda^{k_s})^2|\mathcal{F}_{s-1}] - \mathbb{E}_{k\sim\omega_s}[(\mu^k - \lambda^k)^2] = 0,$$

which implies that $\{Z_s\}_{s\geq 1}$ is a martingale difference sequence. Moreover, it is easy to check that $|Z_s(\lambda)| \leq 4M^2$. Unfortunately, we cannot directly use this martingale property to concentrate the desired term since λ is adaptively chosen as a function of the whole history up to time t. As a solution, we shall use a covering argument on the whole model family $\mathcal{M}$.

Suppose that we have a finite ξ-cover $\bar{\mathcal{M}}_\xi$ of $\mathcal{M}$, i.e., for any $\lambda \in \mathcal{M}$, there exists $\bar{\lambda} \in \bar{\mathcal{M}}_\xi$ such that $\|\lambda - \bar{\lambda}\|_\infty \leq \xi$. For such a couple $(\lambda, \bar{\lambda})$, this implies that, for any $s \geq 1, k \in [K]$,

$$\begin{aligned}\left|(\mu^k - \lambda^k)^2 - (\mu^k - \bar{\lambda}_k)^2\right| &= \left|(\bar{\lambda}_k - \lambda^k)^2 + 2(\mu^k - \bar{\lambda}_k)(\bar{\lambda}_k - \lambda^k)\right| \\ &\leq (\bar{\lambda}_k - \lambda^k)^2 + 2|\mu^k - \bar{\lambda}_k||\bar{\lambda}_k - \lambda^k| \leq \xi^2 + 4M\xi.\end{aligned}$$

Moreover, using this bound in the definition of $Z_s(\lambda)$,

$$\left|\sum_{s=1}^t Z_s(\lambda) - \sum_{s=1}^t Z_s(\bar{\lambda})\right| \leq 2t\xi^2 + 8tM\xi.$$

Let $h(t)$ be some function to be specified later. With some abuse of notation w.r.t. the derivation above, we shall instantiate a different ξ_t-cover for each time step t. Then

$$\begin{aligned}\mathbb{P}\left\{\exists t \geq 1 : \sup_{\lambda\in\mathcal{M}}\left|\sum_{s=1}^t Z_s(\lambda)\right| > h(t)\right\} &= \mathbb{P}\left\{\exists t \geq 1 : \sup_{\lambda\in\mathcal{M}}\inf_{\bar{\lambda}\in\bar{\mathcal{M}}_{\xi_t}}\left|\sum_{s=1}^t Z_s(\lambda) \pm Z_s(\bar{\lambda})\right| > h(t)\right\} \\ &\overset{(a)}{\leq} \mathbb{P}\left\{\exists t \geq 1 : \sup_{\bar{\lambda}\in\bar{\mathcal{M}}_{\xi_t}}\left|\sum_{s=1}^t Z_s(\bar{\lambda})\right| + \sup_{\lambda\in\mathcal{M}}\inf_{\bar{\lambda}\in\bar{\mathcal{M}}_{\xi_t}}\left|\sum_{s=1}^t Z_s(\lambda) - Z_s(\bar{\lambda})\right| > h(t)\right\} \\ &\overset{(b)}{\leq} \mathbb{P}\left\{\exists t \geq 1, \bar{\lambda} \in \bar{\mathcal{M}}_{\xi_t} : \left|\sum_{s=1}^t Z_s(\bar{\lambda})\right| > h(t) - 2t{\xi_t}^2 - 8tM\xi_t\right\} \\ &\overset{(c)}{\leq} \sum_{t=1}^\infty \sum_{\bar{\lambda}\in\bar{\mathcal{M}}_{\xi_t}} \mathbb{P}\left\{\left|\sum_{s=1}^t Z_s(\bar{\lambda})\right| > h(t) - 2t{\xi_t}^2 - 8tM\xi_t\right\},\end{aligned}$$

where (a) follows by the triangle inequality, (b) from the property of the cover, and (c) from the union bound and the fact that the cover is finite. Let $\delta'_t \in (0,1)$. If we choose $h(t) := 2M^2\sqrt{\frac{t}{2}\log(2/\delta'_t)} + 2t{\xi_t}^2 + 8tM\xi_t$, using Azuma's inequality, each probability in the sum above is bounded by δ'_t. Hence, choosing $\delta'_t := \frac{\delta'}{2|\bar{\mathcal{M}}_{\xi_t}|t^2}$,

$$\sum_{t=1}^\infty \sum_{\bar{\lambda}\in\bar{\mathcal{M}}_{\xi_t}} \mathbb{P}\left\{\left|\sum_{s=1}^t Z_s(\bar{\lambda})\right| > h(t) - 2t{\xi_t}^2 - 8tM\xi_t\right\} \leq \sum_{t=1}^\infty \frac{\delta'}{2t^2} \leq \delta',$$

where the last inequality can be verified easily. Therefore, putting everything together, we proved that

$$\mathbb{P}\left\{\exists t \geq 1 : \sup_{\lambda\in\mathcal{M}}\left|\sum_{s=1}^t Z_s(\lambda)\right| > 2M^2\sqrt{\frac{t}{2}\log\frac{4|\bar{\mathcal{M}}_{\xi_t}|t^2}{\delta'}} + 2t{\xi_t}^2 + 8tM\xi_t\right\} \leq \delta'.$$

It only remains to build the cover, compute its size, and specify the value of ξ_t. Recall that each model $\lambda \in \mathcal{M}$ can be written as $\lambda = A\theta' + \eta'$, where $\|\eta'\|_\infty \leq \varepsilon$ and $\|\lambda\|_\infty \leq M$. Using Lemma 28 below, we have that $\|\theta'\|_2 \leq \bar{B} := \frac{M+\varepsilon}{\sqrt{\Gamma(A)}}$. Then, we can build two separate covers for the linear and deviation parts. Specifically, we build a $\xi_t/(2L)$-cover $\bar{\mathcal{M}}_t^{\text{lin}}$ in ℓ_2-norm for the linear part and a $\xi_t/2$-cover $\bar{\mathcal{M}}_t^{\text{dev}}$ in ℓ_∞-norm for the deviation part. Then, we take the full cover as the (finite) set $\bar{\mathcal{M}}_t := \{\bar{\lambda} = A\bar{\theta} + \bar{\eta} : \bar{\theta} \in \bar{\mathcal{M}}_t^{\text{lin}}, \bar{\eta} \in \bar{\mathcal{M}}_t^{\text{dev}}\}$. With this choice, we have that, for any $\lambda = A\theta' + \eta'$, there exits $\bar{\lambda} \in \bar{\mathcal{M}}_t$ such that

$$\|\lambda - \bar{\lambda}\|_\infty = \|A\theta' + \eta' - A\bar{\theta} - \bar{\eta}\|_\infty \leq L\|\theta' - \bar{\theta}\|_2 + \|\eta' - \bar{\eta}\|_\infty \leq \xi_t.$$

Let us compute the size of the cover $\bar{\mathcal{M}}_t$. It is easy to see that this is $|\bar{\mathcal{M}}_t| = |\bar{\mathcal{M}}_t^{\text{lin}}||\bar{\mathcal{M}}_t^{\text{dev}}|$. For the linear one, it is known that the $\xi_t/(2L)$-covering number (in ℓ_2-norm) of a ball in $\mathbb{R}^d$ with radius $\bar{B}$ is at most $(6L\bar{B}/\xi_t)^d$. For the deviation, we can have a $\xi_t/2$ cover in ℓ_∞-norm with at most $\max\{(4\varepsilon/\xi_t)^K, 1\}$ points, where the maximum is to deal with too small values of ε (*e.g.*, $\varepsilon = 0$). Then, the final cover has size at most $|\bar{\mathcal{M}}_t| \le (6\bar{B}L/\xi_t)^d \max\{(4\varepsilon/\xi_t)^K, 1\}$. Setting $\xi_t = 1/t$, we get the desired bound. □

F δ-correctness and sample complexity analysis

F.1 Correctness

We prove Lemma 3 in the main paper, restated below.

Lemma. *Let W_{-1} be the negative branch of the Lambert W function and let $\overline{W}(x) = -W_{-1}(-e^{-x}) \approx x + \log x$. For $\delta \in (0,1)$, define*

$$\beta_{t,\delta}^{\text{uns}} := 2K\overline{W}\left(\frac{1}{2K}\log\frac{2e}{\delta} + \frac{1}{2}\log(8eK\log t)\right), \tag{7}$$

$$\beta_{t,\delta}^{\text{lin}} := \frac{1}{2}\left(4\sqrt{t}\varepsilon + \sqrt{2}\sqrt{1 + \log\frac{1}{\delta} + \left(1 + \frac{1}{\log(1/\delta)}\right)\frac{d}{2}\log\left(1 + \frac{t}{2d}\log\frac{1}{\delta}\right)}\right)^2 . \tag{8}$$

Then, for the choice $\beta_{t,\delta} := \min\{\beta_{t,\delta}^{\text{uns}}, \beta_{t,\delta}^{\text{lin}}\}$, MisLid *is δ-correct.*

Proof. δ-correctness is composed of two properties: stopping in a finite time with probability one and verifying, for all instances $\mu \in \mathcal{M}$, $\mathbb{P}(\hat{S}_m \not\subseteq S^\star(\mu)) \le \delta$. The fact that the stopping time is finite almost surely is a consequence of the sample complexity bound (see further down in this section). We now prove the bound on the probability of error in identification.

We first relate the event that the algorithm does not return a correct answer to a large deviation, by writing that for the algorithm to make a mistake, there must be a time at which the stopping condition is met and $\tilde{\mu}_t$ is in the alternative to μ:

$$\mathbb{P}(\hat{S}_m \not\subseteq S^\star(\mu)) \le \mathbb{P}\left(\exists t \in \mathbb{N},\ \inf_{\lambda \in \Lambda_m(\tilde{\mu}_t)} \|\tilde{\mu}_t - \lambda\|_{D_{N_t}}^2 > 2\beta_{t,\delta} \ \wedge\ \tilde{\mu}_t \in \Lambda_m(\mu)\right).$$

If the two conditions of the right-hand side happen, then $\mu \in \Lambda_m(\tilde{\mu}_t)$ and we get

$$\mathbb{P}(\hat{S}_m \not\subseteq S^\star(\mu)) \le \mathbb{P}\left(\exists t \in \mathbb{N},\ \|\tilde{\mu}_t - \mu\|_{D_{N_t}}^2 > 2\beta_{t,\delta}\right).$$

It then suffices to prove that we have both

$$\mathbb{P}\left(\exists t \in \mathbb{N},\ \frac{1}{2}\|\tilde{\mu}_t - \mu\|_{D_{N_t}}^2 > \beta_{t,\delta}^{\text{lin}}\right) \le \delta/2\,, \tag{9}$$

$$\text{and}\quad \mathbb{P}\left(\exists t \in \mathbb{N},\ \frac{1}{2}\|\tilde{\mu}_t - \mu\|_{D_{N_t}}^2 > \beta_{t,\delta}^{\text{uns}}\right) \le \delta/2\,. \tag{10}$$

The result for (10) is Lemma 11 (and the remark below that lemma stating that it applies to $\tilde{\mu}_t$). We now prove the concentration inequality using the linear term (9).

let $\tilde{\theta}_{t,\varepsilon}$ and $\tilde{\eta}_{t,\varepsilon}$ be parameters for $\tilde{\mu}_t$ with $\|\tilde{\eta}_{t,\varepsilon}\| \le \varepsilon$, which exist since $\tilde{\mu}_t \in \mathcal{M}$. On the other hand, let $\tilde{\theta}_t$ and $\tilde{\eta}_t$ be the orthogonal parameters of $\tilde{\mu}_t$ with respect to N_t.

$$\begin{aligned}
\|\tilde{\mu} - \mu\|_{D_{N_t}} &= \|A(\tilde{\theta}_{t,\varepsilon} - \theta) + \tilde{\eta}_{t,\varepsilon} - \eta\|_{D_{N_t}} \\
&\le \|A(\tilde{\theta}_{t,\varepsilon} - \theta)\|_{D_{N_t}} + \|\tilde{\eta}_{t,\varepsilon} - \eta\|_{D_{N_t}} \\
&= \|\tilde{\theta}_{t,\varepsilon} - \theta\|_{V_t} + \|\tilde{\eta}_{t,\varepsilon} - \eta\|_{D_{N_t}} \\
&\le \|\tilde{\theta}_{t,\varepsilon} - \tilde{\theta}_t\|_{V_t} + \|\tilde{\theta}_t - \theta_t\|_{V_t} + \|\theta_t - \theta\|_{V_t} + \|\tilde{\eta}_{t,\varepsilon} - \eta\|_{D_{N_t}}\,.
\end{aligned}$$

Lemma 6 bounds the first and third terms by $\sqrt{t}\varepsilon$. The last term is bounded by $\sqrt{t}\|\tilde{\eta}_{t,\varepsilon}-\eta\|_\infty \le 2\sqrt{t}\varepsilon$ since both vectors have ℓ_∞ norm bounded by ε.

Finally

$$\mathbb{P}\left(\exists t\in\mathbb{N},\ \frac{1}{2}\left\|\tilde{\mu}_t-\mu\right\|^2_{D_{N_t}} > \beta^{\text{lin}}_{t,\delta}\right) \le \mathbb{P}\left(\exists t\in\mathbb{N},\ \frac{1}{2}\left\|\widehat{\theta}_t-\theta_t\right\|^2_{V_t} > \frac{1}{2}(\sqrt{2\beta^{\text{lin}}_{t,\delta}}-4\sqrt{t}\varepsilon)^2\right) \le \delta/2$$

by Corollary 2.

□

F.2 Restriction to a good event

Assumption. We start by pulling arms deterministically until t_0, such that $V_{t_0} \ge 2L^2 I_d$. See paragraph "Initialization phase" in Subsection 4.1 in the main paper.

Definition of the good event. For $t \ge t_0$ and $k \in [K]$, define

$$\alpha^{\text{lin}}_t := \log(5t^2) + d\log\left(1+\frac{t}{2d}\right),\ \alpha^{\text{uns}}_t := 2K\overline{W}\left(\frac{1}{2K}\log(2e5t^3)+\frac{1}{2}\log(8eK\log t)\right).$$

Consider the following events. Each of these holds with probability at least $1-\frac{1}{5t^2}$ by the indicated concentration result.

1. Concentration of the projected linear part (Corollary 1)
$$\mathcal{E}^1_t := \left\{\forall s\ge t_0 : \frac{1}{2}\left\|\tilde{\theta}_s-\theta_s\right\|^2_{V_s} \le \alpha^{\text{lin}}_t\right\},$$
2. Unstructured concentration of the projected estimator (Lemma 11)
$$\mathcal{E}^2_t := \left\{\forall s\le t : \frac{1}{2}\left\|\tilde{\mu}_s-\mu\right\|^2_{D_{N_s}} \le \alpha^{\text{uns}}_t\right\},$$
3. Martingale concentration for sampling (Lemma 17)
$$\mathcal{E}^3_t := \left\{\sup_{\lambda\in\mathcal{M}}\left|\sum_{s=1}^t Z_s(\lambda)\right| \le r(t)\right\},$$
where $r(t)$ is obtained by setting $\delta' = \frac{1}{5t^2}$ in $r(t,\delta')$ in Lemma 17, which yields
$$r(t) = 2M^2\sqrt{\frac{t}{2}\left(\log(4\times 5t^4)+d\log\frac{6(M+\varepsilon)Lt}{\sqrt{\Gamma(A)}}+K\log\max\{4\varepsilon t,1\}\right)}+2+8M.$$
4. Elliptical potential with sampling (Lemma 12)
$$\mathcal{E}^4_t := \left\{\sum_{s=t_0+1}^t\sum_{k=1}^K \omega^k_s\left\|\phi_k\right\|^2_{V^{-1}_{s-1}} \le \sqrt{2t\log(5t^2)}+d\log\left(1+\frac{t}{d}\right)\right\},$$
5. Elliptical potential with sampling for the unstructured bound (Lemma 16)
$$\mathcal{E}^5_t := \left\{\sum_{s=t_0+1}^t\sum_{k=1}^K \omega^k_s\min\left\{4M^2,\frac{2\alpha^{\text{uns}}_t}{N^k_{s-1}}\right\} \le 4M^2\sqrt{2t\log(5t^2)}+4M^2K+2K\alpha^{\text{uns}}_t\log(et)\right\}.$$

Then, we define the "good" event $\mathcal{E}_t := \bigcap_{i=1}^5 \mathcal{E}^i_t$.

Lemma 18. *For all $t\ge 1$, $\mathbb{P}(\mathcal{E}^c_t) \le 1/t^2$.*

Proof. Apply an union bound by noting that each event $\mathcal{E}^i_t$ fails with probability at most $1/(5t^2)$. □

Lemma 19. *Let $T_0(\delta) \in \mathbb{N}$ be such that for $t \geq T_0(\delta)$, $\mathcal{E}_t \subseteq \{\tau_\delta \leq t\}$. Then $\mathbb{E}[\tau_\delta] \leq T_0(\delta) + 2$.*

Proof. Successively using the definition of $T_0(\delta)$ and Lemma 18:

$$\mathbb{E}[\tau_\delta] = \sum_{t=0}^{+\infty} \mathbb{P}(\tau_\delta > t) \leq T_0(\delta) + \sum_{t=T_0(\delta)}^{+\infty} \mathbb{P}(\mathcal{E}_t^c) \leq T_0(\delta) + \sum_{t=1}^{+\infty} \frac{1}{t^2} \leq T_0(\delta) + 2 \,.$$

□

Consequences of the good event.

Lemma 20. *For $t \geq t_0$ and $k \in [K]$, define*

$$c_t^k := \min\left\{ 8(LK+1)^2\varepsilon^2 + 4\alpha_{t^2}^{\mathrm{lin}} \left\|\phi_k\right\|_{V_t^{-1}}^2 , \frac{2\alpha_{t^2}^{\mathrm{uns}}}{N_t^k}, 4M^2 \right\} ,$$

where we use the convention that $2\alpha^{\mathrm{uns}}/N_t^k = +\infty$ if $N_t^k = 0$. Then under $\mathcal{E}_t$, for all $s \in \{\max\{t_0, \sqrt{t}\}, \dots, t\}$ and $k \in [K]$, $(\tilde{\mu}_s^k - \mu^k)^2 \leq c_s^k$.

Proof. We know that $\frac{1}{2}\left\|\tilde{\theta}_s - \theta_s\right\|_{V_t}^2 \leq \alpha_t^{\mathrm{lin}}$ holds for all $s \geq t_0$ by definition of $\mathcal{E}_t^1$. For $s \geq \max\{t_0, \sqrt{t}\}$ we also have $\alpha_{s^2}^{\mathrm{lin}} \geq \alpha_t^{\mathrm{lin}}$, hence $\frac{1}{2}\left\|\tilde{\theta}_s - \theta_s\right\|_{V_t}^2 \leq \alpha_{s^2}^{\mathrm{lin}}$. Using first $(a+b)^2 \leq 2a^2+2b^2$ then the Cauchy-Schwarz inequality on $(V_s^{-1/2}\phi_k)^\top (V_s^{1/2}(\tilde{\theta}_s - \theta_s))$ and Lemma 5,

$$\begin{aligned}(\tilde{\mu}_s^k - \mu^k)^2 \leq 2(\phi_k^\top(\tilde{\theta}_s - \theta_s))^2 + 2(\tilde{\eta}_s^k - \eta_s^k)^2 &\leq 2\left\|\phi_k\right\|_{V_s^{-1}}^2 \left\|\tilde{\theta}_s - \theta_s\right\|_{V_s}^2 + 8(LK+1)^2\varepsilon^2 \\ &\leq 8(LK+1)^2\varepsilon^2 + 4\alpha_{s^2}^{\mathrm{lin}} \left\|\phi_k\right\|_{V_s^{-1}}^2 \,.\end{aligned}$$

Moreover by definition of $\mathcal{E}_t^2$, for all $s \leq t$, $\frac{1}{2}\left\|\tilde{\mu}_s - \mu\right\|_{D_{N_s}}^2 \leq \alpha_t^{\mathrm{uns}}$. For $s \geq \max\{t_0, \sqrt{t}\}$ we have $\alpha_{s^2}^{\mathrm{uns}} \geq \alpha_t^{\mathrm{uns}}$, hence $\frac{1}{2}\left\|\tilde{\mu}_s - \mu\right\|_{D_{N_s}}^2 \leq \alpha_{s^2}^{\mathrm{uns}}$. Therefore,

$$(\tilde{\mu}_s^k - \mu^k)^2 = (e_k^\top(\tilde{\mu}_s - \mu))^2 \leq \left\|e_k\right\|_{D_{N_s}^{-1}}^2 \left\|\mu - \tilde{\mu}_s\right\|_{D_{N_s}}^2 \leq \frac{2\alpha_{s^2}^{\mathrm{uns}}}{N_s^k}.$$

Finally, $(\tilde{\mu}_s^k - \mu^k)^2 \leq \left\|\tilde{\mu}_s - \mu\right\|_\infty^2 \leq 4M^2$. □

Lemma 21. *For all $t \geq 1$, under the good event $\mathcal{E}_t$,*

$$\forall s \in \{t_0, t_0+1, \dots, t\} : \left\|\tilde{\mu}_s - \mu\right\|_{D_{N_s}}^2 \leq f(t) := 2\min\{\alpha_t^{\mathrm{uns}}, \alpha_t^{\mathrm{lin}} + 2t(LK+1)^2\varepsilon^2\}.$$

Proof. That for all $s \leq t$, $\left\|\tilde{\mu}_s - \mu\right\|_{D_{N_s}}^2 \leq 2\alpha_t^{\mathrm{uns}}$ directly follows from the definition of $\mathcal{E}_t^2$. To see the second inequality, we first decompose the norm on the lefthand-side into its linear and deviation components

$$\left\|\tilde{\mu}_s - \mu\right\|_{D_{N_s}}^2 = \left\|\tilde{\theta}_s - \theta_s\right\|_{V_s}^2 + \left\|R_{N_s} D_{N_s}^{1/2}\tilde{\eta}_s - R_{N_s} D_{N_s}^{1/2}\eta_s\right\|^2 \,.$$

The deviation part can be bounded by $4t(LK+1)^2\varepsilon^2$ for all $s \leq t$ using Lemma 5. The linear part can be bounded by $2\alpha_t^{\mathrm{lin}}$ for all $s \geq t_0$ by the definition of $\mathcal{E}_t^1$. □

F.3 Analysis under a good event

Fix any time step $t \geq t_0$. Suppose that the good event $\mathcal{E}_t$ of Section F.2 holds and the algorithm does not stop at time t. We proceed in different steps.

Stopping rule analysis.

Theorem 4. *If the algorithm does not stop at time t then under $\mathcal{E}_t$, using stopping threshold $\beta_{t,\delta}$ as defined in Lemma 3 in the main paper,*

$$2\beta_{t,\delta} \geq \inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_{W_t}} - h_\delta(t) - r(t) \,.$$

where

- $h_\delta(t) = \sqrt{8\beta_{t,\delta} f(t)} + f(t)$, *with $f(t)$ a bound on $\|\mu - \tilde{\mu}_t\|^2_{D_{N_t}}$ (see Lemma 21)* ,
- $r(t) = 2M^2 \sqrt{\frac{t}{2}\left(\log(4 \times 5t^4) + d \log \frac{6(M+\varepsilon)Lt}{\sqrt{\Gamma(A)}} + K \log \max\{4\varepsilon t, 1\}\right)} + 2 + 8M$,

and $W_t := \sum_{s \leq 1}^t \omega_s$ is the sum over time of the weight vectors played by the learner.

The proof of this theorem is detailed in Steps 1 to 3 below.

Step 1. From $\Lambda_m(\tilde{\mu}_t)$ to $\Lambda_m(\mu)$.

Lemma 22. *For all $\mu, \mu' \in \mathcal{M}$, for any non-negative function $f : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}$ with $f(x, x) = 0$,*

$$\inf_{\lambda \in \Lambda_m(\mu)} f(\mu, \lambda) \geq \inf_{\lambda \in \Lambda_m(\mu')} f(\mu, \lambda) \,.$$

Proof. Either $\Lambda_m(\mu) = \Lambda_m(\mu')$ and the two expressions are equal, or $\Lambda_m(\mu) \neq \Lambda_m(\mu')$. In the second case, $\mu \in \Lambda_m(\mu')$. The right-hand side is then equal to zero, which is lower than the left-hand side since f is non-negative. □

Since the algorithm does not stop at time t, from the stopping rule

$$2\beta_{t,\delta} \geq \inf_{\lambda \in \Lambda_m(\tilde{\mu}_t)} \|\tilde{\mu}_t - \lambda\|^2_{D_{N_t}} \,,$$

where $\Lambda_m(\tilde{\mu}_t)$ is the set of alternative models to $\tilde{\mu}_t$. We change the alternative set over which the minimization is performed using Lemma 22:

$$2\beta_{t,\delta} \geq \inf_{\lambda \in \Lambda_m(\tilde{\mu}_t)} \|\tilde{\mu}_t - \lambda\|^2_{D_{N_t}} \geq \inf_{\lambda \in \Lambda_m(\mu)} \|\tilde{\mu}_t - \lambda\|^2_{D_{N_t}} \,. \tag{11}$$

Step 2. From $\tilde{\mu}_t$ to μ. The next step is to replace the estimated mean $\tilde{\mu}_t$ in the norm with the true mean μ. For all $\lambda \in \mathcal{M}$, using the triangle inequality,

$$\|\tilde{\mu}_t - \lambda\|_{D_{N_t}} \geq \|\mu - \lambda\|_{D_{N_t}} - \|\tilde{\mu}_t - \mu\|_{D_{N_t}} \geq \|\mu - \lambda\|_{D_{N_t}} - \sqrt{f(t)} \,,$$

where the last inequality uses Lemma 21 to concentrate $\|\tilde{\mu}_t - \mu\|^2_{D_{N_t}}$. Using this for the specific choice of $\lambda_t \in \arg\min_{\lambda \in \Lambda_m(\mu)} \|\tilde{\mu}_t - \lambda\|^2_{D_{N_t}}$ in combination with (11), we obtain

$$\left(\sqrt{2\beta_{t,\delta}} + \sqrt{f(t)}\right)^2 \geq \|\mu - \lambda_t\|^2_{D_{N_t}} \quad \Rightarrow \quad 2\beta_{t,\delta} \geq \inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_{N_t}} - h_\delta(t) \,, \tag{12}$$

where $h_\delta(t) := \sqrt{8\beta_{t,\delta} f(t)} + f(t)$ is a sub-linear function of both t and $\log(1/\delta)$.

Step 3. From N_t to W_t. We now show that it is possible to replace N_t with $W_t := \sum_{s=1}^t \omega_s$ in the norm at the price of subtracting another low-order term. Let $Z_s(\lambda) := (\mu^{k_s} - \lambda^{k_s})^2 - \mathbb{E}_{k \sim \omega_s}[(\mu^k - \lambda^k)^2]$. Note that $\|\mu - \lambda\|^2_{D_{N_t}} = \sum_{s=1}^t (\mu^{k_s} - \lambda^{k_s})^2$ and $\|\mu - \lambda\|^2_{D_{W_t}} = \sum_{s=1}^t \|\mu - \lambda\|^2_{D_{\omega_s}} = \sum_{s=1}^t \mathbb{E}_{k \sim \omega_s}[(\mu^k - \lambda^k)^2]$. Therefore, from (12),

$$\begin{aligned}
2\beta_{t,\delta} &\geq \inf_{\lambda \in \Lambda_m(\mu)} \left(\|\mu - \lambda\|^2_{D_{N_t}} - \|\mu - \lambda\|^2_{D_{W_t}} + \|\mu - \lambda\|^2_{D_{W_t}}\right) - h_\delta(t) \\
&= \inf_{\lambda \in \Lambda_m(\mu)} \left(\|\mu - \lambda\|^2_{D_{W_t}} + \sum_{s=1}^t Z_s(\lambda)\right) - h_\delta(t) \\
&\geq \inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_{W_t}} - \sup_{\lambda \in \mathcal{M}} \left|\sum_{s=1}^t Z_s(\lambda)\right| - h_\delta(t) \,.
\end{aligned}$$

Using the good event $\mathcal{E}_t^3$, we can finally write

$$2\beta_{t,\delta} \geq \inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_{W_t}} - h_\delta(t) - r(t)\,, \tag{13}$$

which ends proving Theorem 4.

Sampling rule analysis. Let $H_\mu = \sup_{\omega \in \Delta} \inf_{\lambda \in \Lambda_m(\mu)} \frac{1}{2}\|\mu - \lambda\|^2_{D_\omega}$ (the inverse complexity at μ). In the first part of the sampling rule analysis, we introduce the optimistic estimates $g_t(\omega)$ mentioned in Algorithm 1 in the main paper, which will be used by the learner for ω_t.

Theorem 5. *Let $(\tilde{\mu}_s)_{s \leq t} \in \mathcal{M}^{[t]}$ be estimates such that under $\mathcal{E}_t$, we have a bound c_s^k on $(\tilde{\mu}_s^k - \mu^k)^2$ for all $k \in [K]$ and $s \in [t]$. Then define the* optimistic estimate

$$g_s(\omega) := \sum_{k=1}^K \omega^k \left(\left|\tilde{\mu}_{s-1}^k - \lambda_s^k\right| + \sqrt{c_{s-1}^k} \right)^2 \qquad \textit{where } \lambda_s := \operatorname*{arg\,min}_{\lambda \in \Lambda_m(\tilde{\mu}_{s-1})} \|\tilde{\mu}_{s-1} - \lambda\|^2_{D_{\omega_s}}\,.$$

Under $\mathcal{E}_t$,

$$\inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_{W_t}} \geq \sum_{s=t_0+1}^t g_s(\omega_s) - 4C_t - 4\sqrt{2tC_tH_\mu}\,,$$

with $C_t := \sum_{s=t_0+1}^t \sum_{k=1}^K \omega_s^k c_{s-1}^k$.

The proof of this theorem is detailed in the Steps 4 to 7 below. Once this result is established, we will use the regret property of the learner to exhibit the final bound (Steps 8 to 10).

Step 4. From $\Lambda_m(\mu)$ back to $\Lambda_m(\tilde{\mu}_{s-1})$ for $s \in [t]$. We now start moving from $\inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_{W_t}}$ to the actual gain fed into the online learner at time t. We first need to go back to the estimated set of alternative models at each time $s = 0, \dots, t-1$. We have

$$\inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_{W_t}} = \inf_{\lambda \in \Lambda_m(\mu)} \sum_{s=1}^t \|\mu - \lambda\|^2_{D_{\omega_s}} \geq \sum_{s=1}^t \inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_{\omega_s}} \tag{14}$$

$$\geq \sum_{s=1}^t \inf_{\lambda \in \Lambda_m(\tilde{\mu}_{s-1})} \|\mu - \lambda\|^2_{D_{\omega_s}}\,, \tag{15}$$

where the first inequality follows by the concavity of the infimum, and the second one is an application of Lemma 22.

Step 5. Drop the first rounds. The first t_0 rounds are dedicated to making sure that V_t is sufficiently large (for the partial order on positive definite matrices). Also, our upper bounds on the deviation of $\tilde{\mu}_t^k$ from μ^k are valid from $\max\{t_0, \sqrt{t}\}$. We define $t_0'(t) = \max\{t_0, \sqrt{t}\}$. We drop the corresponding nonnegative terms from the sum to keep only the rounds for which t is large enough:

$$\sum_{s=1}^t \inf_{\lambda \in \Lambda_m(\tilde{\mu}_{s-1})} \|\mu - \lambda\|^2_{D_{\omega_s}} \geq \sum_{s=t_0'(t)+1}^t \inf_{\lambda \in \Lambda_m(\tilde{\mu}_{s-1})} \|\mu - \lambda\|^2_{D_{\omega_s}}\,.$$

Step 6. From μ back to $\tilde{\mu}_{s-1}$ for $s \in [t]$. We can now use the concentration of $\tilde{\mu}_{s-1}$ to replace μ in all terms $\|\mu - \lambda\|^2_{D_{\omega_s}}$ for $s \in [t]$. Let $\lambda_s^\mu := \arg\min_{\lambda \in \Lambda_m(\tilde{\mu}_{s-1})} \|\mu - \lambda\|^2_{D_{\omega_s}}$. Using first the triangle inequality, then the inequality $\|a - b\| \geq \|a\| - \|b\|$ for an ℓ_2 norm in dimension $t - t_0'(t)$,

$$\sqrt{\sum_{s=t_0'(t)+1}^t \|\mu - \lambda_s^\mu\|^2_{D_{\omega_s}}} \geq \sqrt{\sum_{s=t_0'(t)+1}^t \left(\|\tilde{\mu}_{s-1} - \lambda_s^\mu\|_{D_{\omega_s}} - \|\mu - \tilde{\mu}_{s-1}\|_{D_{\omega_s}} \right)^2}$$

$$\geq \sqrt{\sum_{s=t_0'(t)+1}^t \|\tilde{\mu}_{s-1} - \lambda_s^\mu\|^2_{D_{\omega_s}}} - \sqrt{\sum_{s=t_0'(t)+1}^t \|\mu - \tilde{\mu}_{s-1}\|^2_{D_{\omega_s}}}\,.$$

We now remark that $\sum_{s=t_0'(t)+1}^{t} \|\tilde{\mu}_{s-1} - \mu\|^2_{D_{w_s}} \leq C_t$ and get, by the definition of λ_s^μ

$$\sqrt{\sum_{s=t_0'(t)+1}^{t} \|\mu - \lambda_s^\mu\|^2_{D_{\omega_s}}} + \sqrt{C_t} \geq \sqrt{\sum_{s=t_0'(t)+1}^{t} \|\tilde{\mu}_{s-1} - \lambda_s^\mu\|^2_{D_{w_s}}} \tag{16}$$

$$\geq \sqrt{\sum_{s=t_0'(t)+1}^{t} \inf_{\lambda \in \Lambda_m(\tilde{\mu}_{s-1})} \|\tilde{\mu}_{s-1} - \lambda\|^2_{D_{w_s}}} \,. \tag{17}$$

Step 7. Optimistic gains. We now replace the term on the right-hand side in (17) by the optimistic gains fed into the online learner. At time s, we define optimistic estimates

$$g_s(\omega) := \sum_{k=1}^{K} \omega^k \left(\left| \tilde{\mu}_{s-1}^k - \lambda_s^k \right| + \sqrt{c_{s-1}^k} \right)^2 \quad \text{where } \lambda_s := \operatorname*{arg\,min}_{\lambda \in \Lambda_m(\tilde{\mu}_{s-1})} \|\tilde{\mu}_{s-1} - \lambda\|^2_{D_{\omega_s}} \,.$$

Lemma 23. *For all $\omega \in \Delta_K$ and $s \geq t_0'(t)$, $g_s(\omega) \geq \inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_\omega}$.*

Proof. For all $k \in [K]$, $s > t_0'(t)$, and $\lambda \in \mathbb{R}^K$, using Lemma 20 (to write $(\mu^k - \tilde{\mu}_{s-1}^k)^2 \leq c_{s-1}^k$):

$$\begin{aligned}
\left(\mu^k - \lambda^k\right)^2 = \left(\tilde{\mu}_{s-1}^k - \lambda^k + \mu^k - \tilde{\mu}_{s-1}^k\right)^2 &\leq \left(\left|\tilde{\mu}_{s-1}^k - \lambda^k\right| + \left|\mu^k - \tilde{\mu}_{s-1}^k\right|\right)^2 \\
&\leq \left(\left|\tilde{\mu}_{s-1}^k - \lambda^k\right| + \sqrt{c_{s-1}^k} \right)^2 .
\end{aligned}$$

Then, for any $\omega \in \Delta_K$, by noticing that function $f : \lambda \mapsto \sum_k^K \omega^k \left(\left|\tilde{\mu}_{s-1}^k - \lambda^k\right|^2 + 2\left|\tilde{\mu}_{s-1}^k - \lambda^k\right| \sqrt{c_{s-1}^k} \right)$ is nonnegative and that $f(\tilde{\mu}_{s-1}^k) = 0$:

$$\begin{aligned}
g_s(\omega) &:= \sum_{k=1}^{K} \omega^k \left(\left|\tilde{\mu}_{s-1}^k - \lambda_s^k\right| + \sqrt{c_{s-1}^k} \right)^2 \\
&\geq \inf_{\lambda \in \Lambda_m(\tilde{\mu}_{s-1})} \sum_{k=1}^{K} \omega^k \left(\left|\tilde{\mu}_{s-1}^k - \lambda^k\right| + \sqrt{c_{s-1}^k} \right)^2 \\
&= \sum_{k=1}^{K} \omega^k c_{s-1}^k + \inf_{\lambda \in \Lambda_m(\tilde{\mu}_{s-1})} \sum_{k=1}^{K} \omega^k \left(\left|\tilde{\mu}_{s-1}^k - \lambda^k\right|^2 + 2\left|\tilde{\mu}_{s-1}^k - \lambda^k\right| \sqrt{c_{s-1}^k} \right) \\
&\geq \sum_{k=1}^{K} \omega^k c_{s-1}^k + \inf_{\lambda \in \Lambda_m(\mu)} \sum_{k=1}^{K} \omega^k \left(\left|\tilde{\mu}_{s-1}^k - \lambda^k\right|^2 + 2\left|\tilde{\mu}_{s-1}^k - \lambda^k\right| \sqrt{c_{s-1}^k} \right) \\
&\quad \text{(due to Lemma 22)} \\
&= \inf_{\lambda \in \Lambda_m(\mu)} \sum_{k=1}^{K} \omega^k \left(\left|\tilde{\mu}_{s-1}^k - \lambda^k\right| + \sqrt{c_{s-1}^k} \right)^2 \\
&\geq \inf_{\lambda \in \Lambda_m(\mu)} \sum_{k=1}^{K} \omega^k \left(\mu^k - \lambda^k\right)^2 \quad \text{(using the previously derived coordinate-wise majoration)} \\
&= \inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|^2_{D_\omega} \,.
\end{aligned}$$

□

We now prove an upper bound on $g_w(\omega)$, which will be useful later.

Lemma 24. *For all $s \geq t_0$ and all $\omega \in \triangle_K$, $g_s(\omega) \leq 36M^2$*

Proof. Using the definition of $c_t^k, k \in [K], t \geq 0$ in Lemma 20, and $\mu, \lambda_s \in \mathcal{M}$: $g_s(\omega) = \sum_{k=1}^K \omega^k \left(\left| \tilde{\mu}_{s-1}^k - \lambda_s^k \right| + \sqrt{c_{s-1}^k} \right)^2 \leq \sum_{k=1}^K \omega^k \left(\left| \mu^k - \lambda_s^k \right| + 2\sqrt{c_{s-1}^k} \right)^2 \leq \sum_{k=1}^K \omega^k \left(6M\right)^2 = 36M^2$. □

We have proved that the estimates are indeed optimistic in the sense that they are an upper-bound to the value of interest, as mentioned in paragraph "Optimistic gains" in Subsection 4.1 in the main paper. We now bound by how much they overestimate the empirical value.

Lemma 25.

$$\sqrt{\sum_{s=t_0'(t)+1}^{t} \inf_{\lambda \in \Lambda_m(\tilde{\mu}_{s-1})} \|\tilde{\mu}_{s-1} - \lambda\|_{D_{\omega_s}}^2} \geq \sqrt{\sum_{s=t_0'(t)+1}^{t} g_s(\omega_s)} - \sqrt{C_t} \,. \tag{18}$$

Proof. We start by a bound for a single $s \in \mathbb{N}$. Using the triangle inequality for an ℓ_2 norm,

$$\begin{aligned}
\sqrt{g_s(\omega_s)} &= \sqrt{\sum_{k=1}^K \omega_s^k \left(\left| \tilde{\mu}_{s-1}^k - \lambda_s^k \right| + \sqrt{c_{s-1}^k} \right)^2} \\
&\leq \sqrt{\sum_{k=1}^K \omega_s^k \left(\tilde{\mu}_{s-1}^k - \lambda_s^k \right)^2} + \sqrt{\sum_{k=1}^K \omega_s^k c_{s-1}^k} \,.
\end{aligned}$$

Reordering this inequality, we get

$$\inf_{\lambda \in \Lambda_m(\tilde{\mu}_{s-1})} \|\tilde{\mu}_{s-1} - \lambda\|_{D_{\omega_s}}^2 \geq \left(\sqrt{g_s(\omega_s)} - \sqrt{\sum_{k=1}^K \omega_s^k c_{s-1}^k} \right)^2 .$$

Then, summing over $s \in [t_0'(t)+1, t]$ and using $\|a-b\| \geq \|a\| - \|b\|$,

$$\begin{aligned}
\sqrt{\sum_{s=t_0'(t)+1}^{t} \inf_{\lambda \in \Lambda_m(\tilde{\mu}_{s-1})} \|\tilde{\mu}_{s-1} - \lambda\|_{D_{\omega_s}}^2} &\geq \sqrt{\sum_{s=t_0'(t)+1}^{t} \left(\sqrt{g_s(\omega_s)} - \sqrt{\sum_{k=1}^K \omega_s^k c_{s-1}^k} \right)^2} \\
&\geq \sqrt{\sum_{s=t_0'(t)+1}^{t} g_s(\omega_s)} - \sqrt{\sum_{s=t_0'(t)+1}^{t} \sum_{k=1}^K \omega_s^k c_{s-1}^k} \,.
\end{aligned}$$

□

Summary of Steps 4 to 7. Putting together Equations (15), (17) and (18), we proved that under event $\mathcal{E}_t$, for estimates $(\tilde{\mu}_s)_{s \leq t}$ such that we have a bound c_s^k on $(\tilde{\mu}_s^k - \mu^k)^2$ for all $s \in \{t_0'(t)+1, \dots, t\}$ and $k \in [K]$,

$$\sqrt{\inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|_{D_{W_t}}^2} + 2\sqrt{C_t} \geq \sqrt{\sum_{s=t_0'(t)+1}^{t} g_s(\omega_s)} \,.$$

Note that $\inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|_{D_{W_t}}^2 \leq t \max_{\omega \in \Delta_K} \inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|_{D_\omega}^2 = 2tH_\mu$. We then get

$$\inf_{\lambda \in \Lambda_m(\mu)} \|\mu - \lambda\|_{D_{W_t}}^2 \geq \sum_{s=t_0'(t)+1}^{t} g_s(\omega_s) - 4C_t - 4\sqrt{2tC_tH_\mu} \,,$$

which ends proving Theorem 5.

Step 8**. No-regret property.** The first t_0 rounds are used to initialize our algorithm. After that, we use a learner with small regret. We will bound the gains between t_0 and $t_0'(t) = \max\{t_0, \sqrt{t}\}$ by $36M^2$ (see Lemma 24). We use the regret bound of the learner (refer to Definition 2 in the main paper) to get that, for some additional low-order term $C_{\mathcal{L}}(K,B)\sqrt{t}$, and by combining Theorems 4 and 5:

$$\begin{aligned}
2\beta_{t,\delta} &\geq \sum_{s=t_0'(t)+1}^{t} g_s(\omega_s) - h_\delta(t) - r(t) - 4C_t - 4\sqrt{2tC_tH_\mu} \\
&\geq \sum_{s=t_0+1}^{t} g_s(\omega_s) - h_\delta(t) - r(t) - 4C_t - 4\sqrt{2tC_tH_\mu} - \max\{\sqrt{t}-t_0, 0\}36M^2 \\
&\geq \max_{\omega\in\Delta_K} \sum_{s=t_0+1}^{t} g_s(\omega) - h_\delta(t) - r(t) - 4C_t - 4\sqrt{2tC_tH_\mu} - C_{\mathcal{L}}(K,B)\sqrt{t} \\
&\quad - \max\{\sqrt{t}-t_0, 0\}36M^2 \,.
\end{aligned}$$

A specific upper bound on the regret for the learner AdaHedge used in the implementation of MISLID is mentioned in Lemma 27.

Step 9**. From the optimal gain to the lower bound value.** Finally, we can relate the optimal optimistic gain of the learner to the value of the lower bound. Using the optimism (Lemma 23),

$$\max_{\omega\in\Delta_K} \sum_{s=t_0+1}^{t} g_s(\omega) \geq \max_{\omega\in\Delta_K} \sum_{s=t_0+1}^{t} \inf_{\lambda\in\Lambda_m(\mu)} \|\mu-\lambda\|_{D_\omega}^2 = (t-t_0) \underbrace{\max_{\omega\in\Delta_K} \inf_{\lambda\in\Lambda_m(\mu)} \|\mu-\lambda\|_{D_\omega}^2}_{=2H_\mu} .$$

Step 10**. Computing the sample complexity.** We thus have obtained an inequality of the form

$$\begin{aligned}
2\beta_{t,\delta} \geq 2tH_\mu - h_\delta(t) - r(t) - 4C_t - 4\sqrt{2tC_tH_\mu} - C_{\mathcal{L}}(K,B)\sqrt{t} - 2t_0H_\mu \\
- \max\{\sqrt{t}-t_0, 0\}36M^2 \,,
\end{aligned}$$

from which we can obtain the desired sample complexity bound. Remember that

- $\beta_{t,\delta} := \min\left\{\beta_{t,\delta}^{\text{uns}}, \beta_{t,\delta}^{\text{lin}}\right\}$
- $r(t) := M^2\sqrt{2t\left(\log(4\times 5t^4) + d\log\frac{6(M+\varepsilon)Lt}{\sqrt{\Gamma(A)}} + K\log\max\{4\varepsilon t, 1\}\right)} + 2 + 8M$
- $h_\delta(t) := \sqrt{8\beta_{t,\delta}f(t)} + f(t)$
- $f(t) := 2\min\left\{\alpha_t^{\text{uns}}, \alpha_t^{\text{lin}} + 2t(LK+1)^2\varepsilon^2\right\}$

where:

$$\beta^{\mathrm{uns}}_{t,\delta} := 2K\overline{W}\left(\frac{1}{2K}\log\frac{2e}{\delta} + \frac{1}{2}\log(8eK\log t)\right) ,$$

$$\beta^{\mathrm{lin}}_{t,\delta} := \frac{1}{2}\left(4\sqrt{t}\varepsilon + \sqrt{2}\sqrt{1 + \log\frac{1}{\delta} + \left(1 + \frac{1}{\log(1/\delta)}\right)\frac{d}{2}\log\left(1 + \frac{t}{2d}\log\frac{1}{\delta}\right)}\right)^2 ,$$

$$\alpha^{\mathrm{uns}}_t := 2K\overline{W}\left(\frac{1}{2K}\log(14et^3) + \frac{1}{2}\log(8eK\log t)\right) = \beta^{\mathrm{uns}}_{t,1/5t^3} ,$$

$$\alpha^{\mathrm{lin}}_t := \log(5t^2) + d\log\left(1 + \frac{t}{2d}\right) ,$$

$$c^k_t := \min\left\{8(LK+1)^2\varepsilon^2 + 4\alpha^{\mathrm{lin}}_{t^2}\left\|\phi_k\right\|^2_{V_t^{-1}} , \frac{2\alpha^{\mathrm{uns}}_{t^2}}{N^k_t}, 4M^2\right\} ,$$

$$C_t := \sum_{s=t_0+1}^{t}\sum_{k=1}^{K} w^k_s c^k_{s-1} \le 8(LK+1)^2\varepsilon^2 t + 2\alpha^{\mathrm{lin}}_{t^2}\left(\sqrt{2t\log(5t^2)} + d\log\left(1 + \frac{t}{d}\right)\right) ,$$

$$C_t \le 4M^2\sqrt{2t\log(5t^2)} + 4M^2K + 2K\alpha^{\mathrm{uns}}_{t^2}\log(et) .$$

Combining this bound with Lemma 19 proves Theorem 2 in the main paper.

F.4 Using several estimates

If we employ two sets of estimates, with corresponding optimism functions $(g^i_s(\omega))_{i\in\{1,2\}}$ and bounds $c^k_{i,s}$, we get for $i \in \{1,2\}$,

$$\begin{aligned}\inf_{\lambda\in\Lambda_m(\mu)}\left\|\mu-\lambda\right\|^2_{D_{W_t}} &\ge \max_{i\in\{1,2\}}\left(\sum_{s=t_0+1}^{t} g^i_s(\omega_s) - 4C^i_t - 4\sqrt{2tC^i_tH_\mu}\right)\\ &\ge \sum_{s=t'_0(t)+1}^{t}\min_{i\in\{1,2\}} g^i_s(\omega_s) - \min_{i\in\{1,2\}}\left(4C^i_t + 4\sqrt{2tC^i_tH_\mu}\right) ,\end{aligned}$$

where the quantity C^i_t is similarly defined as C_t, with respect to gains g^i_t.

Since the minimum of concave functions is concave, $g_s : \omega \mapsto \min_{i\in\{1,2\}} g^i_s(\omega)$ is concave (which allows the use of a regret-minimizing algorithm, see Subsection F.6). It satisfies the inequality of Lemma 23 and its gradient is the gradient of $g^{i^\star}_s(\omega)$ for $i^\star \in \arg\min_{i\in\{1,2\}} g^i_s(\omega)$.

F.5 Aggressive Optimism

If we are happy with an algorithm which is within a factor 2 of the lower bound for the $\log\frac{1}{\delta}$ term instead of insisting on a factor 1, we can use a different, more aggressive optimism. Take

$$\widehat{g}_s(\omega_s) := 2\sum_{k=1}^{K}\omega^k\left((\tilde{\mu}^k_{s-1} - \lambda^k_s)^2 + c^k_{s-1}\right) \qquad \text{where } \lambda_s := \operatorname*{arg\,min}_{\lambda\in\Lambda_m(\tilde{\mu}_{s-1})}\left\|\tilde{\mu}_{s-1} - \lambda\right\|^2_{D_{\omega_s}} .$$

The main difference is that the term added to $(\tilde{\mu}^k_{s-1} - \lambda^k_s)^2$ is of order c^k_{s-1} instead of $\sqrt{c^k_{s-1}}$. In an unstructured bandit, that means $1/N^k_t$ instead of $1/\sqrt{N^k_t}$. Let us prove the counterpart to Lemma 23 for these new gains:

Lemma 26. *For all $\omega \in \Delta_K$, $\widehat{g}_s(\omega_s) \ge \inf_{\Lambda\in\lambda_m(\mu)}\left\|\mu-\lambda\right\|^2_{D_\omega}$.*

Proof. For all $k \in [K]$ and $\lambda \in \mathbb{R}^K$, using Lemma 20

$$(\mu^k - \lambda^k)^2 \le 2(\tilde{\mu}^k_{s-1} - \lambda^k)^2 + 2(\mu^k - \tilde{\mu}^k_{s-1})^2 \le 2(\tilde{\mu}^k_{s-1} - \lambda^k)^2 + 2c^k_{s-1} .$$

Then, since $\omega \in \Delta_K$

$$2\sum_{k=1}^{K}\omega^k\left((\tilde{\mu}_{s-1}^k-\lambda^k)^2+c_{s-1}^k\right)\geq\sum_{k=1}^{K}\omega^k(\mu^k-\lambda_s^k)^2=\|\mu-\lambda_s\|_{D_\omega}^2\geq\inf_{\Lambda\in\lambda_m(\mu)}\|\mu-\lambda\|_{D_\omega}^2\ .$$

□

Then, using Lemma 26 and the definition of λ_s, we have

$$\begin{aligned}\widehat{g}_s(\omega)-2\inf_{\lambda\in\Lambda_m(\tilde{\mu}_{s-1})}\|\tilde{\mu}_{s-1}-\lambda\|_{D_{\omega_s}}^2&=\sum_{k=1}^{K}\omega_s^k\left[2(\tilde{\mu}_{s-1}^k-\lambda_s^k)^2+2c_{s-1}^k-2(\tilde{\mu}_{s-1}^k-\lambda_s^k)^2\right]\\&=2\sum_{k=1}^{K}\omega_s^k c_{s-1}^k\ .\end{aligned}$$

So now we can prove a counterpart to Step 7 in the proof of Theorem 5:

$$\begin{aligned}&\sum_{s=t_0+1}^{t}\inf_{\lambda\in\Lambda_m(\tilde{\mu}_{s-1})}\|\tilde{\mu}_{s-1}-\lambda\|_{D_{\omega_s}}^2\\&=\frac{1}{2}\sum_{s=t_0+1}^{t}\widehat{g}_s(\omega_s)-\frac{1}{2}\sum_{s=t_0+1}^{t}\left(\widehat{g}_s(\omega_s)-2\inf_{\lambda\in\Lambda_m(\tilde{\mu}_{s-1})}\|\tilde{\mu}_{s-1}-\lambda\|_{D_{\omega_s}}^2\right)\\&\geq\frac{1}{2}\sum_{s=t_0+1}^{t}\widehat{g}_s(\omega_s)-C_t\ .\end{aligned}$$

F.6 Regret of AdaHedge

Lemma 27 ([10])**.** *On the online learning problem with K arms and gains $g_s(\omega):=\sum_{k\in[K]}\omega^k U_s^k$ for $s\in[t]$, AdaHedge, predicting $(\omega_s)_{s\in[t]}$, has regret*

$$R_t:=\max_{\omega\in\Delta_K}\sum_{s=1}^{t}g_s(\omega)-g_s(\omega_s)\leq 2\sigma\sqrt{t\log(K)}+16\sigma(2+\log(K)/3)\ ,$$

$$\text{where }\sigma:=\max_{s\leq t}\left(\max_{k\in[K]}U_s^k-\min_{k\in[K]}U_s^k\right)\ .$$

We recall here the "gradient trick", which we can use to employ AdaHedge on any concave gains.If for any time $t>0$, the loss function ℓ_t at that time is convex, then for all $\omega\in\Delta_K$,

$$\sum_{s=1}^{t}\ell_t(\omega_t)-\ell_t(\omega)\leq\sum_{s=1}^{t}(\omega_t-\omega)^\top\nabla\ell_t(\omega_t)$$

Running a regret-minimizing algorithm with loss $\bar{\ell}_t(\omega)=\omega^\top\nabla\ell_t(\omega_t)$ then leads to a regret bound on ℓ_t.

F.7 Technical tools

Generic bounds on vector norms.

Lemma 28. *Let $\theta\in\mathbb{R}^d,\eta\in\mathbb{R}^K$ be such that $\|\eta\|_\infty\leq\varepsilon$ and $\|A\theta+\eta\|_\infty\leq M$. Then*

$$\|\theta\|_2\leq\frac{M+\varepsilon}{\sqrt{\Gamma(A)}}\ ,$$

where $\Gamma(A):=\max_{\omega\in\Delta_K}\sigma_{\min}\left(\sum_{k=1}^{K}\omega^k\phi_k\phi_k^\top\right)$, where $\sigma_{\min}(M)$ is the minimal singular value of matrix M.

Proof. For $\lambda := A\theta + \eta$ with $\|\eta\|_\infty \leq \varepsilon$ and $\|\lambda\|_\infty \leq M$,

$$\|A\theta\|_\infty = \max_{k\in[K]} \left|\phi_k^\top \theta\right| \geq \|\theta\|_2 \min_{u\in\mathbb{R}^d:\|u\|_2=1} \max_{k\in[K]} \left|\phi_k^\top u\right| \,, \tag{19}$$

using that the value for $\theta/\|\theta\|_2$ is larger than the minimum over $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$. On the other hand, successively using the triangle inequality and the boundedness assumptions,

$$\|A\theta\|_\infty \leq \|A\theta + \eta\|_\infty + \|\eta\|_\infty \leq M + \varepsilon \,. \tag{20}$$

Note also that

$$\min_{u\in\mathbb{R}^d:\|u\|_2=1} \max_{k\in[K]} \left|\phi_k^\top u\right|^2 = \min_{u\in\mathbb{R}^d:\|u\|_2=1} \max_{\omega\in\Delta_K} \|u\|^2_{\left(\sum_{k=1}^K \omega^k \phi_k \phi_k^\top\right)} \geq \underbrace{\max_{\omega\in\Delta_K} \sigma_{\min}\left(\sum_{k=1}^K \omega^k \phi_k \phi_k^\top\right)}_{:=\,\Gamma(A)} \,, \tag{21}$$

where the inequality stems from the min-max theorem (principle for singular values). Finally, by combining the three inequalities (19), (20) and (21), $\|\theta\|_2 \leq \frac{M+\varepsilon}{\sqrt{\Gamma(A)}}$.

□

The term $\Gamma(A)$ depends only on the set of linear features $\{\phi_k\}_{k\in[K]}$. In the unstructured case (where $\phi_k = e_k$), we have $\Gamma(A) = \frac{1}{K}$. However, in a structured case with $d \ll K$, $\Gamma(A)$ can be much smaller. For instance, when A contains the canonical basis of $\mathbb{R}^d$, we have $\Gamma(A) \geq \frac{1}{d}$.

G Experimental evaluation

G.1 Computational architectures

Experiments on simulated datasets (Experiments (A), (B), (C)) were run on a personal computer (processor: Intel Core i7 − 8750H, cores: 12, frequency: 2.20GHz, RAM: 16GB).

Experiment (D) was run on a personal computer (processor: Intel Core i7 − 9700K, cores: 8, frequency: 3.60GHz, RAM: 16GB).

Experiment (E) was run on a internal cluster (processor: Westmere E56xx/L56xx/X56xx (Nehalem−C), cores: 24, frequency: 3.2GHz, RAM: 155GB).

G.2 License for the assets

Experiment (D). The drug repurposing dataset for epilepsy was proposed in [35], and made publicly available under the MIT license.

Experiment (E). The original dataset Last.fm is publicly available online at `https://www.last.fm/` under CC BY-SA 4.0.

Experimental code. The code hosted at `https://github.com/clreda/misspecified-top-m` is under MIT license.

G.3 Extracting representations from real datasets

We describe in detail the procedure we adopted to extract misspecified linear representations from the real-world datasets of Experiment (D) and (E). In both cases, we adopted a very similar procedure based on training neural networks as the one used in [34]. We describe all its steps for the sake of completeness.

Step 1. (Data preprocessing) First, we start from preprocessing the raw data to obtain a dataset containing tuples of the form (ϕ, x), where $\phi \in \mathbb{R}^d$ is an arm feature and $x \in \mathbb{R}$ is a reward. The drug repurposing dataset used in [35] (hosted on their repository) is already available in this form, with a total of 509 arms representing different drugs, $d = 67$ features representing genes, and, for each of

them, 18 reward samples representating the responses of 18 different patients to such drugs. Out of those 509 arms, we filter out those which outcomes are unknown (associated "true" scores are set to 0, according to the file of scores available on the same repository). Then 175 arms (representing either antiepileptics, with score equal to 1, and proconvulsants, with score equal to -1) are left.

On the other hand, the Last.fm dataset is in a different form; it contains information about the music artists listened by each user of the system. As done in [34, Appendix F.4], we first preprocessed the data by keeping only artists listened by at least 120 users and users that listened at least to 10 different artists. We thus obtained $U = 1,322$ users and $A = 103$ artists. The result is a matrix in $\mathbb{R}^{U \times A}$ containing the number of times each user listened to each artist (which we treat as reward). We then extract user-artist features by applying low-rank Singular Value Decomposition on this matrix and keeping only the top 80 singular values. This yields U d-dimensional user features, and A d-dimensional artist features, where $d = 80$. The final user-artist features are the concatenation of the two, which yields a dataset with $U \times A$ tuples $(\phi, x) \in \mathbb{R}^d \times \mathbb{N}$ in our desired form.

Step 2. (Neural-network training) For both datasets, the second step consists in training a neural network to regress from ϕ to x. First, we split the datasets randomly into 80% training set and 20% test set. Then, we train a neural network with two hidden layers of size 256, rectified linear unit activations, and a linear output layer of 8 neurons. We obtain an R^2 score on the test set of 0.92 for the drug repurposing data, and 0.85 for the Last.fm one.

Step 3. (Extracting a linear representation) Finally, we extract a linear model from the trained neural network by taking, for each input $\phi \in \mathbb{R}^d$ in our data, the 8-dimensional features (i.e., activations) computed in the last layer together with the corresponding parameters. When specified, a subset of arm features is considered instead of the whole dataset. Then, in that case, we apply a lossless dimensionality reduction to make sure these features span the whole space. The reduced features are the one we feed into our learning algorithms (Experiment (D): $d = 5$, $K = 10$; Experiment (E.i): $d = 8$, $K = 103$; Experiment (E.ii): $d = 7$, $K = 50$). Moreover, we compute the maximum absolute error of this linear model in predicting the original data, and use that as a proxy for ε.

Note that, since the Last.fm data is in the form of user-artist features and, in our problem, we consider the artists only as arms, the representation we select for our experiments is obtained by choosing a user randomly among the available $U = 1,322$ ones.

Moreover, in Step 3, we apply a dimension reduction procedure on features to ensure the feature matrix is not ill-conditioned, at the cost of increasing the norm of misspecification ε. This is needed in order to reduce the length t_0 of the initialization sequence ; remember that in Appendix D.1 we showed that t_0 is upper-bounded by quantity $d \left\lceil \frac{2L^2}{\Gamma'(A)} \right\rceil$, where $\Gamma'(A) := \min \left\{ \sigma_{\min} \left(\sum_{k \in \mathcal{B}} \phi_k \phi_k^\top \right) \mid \mathcal{B} \text{ barycentric spanner of } A \text{ of size } d \right\}$ crucially relies on the conditioning of A. How much misspecification is required to improve the conditioning of the matrix is an open question (which has also been raised in other recent works [34]). Ideally, one would want to learn a representation of the data which balances those two effects, but we leave such a method to future investigations.

G.4 Numerical results for sample complexity

Table 2: Statistics (mean $\pm$ standard deviation rounded up to the next integer) for Experiment (A). Names are similar to those in the first two leftmost plots of Figure 1. Values are averaged across 500 iterations. LinGapE is not δ-correct in the setting where $\varepsilon = 5$ (with $\delta = 0.05$).

Sample complexity	LinGapE	MISLID
$\varepsilon = 0$	577 ± 348	890 ± 546
$\varepsilon = 5$	553 ± 536	$5,156 \pm 3,629$

Table 3: Statistics (mean $\pm$ standard deviation rounded up to the next integer) for Experiments (D) and (E). Names are similar to those in the plots of Figure 2. Values are averaged across 100 iterations. Note that LinGapE is not δ-correct (with $\delta = 0.05$) in Experiment E.

Sample complexity	LinGapE	MISLID
Experiment D	$21,593 \pm 8,296$	$42,751 \pm 13,942$
Experiment E	$10,907 \pm 4,474$	$289,703 \pm 185,205$

G.5 Tricks to reduce sample and computational complexity on large instances (D) and (E)

In large instances (more particularly on our real-life datasets in Section 5 in the main paper), the number of arms can be large, and the theoretically supported version of the algorithm MISLID might become too slow. Based on our experiments, we have decided to change some parts of the algorithm.

No optimism. As shown in the rightmost plot in Figure 1 in the main paper, empirical gains (i.e., without any optimistic bonus) actually considerably improve sample complexity.

Restriction of the set of arms used in the sampling rule. In order to compute the gains which are fed to the learner, MISLID needs to compute the closest alternative, which implies solving $m(K-m)$ quadratic optimization problems, one for each pair of arms (i, j), with i among the m best arms and j among the $K - m$ worse arms (as defined in Theorem 1 in the main paper). We observed that the majority of arms never realize the minimum over (i, j) of the distance to the alternative, and in hindsight they could be ignored. We mimicked that behavior by only considering a subset of arms at each step. We kept $m + d$ arms in memory, consisting of the recent argmins i, j for the closest alternative model, and sampled d more among the $K - m$ worse arms. The resulting set of at most $m + 2d$ arms is then used to compute the closest alternative. The gain in computational complexity is large when $K \gg d$, since we solve $m(m + 2d)$ minimization problems instead of $m(K - m)$. We don't use that trick to compute the stopping rule, since we would not be guaranteed to preserve δ-correctness.

Geometric grid for testing the stopping rule. Instead of checking the stopping criterion at each learning round of the algorithm, we suggest testing it on a geometric grid (that is, testing it for the first time at t_1, and then retest it at γt_1, then at $\gamma^2 t_1$, etc. where $1 < \gamma \leq 1.3$ in practice), and restrict the computation of the stopping rule to a random subset of arms. In our experiments, we have actually used $\gamma = 1.2$. When using the geometric grid, we can obtain a sample complexity bound of the same form as in Theorem 2 in the main paper, except that $T_0(\delta)$ is replaced by $\gamma T_0(\delta)$.

Together, the sampling and stopping rule changes reduce the time needed to complete a run of the algorithm by a factor 29 on Experiment (D), while increasing the sample complexity by a factor 1.2 (refer to Table 4, comparing algorithmic versions named "AdaHedge" and "Default"). See the middle plot of Figure 3 for a comparison of the sample complexity.

Table 4: Statistics (mean $\pm$ standard deviation rounded up to the next integer) for Experiment (D), with different versions of MISLID. Names are similar to those in the center plot of Figure 3. Values are averaged across 100 iterations.

Per run	AdaHedge	Greedy	Default
Average runtime (in sec.)	69 ± 20	76 ± 178	$1,993 \pm 1,311$
Average sample complexity	$51,965 \pm 15,260$	$52,108 \pm 125,230$	$42,751 \pm 13,943$

We have also tested another learner which is less conservative than AdaHedge, to check if this improves sample complexity (note that we did not show any experiment using this trick in the main paper):

Change of learner. We replace AdaHedge by a Greedy/Follow-The-Leader learner combination for the computation of (ω_t, λ_t).

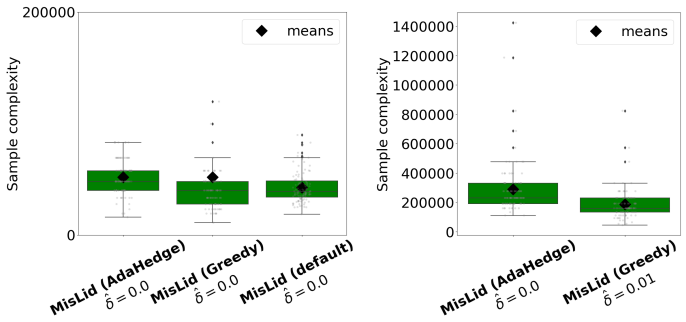

Figure 3: Comparison between default MISLID, modified MISLID using learner AdaHedge, and modified MISLID using learner Greedy (Experiment (D) *(left)*, Experiment (E)). Unfortunately, one outlier in the runs using learner Greedy in Experiment (D), above $1,200,000$ rounds, would prevent the readability of the plot if figured. To overcome this issue, we have cropped out the y-axis above $200,000$ in this plot.

We have run three versions of MISLID on the dataset of Experiment (D): the default MISLID, the modified version with learner AdaHedge, and another modified version with the Greedy learner. We have also launched the latter two on Experiment (E). See Figure 3.