Optimal Exploration is no harder than Thompson Sampling

Zhaoqi Li
University of Washington

Kevin Jamieson
University of Washington

Lalit Jain
University of Washington

Abstract

Given a set of arms $\mathcal{Z} \subset \mathbb{R}^d$ and an unknown parameter vector $\theta_* \in \mathbb{R}^d$, the pure exploration linear bandit problem aims to return $\arg\max_{z\in\mathcal{Z}} z^\top\theta_*$, with high probability through noisy measurements of $x^\top\theta_*$ with $x \in \mathcal{X} \subset \mathbb{R}^d$. Existing (asymptotically) optimal methods require either a) potentially costly projections for each arm $z \in \mathcal{Z}$ or b) explicitly maintaining a subset of $\mathcal{Z}$ under consideration at each time. This complexity is at odds with the popular and simple Thompson Sampling algorithm for regret minimization, which just requires access to a posterior sampling and argmax oracle, and does not need to enumerate $\mathcal{Z}$ at any point. Unfortunately, Thompson sampling is known to be sub-optimal for pure exploration. In this work, we pose a natural question: is there an algorithm that can explore optimally and only needs the same computational primitives as Thompson Sampling? We answer the question in the affirmative. We provide an algorithm that leverages only sampling and argmax oracles and achieves an exponential convergence rate, with the exponent equal to the exponent of the optimal fixed allocation asymptotically. In addition, we show that our algorithm can be easily implemented and performs as well empirically as existing asymptotically optimal methods.

1 INTRODUCTION

The pure exploration bandit problem considers a sequential game between a learner with two sets of arms $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$ and nature. In each round, the learner chooses an arm $x \in \mathcal{X}$ and observes a noisy stochastic reward $y = x^\top\theta_* + \epsilon$ where $\theta_* \in \Theta$ is an unknown parameter vector and ϵ is assumed to be i.i.d Gaussian noise. The goal of the learner is to identify $z_* = \arg\max_{z\in\mathcal{Z}} z^\top\theta_*$ with high probability in a few measurements. The case of $\mathcal{X} = \mathcal{Z}$ is perhaps the most natural case to consider, and has enjoyed a fair amount of attention (Soare et al., 2014; Fiez et al., 2019; Degenne et al., 2020). However, all proposed approaches share a common trait - complexity. Existing optimal algorithms rely on either explicitly enumerating a potentially large subset of $\mathcal{Z}$ or periodically solving a convex optimization program at every iteration. Consequently, it prompts us to question: is such complexity indeed indispensable for reaching asymptotic optimality?

Maintaining our focus on the specific instance where $\mathcal{X} = \mathcal{Z}$, we note that the pure exploration task can be addressed using any readily available regret minimization algorithm. That is, if an algorithm generates a series of plays $\{x_t\}_{t=1}^T$ such that $\max_{x\in\mathcal{X}} \sum_{t=1}^T \langle\theta_*, x - x_t\rangle \le d\sqrt{T}$ then this immediately implies that $\widehat{x}_T$ drawn uniformly from the set $\{x_t\}_{t=1}^T$ is equal to $x_* = \arg\max_{x\in\mathcal{X}}\langle x, \theta_*\rangle$ with constant probability as soon as $T \ge d^2/\Delta_{\min}^2$, where $\Delta_{\min} = \min_{x\in\mathcal{X}, x\neq x_*} \theta_*^\top(x^* - x)$. One popular regret-minimization algorithm is Thompson Sampling (TS). Following its re-emergence from nearly seven decades of relative obscurity, it has rapidly ascended to become the most prevalently applied bandit algorithm in practical scenarios, as per the industrial experience of the authors. We postulate that its popularity is due to (1) its simplicity to implement, (2) its flexibility to encode side-information in its prior, (3) its computational efficiency, and (4) strong empirical performance. The algorithm works by maintaining a distribution p_t over Θ given all observations up to the time t, and then plays $x_t = \arg\max_{x\in\mathcal{X}}\langle x, \theta_t\rangle$ where $\theta_t \sim p_t$. Once $y_t = \langle x_t, \theta_*\rangle + \epsilon_t$ is observed, the distribution is updated and the process repeats. As we can see, TS only relies on the ability to sample from a posterior distribution and compute a maximum inner product (an argmax oracle) - both operations which have been heavily studied and optimized. Unfortunately, TS is known to be sub-optimal for the pure exploration linear bandits problem due to its greedy exploration strategy.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

Indeed, there exist instances of $\mathcal{X}$ and θ_* for which the sample complexity of TS to identify the best arm scales *quadratically* in the optimal sample complexity achieved by other algorithms (Soare et al., 2014). Even for regret minimization, it is know that TS is far from optimal from an instance-dependent perspective (Lattimore and Szepesvari, 2017). But yet, due to its many favorable properties it is still the go-to algorithm in practice.

This paper aims to answer the following fundamental theoretical question: *Is there an algorithm that enjoys asymptotically optimal exploration that does not need to explicitly enumerate $\mathcal{Z}$ and only relies on posterior sampling and an argmax oracle?* We achieve this goal by not striving too far from the Thompson sampling algorithm itself and only assuming access to a sampling oracle and arg-max oracle. In fact, our proposed algorithm can be viewed as a generalization of Top-Two Thompson Sampling for the standard multi-armed bandit game (Russo, 2016) to the richer linear setting. At each iteration t, we maintain a sampling distribution centered at $\widehat{\theta}_t$ (a least squares estimator computed after t samples), and get a sample θ_t whose best arm is different than that of $\widehat{\theta}_t$ using a sampling oracle. Once such a θ_t is found, we update an online learner maintaining a distribution over $\mathcal{X}$ with rewards $\|\theta_t - \widehat{\theta}_t\|^2_{xx^\top}$. We prove that $\mathbb{P}(\widehat{z}_t \neq z_* | \{x_s\}_{s=1}^{t-1})$ decreases at an exponential rate with the exponent of the optimal fixed allocation. We also demonstrate that our method is not only theoretically sound by achieving an optimal sample complexity given oracle access, but is also computationally efficient empirically.

1.1 Problem Setting and Notation

We first define the linear bandit setting. Let $\mathcal{X}, \mathcal{Z} \in \mathbb{R}^d$ be two sets of arms and $\Theta \subset \mathbb{R}^d$ be the parameter space. At time t, we draw an action $x_t \in \mathcal{X}$, and receive the reward $y_t = x_t^\top \theta_* + \epsilon_t$ where $\theta_* \in \Theta$ and ϵ_t is i.i.d. Gaussian noise. The choice of arm x_t at time t is dependent on the filtration generated by $\{(x_s, y_s)\}_{s=1}^{t-1}$; furthermore, we denote the conditional probability given this filtration be $\mathbb{P}_\theta$.

Goal: We are interested in the best-arm identification task, i.e. we would like to find $z_* := \arg\max_{z \in \mathcal{Z}} z^\top \theta_*$ with high probability, while minimizing the number of measurements taken in $\mathcal{X}$.

We make the following assumption on the parameters that we will discuss further in Section 3.1.

Assumption 1. Θ is closed and bounded, with a non-empty interior.

Assumption 2. Assume that $\max_x \|x\|_2 \leq L$.

Assumption 3. Assume that $\text{span}(\mathcal{Z}) \subset \text{span}(\mathcal{X})$ and the optimal arm $z_* \in \mathcal{Z}$ is unique.

Notation. For any matrix $A \in \mathbb{R}^{d \times d}$, we define the norm $\|x\|_A^2 := x^\top A x$. Given a set $\mathcal{S}$, we define the simplex $\triangle_{\mathcal{S}} := \{\lambda \in \mathbb{R}^{|\mathcal{S}|}_{\geq 0} : \sum_{i=1}^{|\mathcal{S}|} \lambda_i = 1\}$. Finally, given a (multivariate) normal distribution $\mathcal{N}(\theta, \Sigma^{-1})$ on $\mathbb{R}^d$ and some set Θ, we define the truncated normal distribution, denoted as $\text{TN}(\theta, \Sigma^{-1}; \Theta)$, to be the normal distribution restricted on Θ. For some $\lambda \in \triangle_{\mathcal{X}}$, we define $A(\lambda) := \sum_{x \in \mathcal{X}} \lambda_x x x^\top$. We define $\Delta_{\max} := \max_{x \in \mathcal{X}} \max_{\theta, \theta' \in \Theta} |x^\top(\theta - \theta')|$. We define the constants used in the algorithm as $C_{3,\ell} = \Delta_{\max} + L^2 \sqrt{d \log(T_\ell \ell^2)}$. The precise definition is in Appendix A.

2 MOTIVATING OUR APPROACH

Among all adaptive algorithms, it is known that for every $\theta_* \in \Theta$ there exists a $\lambda \in \triangle_{\mathcal{X}}$ such that sampling $x_1, x_2, \ldots, \overset{i.i.d.}{\sim} \lambda$ achieves the optimal sample complexity in the fixed confidence setting (Soare et al., 2014; Fiez et al., 2019; Degenne et al., 2020). Specifically, for any $\Theta \subset \mathbb{R}^d$ and $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$ define

$$\tau^* := \max_{\lambda \in \triangle_{\mathcal{X}}} \min_{\theta \in \Theta^c_{z_*}} \tfrac{1}{2} \|\theta - \theta_*\|^2_{A(\lambda)^{-1}} \tag{1}$$

where $\Theta^c_{z_*} = \{\theta \in \Theta : \exists z \in \mathcal{Z}, z^\top \theta \geq z_*^\top \theta\}$. Then it is known that to identify z_* with probability at least $1 - \delta$, the expected sample complexity of any algorithm scales as $(\tau^*)^{-1} \log(2.4/\delta)$. Moreover, sampling according to the λ that achieves the maximum, when paired with an appropriate stopping time, achieves the optimal sample complexity asymptotically. As our setting is more naturally analyzed in the so-called fixed budget setting, we next state a result that can be viewed as a generalization of the result of Russo (2016) originally stated for the multi-armed bandit setting. Note that this is a lower bound similar to Glynn and Juneja (2004) and not a lower bound for the traditional fixed budget setting in multi-armed bandits (Karnin et al., 2013), since we only allow fixed λ not adapting to the observations.

Theorem 2.1. *Fix $\Theta = \mathbb{R}^d$ and any $\theta_* \in \Theta$. For some λ consider a procedure that draws $x_1, \ldots, x_T \sim \lambda$, then observes $y_t = \langle x_t, \theta_* \rangle + \epsilon_t$ for each t with $\epsilon_t \sim \mathcal{N}(0,1)$, and then computes $\widehat{z}_T = \arg\max_{z \in \mathcal{Z}} \langle z, \widehat{\theta}_T \rangle$ where $\widehat{\theta}_T = \arg\min_{\theta \in \Theta} \sum_{t=1}^T \|y_t - \langle \theta, x_t \rangle\|_2^2$. Then for any $\lambda \in \triangle_{\mathcal{X}}$ we have*

$$\limsup_{T \to \infty} -\frac{1}{T} \log\left(\mathbb{P}_{\theta_*, x_t \sim \lambda}(\widehat{z}_T \neq z_*)\right) \leq \tau^*.$$

The quantity τ^* is naturally interpreted from a hypothesis-testing lens. Given a fixed sampling distribution λ, note that $\mathbb{E}_{x \sim \lambda} KL(\mathcal{N}(\theta^\top x, 1) || \mathcal{N}(\theta_*^\top x, 1)) =$

$\frac{1}{2}\|\theta-\theta_*\|^2_{A(\lambda)}$. Thus the min-max problem above aims to construct the distribution λ which maximizes the smallest KL divergence between θ and any alternative with a different best-arm. As noticed by many authors, this can be translated into a game-theoretic language. The max-player chooses a distribution over the set of possible measurements $\mathcal{X}$. At the same time, the min-player chooses an alternative θ whose best arm is not z_* in an attempt to fool the λ-player. This lower bound intuitively suggests a strategy for algorithm designers: devise a sampling method that ensures the resultant allocation aligns with the aforementioned objective.

In this pursuit (discussed extensively in Section 4) the game-theoretic perspective has been directly exploited by several works to give asymptotically optimal algorithms. The approaches of these works differ in detail but are similar in spirit and are motivated by the following oracle strategy that has access to θ_*. At each time, the max-player utilizes a no-regret online learner, such as exponential weights (Bubeck, 2011), to set λ_{t+1} based on an estimate of the best-response of the min-player, namely $\min_{\theta\in\Theta^c_{z_*}}\|\theta-\theta_*\|^2_{A(\lambda_t)}$. This guarantees that

$$\max_{\lambda\in\triangle_{\mathcal{X}}}\min_{\theta\in\Theta^c_{z_*}}\|\theta-\theta_*\|^2_{A(\lambda)}-\sum_{t=1}^{T}\min_{\theta\in\Theta^c_{z_*}}\|\theta-\theta_*\|^2_{A(\lambda_t)}\le o(T)$$

which by a standard Jensen's inequality argument is sufficient to ensure that $\frac{1}{T}\sum_{t=1}^T\lambda_t$ is an approximate solution to the original saddle point problem. Then, the arm x_t pulled is sampled from λ_t at each time (or a deterministic tracking strategy is used).

The main computational challenge in this approach is that obtaining the best-response can be rather involved. The alternative set can be decomposed as a union of intersections of a convex set with a halfspace: $\Theta^c_{z_*}=\cup_{z\neq z_*}\Theta\cap\{\theta\in\mathbb{R}^d: z^\top\theta\ge z_*^\top\theta\}$. Thus computing the best-response involves computing $|\mathcal{Z}|$-many projections onto convex sets. For small values of $|\mathcal{Z}|$, this may be feasible. However, this computation may be onerous if $|\mathcal{Z}|$ is large or the projection step is very expensive, for example, in many combinatorial bandit settings such as shortest path problems in a graph (Chen et al., 2017). As another example, in practical recommendation systems where $\mathcal{Z}$ represents items to be recommended, $|\mathcal{Z}|$ may be in the millions. Thus computing $|\mathcal{Z}|$ many projections under latency constraints may be impossible, even though Thompson Sampling can easily recommend good items (Biswas et al., 2019). In addition, for both settings, there may be no easy closed-form expression for the projection.

Our method is based on the following equivalent formulation of τ^*. By linearizing the min over alternatives with a distribution over $\Theta^c_{z_*}$, we can apply Sion's minimax theorem:

$$\begin{aligned}&\max_{\lambda\in\triangle_{\mathcal{X}}}\inf_{\theta\in\Theta^c_{z_*}}\frac{1}{2}\|\theta-\theta_*\|^2_{A(\lambda)}\\&=\max_{\lambda\in\triangle_{\mathcal{X}}}\min_{p\in\triangle(\Theta^c_{z_*})}\mathbb{E}_{\theta\sim p}\left[\frac{1}{2}\|\theta-\theta_*\|^2_{A(\lambda)}\right]\\&=\min_{p\in\triangle(\Theta^c_{z_*})}\max_{\lambda\in\triangle_{\mathcal{X}}}\mathbb{E}_{\theta\sim p}\left[\frac{1}{2}\|\theta-\theta_*\|^2_{A(\lambda)}\right],\end{aligned}$$

where $\triangle(\Theta^c_{z_*})$ denotes the set of distribution over the alternative set $\Theta^c_{z_*}$. This replaces the projections with an expectation over a distribution on $\Theta^c_{z_*}$. At first glance, the situation may seem worse - we have gone from finitely many projections to needing to maintain a distribution over a potentially infinite set!

However, imagine that Θ is finite and that we solve this saddle-point problem by maintaining a no-regret learner for the max-player as before, while similarly maintaining a no-regret learner for the min-player. Standard results in convex optimization guarantee that the average of the iterates of the two learners converge to a saddle point eventually (Liu and Orabona, 2022). To be more precise, at each round t we draw an $x_t\sim\lambda_t$ and feed the (stochastic) loss $\sum_{\theta\in\Theta^c_{z_*}}p_{t,\theta}\|\theta-\theta_*\|^2_{x_tx_t^\top}$ to the learner for the min-player. Assuming the min-player learner is exponential weights, then the update is

$$p_{t+1,\theta}\propto p_{t,\theta}e^{-\eta\|\theta_*-\theta\|^2_{x_tx_t^\top}}\propto e^{-\eta\|\theta_*-\theta\|^2_{\sum_{s=1}^t x_sx_s^\top}}.$$

where η is an appropriate step-size. Hence, the resulting distribution p_{t+1} is reminiscent of the probability density function of a multivariate normal distribution $N(\theta_*,\eta^{-1}(\sum_{s=1}^t x_sx_s^\top)^{-1})$ restricted to $\Theta^c_{z_*}$. This observation motivates our algorithm - for the min-player we maintain an appropriate normal distribution and at each round, use samples from this distribution to generate a stochastic loss to feed the max-player. *This approach avoids explicitly maintaining $\mathcal{Z}$ or ever needing to compute a projection!* Of course, this discussion has relied on knowledge of θ_* and z_*. In the next section, we explain how our algorithm, PEPS, overcomes these restrictions.

3 BEST ARM IDENTIFICATION THROUGH SAMPLING

Our main method PEPS is presented in Algorithm 1. Given a budget of T samples, we repeatedly sample θ_t utilizing a sampling oracle SAMPLE. We then sample an $x_t\sim\tilde{\lambda}_t$ where $\tilde{\lambda}_t$ is the distribution λ_t maintained by the λ-learner at time t mixed in with a diminishing amount γ_t of the G-optimal distribution λ^G. After playing x_t and observing a reward y_t, PEPS updates both the λ_t and the estimate $\widehat{\theta}_t$ with the covariance.

Algorithm 1 Pure Exploration with Projection-Free Sampling (PEPS)

Input: Finite set of arms $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Z} \subset \mathbb{R}^d$, time horizon T, $\eta_\lambda, \eta_p, \alpha$

1: Define $\lambda^G = \arg\min_{\lambda \in \triangle_\mathcal{X}} \max_{x \in \mathcal{X}} \|x\|^2_{A(\lambda)^{-1}}$, $\lambda_1 = \frac{1}{|\mathcal{X}|}\mathbf{1}$
2: Initialize $V_0 = I$, $S_0 = 0$, $p_1 = N(0, V_0)$, $\widehat{\theta}_1$ arbitrarily
3: **for** $t = 1, 2, \cdots, T$ **do**
4: $\quad \gamma_t = t^{-\alpha}$
5: $\quad$ //Top Two Sampling
6: $\quad$ Compute $\widehat{z}_t = \operatorname{argmax}_{z \in \mathcal{Z}} z^\top \widehat{\theta}_t$
7: $\quad$ Sample $\theta_t = \mathsf{SAMPLE}(\mathrm{TN}(\widehat{\theta}_t, \eta_p^{-1} V_{t-1}^{-1}; \Theta^c_{\widehat{z}_t}))$
8:
9: $\quad$ //Take Sample and Observe Reward
10: $\quad$ Sample $x_t \sim \tilde{\lambda}_t$ where $\tilde{\lambda}_t = (1-\gamma_t)\lambda_t + \gamma_t \lambda^G$
11: $\quad$ Observe $y_t = \langle \theta_*, x_t \rangle + \epsilon_t$ where $\epsilon_t \sim \mathcal{N}(0,1)$
12:
13: $\quad$ //Update
14: $\quad$ Update $V_t = V_{t-1} + x_t x_t^\top$, $S_t = S_{t-1} + x_t y_t$, and $\widehat{\theta}_{t+1} = V_t^{-1} S_t$
15: $\quad$ Update $\lambda_{t+1} \propto \lambda_t e^{\eta_\lambda \tilde{g}_t}$ where $\tilde{g}_{t,x} = \left\| \theta_t - \widehat{\theta}_t \right\|^2_{xx^\top}, \forall x \in \mathcal{X}$
16: **end for**
17: Sample $\tilde{\theta} = \mathsf{SAMPLE}(\mathrm{TN}(\widehat{\theta}_{T+1}, V_T^{-1}; \Theta))$

Output: $\widehat{z}_\ell(\tilde{\theta}) = \arg\max_{z \in \mathcal{Z}} z^\top \tilde{\theta}$

In particular, given samples $\{x_s\}_{s=1}^t$, we let $\widehat{\theta}_{t+1} = V_t^{-1} S_t$ where $V_t = \sum_{s=1}^t x_s x_s^\top$ and $S_t = \sum_{s=1}^t x_s y_s$. Algorithm 1 depends on a finite time horizon T. To ensure that our algorithm is anytime and eventually converges to the optimal sampling scheme, we employ an outer loop Algorithm 2 utilizing a doubling scheme. Before we explain the theoretical guarantees, we first detail some of the aspects of the algorithm.

Updating the sampling distribution for θ_t. Our main innovation is introducing a distribution over $\Theta^c_{\widehat{z}_t}$ from which we can sample over. In particular, in each round, we sample θ_t from $\mathrm{TN}(\widehat{\theta}_t, \eta_p^{-1} V_{t-1}^{-1}; \Theta^c_{\widehat{z}_t})$, which is a *truncated normal distribution* with support $\Theta^c_{\widehat{z}_t}$ (Burkardt, 2014).

Algorithm 2 Doubling trick

Input: Finite set of arms $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Z} \subset \mathbb{R}^d$

1: **for** $\ell = 0, 1, \cdots, L$ **do**
2: $\quad$ Set $T_\ell = 2^\ell$, $\eta_\lambda = \sqrt{\frac{\log|\mathcal{X}|}{C_{3,\ell}^2 T_\ell}}$, $\eta_p = \sqrt{\frac{d \log(T_\ell C_{3,\ell})}{C_{3,\ell}^2 T_\ell}}$, $\alpha = 1/4$
3: $\quad \widehat{z}_\ell = \mathrm{PEPS}(\mathcal{X}, \mathcal{Z}, T_\ell, \eta_\lambda, \eta_p, \alpha)$
4: **end for**

Output: $\widehat{z}_L$

Following the discussion in the Section 2, it is tempting to see this update as a form of continuous exponential weights (Bubeck, 2011). However, this is not quite true since the underlying action set $\Theta^c_{\widehat{z}_t}$ is changing each round. This creates several technical challenges in the proof. Note that similar to previous works, we could have maintained a learner for each $z \in \mathcal{Z}$ (Degenne et al., 2020). However, our approach of maintaining a distribution prevents the need for this additional complexity of enumerating $\mathcal{Z}$.

From the perspective of exponential weights, η_p is a step size: the dependence on d in the numerator comes from the dimension of Θ; and $C_{3,\ell}^2$ is an upper bound on the stochastic loss $\|\theta_t - \widehat{\theta}_t\|^2_{x_t x_t^\top}$ that we guarantee with high probability due to forced exploration and boundedness of Θ.

We have the following regret guarantee on the online min learner. For notational convenience, in this section, for some set $\mathcal{S}$ with nonempty interior, we let $p_t(\mathcal{S}) = \mathrm{TN}(\widehat{\theta}_t, \eta_p^{-1} V_{t-1}^{-1}; \mathcal{S})$ be the truncated normal distribution with support on $\mathcal{S}$.

Lemma 3.1 (informal)**.** *In round T_ℓ of epoch ℓ of Algorithm 2, we have with probability greater than $1 - 1/\ell^2$,*

$$\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta \sim p_t(\Theta^c_{z_*})} \left[\left\| \theta - \widehat{\theta}_t \right\|^2_{x_t x_t^\top} \right] - \inf_{\theta \in \Theta^c_{z_*}} \|\theta - \theta_*\|^2_{V_{T_\ell}} \leq O(d\sqrt{T_\ell} \log(L T_\ell)).$$

Sampling Oracle. Our algorithm involves a sampling oracle that takes samples from a truncated normal distribution.

Definition 3.2 (Sampling oracle (SAMPLE))**.** The oracle $\mathsf{SAMPLE}(p)$ is an algorithm that given some distribution p, returns a sample $\theta \sim p$.

There are various ways to implement this sampling oracle efficiently. The easiest way is to use rejection sampling. In particular, on line 7, for each round t, we repeatedly sample $\theta_t \sim N(\widehat{\theta}_t, \eta_p^{-1} V_{t-1}^{-1})$ until the best-arm of $\arg\max_{z \in \mathcal{Z}} z^\top \theta_t$ is not our current best guess $\widehat{z}_t = \arg\max_{z \in \mathcal{Z}} z^\top \widehat{\theta}_t$, and on line 17 we repeatedly sample $\tilde{\theta} \sim N(\widehat{\theta}_{T+1}, V_T^{-1})$ until $\tilde{\theta} \in \Theta$. Regarding the computation cost of rejection sampling, we suffer from some of the same challenges as Top-two sampling algorithms, which empirically work well in practice (Russo, 2016). From a practical perspective, the rejection sampling step is only computationally costly if it requires many draws from the posterior to find a θ in the alternative $\Theta^c_{\widehat{z}_t}$. However, note that if we draw $O(1/\nu)$

vectors and none of them are in the alternative $\Theta^c_{\widehat{z}_t}$, by Markov's inequality, this arm they all agree on is the best arm with probability $1-\nu$. Thus, as soon as it becomes computationally costly to sample an alternative, the problem is basically solved. We demonstrate empirically that the computational complexity is not at all onerous in Section 5 and Appendix F. Also, we note that our focus is on the query complexity given an effective way to sample, not the complexity of sampling from the distribution itself. Since the sampling oracle only returns one sample at the end, our algorithm still achieves an asymptotically optimal *sample complexity* even if we draw $O(1/\nu)$ vectors inside the oracle.

Moreover, we remark that sampling from truncated normal distributions is a well-explored practice across statistics and machine learning, especially when sampling in a convex set. A variety of efficient methods such as Gibbs and hit-and-run procedures are available for this purpose (Devroye, 1986; Murphy, 2013; Li and Ghosh, 2015; Laddha and Vempala, 2023). In particular, the hit-and-run algorithm ensures one gets a sample in the convex set with probability $1-\nu$ in $O(d^3\log(1/\nu))$ samples in the worst case (Lovász, 1999). Furthermore, novel approaches have improved the efficiency of traditional rejection techniques, especially when dealing with a convex support of the truncated normal distribution (Maatouk and Bay, 2016).

Update for λ_t. To update λ_t, which corresponds to the action of our max-player, we employ an exponential weighted learner (Hedge) over the set of actions $\mathcal{X}$. The reward vector $\tilde{g}_t \in \mathbb{R}^{|\mathcal{X}|}$ is stochastic with expectation $\mathbb{E}\tilde{g}_{t,x} = \mathbb{E}_{\theta\sim p_t(\Theta^c_{\widehat{z}_t})}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}$ conditioning on the history of the algorithm $\{(x_s,y_s,\theta_s)\}_{s=1}^{t-1}$, and is bounded in high probability. We show that if we choose $\alpha=\frac{1}{4}$ and let $\tilde{\Delta}_{\max}$ be an upper bound on the loss function, we have the following regret guarantee:

Lemma 3.3 (informal)**.** *In round T_ℓ of epoch ℓ of Algorithm 2, we have with probability greater than $1-1/\ell^2$,*

$$\max_{\lambda\in\triangle_{\mathcal{X}}}\sum_{t=1}^{T_\ell}\mathbb{E}_{\theta\sim p_t(\Theta^c_{\widehat{z}_t})}\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\lambda)} - \sum_{t=1}^{T_\ell}\mathbb{E}_{\theta\sim p_t(\Theta^c_{\widehat{z}_t})}\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\lambda_t)} \le O\left(\sqrt{(d+\tilde{\Delta}_{\max})T_\ell\log\ell}\right)$$

Forced Exploration with G-optimal Design. To ensure adequate sampling in all directions, in each round we mix in some amount of the G-optimal distribution, denoted as $\lambda^G := \arg\min_{\lambda\in\Delta_{\mathcal{X}}}\max_{x\in\mathcal{X}}\|x\|^2_{A(\lambda)^{-1}}$. This ensures that $\max_{x\in\mathcal{X}}\|\widehat{\theta}_t-\theta\|_{xx^\top}$ is bounded for any $\theta\in\Theta$ and $\widehat{z}_t$ is eventually z_* with probability 1. The rate at which the mixture of this distribution decays as $t^{-\alpha}$, for any $0<\alpha<1/2$, so it has no effect on asymptotic performance. We note that thanks to the implicit anti-concentration properties of sampling θ_t from a multivariate Gaussian, this step is probably unnecessary and just an artifact of the analysis (Agrawal and Goyal, 2017).

Argmax Oracle One advantage of our approach that is most reminiscient of Thompson Sampling is the calculation of $\widehat{z}_t$ at the start of each epoch. In practice, if we have an efficient arg max-oracle, this calculation can be computationally efficient and does not require maintaining $\mathcal{Z}$. By exploiting arg max oracles, we can tractably solve problems like shortest-path and matchings, even in settings where $|\mathcal{Z}|$ is super-exponential in d (Katz-Samuels et al., 2020).

Doubling Trick As presented, the regret guarantees for Lemmas 3.1 and 3.3 require fixed step sizes η_λ,η_p. To overcome this need for a fixed step size, we use a doubling trick and restart the algorithm every 2^ℓ samples (Shalev-Shwartz et al., 2012). We believe the use of the doubling trick is purely a theoretical restriction and a more careful analysis could provide an anytime algorithm with no restarts.

3.1 Theoretical Guarantees

Recall that at the end of each epoch, $\widehat{z}_\ell(\theta) = \arg\max_{z\in\mathcal{Z}} z^\top\theta$ is the optimal answer for some $\theta\sim\pi_\ell$. Our main result is the following guarantee on Algorithm 2.

Theorem 3.4. *With probability* 1*,*

$$\lim_{\ell\to\infty} -\frac{1}{T_\ell}\log\mathbb{P}_{\theta\sim\pi_\ell}(\widehat{z}_\ell(\theta)\neq z_*) = \tau^*,$$

where $\pi_\ell := N(\widehat{\theta}_{T_\ell}, V^{-1}_{T_\ell-1})$ restricted to Θ.

Thus our algorithm guarantees that asymptotically the probability that we do not identify the optimal arm decays at the rate of $e^{-T\tau^*}$, with τ^* being the optimal exponent as given in Theorem 2.1. Such guarantees on the probability of a sampled arm are similar to those in the Bayesian best-arm literature, namely Russo (2016) and Jourdan et al. (2022). In these works, a posterior distribution is maintained and they guarantee that the posterior probability that a non-optimal arm is sampled converges at an exponential rate, with the best possible exponent among all allocation rules. We provide a similar guarantee here for linear bandits. As a remark, this does not directly lead to a bound on the frequentist probability of error, which requires integration of the posterior probability over all randomness in the algorithm. We provide a small sketch of the proof now. A full proof is in Appendix C.

Proof sketch. We say that $a_n \doteq b_n$ if $\frac{1}{n}\log(a_n/b_n) \to 0$ as $n \to \infty$. We focus on a fixed round ℓ of Algorithm 2. Using the fact that the expectation of the empirical log-likelihood ratio (conditioned on the data collected) between θ_* and some $\theta \in \Theta$ is the KL divergence between them, we can show using a Laplace Approximation

$$\mathbb{P}_{\theta\sim\pi_\ell}(\widehat{z}_\ell \neq z_*) \doteq \exp\left(-T_\ell \inf_{\theta\in\Theta^c_{z_*}} \frac{1}{2}\left\|\theta-\theta_*\right\|^2_{A(\bar{e}_{T_\ell})}\right).$$

where $\bar{e}_{T_\ell} = \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} e_{x_t}$. Letting $\bar{p}_{T_\ell} = \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} p_t(\Theta^c_{\widehat{z}_t})$, we have

$$\begin{aligned}
&\max_{\lambda\in\triangle_\mathcal{X}} \mathbb{E}_{\theta\sim\bar{p}_{T_\ell}} \left\|\widehat{\theta}_t-\theta\right\|^2_{A(\lambda)} - \min_{p\in\triangle(\Theta^c_{z_*})} \mathbb{E}_{\theta\sim p}\left\|\widehat{\theta}_t-\theta\right\|^2_{A(\bar{e}_{T_\ell})} \\
&= \max_{\lambda\in\triangle_\mathcal{X}} \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{\widehat{z}_t})}\left\|\widehat{\theta}_t-\theta\right\|^2_{A(\lambda)} \\
&\quad - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{\widehat{z}_t})}\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\lambda_t)} \\
&\qquad\qquad \text{(regret for max learner)} \\
&\quad + \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{\widehat{z}_t})}\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\lambda_t)} \\
&\quad - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{z_*})}\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} \quad \text{(error when } \widehat{z}_t \neq z_*) \\
&\quad + \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{z_*})}\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} - \frac{1}{T_\ell}\inf_{\theta\in\Theta^c_{z_*}}\left\|\theta-\theta_*\right\|^2_{V_{T_\ell}}. \\
&\qquad\qquad \text{(regret for the min learner)}
\end{aligned}$$

The regret guarantees in Lemmas 3.1 and 3.3 ensure the first and third sum are $o(1)$ and so go to 0 as $T_\ell \to \infty$. The fact that $p_t(\Theta^c_{\widehat{z}_t})$ is equal to $p_t(\Theta^c_{z_*})$ for large enough t ensures that the middle term similarly goes to 0. Combining all terms and the fact that $\widehat{\theta}_t$ is close to θ_* guarantees that for any $\epsilon > 0$ there is a sufficiently large ℓ such that $\max_{\lambda\in\triangle_\mathcal{X}} \mathbb{E}_{\theta\sim\bar{p}_{T_\ell}}\left\|\theta_*-\theta\right\|^2_{A(\lambda)} - \min_{p\in\triangle(\Theta^c_{z_*})}\mathbb{E}_{\theta\sim p}\left\|\theta_*-\theta\right\|^2_{A(\bar{e}_{T_\ell})} \leq \epsilon$, which using minimax duality implies that $\inf_{\theta\in\Theta^c_{z_*}} \frac{1}{2}\left\|\theta-\theta_*\right\|^2_{A(\bar{e}_{T_\ell})} \geq \max_{\lambda\in\triangle_\mathcal{X}}\min_{p\in\triangle(\Theta^c_{z_*})}\mathbb{E}_{\theta\sim p}\left[\left\|\theta_*-\theta\right\|^2_{A(\lambda)}\right] - \epsilon$. Since the first term on the right-hand side is τ^*, we have shown that $\inf_{\theta\in\Theta^c_{z_*}} \frac{1}{2}\left\|\theta-\theta_*\right\|^2_{A(\bar{e}_{T_\ell})} \geq \tau^* - \epsilon$. Since by definition $\tau^* \geq \inf_{\theta\in\Theta^c_{z_*}} \frac{1}{2}\left\|\theta-\theta_*\right\|^2_{A(\bar{e}_{T_\ell})}$, choosing $\epsilon \to 0$ concludes the proof that $\mathbb{P}_{\theta\sim\pi_\ell}(\widehat{z}_\ell \neq z_*) \doteq \exp\left(-T_\ell \inf_{\theta\in\Theta^c_{z_*}} \frac{1}{2}\left\|\theta-\theta_*\right\|^2_{A(\bar{e}_{T_\ell})}\right) = \exp(-T_\ell\tau^*)$. □

Remark: Stopping times. Note that we are not providing a guarantee on the expected stopping time for any finite δ. Existing asymptotically optimal approaches which guarantee a finite stopping time in high probability, e.g. Degenne et al. (2020), utilize a generalized log-likelihood-ratio test of the form

$$\max_{z\in\mathcal{Z}} \min_{\theta\in\Theta^c_{\widehat{z}_t}} \left\|\theta-\widehat{\theta}_t\right\|_{V_t} \geq \beta(t,\delta)$$

where $\beta(t,\delta) = O(\sqrt{d\log((T+\|\theta_*\|_2)/\delta)})$ is an anytime confidence bound controlling the deviations of $\|\theta-\widehat{\theta}_t\|_{V_t}$ (Abbasi-Yadkori et al., 2011). As a result, their algorithms saturate the lower bound for an expected stopping time, i.e. $\limsup_{\delta\to\infty}\mathbb{E}[\tau_\delta]/\log(1/\delta) \leq (\tau^*)^{-1}$. Unfortunately, this GLRT stopping rule itself requires a projection onto each element of $\mathcal{Z}$. We leave it as an open question whether an algorithm can be developed which is asymptotically optimal, requires no explicit projection, and has a finite expected stopping time in high probability.

Remark: Bounded assumptions on Θ. We assume Θ is closed and bounded. The boundedness assumption is needed since we would like to control that for each $\theta \in \Theta$, the rewards $x^\top\theta$ to be bounded for all arms $x \in \mathcal{X}$, which is used in our regret analysis for each learner. Learning algorithms such as AdaHedge (De Rooij et al., 2014) avoid the need for bounded rewards and we leave it as a future research direction to remove this condition.

4 RELATED WORK

Pure Exploration Linear Bandits The pure exploration linear bandit problem was introduced in the seminal work of Soare et al. (2014). In recent years, there has been renewed interest in this problem due to its ability to capture many best-arm-identification and pure exploration settings. Following the experimental design approach first considered by Soare et al. (2014), several different algorithmic frameworks were considered (Tao et al., 2018; Xu et al., 2018; Karnin et al., 2013).

One of the first algorithms to achieve matching instance-optimal upper and lower bounds (within logarithmic factors) for the case of $\mathbb{R}^d$ was by Fiez et al. (2019) and depends on an elimination scheme. Shortly after, several works proposed asymptotically optimal algorithms. The first of these methods utilized the track and stop approach given in Jedra and Proutiere (2020), which fully solves the τ^* objective of Equation 1 using a plug-in estimator $\widehat{\theta}_t$ at each round. Due to the computational difficulty of this, several works proposed alternatives that iteratively updated the sampling distribution in each round. This includes the game theoretic viewpoint we utilize first proposed by Degenne et al. (2020, 2019), and a novel modification of Frank-Wolfe by Wang et al. (2021). Other works have

augmented these approaches by providing elimination schemes to reduce the set of alternative $\mathcal{Z}$ that need to be considered each round. Zaki et al. (2022) proposes a hybrid approach combining the elimination from Fiez et al. (2019) and Degenne et al. (2020) to remove the condition that Θ needs to be bounded. Tirinzoni and Degenne (2022) provide an elimination approach where they carefully exploit properties of $\mathcal{Z}$. Finally, we mention that the pure exploration problem has also been considered in the generalized linear bandit (logistic) settings in Kazerouni and Wein (2021) and Jun et al. (2021). Future work could explore extending sampling methods to these settings.

Oracle Based Approaches As discussed before, if $\mathcal{Z}$ is a large or combinatorial set, it may be impossible to maintain and appropriate oracles are needed. Katz-Samuels et al. (2020) considers the linear combinatorial setting for matroid-like classes e.g. shortest-path, top-k, and bipartite matching. By exploiting ideas similar to Fiez et al. (2019), they provide an algorithm utilizing the argmax oracle to achieve near optimal sample complexity. A recent work by Li et al. (2022) reduces optimal policy learning in agnostic contextual bandits to pure exploration and provides a method analogous to Agarwal et al. (2014) which only relies on cost-sensitive classification.

Top Two Methods Our approach is perhaps most reminiscent of the Top-Two Thompson Sampling (TTTS) algorithm for best-arm identification in multi-armed bandits[1] of Russo (2016). Similar to Thompson sampling Russo et al. (2018), TTTS maintains a posterior distribution over the means of the arms, and at each round samples a mean vector from the distribution and chooses the arm with the highest sampled mean. It then continues to sample mean vectors, until one is returned whose highest mean is different from the previous found one. Both arms are then pulled. As discussed in the introduction, our algorithm is similar in spirit - we sample until finding a parameter vector whose best-arm is different from our current estimate and then we utilize these vectors to update our learners. Top-two algorithms for multi-armed bandits perform well in practice and have been extensively studied in Bayesian and frequentist settings under various assumptions on noise (Qin et al., 2017; Shang et al., 2020; Jourdan et al., 2022; Qin and Russo, 2022; Lee et al., 2023). However, they often depend on a parameter β, and only achieve a weaker notion of β-optimality. Our work is the first to propose and analyze an asymptotically optimal Top-two algorithm for the general linear bandit setting. We remark that the LinGapE algorithm (Xu et al., 2018) also uses a top-two approach and tends to perform well empricially, however it is unknown whether it is asymptotically optimal.

Online Learning and Thompson Sampling Finally we remark that the connection between Thompson Sampling and online learning has been previously explored in the early work of Li (2013). This work focuses on the regret setting. Other works in the regret setting have explored connections between information-theoretic analysis of Thompson sampling and online stochastic mirror descent algorithms (Lattimore and Gyorgy, 2021; Zimmert and Lattimore, 2019). We hope that our work provides a strong step in this direction for the structured pure exploration literature.

5 EXPERIMENTS

In the following, we provide some preliminary experiments to demonstrate the performance of Algorithm 1. Note that the contribution of this paper is primarily theoretical - our goal is to demonstrate that asymptotically optimal algorithms for pure exploration can rely purely on sampling oracles. We hope that the preliminary experiments we provide encourage further exploration of this line of thinking and lead to algorithms that can be as easy to apply as Thompson sampling in practice.

With this in mind, we ran the following modification of some of the algorithms of the previous section. Firstly, we eschewed the doubling trick and instead just ran `PEPS` directly for a fixed horizon side T. Secondly, for the max-learner we made use of AdaHedge which is able to use an adaptive step size. Finally, we set $\eta_p = 1$. Though our algorithm only has theoretical guarantees over a bounded set Θ, we believe that this is primarily a limitation of our analysis and so we set $\Theta = \mathbb{R}^d$. We also remove the forced G-optimal exploration for the same reason. For the sampling oracle, we use rejection sampling method because of its simplicity. We demonstrate empirically that the computation cost is not onerous. We plot the number of rejection steps used each round along with clock time per iteration for our method in Appendix F. We also see that our method is running faster than the benchmark LinGame especially when the number of arms is large in Table 3 in Appendix F. Further details on our experimental setup and additional evaluations are also in Appendix F.

The main algorithms we compare to are Thompson Sampling (Russo et al., 2018), LinGame (Degenne et al., 2020), and LinGapE Xu et al. (2018). LinGame is based on the two-player game strategy with best-response detailed in Section 2. For a fair comparison, we run LinGame and LinGapE without stopping. The goal of

[1] i.e. the arms are standard basis vectors $\mathcal{X} = \mathcal{Z} = \{e_1, \cdots, e_d\} \in \mathbb{R}^d$ and $\Theta = [0,1]^d$

	Soare's instance (Soare et al., 2014)			Sphere			TopK		
δ	0.1	0.05	0.01	0.1	0.05	0.01	0.2	0.1	0.05
PEPS	1027	1606	3284	294	476	794	7326	14188	22518
LinGame	828	1500	2688	186	282	638	8838	29963	>30000
LinGapE	708	1141	2281	316	433	690	7096	20570	>30000
Oracle	766	1232	2576	243	328	473	17363	>30000	>30000
TS	>5000	>5000	>5000	431	1046	2176	N/A	N/A	N/A

Table 1: The number of samples needed for $\mathbb{P}_{\theta\sim\pi_\ell}(\widehat{z}_\ell = z_*) > 1-\delta$ for various algorithms

our experiments was to demonstrate that sampling and no-projection algorithms can be competitive against algorithms that explicitly project. From this perspective, we did not consider algorithms that eliminate. For a more extensive empirical comparison of existing algorithms, please see Tirinzoni and Degenne (2022). We also include an oracle strategy that pulls arms from the allocation derived from the lower bound.

In summary, our algorithm achieves a similar performance compared to LinGame and LinGapE while beating LinTS in Soare and Sphere instances. For Top-k instance, our algorithm beats LinGame, LinTS, and LinGapE. Note that our algorithm is the first algorithm that relies purely on just sampling oracles and our theoretical analysis is only asymptotic, the experimental results are satisfactory since they show that our algorithm works decently well in practice. Now we detail the setting for each instance.

Soare's Instance (Soare et al., 2014). The first instance we consider is the standard benchmark linear bandit instance described in Soare et al. (2014). In this instance, the arm set $\mathcal{X} \subset \mathbb{R}^2$ with $|\mathcal{X}| = 3$. The first two arms are $x_1 = e_1, x_2 = e_2 \subset \mathbb{R}^2$, the canonical basis vectors, and an informative arm $x_3 = (\cos(\omega), \sin(\omega))$. The true parameter is $\theta_* = (1, 0) \in \mathbb{R}^d$.

In this problem, the optimal arm is always x_1. However, when the angle ω is small, it becomes challenging to distinguish the interfering arm x_{d+1} from x_1. An effective sampling strategy would pull arm x_2 instead of x_1 to reduce uncertainty between x_1 and x_{d+1} effectively. However, Thompson sampling will tend to pull x_1, which will take much longer to distinguish between the two competing arms. The experiments were carried out on a problem instance with $d = 2$ and $\omega = 0.1$. Our algorithm achieves a similar performance compared with LinGame and LinGapE while beats LinTS.

Sphere. Following Tao et al. (2018) and Degenne et al. (2020), we also consider a linear bandit instance where the arm set $\mathcal{X} \subset B^d := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ is randomly drawn from a unit sphere of dimension

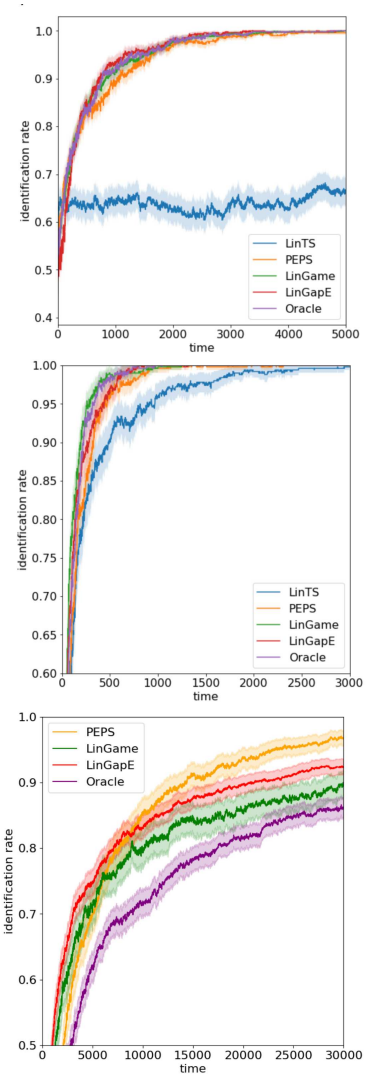

Figure 1: Best-arm identification rate for PEPS, LinGame (Degenne et al., 2020), LinGapE (Xu et al., 2018), Thompson sampling, and fixed weight strategy under three instances: Soare instance with $\omega = 0.1$, sphere instance with $d = 6$ and $|\mathcal{X}| = 20$, and Top-k instance with $d = 12$ and $k = 3$, with 500 repetitions for each instance. Confidence intervals with plus or minus two standard errors are shown.

d. For the true parameter, we select the two arms, x and x', that are closest to each other, and define $\theta_* = x + 0.01(x' - x)$, ensuring that x is the best arm. In our experiment, we run the three algorithms on a problem instance with $d = 6$ and $|\mathcal{X}| = 20$. As we can see, our algorithm still outperforms Thompson sampling and is competitive with LinGame and LinGapE.

Top-k. The third instance we consider is the top-k combinatorial bandit problem where the goal is to identify the top-k means. In the linear setting, this can be expressed as $\mathcal{X} = \{e_1, \cdots, e_d\} \subset \mathbb{R}^d$ and $\mathcal{Z} = \{e_{i_1} + \cdots + e_{i_k} : i_1, \cdots, i_k \in \binom{[d]}{k}\} \subset \mathbb{R}^d$, i.e. $\mathcal{X}$ is the standard basis and $\mathcal{Z}$ is the set of indicator vectors of subsets of size k. Then, the best arm in this new arm set $\mathcal{Z}$ corresponds to the top-k arms in $\mathcal{X}$, which is the goal of top-k identification. Then we run BAI algorithms on this new arm set. We take $\theta = [1, .95, .90, \cdots, 1 - .05i, \cdots] \in \mathbb{R}^d$. As we can see, our algorithm outperforms LinGame and LinGapE in this instance.

We also present Table 1 describing the number of samples needed to reach a $1-\delta$ idenfication rate for various δ values. Note that we do not run Thompson sampling for the Top-k instance (it is not defined when $\mathcal{X} \neq \mathcal{Z}$ so we put N/A there), and $> n$ in the table means that the algorithm fails to achieve $1 - \delta$ for the n iterations we run in the experiment. We can see that our algorithm, PEPS, achieves an $1 - \delta$ best-arm identification probability for all δ in all instances, with a rate similar to LinGame, outperforming LinTS in all three instances.

6 CONCLUSION

In this paper, we present the first sampling-based projection-free algorithm for pure exploration in linear bandits. Our algorithm only relies on a sampling oracle and an argmax oracle, so our algorithm is tractable in various settings. We show that our algorithm is asymptotically optimal in the sense that the probability that we do not identify the optimal arm decays exponentially with the optimal rate for a fixed allocation. We provide experiments demonstrating that our algorithm beats Thompson sampling and has competitive performance against benchmark algorithms such as LinGame (Degenne et al., 2020) in various problem instances. Our current approach has various limitations: for example, we need to assume that Θ is bounded. However, we hope that this work opens a line of investigation into better sampling-based algorithms for effective exploration.

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

S. Agrawal and N. Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

A. Biswas, T. T. Pham, M. Vogelsong, B. Snyder, and H. Nassif. Seeker: Real-time interactive search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2875, 2019.

S. Bubeck. Introduction to online optimization. *Lecture notes*, 2:1–86, 2011.

J. Burkardt. The truncated normal distribution. *Department of Scientific Computing Website, Florida State University*, 1:35, 2014.

L. Chen, A. Gupta, J. Li, M. Qiao, and R. Wang. Nearly optimal sampling algorithms for combinatorial pure exploration. In *Conference on Learning Theory*, pages 482–534. PMLR, 2017.

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

R. Degenne, W. M. Koolen, and P. Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019.

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

L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, NY, USA, 1986.

T. Fiez, L. Jain, K. G. Jamieson, and L. Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

P. Glynn and S. Juneja. A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference, 2004.*, volume 1. IEEE, 2004.

Y. Jedra and A. Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

M. Jourdan, R. Degenne, D. Baudry, R. de Heide, and E. Kaufmann. Top two algorithms revisited. *arXiv preprint arXiv:2206.05979*, 2022.

K.-S. Jun, L. Jain, B. Mason, and H. Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *International Conference on Machine Learning*, pages 5148–5157. PMLR, 2021.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.

J. Katz-Samuels, L. Jain, K. G. Jamieson, et al. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *Advances in Neural Information Processing Systems*, 33:10371–10382, 2020.

A. Kazerouni and L. M. Wein. Best arm identification in generalized linear bandits. *Operations Research Letters*, 49(3):365–371, 2021.

A. Laddha and S. S. Vempala. Convergence of gibbs sampling: coordinate hit-and-run mixes fast. *Discrete & Computational Geometry*, pages 1–20, 2023.

T. Lattimore and A. Gyorgy. Mirror descent and the information ratio. In *Conference on Learning Theory*, pages 2965–2992. PMLR, 2021.

T. Lattimore and C. Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

J. Lee, J. Honda, and M. Sugiyama. Thompson exploration with best challenger rule in best arm identification. *arXiv preprint arXiv:2310.00539*, 2023.

L. Li. Generalized thompson sampling for contextual bandits. *arXiv preprint arXiv:1310.7163*, 2013.

Y. Li and S. K. Ghosh. Efficient sampling methods for truncated multivariate normal and student-t distributions subject to linear inequality constraints. *Journal of Statistical Theory and Practice*, 9(4):712–732, 2015.

Z. Li, L. Ratliff, K. G. Jamieson, L. Jain, et al. Instance-optimal pac algorithms for contextual bandits. *Advances in Neural Information Processing Systems*, 35: 37590–37603, 2022.

M. Liu and F. Orabona. On the initialization for convex-concave min-max problems. In *International Conference on Algorithmic Learning Theory*, pages 743–767. PMLR, 2022.

L. Lovász. Hit-and-run mixes fast. *Mathematical programming*, 86:443–461, 1999.

H. Maatouk and X. Bay. A new rejection sampling method for truncated multivariate gaussian random variables restricted to convex sets. In *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, Belgium, April 2014*, pages 521–530. Springer, 2016.

K. P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

C. Qin and D. Russo. Adaptivity and confounding in multi-armed bandit experiments. *arXiv preprint arXiv:2202.09036*, 2022.

C. Qin, D. Klabjan, and D. Russo. Improving the expected improvement algorithm. *Advances in Neural Information Processing Systems*, 30, 2017.

D. Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418. PMLR, 2016.

D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11 (1):1–96, 2018.

S. Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

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

M. Soare, A. Lazaric, and R. Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

C. Tao, S. Blanco, and Y. Zhou. Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, pages 4877–4886. PMLR, 2018.

A. Tirinzoni and R. Degenne. On elimination strategies for bandit fixed-confidence identification. *arXiv preprint arXiv:2205.10936*, 2022.

P.-A. Wang, R.-C. Tzeng, and A. Proutiere. Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821, 2021.

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

M. Zaki, A. Mohan, and A. Gopalan. Improved pure exploration in linear bandits with no-regret learning. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 3709–3715. International Joint Conferences on Artificial Intelligence, 2022.

J. Zimmert and T. Lattimore. Connections between mirror descent, thompson sampling and the information ratio. *Advances in Neural Information Processing Systems*, 32, 2019.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Contents

A NOTATIONS AND GENERAL DESCRIPTION

In the following, we let the index t, $1 \le t \le T_\ell$ denote the timestep in round ℓ for any ℓ. Throughout this section we will make use of the filtration $\mathcal{F}_t = \{(x_s, \theta_s, y_s)\}_{s=1}^{t-1}$ defined in any round. The table below summarizes the notations used in the proof.

$\bar{p}_{T_\ell} = \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} p_t$	Average of p at the end of round ℓ
$\bar{e}_{T_\ell} = \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} e_{x_t}$	Empirical probability of arms pulled at the end of round ℓ
$\pi_\ell \sim N(\widehat{\theta}_{T_\ell+1}, \eta_p^{-1} V_{T_\ell}^{-1})$ restricted on Θ	The distribution θ is sampled from at the end of round ℓ
$\Delta_{\min} = \min_{x \neq x^*}(x^* - x)^\top \theta^*$	minimum gap
$T_2(\ell) = \max_{x\in\mathcal{X}} \left(\frac{6\sqrt{\log(\lvert\mathcal{X}\rvert T_\ell \ell^2)}}{\lambda_x^G}\right)^4$	a time after which each arm gets sufficiently number of pulls
$T_0(\ell) = \max\left\{\left(\frac{d\beta(t,\ell^2)\max_{z\in\mathcal{Z}}\lVert z\rVert_1}{\Delta_{\min}}\right)^{4/3}, T_2(\ell)+1\right\}$	a time after which we have $\widehat{z}_t = z_*$ with high probability
$\ell_0 := \min\{\ell : T_\ell \ge T_0(\ell)^{3/2}\}$	minimum round number such that we have guarantee of convergence with high probability
L	upper bound on $\max_{x\in\mathcal{X}} \lVert x\rVert_2$
B	upper bound on $\lVert\theta_*\rVert_2$
$B_{\mathcal{X}}$	$\max_{x\in\mathcal{X}}\max_{\theta\in\Theta} x^\top\theta$
$\Delta_{\max}$	$\max_{x\in\mathcal{X}}\max_{\theta,\theta'\in\Theta} \lvert x^\top(\theta-\theta')\rvert$
$\beta(t,1/\delta) = B + \sqrt{2\log(1/\delta) + d\log\left(\frac{d+tL^2}{d}\right)}$	anytime confidence bound for $\left\lVert\widehat{\theta}_t - \theta^*\right\rVert^2_{V_{t-1}}$
$C_{1,\ell} = \Delta_{\max} + L^2\beta(T_\ell, \ell^2)$	an upper bound on $\max_{x\in\mathcal{X}}\max_{t\le T_\ell}\lvert\langle x, \widehat{\theta}_t\rangle\rvert$
$C_{3,\ell} = B_{\mathcal{X}} + \Delta_{\max} + L^2\beta(T_\ell,\ell^2)$	an upper bound on $\max_{x\in\mathcal{X}}\max_{\theta\in\Theta}\max_{t\le T_\ell}\lvert\langle x, \theta - \widehat{\theta}_t\rangle\rvert$

Table 2: Table of constants and upper bounds used in the proof

Let $N_{t,x}$ denote the number of times arm x gets pulled at time t. We then define several good events needed to guarantee the performance of PEPS at round ℓ.

$$\mathcal{E}_{1,\ell} = \bigcap_{t=1}^{T_\ell}\left\{\left\lVert\widehat{\theta}_t - \theta^*\right\rVert^2_{V_{t-1}} \le \beta(t,\ell^2)\right\},$$

$$\mathcal{E}_{2,\ell} = \bigcap_{t=1}^{T_\ell}\left\{\max_{x\in\mathcal{X}} \lvert x^\top\widehat{\theta}_t\rvert \le C_{1,\ell}\right\},$$

$$\mathcal{E}_{3,\ell} = \bigcap_{t\ge T_2}^{T_\ell}\bigcap_{x\in\mathcal{X}} \mathcal{G}_{t,x} \text{ where } \mathcal{G}_{t,x} = \{V_t \ge t^{3/4}A(\lambda^G)\}, \forall t \ge T_2, x\in\mathcal{X}$$

$$\mathcal{E}_{4,\ell} = \cup_{t\ge T_0}\mathbf{1}\{\widehat{z}_t = z_*\}$$

Throughout the proof we also define for some random variable $x \in \mathcal{X}$ with $x \sim p$ and some function $f(x)$,

$$\mathbb{F}_{x\sim p}[f(x)] = \sum_{x\in\mathcal{X}} p_x f(x).$$

The rest of the supplement is organized as follows. In Section B, we present a proof of the lower bound stated in Theorem 2.1. Section F provides more experimental results.

In Section C, we prove the main theorem (Theorem 3.4) stated in the paper by combining a saddle-point convergence argument with a guarantee on the likelihood ratio. We tackle the latter in Section C.1, where we provide we relate the empirical probability of finding the best-arm at the end of a round of PEPS to the

likelihood ratio. In Section C.2, we show the saddle point approximation and provide a guarantee on how well τ^* is approximated after one round of PEPS. This argument depends on

- Section C.3 and C.4 which provide regret guarantees on the max and min learners.
- Section C.5 provides lemmas bounding terms related to the approximation error of $\widehat{\theta}_{T_\ell}$ to θ^*.
- Section C.6 formally shows that after certain rounds each arm gets enough samples.
- Section D shows that good events needed to guarantee performance of PEPS happen with high probability.

Finally, Section E provides some technical lemmas used in the proof.

B PROOF OF THEOREM 2.1

Theorem B.1. *Fix $\Theta = \mathbb{R}^d$ and any $\theta_* \in \Theta$. For some λ consider a procedure that draws $x_1, \dots, x_T \sim \lambda$, then observes $y_t = \langle x_t, \theta_* \rangle + \epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0,1)$, and then computes $\widehat{z}_T = \arg\max_{z \in \mathcal{Z}} \langle z, \widehat{\theta}_T \rangle$ where $\widehat{\theta}_T = \arg\min_{\theta \in \Theta} \sum_{t=1}^T \|y_t - \langle \theta, x_t \rangle\|_2^2$. Then for any $\lambda \in \triangle_{\mathcal{X}}$ we have*

$$\limsup_{T \to \infty} -\frac{1}{T} \log\Big(\mathbb{P}_{\theta_*, x_t \sim \lambda}(\widehat{z}_T \neq z_*)\Big) \leq \tau^*.$$

Proof. Assume that $\{z - z_*\}_{z \in \mathcal{Z}}$ span $\mathbb{R}^d$. Otherwise, discard the components of $\mathcal{X}$ and θ_* that are orthogonal to the span of $\{z - z_*\}_{z \in \mathcal{Z}}$ and reparameterize in the subspace spanned by $\{z - z_*\}_{z \in \mathcal{Z}}$. We can then work in this reparameterized space, so without loss of generality we can assume $\{z - z_*\}_{z \in \mathcal{Z}}$ span $\mathbb{R}^d$.

Furthermore, assume that $\mathcal{X}$ spans $\mathbb{R}^d$. If this were not true, then there could be a component of θ_* that is orthogonal to the span of $\mathcal{X}$ which makes z_* not identifiable since we assumed $\{z - z_*\}_{z \in \mathcal{Z}}$ spans $\mathbb{R}^d$. That is, if $\theta_*^\perp$ is the projection of θ_* onto the subspace orthogonal to the span of $\mathcal{X}$, then $\langle z - z_*, \theta_*^\perp \rangle$ could be arbitrarily large but no measurement could detect $\theta_*^\perp$.

Putting the two assumptions together, we conclude that there exists a $\lambda \in \triangle_{\mathcal{X}}$ such that $A(\lambda) \succ 0$ (equivalently, $\lambda_{\min}(A(\lambda)) > 0$) and $\max_{z \in \mathcal{Z}} \|z - z_*\|_{A(\lambda)^{-1}} < \infty$. Fix any λ satisfying such conditions. Define the event $G_\lambda = \{\sum_{t=1}^T x_t x_t \succeq A(\lambda) T (1 - g_{\lambda,T})\}$ for some $g_{\lambda,T} = o(T)$ sequence to be defined next.

By applying matrix Chernoff to the random matrices $\{\frac{1}{T} A(\lambda)^{-1} x_t x_t^\top\}_t$ we have for any $\epsilon \in [0,1)$ that

$$\mathbb{P}\Big(\frac{1}{T} \sum_{t=1}^T x_t x_t^\top \succeq A(\lambda)(1 - \epsilon)\Big) \geq 1 - d \exp(-\epsilon^2 / 2R)$$

where $R = \max_t \lambda_{\max}(\frac{1}{T} A(\lambda)^{-1} x_t x_t^\top)$. Observe that

$$\begin{aligned} \lambda_{\max}(\frac{1}{T} A(\lambda)^{-1} x_t x_t^\top) &\leq \|\frac{1}{T} A(\lambda)^{-1} x_t x_t^\top\|_2 \\ &\leq L^2 / \lambda_{\min}(A(\lambda)) T. \end{aligned}$$

So taking $\epsilon = g_{\lambda,T} = \sqrt{\frac{2L^2 \lambda_{\min}(A(\lambda))^{-1} \log(dT)}{T}}$ we have that $\mathbb{P}(G_\lambda) \geq 1 - 1/T$ whenever $g_{\lambda,T} < 1$ which holds for sufficiently large T.

Now, for any $\{x_t\}_{t=1}^T$ that span $\mathbb{R}^d$ (will be guaranteed by event G_λ) we have that

$$\begin{aligned}\widehat{\theta}_T &= \arg\min_{\theta\in\Theta} \sum_{t=1}^T \|y_t - \langle \theta, x_t\rangle\|_2^2 \\ &= \Big(\sum_{t=1}^T x_t x_t^\top\Big)^{-1} \sum_{t=1}^T x_t y_t \\ &= \theta_* + \Big(\sum_{t=1}^T x_t x_t^\top\Big)^{-1} \sum_{t=1}^T x_t \epsilon_t \\ &= \theta_* + \Big(\sum_{t=1}^T x_t x_t^\top\Big)^{-1/2} \eta\end{aligned}$$

where the last line holds with inequality in distribution for $\eta \sim \mathcal{N}(0, I_d)$. We conclude that for any z that $\langle \widehat{\theta}_T - \theta_*, z - z_*\rangle$ is a zero-mean Gaussian random variable with variance

$$\begin{aligned}\sigma_{z,\lambda}^2 :=& \mathbb{E}[\langle \widehat{\theta}_T - \theta_*, z - z_*\rangle^2] \\ =& \mathbb{E}[\langle \Big(\sum_{t=1}^T x_t x_t^\top\Big)^{-1/2} \eta, z - z_*\rangle^2] \\ =& (z - z_*)^\top \Big(\sum_{t=1}^T x_t x_t^\top\Big)^{-1} (z - z_*).\end{aligned}$$

Thus, on G_λ we have that $\sigma_{z,\lambda}^2 \le \frac{1}{T(1-g_{\lambda,T})}\|z - z_*\|_{A(\lambda)^{-1}}^2$.

Consequently,

$$\begin{aligned}\mathbb{P}_{\theta_*}(\widehat{z}_T \neq z_*) &= \mathbb{P}_{\theta_*}\Big(\bigcup_{z\in\mathcal{Z}\setminus z_*} \{\widehat{z}_T = z, z \neq z_*\}\Big) \\ &\ge \max_{z\in\mathcal{Z}\setminus z_*} \mathbb{P}_{\theta_*}(\widehat{z}_T = z, z \neq z_*) \\ &= \max_{z\in\mathcal{Z}\setminus z_*} \mathbb{P}_{\theta_*}(\langle \widehat{\theta}_T, z - z_*\rangle \ge 0) \\ &= \max_{z\in\mathcal{Z}\setminus z_*} \mathbb{P}_{\theta_*}(\langle \widehat{\theta}_T - \theta_*, z - z_*\rangle \ge \langle \theta_*, z - z_*\rangle) \\ &\ge \max_{z\in\mathcal{Z}\setminus z_*} \mathbb{E}_{\{x_t\}\sim\lambda}\mathbb{E}_{\theta_*}\big[\mathbf{1}\{G_\lambda\}\mathbf{1}\big\{\langle \widehat{\theta}_T - \theta_*, z - z_*\rangle \ge \langle \theta_*, z - z_*\rangle\big\}\big|\{x_t\}\big]\Big] \\ &= \max_{z\in\mathcal{Z}\setminus z_*} \mathbb{P}_{\{x_t\}\sim\lambda}(G_\lambda)\mathbb{P}_{\eta_1\sim\mathcal{N}(0,1)}\big(\eta_1\sigma_{z,\lambda} \ge \langle \theta_*, z - z_*\rangle\big).\end{aligned}$$

Using the fact that

$$\mathbb{P}_{\eta_1\sim\mathcal{N}(0,1)}\big(\eta_1 \ge s\big) = \int_{x=s}^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx > (\frac{1}{s} - \frac{1}{s^3})\frac{1}{\sqrt{2\pi}} e^{-s^2/2}$$

for positive s, we conclude that

$$\begin{aligned}&\mathbb{P}_{\theta_*}(\widehat{z}_T \neq z_*) \\ &\ge \max_{z\in\mathcal{Z}\setminus z_*} \mathbb{P}_{\{x_t\}\sim\lambda}(G_\lambda)\mathbb{P}_{\eta_1\sim\mathcal{N}(0,1)}\big(\eta_1\sigma_{z,\lambda} \ge \langle \theta_*, z - z_*\rangle\big) \\ &\ge \mathbf{1}\{g_{\lambda,T} < 1\}(1 - \frac{1}{T}) \max_{z\in\mathcal{Z}\setminus z_*} (\frac{\sigma_{z,\lambda}}{\langle \theta_*, z - z_*\rangle} - \frac{\sigma_{z,\lambda}^3}{\langle \theta_*, z - z_*\rangle^3})\frac{1}{\sqrt{2\pi}} e^{-\frac{\langle \theta_*, z - z_*\rangle^2}{\sigma_{z,\lambda}^2}/2} \\ &\ge \max_{z\in\mathcal{Z}\setminus z_*} \mathbf{1}\{g_{\lambda,T} < 1, \tfrac{\langle \theta_*, z - z_*\rangle^2}{\sigma_{z,\lambda}^2} \ge 2\}(1 - \frac{1}{T})\frac{\sigma_{z,\lambda}}{\langle \theta_*, z - z_*\rangle}\frac{1}{\sqrt{8\pi}} e^{-\frac{\langle \theta_*, z - z_*\rangle^2}{\sigma_{z,\lambda}^2}/2} \\ &\ge \max_{z\in\mathcal{Z}\setminus z_*} \mathbf{1}\{g_{\lambda,T} < 1, \tfrac{T(1-g_{\lambda,T})\langle \theta_*, z - z_*\rangle^2}{\|z - z_*\|_{A(\lambda)^{-1}}^2} \ge 2\}(1 - \frac{1}{T})\frac{\|z - z_*\|_{A(\lambda)^{-1}}^2}{T(1-g_{\lambda,T})\langle \theta_*, z - z_*\rangle^2}\frac{1}{\sqrt{8\pi}} e^{-\frac{T(1-g_{\lambda,T})\langle \theta_*, z - z_*\rangle^2}{\|z - z_*\|_{A(\lambda)^{-1}}^2}/2}.\end{aligned}$$

Thus, because $g_{\lambda,T} = o(T)$ and $\frac{\|z-z_*\|^2_{A(\lambda)^{-1}}}{\langle\theta_*, z-z_*\rangle^2} < \infty$ we have that

$$\begin{aligned}\limsup_{T\to\infty} -\frac{1}{T}\log\left(\mathbb{P}_{\theta_*,x_t\sim\lambda}(\widehat{z}_T \neq z_*)\right) &\leq \frac{\langle\theta_*, z-z_*\rangle^2}{\|z-z_*\|^2_{A(\lambda)^{-1}}}/2 \\ &= \min_{\theta\in\Theta^c_{z_*}} \|\theta-\theta_*\|^2_{A(\lambda)}/2 \\ &\leq \max_{\lambda\in\triangle_{\mathcal{X}}} \min_{\theta\in\Theta^c_{z_*}} \|\theta-\theta_*\|^2_{A(\lambda)}/2 = \tau^*\end{aligned}$$

where the second line uses the fact that $\Theta = \mathbb{R}^d$. □

C PROOF OF THEOREM 3.4

Theorem C.1. *Under Algorithm 1 and 2 and Assumption 1, we have the sampling distribution satisfies with probability 1,*

$$\lim_{\ell\to\infty} -\frac{1}{T_\ell}\log\pi_\ell(\Theta^c_{z_*}) = \tau^*.$$

Proof. By Theorem C.2, we have that for $\ell \geq \ell_0$, $\mathbb{P}(\mathcal{E}^c_\ell) \leq \frac{5}{\ell^2}$. Also, since $T_\ell = 2^\ell$, and $T_0(\ell)$ only scales logarithmically in ℓ, so $\ell_0 < \infty$. Therefore, $\sum_{\ell=1}^\infty \mathbb{P}(\mathcal{E}_\ell) < \infty$. By Borel-Cantelli, we have

$$\mathbb{P}\left(\limsup_{n\to\infty} \mathcal{E}^c_\ell\right) = 0.$$

Note that $\limsup_{\ell\to\infty}\mathcal{E}_\ell = \bigcap_{\ell=1}^\infty\bigcup_{k=\ell}^\infty \mathcal{E}_k$, this implies that the probability that infinitely many of them occur is zero, which means that $\mathcal{E}_\ell$ eventually holds for sufficiently large ℓ with probability 1. However, under $\mathcal{E}_\ell$ we have

$$\begin{aligned}\pi_\ell(\Theta^c_{z_*}) &= \frac{\int_{\Theta^c_{z_*}}\pi_\ell(\theta)d\theta}{\int_\Theta \pi_\ell(\theta)d\theta} = \frac{\int_{\Theta^c_{z_*}}\pi_\ell(\theta)/\pi_\ell(\theta^*)d\theta}{\int_\Theta \pi_\ell(\theta)/\pi_\ell(\theta^*)d\theta} \\ &\doteq \frac{\int_{\Theta^c_{z_*}} e^{-\frac{T_\ell}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})}}d\theta}{\int_\Theta e^{-\frac{T_\ell}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})}}d\theta} && \text{(by } \mathcal{E}_\ell\text{)} \\ &\doteq e^{-T_\ell \inf_{\theta\in\Theta^c_{z_*}}\frac{1}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})}}. && \text{(Lemma E.2 and } \inf_{\theta\in\Theta}\|\theta-\theta^*\|^2_{A(\lambda)} = 0 \text{ for any } \lambda\text{)}\end{aligned}$$

This implies that there exists some $\epsilon'_\ell \to 0$ such that

$$\left|-\frac{1}{T_\ell}\log\pi_\ell(\Theta^c_{z_*}) - \inf_{\theta\in\Theta^c_{z_*}}\frac{1}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})}\right| \leq \epsilon'_\ell.$$

Under $\mathcal{E}_{6,\ell}$, there exists some sequence $\epsilon_\ell \to 0$ such that

$$\tau^* - \inf_{\theta\in\Theta^c_{z_*}}\frac{1}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})} \leq \epsilon_\ell.$$

Since

$$\tau^* = \max_{\lambda\in\triangle_{\mathcal{X}}}\inf_{\theta\in\Theta^c_{z_*}}\frac{1}{2}\|\theta-\theta^*\|^2_{A(\lambda)} \geq \inf_{\theta\in\Theta^c_{z_*}}\frac{1}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})},$$

combining the above three displays, we have under $\mathcal{E}_\ell$,

$$\left|-\frac{1}{T_\ell}\log\pi_\ell(\Theta^c_{z_*}) - \tau^*\right| \leq \epsilon_\ell + \epsilon'_\ell,$$

where $\epsilon_\ell + \epsilon'_\ell \to 0$ as $\ell\to\infty$. Combining this with the fact that $\mathbb{P}\left(\limsup_{\ell\to\infty}\mathcal{E}_\ell\right) = 0$, we have with probability 1,

$$\lim_{\ell\to\infty} -\frac{1}{T_\ell}\log\pi_\ell(\Theta^c_{z_*}) = \tau^*.$$

□

Theorem C.2. *In round ℓ for $\ell \geq \ell_0$, define*

$$\mathcal{E}_{5,\ell} = \left\{ \sup_{\theta \in \Theta} \frac{1}{T_\ell} \left| \log \frac{\pi_{T_\ell}(\theta^*)}{\pi_{T_\ell}(\theta)} - \frac{T_\ell}{2} \|\theta - \theta^*\|^2_{A(\bar{e}_{T_\ell})} \right| \leq \kappa_\ell \right\}$$
$$\mathcal{E}_{6,\ell} = \left\{ \left| \max_{\lambda \in \triangle_\mathcal{X}} \inf_{\theta \in \Theta^c_{z_*}} \frac{1}{2} \|\theta - \theta^*\|^2_{A(\lambda)} - \inf_{\theta \in \Theta^c_{z_*}} \frac{1}{2} \|\theta - \theta^*\|^2_{A(\bar{e}_{T_\ell})} \right| \leq \epsilon_\ell \right\}$$

with $\epsilon_\ell \to 0$ and $\kappa_\ell \to 0$ as $\ell \to \infty$. Define $\mathcal{E}_\ell = \mathcal{E}_{5,\ell} \cap \mathcal{E}_{6,\ell}$. Then $\mathbb{P}(\mathcal{E}_\ell) \geq 1 - 5/\ell^2$.

Proof. We first summarize the guarantees for the probabilities of events below. For $\ell \geq \ell_0$, we have

- from Lemma C.4, we have that $\mathbb{P}(\mathcal{E}_{6,\ell} | \mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}) \geq 1 - 1/\ell^2$ with choice of $\epsilon_\ell = O(T_\ell^{-1/4})$;
- from Lemma D.1, $\mathbb{P}(\mathcal{E}_{1,\ell}) \geq 1 - 1/\ell^2$;
- by Lemma D.2, $\mathcal{E}_{2,\ell}$ is true under $\mathcal{E}_{3,\ell} \cap \mathcal{E}_{1,\ell}$;
- by Lemma C.16, $\mathbb{P}(\mathcal{E}_{4,\ell} | \mathcal{E}_{1,\ell}) \geq 1 - 1/\ell^2$;
- by Lemma C.3 with $\kappa_\ell = O(T_\ell^{-1/2})$, $\mathbb{P}(\mathcal{E}_{5,\ell}) \geq 1 - 1/\ell^2$;
- by Lemma C.14, $\mathbb{P}(\mathcal{E}_{3,\ell}) \geq 1 - 1/\ell^2$.

Note that $\mathcal{E}_\ell \supset \mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell} \cap \mathcal{E}_{5,\ell} \cap \mathcal{E}_{6,\ell}$, and so

$$\begin{aligned}
\mathcal{E}_\ell^c &\subset \mathcal{E}_{1,\ell}^c \cup \mathcal{E}_{2,\ell}^c \cup \mathcal{E}_{3,\ell}^c \cup \mathcal{E}_{4,\ell}^c \cup \mathcal{E}_{5,\ell}^c \cup \mathcal{E}_{6,\ell}^c \\
&= \mathcal{E}_{1,\ell}^c \cup (\mathcal{E}_{2,\ell}^c \cap \mathcal{E}_{1,\ell} \cap \mathcal{E}_{3,\ell}) \cup \mathcal{E}_{3,\ell}^c \cup (\mathcal{E}_{4,\ell}^c \cap \mathcal{E}_{1,\ell}) \cup \mathcal{E}_{5,\ell}^c \cup (\mathcal{E}_{6,\ell}^c \cap \mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}).
\end{aligned}$$

Therefore, for $\ell \geq \ell_0$,

$$\begin{aligned}
&\mathbb{P}(\mathcal{E}_\ell^c) \\
&\leq \mathbb{P}(\mathcal{E}_{1,\ell}^c) + \mathbb{P}(\mathcal{E}_{2,\ell}^c \cap \mathcal{E}_{1,\ell} \cap \mathcal{E}_{3,\ell}) + \mathbb{P}(\mathcal{E}_{3,\ell}^c) + \mathbb{P}(\mathcal{E}_{4,\ell}^c \cap \mathcal{E}_{1,\ell}) + \mathbb{P}(\mathcal{E}_{5,\ell}^c) + \mathbb{P}(\mathcal{E}_{6,\ell}^c \cap \mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}) \\
&\leq \mathbb{P}(\mathcal{E}_{1,\ell}^c) + \mathbb{P}(\mathcal{E}_{2,\ell}^c | \mathcal{E}_{1,\ell} \cap \mathcal{E}_{3,\ell}) \mathbb{P}(\mathcal{E}_{1,\ell} \cap \mathcal{E}_{3,\ell}) + \mathbb{P}(\mathcal{E}_{3,\ell}^c) + \mathbb{P}(\mathcal{E}_{4,\ell}^c | \mathcal{E}_{1,\ell}) \mathbb{P}(\mathcal{E}_{1,\ell}) \\
&\quad + \mathbb{P}(\mathcal{E}_{5,\ell}^c) + \mathbb{P}(\mathcal{E}_{6,\ell}^c | \mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}) \mathbb{P}(\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}) \\
&\leq \mathbb{P}(\mathcal{E}_{1,\ell}^c) + \mathbb{P}(\mathcal{E}_{2,\ell}^c | \mathcal{E}_{1,\ell} \cap \mathcal{E}_{3,\ell}) + \mathbb{P}(\mathcal{E}_{3,\ell}^c) + \mathbb{P}(\mathcal{E}_{4,\ell}^c | \mathcal{E}_{1,\ell}) + \mathbb{P}(\mathcal{E}_{5,\ell}^c) + \mathbb{P}(\mathcal{E}_{6,\ell}^c | \mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}) \\
&\leq \frac{5}{\ell^2}.
\end{aligned}$$

Therefore, $\mathbb{P}(\mathcal{E}_\ell) \geq 1 - \frac{5}{\ell^2}$. □

C.1 Guarantees on the Likelihood Ratio

Lemma C.3. *We have with probability at least $1 - 1/\ell^2$,*

$$\sup_{\theta \in \Theta} \frac{1}{T_\ell} \left| \log \frac{\pi_\ell(\theta)}{\pi_\ell(\theta^*)} - \frac{T_\ell}{2} \|\theta - \theta^*\|^2_{A(\bar{e}_{T_\ell})} \right| \leq \Delta_{\max} \sqrt{\frac{2d \log\left(\frac{(d + T_\ell L^2)\ell^2}{d}\right)}{T_\ell}}.$$

Which implies that $\frac{\pi_\ell(\theta)}{\pi_\ell(\theta^*)} \doteq e^{-T_\ell \|\theta - \theta^*\|^2_{A(\bar{e}_{T_\ell})}}$.

Proof. Throughout the following we set $T := T_\ell$. Recall that $\pi_\ell(\theta) = \mathcal{N}(\widehat{\theta}_{T+1}, V_T^{-1})$ restricted on Θ, which means that for each $\theta \in \Theta$,

$$\pi_\ell(\theta) = \frac{\exp\left(-\frac{1}{2} \left\| \theta - \widehat{\theta}_{T+1} \right\|^2_{V_T}\right)}{\int_\Theta \exp\left(-\frac{1}{2} \left\| \theta' - \widehat{\theta}_{T+1} \right\|^2_{V_T}\right) d\theta'}.$$

Since the denominator is independent of θ, this means that

$$\frac{\pi_\ell(\theta)}{\pi_\ell(\theta^*)} = \exp\left(-\frac{1}{2}\left(\left\|\theta - \widehat{\theta}_{T+1}\right\|_{V_T}^2 - \left\|\theta^* - \widehat{\theta}_{T+1}\right\|_{V_T}^2\right)\right)$$

where

$$\begin{aligned}
&\left\|\theta^* - \widehat{\theta}_{T+1}\right\|_{V_T}^2 - \left\|\theta - \widehat{\theta}_{T+1}\right\|_{V_T}^2 \\
&= \|\theta^*\|_{V_T}^2 - 2(\theta^*)^\top V_T \widehat{\theta} + \left\|\widehat{\theta}_{T+1}\right\|_{V_T}^2 - \left\|\widehat{\theta}_{T+1}\right\|_{V_T}^2 + 2(\widehat{\theta}_{T+1})^\top V_T \theta - \|\theta\|_{V_T}^2 \\
&= \|\theta^*\|_{V_T}^2 - 2\left(\theta^*\right)^\top V_T \left(\theta^* + V_T^{-1}\sum_{s=1}^T \epsilon_s x_s\right) + 2\theta^\top V_T \left(\theta^* + V_T^{-1}\sum_{s=1}^T \epsilon_s x_s\right) - \|\theta\|_{V_T}^2 \\
&= \|\theta^*\|_{V_T}^2 - 2\left\|\theta^*\right\|_{V_T}^2 - 2(\theta^*)^\top \left(\sum_{s=1}^T \epsilon_s x_s\right) + 2(\theta^*)^\top V_T \theta + 2\theta^\top \left(\sum_{s=1}^T \epsilon_s x_s\right) - \|\theta\|_{V_T}^2 \\
&= -\left\|\theta^* - \theta\right\|_{V_T}^2 - 2\left\langle \theta^* - \theta, \sum_{s=1}^T \epsilon_s x_s \right\rangle \\
&= -\left\|\theta^* - \theta\right\|_{V_T}^2 - 2\sum_{s=1}^T \epsilon_s x_s^\top (\theta^* - \theta).
\end{aligned}$$

Note that

$$\begin{aligned}
\sum_{s=1}^T \epsilon_s x_s^\top (\theta^* - \theta) &= \sum_{s=1}^T \epsilon_s x_s^\top V_T^{-1/2} V_T^{1/2} (\theta^* - \theta) \\
&\leq \left\|\sum_{s=1}^T \epsilon_s x_s\right\|_{V_T^{-1}} \left\|\theta^* - \theta\right\|_{V_T}.
\end{aligned}$$

Note that

$$\left\|\theta^* - \theta\right\|_{V_T} = \sqrt{\left\|\theta^* - \theta\right\|_{V_T}^2} = \sqrt{\sum_{t=1}^T (x_t^\top (\theta^* - \theta))^2} \leq \Delta_{\max}\sqrt{T},$$

and since $\mathbb{E}[\epsilon_s x_s | \mathcal{F}_{s-1}] = 0$ for all s, $\epsilon_s x_s$ is a vector-valued martingale. Then by Theorem 1 of Abbasi-Yadkori et al. (2011), with probability greater than $1-\delta$,

$$\left\|\sum_{s=1}^T \epsilon_s x_s\right\|_{V_T^{-1}} \leq \sqrt{2d\log\left(\frac{d+TL^2}{d\delta}\right)}$$

so with probability $1-\delta$,

$$\left\|\sum_{s=1}^T \epsilon_s x_s\right\|_{V_T^{-1}} \left\|\theta^* - \theta\right\|_{V_T} \leq \Delta_{\max}\sqrt{T}\sqrt{2d\log\left(\frac{d+TL^2}{d\delta}\right)}.$$

so for any $\theta \in \Theta$,

$$\left|\left(\left\|\theta - \widehat{\theta}_{T+1}\right\|_{V_T}^2 - \left\|\theta^* - \widehat{\theta}_{T+1}\right\|_{V_T}^2\right) - \left\|\theta^* - \theta\right\|_{V_T}^2\right| \leq \Delta_{\max}\sqrt{T}\sqrt{2d\log\left(\frac{d+TL^2}{d\delta}\right)},$$

which means that

$$\left|\log\frac{\pi_\ell(\theta^*)}{\pi_\ell(\theta)} - \frac{T}{2}\left\|\theta - \theta^*\right\|_{A(\bar{e}_T)}^2\right| \leq \Delta_{\max}\sqrt{T}\sqrt{2d\log\left(\frac{d+TL^2}{d\delta}\right)}.$$

Taking a supremum over $\theta \in \Theta$ on both sides and taking $\delta = \frac{1}{\ell^2}$ gives the result.

□

C.2 Guarantee on Saddle-Point Convergence of PEPS in Round ℓ

In this section, we present a key result to this proof, which shows that as round ℓ gets large, the distribution from PEPS achieves the optimal allocation deduced by τ^*. Fix a round ℓ. At iteration t, let $\tilde{\lambda}_t$ denote the sampling distribution of x_t . The result is stated in the following lemma. In the proof, we decompose the difference into several terms and argue about each piece in subsequent sections.

Lemma C.4 (Guarantee for PEPS). *On $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}$, for $\ell > \ell_0$ then at the end of epoch ℓ, we have with probability at least $1 - \frac{1}{\ell^2}$,*

$$\tau^* - \inf_{\theta \in \Theta^c_{z_*}} \left[\frac{1}{2} \left\| \theta^* - \theta \right\|^2_{A(\bar{e}_{T_\ell})} \right] \leq \epsilon_\ell$$

for a sequence $\epsilon_\ell \to 0$ as $\ell \to \infty$.

Proof. Recall the definition of $\bar{p}_{T_\ell}$ and $\bar{e}_{T_\ell}$ in Section A. We first show that there exists some ϵ_ℓ that goes to zero as $\ell \to \infty$ such that under $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}$, for $\ell > \ell_0$,

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{F}_{\theta \sim \bar{p}_{T_\ell}} \left[\frac{1}{2} \left\| \theta^* - \theta \right\|^2_{A(\lambda)} \right] - \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} \left[\frac{1}{2} \left\| \theta^* - \theta \right\|^2_{A(\bar{e}_{T_\ell})} \right] \leq \epsilon_\ell.$$

We have

$$\begin{aligned}
&\max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{F}_{\theta \sim \bar{p}_{T_\ell}} \left[\left\| \theta^* - \theta \right\|^2_{A(\lambda)} \right] - \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} \left[\left\| \theta^* - \theta \right\|^2_{A(\bar{e}_{T_\ell})} \right] \\
&= \max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{F}_{\theta \sim \bar{p}_{T_\ell}} \left[\left\| \theta^* - \theta \right\|^2_{A(\lambda)} \right] - \inf_{\theta \in \Theta^c_{z_*}} \left\| \theta^* - \theta \right\|^2_{A(\bar{e}_{T_\ell})} \\
&= \max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{F}_{\theta \sim \bar{p}_{T_\ell}} \left[\left\| \theta^* - \theta \right\|^2_{A(\lambda)} \right] - \frac{1}{T_\ell} \inf_{\theta \in \Theta^c_{z_*}} \left\| \theta - \widehat{\theta}_{T_\ell + 1} \right\|^2_{V_{T_\ell}} + C''_{T_\ell} \\
&= \max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{F}_{\theta \sim \bar{p}_{T_\ell}} \left[\left\| \theta^* - \theta \right\|^2_{A(\lambda)} \right] - \max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\left\| \widehat{\theta}_t - \theta \right\|^2_{A(\lambda)} \right] && \text{(S1. } C'_{T_\ell}\text{)} \\
&\quad + \max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\left\| \widehat{\theta}_t - \theta \right\|^2_{A(\lambda)} \right] - \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\left\| \theta - \widehat{\theta}_t \right\|^2_{A(\tilde{\lambda}_t)} \right] && \text{(S2. regret for max learner)} \\
&\quad + \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\left\| \theta - \widehat{\theta}_t \right\|^2_{A(\tilde{\lambda}_t)} \right] - \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim \tilde{p}_t} \left[\left\| \theta - \widehat{\theta}_t \right\|^2_{A(\tilde{\lambda}_t)} \right] && \text{(S3.)} \\
&\quad + \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim \tilde{p}_t} \left[\left\| \theta - \widehat{\theta}_t \right\|^2_{A(\tilde{\lambda}_t)} \right] - \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim \tilde{p}_t} \left[\left\| \theta - \widehat{\theta}_t \right\|^2_{x_t x_t^\top} \right] && \text{(S4.)} \\
&\quad + \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim \tilde{p}_t} \left[\left\| \theta - \widehat{\theta}_t \right\|^2_{x_t x_t^\top} \right] - \frac{1}{T_\ell} \inf_{\theta \in \Theta^c_{z_*}} \left\| \theta - \widehat{\theta}_{T_\ell + 1} \right\|^2_{V_{T_\ell}} && \text{(S5. regret for the min learner)} \\
&+ C''_{T_\ell},
\end{aligned}$$

where we define

$$\begin{aligned}
C'_{T_\ell} &:= \max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{F}_{\theta \sim \bar{p}_{T_\ell}} \left[\left\| \theta^* - \theta \right\|^2_{A(\lambda)} \right] - \max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\left\| \widehat{\theta}_t - \theta \right\|^2_{A(\lambda)} \right] \\
C''_{T_\ell} &= \frac{1}{T_\ell} \inf_{\theta \in \Theta^c_{z_*}} \left\| \theta - \widehat{\theta}_{T_\ell + 1} \right\|^2_{V_{T_\ell}} - \inf_{\theta \in \Theta^c_{z_*}} \left\| \theta^* - \theta \right\|^2_{A(\bar{e}_{T_\ell})}.
\end{aligned}$$

We now handle each term separately by referring to the lemma which provides a guarantee.

- **(S1)** By Lemma C.10, under $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell}$, for $T_\ell \geq T_2(\ell)$,

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{F}_{\theta \sim \bar{p}_{T_\ell}} \left[\|\theta^* - \theta\|^2_{A(\lambda)} \right] - \max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\left\| \widehat{\theta}_t - \theta \right\|^2_{A(\lambda)} \right]$$
$$\leq \frac{T_2(\ell) L^2 \beta(T_2(\ell), \ell^2)}{T_\ell} + 4d\beta(T_\ell, \ell^2) T_\ell^{-3/4},$$

  so for $T_\ell \geq T_2(\ell)^{3/2}$, we have the above is upper bounded by

$$O(L^2 \beta(T_2(\ell), \ell^2) T_\ell^{-1/2} + 4d\beta(T_\ell, \ell^2) T_\ell^{-3/4});$$

- **(S2)** By Lemma C.5, we have with probability $1 - 1/(3\ell^2)$ conditioned on $\mathcal{E}_{2,\ell}$

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \lambda} \left\| \theta - \widehat{\theta}_t \right\|^2_{xx^\top} - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \tilde{\lambda}_t} \left\| \theta - \widehat{\theta}_t \right\|^2_{xx^\top}$$
$$\leq 2C_{3,\ell}^2 \sqrt{\log |\mathcal{X}| T_\ell} + \sqrt{2C_{3,\ell}^2 T_\ell |\mathcal{X}| \log(T_\ell \ell^2)} + 2C_{3,\ell}^2 \sum_{t=1}^{T_\ell} \gamma_t,$$

  so with a choice of $\gamma_t = t^{-\alpha}$ with $\alpha = 1/4$,

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \lambda} \left\| \theta - \widehat{\theta}_t \right\|^2_{xx^\top} - \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \tilde{\lambda}_t} \left\| \theta - \widehat{\theta}_t \right\|^2_{xx^\top}$$
$$\leq C_{3,\ell}^2 \sqrt{\log |\mathcal{X}|} T_\ell^{-1/2} + \sqrt{2C_{3,\ell}^2 \log \ell^2} T_\ell^{-1/2} + \sqrt{2C_{3,\ell}^2 |\mathcal{X}| \log(3T\ell^2)} T_\ell^{-1/2} + 2C_{3,\ell}^2 T_\ell^{-1/4}$$

- **(S3)** By Lemma C.12, we have conditioned on $\mathcal{E}_{4,\ell} \cap \mathcal{E}_{1,\ell}$ for $\ell \geq \ell_0$,

$$\frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\left\| \theta - \widehat{\theta}_t \right\|^2_{A(\tilde{\lambda}_t)} \right] - \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim \tilde{p}_t} \left[\left\| \theta - \widehat{\theta}_t \right\|^2_{A(\tilde{\lambda}_t)} \right] \leq \frac{2C_{3,\ell}^2 T_0(\ell)}{T_\ell}$$

  for $T_\ell \geq T_0(\ell)^{3/2}$, we have the above is bounded by $2C_{3,\ell}^2 T_\ell^{-1/2}$;

- **(S4)** By Lemma C.8, we have with probability $1 - 1/(3\ell^2)$, conditioned on $\mathcal{E}_{2,\ell}$,

$$\frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim \tilde{p}_t} \left[\left\| \theta - \widehat{\theta}_t \right\|^2_{A(\tilde{\lambda}_t)} \right] - \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim \tilde{p}_t} \left[\left\| \theta - \widehat{\theta}_t \right\|^2_{x_t x_t^\top} \right] \leq \sqrt{\frac{2C_{1,\ell} \log \ell^2}{T_\ell}}$$

- **(S5)** By Lemma C.7, we have with probability $1 - 1/(3\ell^2)$, conditioned on $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell}$,

$$\frac{1}{T_\ell} \left[\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta \sim \tilde{p}_t} \left[\left\| \theta - \widehat{\theta}_t \right\|^2_{x_t x_t^\top} \right] - \inf_{\theta \in \Theta^c_{z_*}} \left\| \theta - \widehat{\theta}_{T_\ell + 1} \right\|^2_{V_{T_\ell}} \right]$$
$$\leq \sqrt{\frac{C_{3,\ell}^2 d \log(T_\ell C_{1,\ell})}{T_\ell}} + C_{3,\ell} \sqrt{\frac{2d\beta(T_\ell, \ell^2)}{T_\ell} \log\left(\frac{d + T_\ell L^2}{d} \right)} + C_{3,\ell} \sqrt{\frac{2\log(\ell^2)}{T_\ell}}.$$

- (C''_{T_ℓ}) By Lemma C.11, conditioned on $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell}$, we have

$$\left| \frac{1}{T_\ell} \inf_{\theta \in \Theta^c_{z_*}} \left\| \theta - \widehat{\theta}_{T_\ell + 1} \right\|^2_{V_{T_\ell}} - \frac{1}{T_\ell} \inf_{\theta \in \Theta^c_{z_*}} \|\theta^* - \theta\|^2_{V_{T_\ell}} \right| \leq (C_{3,\ell} + \Delta_{\max}) \sqrt{\frac{\beta(T_\ell, \ell^2)}{T_\ell}}$$

Add them altogether, we get that with probability greater than $1 - 1/\ell^2$ on $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{4,\ell}$

$$\begin{aligned}
&\max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{E}_{\theta \sim \bar{p}_{T_\ell}} \left[\|\theta^* - \theta\|^2_{A(\lambda)} \right] - \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{E}_{\theta \sim p} \left[\|\theta^* - \theta\|^2_{A(\bar{\lambda}_{T_\ell})} \right] \\
&\leq L^2 \beta(T_2(\ell), \ell^2) T_\ell^{-1/2} + 4d\beta(T_\ell, \ell^2) T_\ell^{-3/4} \\
&\quad + C_{3,\ell}^2 \sqrt{\log|\mathcal{X}|} T_\ell^{-1/2} + \sqrt{2C_{3,\ell}^2 \log \ell^2} T_\ell^{-1/2} + \sqrt{2C_{3,\ell}^2 |\mathcal{X}| \log(3T\ell^2)} T_\ell^{-1/2} + C_{3,\ell}^2 T_\ell^{-1/4} \\
&\quad + 2C_{3,\ell}^2 T_\ell^{-1/2} + \sqrt{\frac{2C_{1,\ell} \log \ell^2}{T_\ell}} \\
&\quad + \sqrt{\frac{C_{3,\ell}^2 d \log(T_\ell C_{1,\ell})}{T_\ell}} + C_{3,\ell} \sqrt{\frac{2d\beta(T_\ell, \ell^2)}{T_\ell} \log\left(\frac{d + T_\ell L^2}{d} \right)} + C_{3,\ell} \sqrt{\frac{2\log(\ell^2)}{T_\ell}} \\
&\quad + (C_{3,\ell} + \Delta_{\max}) \sqrt{\frac{\beta(T_\ell, \ell^2)}{T_\ell}}.
\end{aligned}$$

Note that each term approaches zero as $T_\ell \to \infty$. By the choice of $T_\ell = 2^\ell$ in the algorithm, this implies that there exists some $\epsilon_\ell > 0$ with $\epsilon_\ell \to 0$ as $\ell \to \infty$ such that for each ℓ,

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{F}_{\theta \sim \bar{p}_{T_\ell}} \left[\|\theta^* - \theta\|^2_{A(\lambda)} \right] - \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} \left[\|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})} \right] \leq \epsilon_\ell. \tag{2}$$

Now we show how this result leads to the saddle point convergence. Note that

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{F}_{\theta \sim \bar{p}_{T_\ell}} \left[\|\theta^* - \theta\|^2_{A(\lambda)} \right] \geq \max_{\lambda \in \triangle_{\mathcal{X}}} \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} [\|\theta^* - \theta\|^2_{A(\lambda)}] \geq \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} \left[\|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})} \right],$$

so using Equation 2 we have

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} [\|\theta^* - \theta\|^2_{A(\lambda)}] - \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} \left[\|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})} \right] \leq \epsilon_\ell.$$

However, note that

$$\min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} \left[\|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})} \right] = \inf_{\theta \in \Theta^c_{z_*}} \|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})}$$

and $\max_{\lambda \in \triangle_{\mathcal{X}}} \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} [\|\theta^* - \theta\|^2_{A(\lambda)}] = \tau^*$, we have shown that

$$\tau^* - \inf_{\theta \in \Theta^c_{z_*}} \|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})} < \epsilon_\ell.$$

□

C.3 Guarantees on the max-learner

In this section, we show that the max-learner gets sublinear regret as ℓ gets large. The key idea is that we mix a diminishing amount of G-optimal distribution each round, and we show that by its diminishing nature, the mixing of G-optimal distribution keeps the regret sublinear.

Lemma C.5. *Under $\mathcal{E}_{\ell,2}$, with the choice of $\eta_\lambda = \sqrt{\frac{\log|\mathcal{X}|}{C_{3,\ell}^4 T}}$, we have with probability greater than $1 - 1/\ell^2$,*

$$\begin{aligned}
&\max_{\lambda \in \triangle_{\mathcal{X}}} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \lambda} \left\| \theta - \widehat{\theta}_t \right\|^2_{xx^\top} - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \tilde{\lambda}_t} \left\| \theta - \widehat{\theta}_t \right\|^2_{xx^\top} \\
&\leq 2C_{3,\ell}^2 \sqrt{\log|\mathcal{X}| T_\ell} + \sqrt{2C_{3,\ell}^2 T_\ell |\mathcal{X}| \log(T_\ell \ell^2)} + 2C_{3,\ell}^2 \sum_{t=1}^{T_\ell} \gamma_t.
\end{aligned}$$

Proof. We first show that the statement is true for some fixed λ, i.e. we would like to show that with probability $1-\delta$,

$$\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\tilde{\lambda}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]$$
$$\le C_{3,\ell}^2\sqrt{\log|\mathcal{X}|T_\ell} + \sqrt{2C_{3,\ell}^2 T_\ell \log(1/\delta)} + 2C_{3,\ell}^2\sum_{t=1}^{T_\ell}\gamma_t.$$

Let $\mathcal{F}_{t-1}$ be the history up to time t. Then for any fixed λ,

$$\mathbb{E}_{\theta_t}[\mathbb{F}_{x\sim\lambda}[\|\widehat{\theta}_t-\theta_t\|^2_{xx^\top}]|\mathcal{F}_{t-1}] = \mathbb{F}_{\theta\sim p_t, x\sim\lambda}[\|\widehat{\theta}_t-\theta\|^2_{xx^\top}].$$

Thus, setting

$$X_t = \mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \mathbb{F}_{x\sim\tilde{\lambda}_t,\theta\sim p_t}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}$$
$$-\left[\mathbb{F}_{x\sim\lambda}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \mathbb{F}_{x\sim\lambda,\theta\sim p_t}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]$$

we see that the X_t form a Martingale difference sequence, i.e. $\mathbb{E}[X_t|\mathcal{F}_{t-1}]=0$. Note that for any $\theta\in\Theta$,

$$\mathbb{F}_{x\sim\lambda_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]$$
$$= \mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] + \gamma_t\left(\mathbb{F}_{x\sim\lambda_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \mathbb{F}_{x\sim\lambda^G}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]\right),$$

Since under $\mathcal{E}_{2,\ell}$, we have for any $x\in\mathcal{X}$, $\theta\in\Theta$, any $t\le T_\ell$, $\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\le C_{3,\ell}^2$, we have for any $\theta\in\Theta$,

$$\mathbb{F}_{x\sim\lambda_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] \le \mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] + 2C_{3,\ell}^2\gamma_t.$$

Then we have

$$\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\tilde{\lambda}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]$$
$$= \sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\lambda}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right]$$
$$-\left[\sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\lambda}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right]\right.$$
$$\left.-\left[\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} - \sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\tilde{\lambda}_t,\theta\sim p_t}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]\right]$$
$$\le \sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\lambda}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\lambda_t}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right]$$
$$-\left[\sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\lambda}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right]\right.$$
$$\left.-\left[\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} - \sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\tilde{\lambda}_t,\theta\sim p_t}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]\right] + 2C_{3,\ell}^2\sum_{t=1}^{T_\ell}\gamma_t \tag{3}$$

Note that

$$\sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\lambda}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right]-\sum_{t=1}^{T_\ell}\mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right]$$
$$-\left[\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\lambda}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}-\sum_{t=1}^{T_\ell}\mathbb{F}_{x\sim\tilde{\lambda}_t,\theta\sim p_t}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]=\sum_{t=1}^{T_\ell}X_t.$$

We know that under $\mathcal{E}_{2,\ell}$, we have for any $x\in\mathcal{X}$, $\theta\in\Theta$, any $t\le T_\ell$, $\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\le C^2_{3,\ell}$. Then, for any t, $|X_t|\le 4C^2_{3,\ell}$, so by Azuma-Hoeffding, with probability $1-\delta$, $\sum_{t=1}^{T_\ell}X_t\le\sqrt{8C^2_{3,\ell}T_\ell\log(1/\delta)}$. Plugging the above and Lemma C.6 in Equation 3 gives us

$$\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\lambda}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]-\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\tilde{\lambda}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]$$
$$\le C^2_{3,\ell}\sqrt{\log|\mathcal{X}|T_\ell}+\sqrt{2C^2_{3,\ell}T_\ell\log(1/\delta)}+2C^2_{3,\ell}\sum_{t=1}^{T_\ell}\gamma_t.$$

This result holds for any λ, but in particular we want it to hold for the λ which maximizes the reward, so we perform a covering argument on λ.

We take an ϵ-cover $\mathcal{S}_\epsilon$ of $\triangle_\mathcal{X}$ in $\|\cdot\|_1$. Then, we know that for any $\lambda\in\triangle_\mathcal{X}$, there is some $\lambda'\in\mathcal{S}_\epsilon$ such that $\|\lambda-\lambda'\|_1\le\epsilon$. Let $w_t(\lambda):=\mathbb{F}_{\theta\sim p_t,x\sim\lambda}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}$. Then, note that for any t and $\lambda_1,\lambda_2\in\triangle_\mathcal{X}$,

$$\begin{aligned}w(\lambda_1)-w(\lambda_2)&=\mathbb{F}_{\theta\sim p_t,x\sim\lambda_1}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}-\mathbb{F}_{\theta\sim p_t,x\sim\lambda_2}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\\&=\mathbb{F}_{\theta\sim p_t}\sum_x([\lambda_1]_x-[\lambda_2]_x)(x^\top(\theta-\widehat{\theta}_t))^2\\&\le C^2_{3,\ell}\mathbb{F}_{\theta\sim p_t}\sum_x([\lambda_1]_x-[\lambda_2]_x)\\&=C^2_{3,\ell}\|\lambda_1-\lambda_2\|_1,\end{aligned}$$

so $w_t(\lambda)$ is $C^2_{3,\ell}$-Lipschitz for any t. Then, assuming that $\bar{\lambda}\in\triangle_\mathcal{X}$ satisfies that

$$\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\bar{\lambda}}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}=\max_{\lambda\in\triangle_\mathcal{X}}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\lambda}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top},$$

we can find some $\lambda_0\in\mathcal{S}_\epsilon$ such that $\left\|\lambda_0-\bar{\lambda}\right\|\le\epsilon$, so by Lipschitzness of w_t for any t, we have

$$\begin{aligned}&\max_{\lambda\in\triangle_\mathcal{X}}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\lambda}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}-\max_{\lambda\in\mathcal{S}_\epsilon}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\lambda}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\\&=\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\bar{\lambda}}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}-\max_{\lambda\in\mathcal{S}_\epsilon}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\lambda}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\\&\le\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\bar{\lambda}}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}-\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\lambda_0}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\\&\le C^2_{3,\ell}T_\ell\epsilon.\end{aligned}$$

Also, let $K=|\mathcal{X}|$. Denote B_1^K as the l_1 ball with dimension K. We know that for $\epsilon\le 1$, $N(B_1^K,\|\cdot\|_1,\epsilon)\le\left(\frac{3}{\epsilon}\right)^K$. Since $\triangle_\mathcal{X}\subset B_1^K$, we have the covering number

$$N(\triangle_\mathcal{X},\|\cdot\|_1,\epsilon)\le N(B_1^k,\|\cdot\|_1,\epsilon)\le\left(\frac{3}{\epsilon}\right)^K.$$

Therefore, $|\mathcal{S}_\epsilon| \le \left(\frac{3}{\epsilon}\right)^K$. By union bounding over all $\lambda \in \mathcal{S}_\epsilon$, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\max_{\lambda\in\mathcal{S}_\epsilon} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda} \left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda_t} \left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} \\
&\le C_{3,\ell}^2\sqrt{\log|\mathcal{X}|T_\ell} + \sqrt{2C_{3,\ell}^2 T_\ell \log(1/(\delta|\mathcal{S}_\epsilon|))} + 2C_{3,\ell}^2 \sum_{t=1}^{T_\ell}\gamma_t \\
&\le C_{3,\ell}^2\sqrt{\log|\mathcal{X}|T_\ell} + \sqrt{2C_{3,\ell}^2 T_\ell |\mathcal{X}| \log(3/(\epsilon\delta))} + 2C_{3,\ell}^2 \sum_{t=1}^{T_\ell}\gamma_t.
\end{aligned}$$

Combining two displays gives us

$$\begin{aligned}
&\max_{\lambda\in\triangle_\mathcal{X}} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda} \left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda_t} \left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} \\
&\le C_{3,\ell}^2\sqrt{\log|\mathcal{X}|T_\ell} + \sqrt{2C_{3,\ell}^2 T_\ell |\mathcal{X}| \log(3/(\delta\epsilon))} + 2C_{3,\ell}^2 \sum_{t=1}^{T_\ell}\gamma_t + C_{3,\ell}^2 T_\ell \epsilon.
\end{aligned}$$

Taking $\epsilon = 1/\sqrt{T_\ell}$ and $\delta = 1/\ell^2$ gives us the result. □

Lemma C.6. *Under $\mathcal{E}_{2,\ell}$, with the choice of $\eta = \sqrt{\frac{\log|\mathcal{X}|}{C_{3,\ell}^4 T_\ell}}$, we have for any λ,*

$$\sum_{t=1}^{T_\ell} \left\|\theta_t-\widehat{\theta}_t\right\|^2_{A(\lambda)} - \sum_{t=1}^{T_\ell} \left\|\theta_t-\widehat{\theta}_t\right\|^2_{A(\lambda_t)} \le C_{3,\ell}^2\sqrt{\log|\mathcal{X}|T_\ell}.$$

Proof. Let $\ell_t(\lambda) = -\left\|\theta_t-\widehat{\theta}_t\right\|^2_{A(\lambda)}$. Then we have

$$[\nabla_\lambda \ell_t(\lambda_t)]_x = -\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top} = \tilde{g}_{t,x}.$$

Since

$$\max_{t\in[T_\ell]} \|\tilde{g}_t\|_\infty = \max_{t\in[T_\ell], x\in\mathcal{X}} \left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top} \le C_{3,\ell}^2,$$

by the guarantee of exponentiated gradient algorithm Orabona (2019), we have that for any λ,

$$\sum_{t=1}^{T_\ell} [\ell_t(\lambda_t) - \ell_t(\lambda)] \le \frac{\log|\mathcal{X}|}{\eta} + \frac{\eta T_\ell}{2} C_{3,\ell}^4.$$

Plugging in the definition of $\ell_t(\lambda)$, we have

$$\sum_{t=1}^{T_\ell} \left\|\theta_t-\widehat{\theta}_t\right\|^2_{A(\lambda)} - \sum_{t=1}^{T_\ell} \left\|\theta_t-\widehat{\theta}_t\right\|^2_{A(\lambda_t)} \le \frac{\log|\mathcal{X}|}{\eta} + \frac{\eta T_\ell}{2} C_{3,\ell}^4.$$

Choosing $\eta = \sqrt{\frac{\log|\mathcal{X}|}{C_{3,\ell}^4 T_\ell}}$, we have

$$\sum_{t=1}^{T_\ell} \left\|\theta_t-\widehat{\theta}_t\right\|^2_{A(\lambda)} - \sum_{t=1}^{T_\ell} \left\|\theta_t-\widehat{\theta}_t\right\|^2_{A(\lambda_t)} \le C_{3,\ell}^2\sqrt{\log|\mathcal{X}|T_\ell}.$$

□

C.4 Guarantees on the min-learner

In this section, we show that the min-learner gets sublinear regret as ℓ gets large. For the min learner, we see that the update for the sampling distribution is very similar to the continuous exponential weights updates Bubeck (2011). The difference between our setting and continuous exponential weights is that the space $\Theta^c_{\widehat{z}_t}$ is changing each time, so we potentially have a changing action space each time. To overcome this challenge, we first analyze the regret guarantee when we assume access to the true alternative in Lemma C.7, and use Lemma C.16 to argue that the estimate $\Theta^c_{\widehat{z}_t}$ is good enough. We state the following guarantee for the min-learner.

Lemma C.7. *On event $\mathcal{E}_{\ell,1} \cap \mathcal{E}_{\ell,2}$, with probability $1 - 1/\ell^2$,*

$$\frac{1}{T_\ell}\left[\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top}\right] - \inf_{\theta\in\Theta^c_{z_*}}\left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right]$$
$$\le \sqrt{\frac{C_{3,\ell}^2 d\log(T_\ell C_{1,\ell})}{T_\ell}} + C_{3,\ell}\sqrt{\frac{2d\beta(T_\ell,\ell^2)}{T_\ell}\log\left(\frac{d+T_\ell L^2}{d}\right)} + C_{3,\ell}\sqrt{\frac{2\log(\ell^2)}{T_\ell}}.$$

Proof. We begin by a bound that will be useful in our exponential weights analogy. At iteration t, we apply Hoeffding's lemma with the following upper bound given $\mathcal{E}_{\ell,1} \cap \mathcal{E}_{\ell,2}$ and Lemma E.1,

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} + \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta-\widehat{\theta}_t\right\|^2_{V_t}\right] \\
&\le C_{3,\ell}^2 + \mathbb{E}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_{t-1}} - \left\|\theta-\widehat{\theta}_t\right\|^2_{V_{t-1}}\right] && (\mathcal{E}_{\ell,2}) \\
&\le C_{3,\ell}^2 + 2C_{3,\ell}(C_{1,\ell}+1) && (\text{Lemma E.1}) \\
&\le 4C_{3,\ell}^2.
\end{aligned}$$

At round $t > 1$, we define $W_t = \int_{\theta\in\Theta^c_{z_*}} \exp\left(-\eta_p\left\|\theta-\widehat{\theta}_t\right\|^2_{V_{t-1}}\right)d\theta$ and W_1 being a uniform distribution on $\Theta^c_{z_*}$. Then

$$\begin{aligned}
&\log\frac{W_{t+1}}{W_t} \\
&= \log\frac{\int_{\theta\in\Theta^c_{z_*}}\exp\left(-\eta_p\left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t}\right)d\theta}{\int_{\theta\in\Theta^c_{z_*}}\exp\left(-\eta_p\left\|\theta-\widehat{\theta}_t\right\|^2_{V_{t-1}}\right)d\theta} \\
&= \log\frac{\int_{\theta\in\Theta^c_{z_*}}\exp\left(-\eta_p\left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \eta_p\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} + \eta_p\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} + \eta_p\left\|\theta-\widehat{\theta}_t\right\|^2_{V_{t-1}} - \eta_p\left\|\theta-\widehat{\theta}_t\right\|^2_{V_{t-1}}\right)d\theta}{\int_{\theta\in\Theta^c_{z_*}}\exp\left(-\eta_p\left\|\theta-\widehat{\theta}_t\right\|^2_{V_{t-1}}\right)d\theta} \\
&= \log\frac{\int_{\theta\in\Theta^c_{z_*}}\exp\left(-\eta_p\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} - \eta_p\left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} + \eta_p\left\|\theta-\widehat{\theta}_t\right\|^2_{V_t} - \eta_p\left\|\theta-\widehat{\theta}_t\right\|^2_{V_{t-1}}\right)d\theta}{\int_{\theta\in\Theta^c_{z_*}}\exp\left(-\eta_p\left\|\theta-\widehat{\theta}_t\right\|^2_{V_{t-1}}\right)d\theta} \\
&\le -\eta_p\mathbb{E}_{\theta\sim p_t(\Theta^c_{z_*})}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} + \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta-\widehat{\theta}_t\right\|^2_{V_t}\right] + \frac{\eta_p^2\cdot 4C_{3,\ell}^2}{8}
\end{aligned}$$

where the inequality follows from the Hoeffding inequality $\ln\mathbb{E}e^{sX} \le s\mathbb{E}X + \frac{s^2(a-b)^2}{8}$. By telescoping, we have

$$\begin{aligned}
\log\frac{W_{T_\ell+1}}{W_1} &= \ln\frac{W_{T_\ell+1}}{W_{T_\ell}} + \ln\frac{W_{T_\ell}}{W_{T_\ell-1}} + \cdots + \ln\frac{W_2}{W_1} \\
&\le -\eta_p\sum_{t=1}^{T_\ell}\mathbb{E}_{\theta\sim p_t(\Theta^c_{z_*})}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} + \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta-\widehat{\theta}_t\right\|^2_{V_t}\right] + \frac{T_\ell\eta_p^2 C_{3,\ell}^2}{2}.
\end{aligned}$$

On the other hand, let $\tilde{\theta} = \arg\inf_{\theta\in\Theta^c_{z_*}} \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}$. Let $w_t(\theta) = \exp\left(-\eta_p \left\|\theta - \widehat{\theta}_t\right\|^2_{V_{t-1}}\right)$. Let $\mathcal{N}_\gamma := \{(1-\gamma)\tilde{\theta} + \gamma\theta, \theta \in \Theta^c_{z_*}\}$ for $\gamma > 0$ that we choose later. We have

$$
\begin{aligned}
\log\frac{W_{T_\ell+1}}{W_1} &= \log\left(\frac{\int_{\theta\in\Theta^c_{z_*}} \exp\left(-\eta_p \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right) d\theta}{\int_{\theta\in\Theta^c_{z_*}} 1 d\theta}\right) \\
&\geq \log\left(\frac{\int_{\theta\in\mathcal{N}_\gamma} \exp\left(-\eta_p \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right) d\theta}{\int_{\theta\in\Theta^c_{z_*}} 1 d\theta}\right) \\
&\geq \log\left(\frac{\int_{\theta\in\gamma\Theta^c_{z_*}} \exp\left(-\eta_p \left\|(1-\gamma)\tilde{\theta} + \theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right) d\theta}{\int_{\theta\in\Theta^c_{z_*}} 1 d\theta}\right) \\
&= \log\left(\frac{\int_{\theta\in\Theta^c_{z_*}} \gamma^d \exp\left(-\eta_p \left\|(1-\gamma)\tilde{\theta} + \gamma\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right) d\theta}{\int_{\theta\in\Theta^c_{z_*}} 1 d\theta}\right) \\
&= \log\left(\frac{\int_{\theta\in\Theta^c_{z_*}} \gamma^d \exp\left(-\eta_p \left\|(1-\gamma)\tilde{\theta} + \gamma\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right) d\theta}{\int_{\theta\in\Theta^c_{z_*}} 1 d\theta}\right) \\
&\geq \log\left(\frac{\int_{\theta\in\Theta^c_{z_*}} \gamma^d \exp\left(-\eta_p \left((1-\gamma)\left\|\tilde{\theta} - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} + \gamma\left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right)\right) d\theta}{\int_{\theta\in\Theta^c_{z_*}} 1 d\theta}\right) \\
&\geq \log\left(\frac{\int_{\theta\in\Theta^c_{z_*}} \gamma^d \exp\left(-\eta_p \left((1-\gamma)\left\|\tilde{\theta} - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} + \gamma\left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right)\right) d\theta}{\int_{\theta\in\Theta^c_{z_*}} 1 d\theta}\right) \\
&\geq \log\left(\frac{\int_{\theta\in\Theta^c_{z_*}} \gamma^d \exp\left(-\eta_p \left(\left\|\tilde{\theta} - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} + \gamma T_\ell C_{1,\ell}\right)\right) d\theta}{\int_{\theta\in\Theta^c_{z_*}} 1 d\theta}\right) \\
&= -\eta_p \inf_{\theta\in\Theta^c_{z_*}} \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} + d\log\gamma - \eta_p\gamma T_\ell C_{1,\ell}.
\end{aligned}
$$

where the last inequality follows from the fact that for any $\theta \in \Theta$,

$$
\left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} = \sum_{t=1}^{T_\ell} (x_t^\top(\theta - \widehat{\theta}_{T_\ell+1})^2 \leq T_\ell C^2_{3,\ell}
$$

under $\mathcal{E}_{2,\ell}$. Combining the two displays gives us

$$
\begin{aligned}
&-\eta_p \inf_{\theta\in\Theta^c_{z_*}} \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} + d\log\gamma - \eta_p\gamma T_\ell C_{1,\ell} \\
&\leq -\eta_p \sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{z_*})} \left[\left\|\theta - \widehat{\theta}_t\right\|^2_{x_t x_t^\top} + \left\|\theta - \widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta - \widehat{\theta}_t\right\|^2_{V_t}\right] + \frac{T_\ell \eta_p^2 C^2_{3,\ell}}{2}.
\end{aligned}
$$

Rearranging, we have

$$\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{z_*})}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} + \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta-\widehat{\theta}_t\right\|^2_{V_t}\right] - \inf_{\theta\in\Theta^c_{z_*}} \left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}$$
$$\le \frac{\eta_p C_{3,\ell}^2 T_\ell}{2} + \frac{d\log(1/\gamma)}{\eta_p} + \gamma T_\ell C_{1,\ell}.$$

By choosing $\gamma = \frac{1}{T_\ell C_{1,\ell}}$ and $\eta_p = \sqrt{\frac{d\log(T_\ell C_{1,\ell})}{C_{3,\ell}^2 T_\ell}}$, we have

$$\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{z_*})}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} + \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta-\widehat{\theta}_t\right\|^2_{V_t}\right] - \inf_{\theta\in\Theta^c_{z_*}} \left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} \le \sqrt{T_\ell C_{3,\ell}^2 d\log(T_\ell C_{1,\ell})},$$

so

$$\frac{1}{T_\ell}\left[\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{z_*})}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} + \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta-\widehat{\theta}_t\right\|^2_{V_t}\right] - \inf_{\theta\in\Theta^c_{z_*}} \left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right] \le \sqrt{\frac{C_{3,\ell}^2 d\log(T_\ell C_{1,\ell})}{T_\ell}}.$$

In other words,

$$\frac{1}{T_\ell}\left[\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim \tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top}\right] - \inf_{\theta\in\Theta^c_{z_*}} \left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right]$$
$$\le \sqrt{\frac{C_{3,\ell}^2 d\log(T_\ell C_{1,\ell})}{T_\ell}} + \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta-\widehat{\theta}_t\right\|^2_{V_t}\right].$$

By Lemma C.9, we have with probability $1-1/\ell^2$,

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta-\widehat{\theta}_t\right\|^2_{V_t}\right] \le C_{3,\ell}\sqrt{\frac{2d\beta(T_\ell,\ell^2)}{T_\ell}\log\left(\frac{d+T_\ell L^2}{d}\right)} + C_{3,\ell}\sqrt{\frac{2\log(\ell^2)}{T_\ell}}.$$

Combining the above two displays gives us with probability $1-1/\ell^2$,

$$\frac{1}{T_\ell}\left[\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim \tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top}\right] - \inf_{\theta\in\Theta^c_{z_*}} \left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right]$$
$$\le \sqrt{\frac{C_{3,\ell}^2 d\log(T_\ell C_{1,\ell})}{T_\ell}} + C_{3,\ell}\sqrt{\frac{2d\beta(T_\ell,\ell^2)}{T_\ell}\log\left(\frac{d+T_\ell L^2}{d}\right)} + C_{3,\ell}\sqrt{\frac{2\log(\ell^2)}{T_\ell}}.$$

□

C.5 Approximation Guarantees

In this section, we present several technical lemmas bounding the terms related to the approximation error of $\widehat{\theta}_t$ to θ^* in each iteration t. More specifically, these lemmas show upper bound on the terms in the decomposition in the proof of lemma C.4.

Lemma C.8 (S4). *Under $\mathcal{E}_{2,\ell}$, with probability $1-1/\ell^2$,*

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right] - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top}\right] \le \sqrt{\frac{2C_{1,\ell}\log\ell^2}{T_\ell}}.$$

Proof. Define $M_t = \mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top}\right]$. Note that

$$\mathbb{E}_{x_t}[M_t|\mathcal{F}_{t-1}] = \mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right],$$

so $\tilde{M}_t = M_t - \mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]$ is a mean-zero martingale. Also, under $\mathcal{E}_{2,\ell}$, $|M_t| \le C_{1,\ell}$, then by Azuma-Hoeffding, we have with probability at least $1-\frac{1}{\ell^2}$, $\sum_{t=1}^{T_\ell}\tilde{M}_t \le \sqrt{2C_{1,\ell}T_\ell\log\ell^2}$, so

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right] - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_tx_t^\top}\right] \le \sqrt{\frac{2C_{1,\ell}\log\ell^2}{T_\ell}}.$$

□

Lemma C.9 (C_{T_ℓ})**.** *Under $\mathcal{E}_{1,\ell}\cap\mathcal{E}_{2,\ell}$, with probability $1-1/\ell^2$,*

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta-\widehat{\theta}_t\right\|^2_{V_t}\right] \le C_{3,\ell}\sqrt{\frac{2d\beta(T_\ell,\ell^2)}{T_\ell}\log\left(\frac{d+T_\ell L^2}{d}\right)} + C_{3,\ell}\sqrt{\frac{2\log(\ell^2)}{T_\ell}}.$$

Proof. We first consider some round t and some θ. By Lemma E.1,

$$\left\|\theta-\widehat{\theta}_t\right\|^2_{V_{t-1}} - \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_{t-1}} \le 2C_{3,\ell}(y_t - x_t^\top\widehat{\theta}_t).$$

Therefore,

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{V_{t-1}} - \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_{t-1}}\right] \le \frac{2C_{3,\ell}}{T_\ell}\sum_{t=1}^{T_\ell}(y_t - x_t^\top\widehat{\theta}_t). \tag{4}$$

Now, note that

$$\begin{aligned} y_t - x_t^\top\widehat{\theta}_t &= x_t^\top(\theta^* - \widehat{\theta}_t) + \epsilon_t \\ &\le \|x_t\|_{V_{t-1}^{-1}}\left\|\theta^*-\widehat{\theta}_t\right\|_{V_{t-1}} + \epsilon_t \\ &\le \|x_t\|_{V_{t-1}^{-1}}\sqrt{\beta(t,\ell^2)} + \epsilon_t. && \text{(by } \mathcal{E}_{1,\ell}\text{)} \end{aligned}$$

Note that since $\epsilon_t \sim N(0,1)$ is 1-subGaussian, by Azuma-Hoeffding, we have with probability $1-1/\ell^2$,

$$\sum_{t=1}^{T_\ell}\epsilon_t \le \sqrt{2T_\ell\log(\ell^2)}.$$

By summing it from 1 to T_ℓ, we have under $\mathcal{E}_{1,\ell}$, with probability $1-1/\ell^2$,

$$\begin{aligned} \sum_{t=1}^{T_\ell}(y_t - x_t^\top\widehat{\theta}_t) &\le \sum_{t=1}^{T_\ell}\sqrt{\beta(t,\ell^2)}\,\|x_t\|_{V_{t-1}^{-1}} + \sum_{t=1}^{T_\ell}\epsilon_t \\ &\le \sum_{t=1}^{T_\ell}\sqrt{\beta(t,\ell^2)}\,\|x_t\|_{V_{t-1}^{-1}} + \sqrt{2T_\ell\log(\ell^2)} \\ &\le \sqrt{T_\ell\sum_{t=1}^{T_\ell}\beta(t,\ell^2)\,\|x_t\|^2_{V_{t-1}^{-1}}} + \sqrt{2T_\ell\log(\ell^2)} && \text{(by Cauchy-Schwarz)} \\ &\le \sqrt{T_\ell\beta(T_\ell,\ell^2)\sum_{t=1}^{T_\ell}\|x_t\|^2_{V_{t-1}^{-1}}} + \sqrt{2T_\ell\log(\ell^2)} && \text{(by Cauchy-Schwarz)} \\ &\le \sqrt{T_\ell\beta(T_\ell,\ell^2)2d\log\left(\frac{d+T_\ell L^2}{d}\right)} + \sqrt{2T_\ell\log(\ell^2)}. \\ & && \text{(by Elliptical potential lemma (Abbasi-Yadkori et al., 2011))} \end{aligned}$$

Plugging this in Equation 4 gives the result. □

Lemma C.10 (C'_{T_ℓ}). *Under $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell}$, we have*

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\|\theta^* - \theta\|^2_{A(\lambda)} \right] - \max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\left\| \widehat{\theta}_t - \theta \right\|^2_{A(\lambda)} \right]$$
$$\leq \frac{T_2(\ell) L^2 \beta(T_2(\ell), \ell^2)}{T_\ell} + 4d\beta(T_\ell, \ell^2) T_\ell^{-3/4}.$$

Proof. We have

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\|\theta^* - \theta\|^2_{A(\lambda)} \right] - \max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\left\| \widehat{\theta}_t - \theta \right\|^2_{A(\lambda)} \right]$$
$$\leq \max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\|\theta^* - \theta\|^2_{A(\lambda)} - \left\| \widehat{\theta}_t - \theta \right\|^2_{A(\lambda)} \right].$$

We fix some θ and λ. Note that

$$\begin{aligned}
&\|\theta^* - \theta\|^2_{A(\lambda)} - \left\| \widehat{\theta}_t - \theta \right\|^2_{A(\lambda)} \\
&= (\theta^* + \widehat{\theta}_t - 2\theta)^\top A(\lambda)(\theta^* - \widehat{\theta}_t) \\
&= \sum_{x \in \mathcal{X}} \lambda_x (\theta^* + \widehat{\theta}_t - 2\theta)^\top x x^\top (\theta^* - \widehat{\theta}_t) \\
&\leq \max_{x \in \mathcal{X}} (\theta^* + \widehat{\theta}_t - 2\theta)^\top x x^\top (\theta^* - \widehat{\theta}_t) \\
&\leq (C_{3,\ell} + \Delta_{\max}) \max_{x \in \mathcal{X}} x^\top (\theta^* - \widehat{\theta}_t).
\end{aligned}$$

Therefore,

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\|\theta^* - \theta\|^2_{A(\lambda)} - \left\| \widehat{\theta}_t - \theta \right\|^2_{A(\lambda)} \right] \leq (C_{3,\ell} + \Delta_{\max}) \max_{x \in \mathcal{X}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \left\langle \widehat{\theta}_t - \theta^*, x \right\rangle. \tag{5}$$

By Lemma C.15, under $\mathcal{E}_{3,\ell} \cap \mathcal{E}_{1,\ell}$, for any $t \geq T_2(\ell) + 1$, we have for any $x \in \mathcal{X}$,

$$\left\langle x, \widehat{\theta}_t - \theta^* \right\rangle \leq \frac{d}{t^{3/4}} \beta(t, \ell^2).$$

Also, by Lemma D.2, under $\mathcal{E}_{1,\ell}$, we have for any $t \geq 1$,

$$\left\langle x, \widehat{\theta}_t - \theta^* \right\rangle \leq L^2 \beta(t, \ell^2).$$

Therefore,

$$\begin{aligned}
&\max_{x \in \mathcal{X}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \left\langle \widehat{\theta}_t - \theta^*, x \right\rangle \\
&\leq \max_{x \in \mathcal{X}} \frac{1}{T_\ell} \left[\sum_{t=1}^{T_2(\ell)} \left\langle \widehat{\theta}_t - \theta^*, x \right\rangle + \sum_{t=T_2(\ell)+1}^{T_\ell} \left\langle \widehat{\theta}_t - \theta^*, x \right\rangle \right] \\
&\leq \frac{1}{T_\ell} \left[T_2(\ell) L^2 \beta(T_2(\ell), \ell^2) + \sum_{t=T_2(\ell)+1}^{T_\ell} \frac{d}{t^{3/4}} \beta(t, \ell^2) \right] && \text{(by Lemma D.2 and C.15)} \\
&\leq \frac{1}{T_\ell} \left[T_2(\ell) L^2 \beta(T_2(\ell), \ell^2) + d\beta(T_\ell, \ell^2) \int_{t=T_2(\ell)}^{T_\ell} t^{-3/4} dt \right] \\
&= \frac{1}{T_\ell} \left[T_2(\ell) L^2 \beta(T_2(\ell), \ell^2) + d\beta(T_\ell, \ell^2)(4T_\ell^{1/4} - 4T_2(\ell)^{1/4}) \right] \\
&\leq \frac{T_2(\ell) L^2 \beta(T_2(\ell), \ell^2)}{T_\ell} + 4d\beta(T_\ell, \ell^2) T_\ell^{-3/4}.
\end{aligned}$$

Plugging this in Equation 5 gives us

$$\max_{\lambda\in\triangle_{\mathcal{X}}}\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t}\left[\|\theta^*-\theta\|^2_{A(\lambda)}-\left\|\widehat{\theta}_t-\theta\right\|^2_{A(\lambda)}\right]\le\frac{T_2(\ell)L^2\beta(T_2(\ell),\ell^2)}{T_\ell}+4d\beta(T_\ell,\ell^2)T_\ell^{-3/4}.$$

□

Lemma C.11 (C''_{T_ℓ})**.** *Assume that Θ is closed. Then, we have under $\mathcal{E}_{1,\ell}\cap\mathcal{E}_{2,\ell}$,*

$$\left|\frac{1}{T_\ell}\inf_{\theta\in\Theta^c_{z_*}}\left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}-\frac{1}{T_\ell}\inf_{\theta\in\Theta^c_{z_*}}\|\theta^*-\theta\|^2_{V_{T_\ell}}\right|\le(C_{3,\ell}+\Delta_{\max})\sqrt{\frac{\beta(T_\ell,\ell^2)}{T_\ell}}.$$

Proof. Let $\theta_1:=\arg\inf_{\theta\in\Theta^c_{z_*}}\left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}$ and $\theta_2:=\arg\inf_{\theta\in\Theta^c_{z_*}}\|\theta-\theta^*\|^2_{V_{T_\ell}}$. We have

$$\begin{aligned}
&\inf_{\theta\in\Theta^c_{z_*}}\left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}-\inf_{\theta\in\Theta^c_{z_*}}\|\theta^*-\theta\|^2_{V_{T_\ell}}\\
&\le\left\|\widehat{\theta}_{T_\ell+1}-\theta_2\right\|^2_{V_{T_\ell}}-\|\theta^*-\theta_2\|^2_{V_{T_\ell}}\\
&=\left(\left\|\widehat{\theta}_{T_\ell+1}-\theta_2\right\|_{V_{T_\ell}}-\|\theta^*-\theta_2\|_{V_{T_\ell}}\right)\left(\left\|\widehat{\theta}_{T_\ell+1}-\theta_2\right\|_{V_{T_\ell}}+\|\theta^*-\theta_2\|_{V_{T_\ell}}\right).\\
&\le\left\|\widehat{\theta}_{T_\ell+1}-\theta_*\right\|_{V_{T_\ell}}\left(\left\|\widehat{\theta}_{T_\ell+1}-\theta_2\right\|_{V_{T_\ell}}+\|\theta^*-\theta_2\|_{V_{T_\ell}}\right).
\end{aligned}$$

Note that under $\mathcal{E}_{2,\ell}$,

$$\begin{aligned}
\left\|\widehat{\theta}_{T_\ell+1}-\theta_1\right\|_{V_{T_\ell}}&=\sqrt{\sum_{t=1}^{T_\ell}(x_t^\top(\widehat{\theta}_{T_\ell+1}-\theta_1))^2}\le C_{3,\ell}\sqrt{T_\ell};\\
\|\theta^*-\theta_2\|_{V_{T_\ell}}&=\sqrt{\sum_{t=1}^{T_\ell}(x_t^\top(\theta^*-\theta_2))^2}\le\Delta_{\max}\sqrt{T_\ell}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\inf_{\theta\in\Theta^c_{z_*}}\left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}-\inf_{\theta\in\Theta^c_{z_*}}\|\theta^*-\theta\|^2_{V_{T_\ell}}\\
&\le(C_{3,\ell}+\Delta_{\max})\sqrt{T_\ell}\left\|\widehat{\theta}_{T_\ell+1}-\theta^*\right\|_{V_{T_\ell}}\\
&\le(C_{3,\ell}+\Delta_{\max})\sqrt{T_\ell\beta(T_\ell,\ell^2)}. &\text{(by } \mathcal{E}_{1,\ell})
\end{aligned}$$

□

We use the above lemma to bound the term that relates $\tilde{p}_t$ to p_t.

Lemma C.12 ($\tilde{p}_t$ to p_t)**.** *Under $\mathcal{E}_{2,\ell}\cap\mathcal{E}_{4,\ell}$ for $T_\ell\ge T_0$,*

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]-\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]\le\frac{2C_{3,\ell}^2T_0(\ell)}{T_\ell}.$$

Proof. Note that $\tilde{p}_t = p_t$ under $\mathcal{E}_{4,\ell}$,

$$\begin{aligned}
&\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\left(\mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]-\mathbb{F}_{\theta\sim \tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]\right)\\
&=\frac{1}{T_\ell}\sum_{t=1}^{T_0(\ell)}\left(\mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]-\mathbb{F}_{\theta\sim \tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]\right)\\
&\quad+\frac{1}{T_\ell}\sum_{t=T_0(\ell)+1}^{T_\ell}\left(\mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]-\mathbb{F}_{\theta\sim \tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]\right)\\
&=\frac{1}{T_\ell}\sum_{t=1}^{T_0(\ell)}\left(\mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]-\mathbb{F}_{\theta\sim \tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]\right).
\end{aligned}$$

Since for any $\theta\in\Theta$, under $\mathcal{E}_{2,\ell}$,

$$\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}=\sum_{x\in\mathcal{X}}\tilde{\lambda}_{t,x}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\le\max_{x\in\mathcal{X}}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\le C^2_{3,\ell},$$

we have

$$\frac{1}{T_\ell}\sum_{t=1}^{T_0(\ell)}\left(\mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]-\mathbb{F}_{\theta\sim \tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]\right)\le\frac{2C^2_{3,\ell}T_0(\ell)}{T_\ell}.$$

□

C.6 Guarantees on sampling and learning the estimate

In this section we provide some general guarantees on sampling together with a threshold after which each arm gets enough samples and . Consider a setting where at each time we receive a distribution $\tilde{\lambda}=(1-\gamma_t)\lambda_t+\gamma_t P$ for a fixed distribution P.

Lemma C.13. *Fix a distribution P on $\mathcal{X}$ with full support. On an event that is true with probability greater than $1-\delta$, for any $0<\alpha<1/2$ there exists a $T_1:=T_1(\alpha,\delta,T)$ such that for any $t\ge T_1$,*

$$V_t\ge\frac{c}{1-\alpha}A(P)t^{1-\alpha}.$$

Proof. Fix $x\in\mathcal{X}$, let $N_{t,x}=\sum_{s=1}^t Z_s$ where $Z_s=1$ if $x_s=x$ else 0. Then, $V_t=\sum_{x\in\mathcal{X}}\sum_{s=1}^t Z_s xx^\top$. We assume that $\gamma_s=1/s^\alpha, s\ge 1$.

Note that $\mathbb{P}(Z_s = 1|\mathcal{F}_{s-1}) = (1-\gamma_s)\lambda_{s,x} + \gamma_s P_x$. So for $t > 1$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{s=1}^{t} Z_s \le cP_x \sum_{s=1}^{t}\gamma_s\right) &= \mathbb{P}\left(\sum_{s=1}^{t} Z_s - (1-\gamma_s)\lambda_{s,x} - \gamma_s P_x \le \sum_{s=1}^{t} cP_x\gamma_s - (1-\gamma_s)\lambda_{s,x} - \gamma_s P_x\right) \\
&= \mathbb{P}\left(\sum_{s=1}^{t} Z_s - (1-\gamma_s)\lambda_{s,x} - \gamma_s P_x \le \sum_{s=1}^{t} (c-1)P_x\gamma_s - (1-\gamma_s)\lambda_{s,x}\right) \\
&\le \mathbb{P}\left(\sum_{s=1}^{t} Z_s - (1-\gamma_s)\lambda_{s,x} - \gamma_s P_x \le \sum_{s=1}^{t} (c-1)P_x\gamma_s\right) \\
&\le \mathbb{P}\left(\sum_{s=1}^{t} Z_s - (1-\gamma_s)\lambda_{s,x} - \gamma_s P_x \le -\sum_{s=1}^{t} (1-c)P_x\gamma_s\right) \\
&\le \exp\left(-\frac{1}{t}\left(\sum_{s=1}^{t}(1-c)P_x\gamma_s\right)^2\right) && \text{(Azuma-Hoeffding)} \\
&= \exp\left(-\left(\frac{(1-c)P_x}{\sqrt{t}}\sum_{s=1}^{t}\gamma_s\right)^2\right) \\
&\le \exp\left(-\left(\frac{(1-c)P_x}{\sqrt{t}}\frac{t^{1-\alpha}-1}{1-\alpha}\right)^2\right) && \left(\textstyle\sum_{s=1}^{t}\frac{1}{s^\alpha} \ge \frac{t^{1-\alpha}-1}{1-\alpha}\right) \\
&\le \exp\left(-\left((1-c)P_x\frac{t^{1/2-\alpha}-t^{-1/2}}{1-\alpha}\right)^2\right) \\
&\le \exp\left(-\left(\frac{(1-c)P_x}{2(1-\alpha)}t^{1/2-\alpha}\right)^2\right) && \left(t^{1/2-\alpha}-t^{-1/2} > \tfrac{1}{2}t^{1/2-\alpha},\ t\ge 2\right) \\
&\le \exp\left(-\left(\frac{(1-c)P_x}{2(1-\alpha)}\right)^2 t^{1-2\alpha}\right)
\end{aligned}$$

This implies that with the sequence $\gamma_s = 1/s^\alpha, \alpha < 1/2$ (to ensure $1-2\alpha > 0$), with probability greater than $1-\delta$ we have

$$N_{t,x} = \sum_{s=1}^{t} Z_s \ge cP_x\sum_{s=1}^{t}\gamma_s \ge \frac{cP_x}{1-\alpha}(t^{1-\alpha}-1) \quad \text{whenever} \quad t \ge \left(\frac{2(1-\alpha)\sqrt{\log(1/\delta)}}{(1-c)P_x}\right)^{\frac{2}{1-2\alpha}}.$$

□

The lemma below states that there exists some time T_2 such that all the arms get enough samples.

Lemma C.14. *For $T_2(\ell) = \max_{x\in\mathcal{X}}\left(\frac{6\sqrt{\log(|\mathcal{X}|T_\ell\ell^2)}}{\lambda_x^G}\right)^4$, we have*

$$\mathbb{P}\left(\mathcal{E}_{3,\ell}\right) \ge 1 - 1/\ell^2.$$

Proof. By Lemma C.13 with a choice of $c = 1-\alpha$, $\alpha = \frac{1}{4}$, $\delta = \frac{1}{|\mathcal{X}|T_\ell\ell^2}$, and $P = \lambda^G$, we have for any $t \ge \left(\frac{2(1-\alpha)\sqrt{\log(1/\delta)}}{(1-c)P_x}\right)^{\frac{2}{1-2\alpha}} = \left(\frac{6\sqrt{\log(|\mathcal{X}|T_\ell\ell^2)}}{\lambda_x^G}\right)^4$, we have $\mathbb{P}(V_t \ge t^{3/4}A(\lambda^G)) \ge 1 - \frac{1}{|\mathcal{X}|T_\ell\ell^2}$. Let $T_2(\ell) := \max_{x\in\mathcal{X}}\left(\frac{6\sqrt{\log(|\mathcal{X}|T_\ell\ell^2)}}{\lambda_x^G}\right)^4$, union bounding for $t \in [T_2, T_\ell]$ and $x \in \mathcal{X}$ gives the result. □

Lemma C.15. *Under $\mathcal{E}_{3,\ell}\cap\mathcal{E}_{1,\ell}$, for any $t \ge T_2(\ell)+1$, we have for any $x \in \mathcal{X}$,*

$$\left\langle x, \widehat{\theta}_t - \theta^*\right\rangle \le \frac{d}{t^{3/4}}\beta(t,\ell^2).$$

Proof. Let $N_{t,x}$ be the number of times arm x gets pulled at round t. By Lemma C.14, for $t \geq T_2(\ell)+1$, under $\mathcal{E}_{3,\ell}$, we have

$$V_{t-1} = \sum_{x\in\mathcal{X}} N_{t-1,x} x x^\top \geq t^{3/4} A(\lambda^G).$$

Therefore, for any $x \in \mathcal{X}$,

$$\|x\|^2_{V_{t-1}^{-1}} \leq \frac{1}{t^{3/4}} \|x\|^2_{A(\lambda^G)^{-1}} \leq \frac{d}{t^{3/4}}$$

by Kiefer-Wolfowitz. Therefore, under $\mathcal{E}_{1,\ell}$, for any $x \in \mathcal{X}$,

$$\begin{aligned}
\left\langle x, \widehat{\theta}_t - \theta^* \right\rangle &\leq \|x\|^2_{V_{t-1}^{-1}} \left\|\widehat{\theta}_t - \theta^*\right\|^2_{V_{t-1}} \\
&\leq \frac{d}{t^{3/4}} \left\|\widehat{\theta}_t - \theta^*\right\|^2_{V_{t-1}} \\
&\leq \frac{d}{t^{3/4}} \beta(t,\ell^2).
\end{aligned}$$

□

The following lemma provides a guarantee that we eventually finds z_*.

Lemma C.16. *For* $T_0(\ell) = \max\left\{\left(\frac{d\beta(T_\ell,\ell^2)\max_{z\in\mathcal{Z}}\|z\|_1}{\Delta_{\min}}\right)^{4/3}, T_2(\ell)+1\right\}$, *we have* $\mathbb{P}(\mathcal{E}_{4,\ell}|\mathcal{E}_{1,\ell}\cap\mathcal{E}_{3,\ell}) \geq 1-1/\ell^2$.

Proof. By Lemma C.15, we know that for any $t \geq T_2(\ell)+1$, under $\mathcal{E}_{1,\ell}\cap\mathcal{E}_{3,\ell}$ we have for any $x \in \mathcal{X}$,

$$\left\langle x, \widehat{\theta}_t - \theta^* \right\rangle \leq \frac{d}{t^{3/4}} \beta(t,\ell^2).$$

Since the span of $\mathcal{Z}$ is in the subset of $\mathcal{X}$, for any $z \in \mathcal{Z}$, we write $z_* - z = \sum_{x\in\mathcal{X}} \alpha_{z,x} x$. Then

$$\begin{aligned}
(z_*-z)^\top(\theta_* - \widehat{\theta}_t) &= \sum_{x\in\mathcal{X}} \alpha_{z,x} x^\top(\theta_* - \widehat{\theta}_t) \\
&\leq \sum_{x\in\mathcal{X}} \alpha_{z,x} \frac{d}{t^{3/4}} \beta(t,\ell^2) \\
&\leq \max_{z\in\mathcal{Z}} \|z\|_1 \frac{d}{t^{3/4}} \beta(t,\ell^2).
\end{aligned}$$

Then, for any $t > \left(\frac{d\beta(t,\ell^2)\max_{z\in\mathcal{Z}}\|z\|_1}{\Delta_{\min}}\right)^{4/3}$, we have

$$\max_{z\in\mathcal{Z}} \|z\|_1 \frac{d}{t^{3/4}} \beta(t,\ell^2) < \Delta_{\min},$$

which implies that for any z,

$$\begin{aligned}
&(z_*-z)^\top(\theta_* - \widehat{\theta}_t) < \Delta_{\min} \\
\Rightarrow &(z_*-z)^\top(\widehat{\theta}_t - \theta_*) > -\Delta_{\min} \\
\Rightarrow &(z_*-z)^\top \widehat{\theta}_t > 0,
\end{aligned}$$

which implies that $\widehat{z}_t = z_*$. □

D BOUNDS AND EVENTS THAT HOLD TRUE EACH ROUND

The following lemma states an anytime confidence bound for the least-squares estimator. It is a restatement of Theorem 20.5 of Lattimore and Szepesvári (2020) in our setting.

Lemma D.1 ($\mathcal{E}_{1,\ell}$)**.** *With probability $1 - 1/\ell^2$, for all t, we have*

$$\left\|\widehat{\theta}_t - \theta^*\right\|_{V_{t-1}}^2 \leq B + \sqrt{2\log(\ell^2) + d\log\left(\frac{d+tL^2}{d}\right)}.$$

Proof. Follows from Theorem 20.5 of Lattimore and Szepesvári (2020). □

Lemma D.2 ($\mathcal{E}_{2,\ell}$)**.** *Under $\mathcal{E}_{1,\ell}$, we have for any $x \in \mathcal{X}$ and any $t \in [1, T_\ell]$, $\left\langle x, \widehat{\theta}_t \right\rangle \leq \Delta_{\max} + L^2\beta(T_\ell, \ell^2)$.*

Proof. For any $x \in \mathcal{X}$,

$$\begin{aligned}
\left\langle x, \widehat{\theta}_t \right\rangle &= \langle x, \theta^* \rangle + \left\langle x, \widehat{\theta}_t - \theta^* \right\rangle \\
&\leq \Delta_{\max} + \|x\|_{V_{t-1}^{-1}}^2 \left\|\widehat{\theta}_t - \theta^*\right\|_{V_{t-1}}^2 \\
&\leq \Delta_{\max} + \|x\|_{V_{t-1}^{-1}}^2 \beta(t, \ell^2). && \text{(under } \mathcal{E}_{1,\ell}\text{)}
\end{aligned}$$

Since we have

$$V_{t-1} = V_0 + \sum_{s=1}^{t-1} x_s x_s^\top,$$

for $V_0 = I$, we have the minimum eigenvalue $\sigma_{\min}(V_{t-1}) \geq \sigma_{\min}(V_0) + \sigma_{\min}\left(\sum_{s=1}^{t-1} x_s x_s^\top\right) \geq 1$, so

$$\sigma_{\max}(V_{t-1}^{-1}) = \frac{1}{\sigma_{\min}(V_{t-1})} \leq 1,$$

which implies that

$$\max_{x \in \mathcal{X}} \|x\|_{V_{t-1}^{-1}}^2 \leq \sigma_{\max}(V_{t-1}^{-1}) \max_{x \in \mathcal{X}} \|x\|_2^2 \leq L^2.$$

Therefore,

$$\left\langle x, \widehat{\theta}_t \right\rangle \leq \Delta_{\max} + L^2\beta(t, \ell^2) \leq \Delta_{\max} + L^2\beta(T_\ell, \ell^2).$$

□

E TECHNICAL LEMMAS

Lemma E.1 (Recursive Least Squares Guarantee)**.** *In any round ℓ, conditional on event $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell}$, for any $\theta \in \Theta$ and any $t \in [1, T_\ell]$ we have*

$$\left\|\theta - \widehat{\theta}_{t+1}\right\|_{V_t}^2 - \left\|\theta - \widehat{\theta}_t\right\|_{V_t}^2 \leq 2C_{3,\ell}(y_t - x_t^\top \widehat{\theta}_t) \leq 2C_{3,\ell}(C_{1,\ell} + 1),$$

assuming that all rewards are bounded in $[-1, 1]$.

Proof. We first consider some round t and some θ. Note that $\widehat{\theta}_t = V_t^{-1} X_t^\top Y_t$. Then

$$\begin{aligned}
\widehat{\theta}_{t+1} &= (V_{t-1} + x_t x_t^\top)^{-1}(X_{t-1}^\top Y_{t-1} + x_t y_t) \\
&= \left(V_{t-1}^{-1} - \frac{V_{t-1}^{-1} x_t x_t^\top V_{t-1}^{-1}}{1 + x_t^\top V_{t-1}^{-1} x_t}\right)(X_{t-1}^\top Y_{t-1} + x_t y_t) \\
&= \widehat{\theta}_t - \frac{V_{t-1}^{-1} x_t x_t^\top \widehat{\theta}_t}{1 + x_t^\top V_{t-1}^{-1} x_t} + V_{t-1}^{-1} x_t y_t - \frac{V_{t-1}^{-1} x_t x_t^\top V_{t-1}^{-1} x_t y_t}{1 + x_t^\top V_{t-1}^{-1} x_t} \\
&= \widehat{\theta}_t - \frac{V_{t-1}^{-1} x_t x_t^\top \widehat{\theta}_t}{1 + x_t^\top V_{t-1}^{-1} x_t} + \frac{V_{t-1}^{-1} x_t y_t (1 + x_t^\top V_{t-1}^{-1} x_t) - x_t^\top V_{t-1}^{-1} x_t V_{t-1}^{-1} x_t y_t}{(1 + x_t^\top V_{t-1}^{-1} x_t)} \\
&= \widehat{\theta}_t - \frac{V_{t-1}^{-1} x_t x_t^\top \widehat{\theta}_t}{1 + x_t^\top V_{t-1}^{-1} x_t} + \frac{V_{t-1}^{-1} x_t y_t}{(1 + x_t^\top V_{t-1}^{-1} x_t)} \\
&= \widehat{\theta}_t + \frac{V_{t-1}^{-1} x_t (y_t - x_t^\top \widehat{\theta}_t)}{1 + x_t^\top V_{t-1}^{-1} x_t}
\end{aligned}$$

Hence

$$\widehat{\theta}_{t+1} - \widehat{\theta}_t = \frac{V_{t-1}^{-1} x_t}{1 + x_t^\top V_{t-1}^{-1} x_t}(y_t - x_t^\top \widehat{\theta}_t)$$

and

$$\begin{aligned}
V_t(\widehat{\theta}_{t+1} - \widehat{\theta}_t) &= \frac{V_t V_{t-1}^{-1} x_t}{1 + x_t^\top V_{t-1}^{-1} x_t}(y_t - x_t^\top \widehat{\theta}_t) \\
&= \frac{(I + x_t x_t^\top V_{t-1}^{-1}) x_t}{1 + x_t^\top V_{t-1}^{-1} x_t}(y_t - x_t^\top \widehat{\theta}_t) \\
&= \frac{x_t (1 + x_t^\top V_{t-1}^{-1} x_t)}{1 + x_t^\top V_{t-1}^{-1} x_t}(y_t - x_t^\top \widehat{\theta}_t) \\
&= (y_t - x_t^\top \widehat{\theta}_t) x_t
\end{aligned}$$

Then

$$\begin{aligned}
&\left\|\theta - \widehat{\theta}_{t+1}\right\|_{V_t}^2 - \left\|\theta - \widehat{\theta}_t\right\|_{V_t}^2 \\
&= (\widehat{\theta}_{t+1} - \widehat{\theta}_t)^\top V_t (\widehat{\theta}_{t+1} + \widehat{\theta}_t - 2\theta) \\
&= (y_t - x_t^\top \widehat{\theta}_t) x_t^\top (\widehat{\theta}_{t+1} + \widehat{\theta}_t - 2\theta) \\
&\le 2C_{3,\ell}(y_t - x_t^\top \widehat{\theta}_t) \\
&\le 2C_{3,\ell}(C_{1,\ell} + 1)
\end{aligned}$$

assuming all rewards are bounded by 1. □

Lemma E.2. *For any open set $\tilde{\Theta} \subset \Theta$, we have*

$$\int_{\tilde{\Theta}} \exp\left(-\frac{T_\ell}{2}\left(\|\theta^* - \theta\|_{A(\bar{e}_{T_\ell})}^2\right)\right) d\theta \doteq \exp\left(-\frac{T_\ell}{2} \inf_{\theta \in \tilde{\Theta}} \|\theta^* - \theta\|_{A(\bar{e}_{T_\ell})}^2\right).$$

Proof. The following argument is inspired by an analogous one in Lemma 11 of Russo (2016). Let $\iota_\ell := \int_{\tilde{\Theta}} \exp\left(-\frac{T_\ell}{2}\|\theta^* - \theta\|_{A(\bar{e}_{T_\ell})}^2\right) d\theta$ and $W_{T_\ell}(\theta) := \frac{1}{2}\|\theta^* - \theta\|_{A(\bar{e}_{T_\ell})}^2$. Also, let $\tilde{\theta}_\ell \in \text{closure}(\tilde{\Theta})$ be a point that attains the infimum, i.e.

$$\tilde{\theta}_\ell := \arg\inf_{\theta \in \tilde{\Theta}} \|\theta^* - \theta\|_{A(\bar{e}_{T_\ell})}^2.$$

Such a point must exist by the continuity of $W_{T_\ell}(\theta)$ and closure$(\tilde{\Theta})$ being compact. Then, we first observe that

$$\int_{\tilde{\Theta}} \exp\left(-\frac{T_\ell}{2}\left\|\theta^* - \theta\right\|^2_{A(\bar{e}_{T_\ell})}\right) d\theta \leq \mathrm{Vol}(\tilde{\Theta}) \exp\left(-\frac{T_\ell}{2}\left\|\theta^* - \tilde{\theta}_\ell\right\|^2_{A(\bar{e}_{T_\ell})}\right),$$

so

$$\limsup_{\ell\to\infty} \frac{1}{T_\ell}\log(\iota_\ell) + W_{T_\ell}(\tilde{\theta}_\ell) \leq 0.$$

Second, we fix some arbitrary $\epsilon > 0$. Note that for any $\theta, \theta' \in \Theta$,

$$\begin{aligned}
|W_{T_\ell}(\theta) - W_{T_\ell}(\theta')| &= \frac{1}{2}\left(\|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})} - \|\theta^* - \theta'\|^2_{A(\bar{e}_{T_\ell})}\right) \\
&= \frac{1}{2}\left((2\theta^* - \theta - \theta')^\top A(\bar{e}_{T_\ell})(\theta - \theta')\right) \\
&= \frac{1}{2T_\ell}\sum_{t=1}^{T_\ell}\left((2\theta^* - \theta - \theta')^\top x_t x_t^\top(\theta - \theta')\right) \\
&\leq \Delta_{\max}\max_{x\in\mathcal{X}} x^\top(\theta - \theta') \\
&\leq \Delta_{\max}\max_{x\in\mathcal{X}}\|x\|_2\|\theta - \theta'\|_2 \\
&\leq L\Delta_{\max}\|\theta - \theta'\|_2 .
\end{aligned}$$

Then, there exists $\delta > 0$ such that

$$\|\theta - \theta'\|_2 < \delta \Rightarrow |W_{T_\ell}(\theta) - W_{T_\ell}(\theta')| < \epsilon.$$

Then, we take a δ-cover of Θ with $\|\cdot\|_2$, and intersect them with $\tilde{\Theta}$, and denote the resulting cover as $\mathcal{O}$. Then, $\tilde{\theta}_\ell \in O$ for some $O \in \mathcal{O}$. Since we know that $\mathrm{Vol}(O) > 0$ for any $O \in \mathcal{O}$, we have

$$\iota_\ell \geq \int_O \exp\left(-T_\ell W_{T_\ell}(\theta)\right) d\theta \geq \mathrm{Vol}(O)\exp\left(-T_\ell\left(W_{T_\ell}\left(\tilde{\theta}_\ell\right) - \epsilon\right)\right).$$

Taking logarithm on both sides implies that

$$\frac{1}{T_\ell}\log\left(\iota_\ell\right) + W_{T_\ell}\left(\tilde{\theta}_\ell\right) \geq \frac{\mathrm{Vol}(O)}{T_\ell} - \epsilon \to -\epsilon.$$

Since we choose $\epsilon > 0$ arbitrarily, we have

$$\liminf_{\ell\to\infty} \frac{1}{T_\ell}\log\left(\iota_\ell\right) + W_{T_\ell}\left(\tilde{\theta}_\ell\right) \geq 0.$$

Therefore, $\lim_{\ell\to\infty} \frac{1}{T_\ell}\log\left(\iota_\ell\right) + W_{T_\ell}\left(\tilde{\theta}_\ell\right) = 0$ and the statement follows. □

F SUPPLEMENTARY PLOTS

In this section, we present more supplementary plots. All experiments in the main text and supplement are run on a computing cluster with 64 AMD EPYC 7302 16-Core Processor (1500 MHz) with 1TB of RAM. For LinGame, LinGapE, and Oracle algorithms, we directly use the existing implementation from Tirinzoni and Degenne (2022) with the open-source GitHub link: https://github.com/AndreaTirinzoni/bandit-elimination.

We demonstrate that the computational cost of our algorithm is not heavy. We first plot the average number of rejection samples taken to get some $\theta \in \Theta^c_{\widehat{z}_t}$ in the alternative and the running time for our algorithm to demonstrate the computation cost rejection sampling takes. Figures 2 and 3 show the result. By comparing Figure 2 with Figure 1, we see that the number of rejection samples needed to get some $\theta \in \Theta^c_{\widehat{z}_t}$ is generally less than 30 until $\delta < 0.01$. This shows that the computational burden for rejection sampling is generally not large unless we have basically solved the problem. Also, we can see from Figure 3 that the running time per iteration is generally very small, which means our algorithm runs very fast.

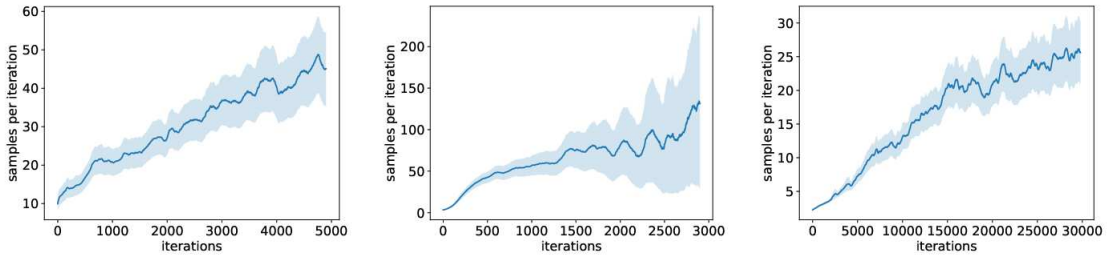

Figure 2: Average number of rejection samples taken until finding some $\theta \in \Theta^c_{\widehat{z}_t}$

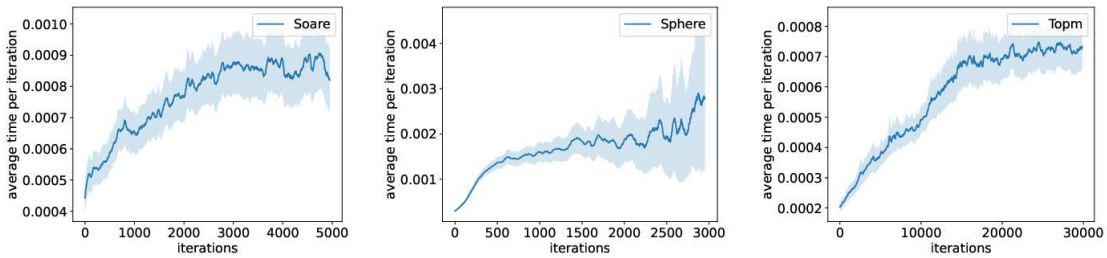

Figure 3: Average clock time per iteration for PEPS under three scenarios

To make a clear comparison of the sampling part in our method with computing the best alternative step in LinGame, we implemented our algorithm, PEPS, in Julia and compared its clock time to existing LinGame implementations on a sphere instance with varying arm numbers, denoted as K. We run both algorithms for a fixed budget of 1000 iterations across 100 trials and compute the average clock time per iteration. We assessed both methods for $K = 50, 200, 1000, 5000, 10000, 20000$, with results presented in milliseconds. Table 3 shows the results. We can see that our method consistently running faster than the benchmark LinGame, particularly as the number of arms increases. This distinction becomes especially significant when $K = 10000$ and $K = 20000$, which corresponds to the case that calculating the best alternative is expensive. Therefore, our method maintains efficiency even in scenarios when computing the alternative is really expensive.

	$K = 50$	$K = 200$	$K = 1000$	$K = 5000$	$K = 10000$	$K = 20000$
PEPS	0.132	0.484	0.681	3.770	6.710	17.110
LinGame	0.152	0.596	3.265	18.610	46.762	126.683

Table 3: Average clock time per iteration for PEPS and LinGame under the sphere instance with $d = 6$ and various number of arms K. Numbers are displayed in milliseconds.