One Shot Inverse Reinforcement Learning for Stochastic Linear Bandits

author names withheld

Editor: Under Review for ALT 2024

Abstract

The paradigm of inverse reinforcement learning (IRL) is used to specify the reward function of an agent purely from its actions and is critical for value alignment and AI safety. Motivated by the need for IRL in large action spaces with limited data, we consider as a first step the problem of learning from a single sequence of actions (i.e., a demonstration) of a stochastic linear bandit algorithm. In general, IRL is made difficult by the lack of access to the rewards seen by the demonstrators. When the demonstrator employs the Phased Elimination algorithm, we develop a simple inverse learning procedure that uses the demonstrator's evolving behavior. We establish guarantees on the performance of our estimator, showing that the linear reward function can be estimated consistently in the time horizon with just a *single* demonstration. In particular, we guarantee that our inverse learner approximates the true reward parameter within a $T^{\frac{1-\omega}{2\omega}}$ error, where T is the number of samples generated by the demonstrator and ω is an action set dependent constant. We complement this result with an information-theoretic lower bound for any inverse learning procedure. Our guarantees are corroborated using simulations on synthetic data and a demonstration constructed from the MovieLens dataset.

Keywords: Inverse Reinforcement Learning, Stochastic Linear Bandits, Phased Elimination

1. Introduction

Using data-driven learning algorithms to design agents that interact with their environment has achieved great success in various fields ranging from robotics and video game playing to language models. For these learned algorithms, reward specification is important for building safe machine-learning systems that align with the goals of the human designer (Amodei et al., 2016). However, alignment with designers' goals using hand-specified rewards is difficult and often mispecified (Anderson, 2001; MacGlashan and Littman, 2015). Inverse Reinforcement Learning(IRL) (Abbeel and Ng, 2004; Ho and Ermon, 2016; Gershman, 2016; Fu et al., 2017; Jacq et al., 2019; Geng et al., 2020) is a well-established paradigm that circumvents the need for explicit reward specification and instead infers a reward function from demonstrations. In IRL, an inverse learner *only* observes the actions of a learned agent and then estimates the environment's reward function. The traditional IRL paradigm assumes that a demonstration consists of a roll-out of the optimal policy (Ng et al., 2000; Abbeel and Ng, 2004) or randomized variants (Ziebart et al., 2008; Ramachandran and Amir, 2007). This paradigm has several limitations, including an often poor sample complexity, i.e. requiring multiple demonstrations. More crucially, even under simple scenarios (tabular RL/bandits), relying purely on demonstrations of an optimal policy can lead to a fundamental *identifability issue*; that is, more than one reward function explaining the demonstrator's actions. Such identifiability issues have been known since the early literature on IRL (Ng et al., 2000; Abbeel and Ng, 2004) and persist even with infinite demonstrations.

The *inverse bandit paradigm*, introduced by Guo et al. (2021), resolves both reward identifiability and sample complexity issues, albeit in the much simpler setting of stochastic multi-armed bandits

(MAB). They show that it is possible to accurately estimate the reward structure by observing a *single online demonstration* of a low-regret bandit algorithm. In particular, they observe the demonstrator's behavior (i.e., the arms it picks) *en route* to optimality and critically utilize the temporal information in online bandit learning to circumvent identifiability issues and the requirement of multiple demonstrations.

The question that motivates this paper is whether learning from a single demonstration in a similar manner is possible for more complex decision-making scenarios. In particular, we are interested in estimating the reward structure in the stochastic linear bandit setting by observing a single demonstration from a low-regret algorithm. This setting in itself is much more challenging — the ideas from Guo et al. (2021) critically utilize the independence of reward distributions in the MAB setting in multiple steps of the algorithm design and analysis and do not generalize to even the linear bandit case, which has highly structured rewards across actions.

In this paper, we show that it is indeed possible to estimate the linear reward parameter consistently in the time horizon from a single demonstration of the *phased elimination algorithm* (Lattimore and Szepesvári, 2020)[1]. To do so, we construct a simple inverse learning algorithm that uses an entirely different idea from the one in Guo et al. (2021). Our algorithm selectively picks a small set of actions from the last epoch of the phased elimination demonstrator and forms a least squares estimate of the reward parameter. The actions are carefully selected to guarantee consistent estimation in the time horizon. Concretely, given an assumption on the density and "smoothness" of the action set (c.f. Assumption 4.1), we verify that our inverse learner with a *single* demonstration estimates reward within an error of $T^{-\left(\frac{\omega-1}{2\omega}\right)}$ where T is the length of the demonstration, and $\omega \in [1, \infty)$ is a constant dependent on the "smoothness" of the action set). We also provide examples of action sets for which these assumptions are reasonable. In addition to the theory, we demonstrate the accuracy of our inverse learner on synthetic as well as semi-synthetic data.

Contributions Our main contributions are listed in more detail below.

- We develop a simple inverse estimator of the reward parameter for a stochastic linear bandit instance ($\theta^* \in \mathbb{R}^d$) from a single demonstration of the phased elimination algorithm (Lattimore and Szepesvári, 2020). Our estimator performs a least-squares estimate using d arms from the last phase of elimination and an estimate of the suboptimality gap of arms eliminated in the last phase. In Theorem 2, we prove an upper bound in the estimation error on the order of $T^{\frac{1-\omega}{2\omega}}$ where T is the time horizon of the forward algorithm and ω is a "smoothness" constant on the action set ranging from $\omega \in [1, \infty)$ (c.f. Assumption 4.1).

- In Theorem 3, we prove an information-theoretic lower bound on the optimal inverse estimator estimation error of $\sqrt{\frac{d}{T}}$, showing that as the action set gets "smoother" around the optimal arm, i.e. $\omega \to \infty$, our inverse estimator becomes close to information-theoretically optimal.

- We empirically evaluate our inverse learning algorithm on synthetic and semi-synthetic data. We first validate the low estimation error of our algorithm over the commonly used action sets consisting of the $\ell_1, \ell_2,$ and ℓ_5 ball. We also consider an application involving linear bandit algorithms for a recommender system on the MovieLens data (Zhu and Kveton, 2022). In particular, we model the problem of predicting the user's "preference vector" as an inverse

1. Note that this is a natural generalization of successive-arm-elimination (Even-Dar et al., 2006) to linear bandits.

linear stochastic bandit problem. We demonstrate that our inverse algorithm can efficiently predict the reward parameter of a user by observing the movies chosen by the recommender system. This could have downstream relevance in predicting the user's preference for movies not seen by the recommender system.

Outline of paper We first provide a brief discussion of the most closely related work in Section 2, and then provide basic background for the stochastic linear bandits problem and phased elimination in Section 3. Section 4 discusses the methodology and proof outline of our main results, Section 5 states the information-theoretic lower bound, and we present our experiments in Section 6. We conclude with a discussion and future work in Section 7.

2. Related work

We organize our related work along two verticals: low-regret algorithms for stochastic linear bandits interchangeably referred to as a *forward algorithm*, and inverse algorithms for reinforcement learning.

2.1. Stochastic Linear Bandits

The setting of stochastic linear bandits was first analyzed by Abe and Long (1999); since then, several algorithms have been proposed that achieve a regret bound of $\mathcal{O}(d\sqrt{T})$ for infinite action sets, and $\widetilde{\mathcal{O}}(\sqrt{dT})$ for finite action sets, e.g. (Dani et al., 2008; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Valko et al., 2014). In both cases, these upper bounds are matched by information-theoretic lower bounds (Lattimore and Szepesvári, 2020). In this paper, we assume that the demonstrator is the Phased Elimination algorithm in Lattimore and Szepesvári (2020); Valko et al. (2014); Esfandiari et al. (2019), which also achieves the optimal $\mathcal{O}(\sqrt{dT})$ regret bound for stochastic linear bandits with a finite action set. This algorithm is spiritually related to the successive-arm-elimination (SAE) algorithm (Even-Dar et al., 2006), which admitted a straightforward analysis for the inverse learning problem in the MAB setting (Guo et al., 2021). However, the phased-elimination algorithm has key differences, including the non-uniform sampling scheme among active arms in each epoch and a doubling in epoch length in each increment. The doubling of epochs, which is not the case for SAE, turns out to be particularly challenging to deal with in inverse estimation but, at the same time, is essential to proving sublinear regret in structured bandit settings.

2.2. Inverse Reinforcement Learning

The original works on IRL (Ng et al., 2000; Abbeel and Ng, 2004) noted an identifiability issue in the reward function from an optimal demonstration that cannot be resolved except in special cases involving additional structure on the reward or additional side information (Gershman, 2016; Amin et al., 2017; Fu et al., 2017; Geng et al., 2020). Assuming randomized variants of the optimal policy (e.g. max-entropy IRL (Ziebart et al., 2008), Bayesian IRL (Ramachandran and Amir, 2007)) can partially alleviate this identifiability issue, but only in special cases. Aside from the aforementioned inverse bandit paradigm (Guo et al., 2021), the works of Gao et al. (2018) and Jacq et al. (2019) also introduced a paradigm of "learning from learners" but used optimization instead of bandit learning for the demonstration and still require several demonstrations. More recently, Hüyük et al. (2022) studied the general problem of IRL for evolving demonstrators by treating the forward algorithm's reward parameter as a Gaussian posterior and using approximate Bayesian inference to infer the learned reward function. In particular, they model the evolution of the forward algorithm as a Gaussian

process but do not utilize any particular structure of the reward function. It is worth noting that there are distinct objectives for learning from demonstrations that can be far easier than IRL; for example, imitation learning (Ho and Ermon, 2016) or apprenticeship learning (Abbeel and Ng, 2004). These tasks usually do not suffer from the same identifiability issues as IRL.

3. Problem Formulation

We now discuss the basic setup for the inverse linear bandit problem. In Section 3.1, we discuss preliminaries for the stochastic linear bandit problem, and in Section 3.2, we describe the *forward algorithm* that we assume the demonstrator will use, i.e. the phased-elimination algorithm (Lattimore and Szepesvári, 2020; Valko et al., 2014). We then formalize the inverse linear bandit problem and our desired estimation error guarantee in Section 3.3.

3.1. Preliminaries on stochastic bandits

Our environment is defined as a structured, parameterized bandit instance $\mathcal{M} = (\theta^*, \mathcal{A})$, where θ^* parameterizes the reward function of the environment and $\mathcal{A}$ is a finite set of actions the forward algorithm may take while interacting with the environment. A *forward algorithm* sequentially interacts with this environment over T rounds. At round t, the algorithm chooses an action from the action set[2], $a_t \in \mathcal{A}$ and receives a reward given by

$$x_t := G_{\theta^*}(a_t) + \eta_t,$$

where $G_{\theta^*}(a)$ is the mean reward function parameterized by θ^* and η_t denotes noise, which we assume to be zero-mean and 1-sub-Gaussian. The forward algorithm repeats this procedure for T steps. The main property that we desire from the forward algorithm is to minimize *pseudo-regret*, defined as

$$R_T = \sum_{t=1}^{T} \max_{a \in \mathcal{A}} G_{\theta^*}(a) - G_{\theta^*}(a_t).$$

As is standard in the bandit literature (Lattimore and Szepesvári, 2020), we desire in particular that $R_T = \widetilde{o}(T)$, i.e. sublinear regret in the total number of rounds T. We consider the special case of the stochastic linear bandit for this work. Here, $\mathcal{A} \subset \mathbb{R}^d$ and $\theta^* \in \mathbb{R}^d$, and the mean reward function is defined as $G_{\theta^*}(a) = \langle a, \theta^* \rangle$.

3.2. The forward algorithm: phased elimination

Inspired by the relative simplicity of the inverse error analysis of the *successive-arm-elimination* algorithm (Even-Dar et al., 2006) for stochastic multi-armed bandits presented in Guo et al. (2021), we will assume that the forward algorithm uses its natural counterpart for the linear bandit problem, which is commonly called *phased elimination* (Lattimore and Szepesvári, 2020; Valko et al., 2014). While not as popular in practice as LinUCB (Abbasi-Yadkori et al., 2011) and linear Thompson sampling (Agrawal and Goyal, 2013), the phased elimination satisfies a similar (optimal) sublinear regret guarantee, given by $R_T = \widetilde{\mathcal{O}}(\sqrt{dT \log |\mathcal{A}|})$. It has found particular use in bandit instances on smooth functions on a graph (Valko et al., 2014).

2. Note that the algorithm has access to the prior history $\mathcal{F}_{t-1} := \{a_1, x_1, a_2, x_2, \ldots, a_{t-1}, x_{t-1}\}$ and can use this history as input to decide an action a_t at round t.

To keep the paper self-contained, we briefly recap the salient properties of the phased elimination algorithm, which we also formally define in Algorithm 1. At a high level, the algorithm operates in phases that increase in length and eliminates a subset of arms at the end of each phase. Consider a phase $\ell \geq 1$ and denote the set of active arms at the beginning of phase ℓ by $\mathcal{A}_\ell$. The algorithm first picks a *G-optimal design* $\{\pi(a)\}_{a \in \mathcal{A}_\ell}$, whose definition is reproduced below.

Definition 1 ***[G-optimal design** (Lattimore and Szepesvári, 2020)] A* G-optimal design *for an action set $\mathcal{A}$ is a function $\pi : \mathcal{A} \to \mathbb{R}$ that maximizes $f(\pi) = \log(\det(V(\pi)))$ such that $\sum_{a \in \mathcal{A}} \pi(a) = 1$ where $V(\pi) = \sum_{a \in \mathcal{A}} n_\ell(a) a a^T$ where $n_\ell(a) = \left\lceil \frac{2d\pi_\ell(a)}{\nu_\ell^2} \log\left(\frac{k\ell(\ell+1)}{\delta}\right) \right\rceil$.*

Algorithm 1: Phased Elimination

Input : δ (probability parameters), T (number of samples), $\{\nu_1, \ldots, \nu_L\}$ (error parameters)

Result: $a_1, \ldots, a_T$

$\ell \leftarrow 0$

$\mathcal{A}_1 \leftarrow \mathcal{A}$

while *Number of actions* $\leq T$ **do**

  $\varepsilon_\ell \leftarrow 2^{-\ell}$

  $\pi_\ell \leftarrow$ G-Optimal design of $\mathcal{A}_\ell$ with δ and ν_ℓ

  $N_\ell \leftarrow 0$

  Play each action $a \in \mathcal{A}_\ell$ each $n_\ell(a) = \left\lceil \frac{2d\pi_\ell(a)}{\nu_\ell^2} \log\left(\frac{k\ell(\ell+1)}{\delta}\right) \right\rceil$ times

  $V_\ell \leftarrow \sum_{a \in \mathcal{A}_\ell} n_\ell(a) a a^T$

  $\theta_\ell \leftarrow V_\ell^{-1} \sum_{t=t_\ell}^{t_\ell + T_\ell} a_t x_t$

  $\mathcal{A}_{\ell+1} \leftarrow \{a \in \mathcal{A}_\ell \text{ s.t. } \max_{b \in \mathcal{A}_\ell}(\langle \theta_l, b - a \rangle) \leq 2\varepsilon_l\}$

  $\ell \leftarrow \ell + 1$

end

It then plays each arm $a \in \mathcal{A}$ exactly $\left\lceil \frac{2d\pi_\ell(a)}{\nu_\ell^2} \log\left(\frac{k\ell(\ell+1)}{\delta}\right) \right\rceil$ times, where δ denotes the allowed probability of failure and ν_ℓ is a error parameter. At the end of phase ℓ, the algorithm uses the observed rewards in phase ℓ alone to construct an ordinary-least-squares estimate of the reward parameter, denoted by θ_ℓ, and eliminates all sufficiently suboptimal arms (in a similar sense to SAE, but with confidence widths given by the structure of the linear model, detailed in Lemma C.1). As long as $\nu_\ell \leq \epsilon_\ell$, this algorithm is known to achieve the regret bound $R_T = \mathcal{O}\left(\sqrt{dT \log\left(\frac{|\mathcal{A}| \log(T))}{\delta}\right)}\right)$, which is optimal for finite action sets.

3.3. The inverse linear bandit problem

We now define the inverse linear bandits problem. The inverse learner is assumed to have access to the sequence of actions $(a_1, \ldots, a_T)$ and the action sets at each phase $(\mathcal{A}_1, \ldots, \mathcal{A}_L)$ from a *single demonstration* of the phased elimination algorithm defined in Section 3.2. As in Guo et al. (2021), we also assume access to the best reward $\mu^* = \max_{a \in \mathcal{A}} \langle a, \theta^* \rangle$ as well as the optimal arm $a^* = \arg\max_{a \in \mathcal{A}} \langle a, \theta^* \rangle$ to avoid degenerative identifiability issues. Our goal is to construct an estimate $\hat{\theta}$ with small relative error to the true reward parameter θ^*, defined as $\frac{\|\hat{\theta} - \theta^*\|_2}{\|\theta^*\|_2}$. We also make the following assumptions on the forward algorithm.

Assumption 3.1

1. *The total number of phases executed by our forward algorithm is upper bounded by $L \leq \bar{L}$ where $\bar{L}$ is a natural number.*

2. *The error parameter at each phase $\nu_\ell = \iota\epsilon_\ell$ is chosen such that $0 < \iota < 1$.*

4. Methodology and main result

In this section, we describe the methodology for our inverse learning approach. We first define some notation specific to this section. We will define the two-dimensional subspace spanned by two vectors u and v as $\text{span}(u, v)$. Moreover, we will define the conditioning of a set as the condition number of the matrix formed where each row of the matrix is an element of the set. That is, for a set of vectors $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$, we define

$$\text{cond}(\mathcal{C}) = \text{cond}\left(\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}\right).$$

The goal of the inverse learner is to learn the environment's true reward parameter θ^*. As mentioned in the introduction, a core challenge in the linear bandit setting is the shared structure across arms — pulling an arm a will change the forward algorithm's estimates of all arms $a' \neq a$, rendering an out-of-the-box approach from Guo et al. (2021) infeasible. At the same time, this shared structure could help estimate θ^* if one could reliably estimate the rewards of a large and "well-behaved" subset of actions. To be more concrete, suppose that we had an oracle where for any arm a in $\mathcal{A}^e \subset \mathcal{A}$, we knew its exact mean reward $G_{\theta^*}(a)$. In this case, the optimal estimator of θ^* would minimize the least-squared error between the rewards and the arms, i.e.

$$\hat{\theta} = \arg\min \sum_{a \in \mathcal{A}^e} \left(R_{\theta^*}(a) - \langle a, \hat{\theta}\rangle\right)^2. \tag{1}$$

At a high level, our inverse learner (Algorithm 2) proceeds in three steps: a) constructing a specific action subset $\mathcal{A}^e$ (Steps 4-6 of Algorithm 2), b) estimating the reward $G_{\theta^*}(a)$ for each $a \in \mathcal{A}^d$ (Step 8 of Algorithm 2), and c) computing the LS estimate of θ^* from these reward samples as observations (also Step 8 of Algorithm 2). Note that the demonstrator chooses δ and ι, which are inputs to the algorithm. Equation (1) suggests that, with near-perfect access to mean rewards, one might want to select the subset $\mathcal{A}^e$ to be as large as possible. However, this is misleading reasoning for several reasons: first, different arms are pulled an unequal number of times due to elimination, meaning that the mean rewards of certain arms can be estimated more reliably than others; second, selecting arms that are too close to each other (i.e., arms a, a' for which $\|a - a'\|_2$ is too small) would result in the estimation error blowing up due to poor conditioning.

The crux of our methodology lies in cleverly designing the action subset $\mathcal{A}^e$ to adequately control the estimation error that arises due to finite samples of each arm as well as the condition number of the design matrix in Equation (1). We now describe each of the steps in Algorithm 2 in more detail and finally present our main result on the estimation error of θ^*.

4.1. Construction of action subset $\mathcal{A}^e$

We now describe the first part of the algorithm (Steps 4-6 in Algorithm 2), which constructs the action subset $\mathcal{A}^e$. First, we select arms only from the last eliminated set, i.e. $\mathcal{A}^e \subset (\mathcal{A}_L \setminus \mathcal{A}_{L-1})$,

Algorithm 2: Inverse Estimator

Data: $(\mathcal{A}_1, \dots, \mathcal{A}_L)$
Result: $\hat{\theta}$
$\mathcal{A}^e = \{\}$
$\beta := (3(1-\iota)\epsilon_\ell)^{\frac{1}{\omega}}$
for $i \in [d]$ **do**
  if $\exists a \in \mathcal{A}$ *s.t.* $\tau(a,i) \geq \beta, \text{dist}(a,i) \leq \gamma, 2(1-\iota)\epsilon_\ell \leq \langle \theta^*, a^* - a \rangle \leq 4(1-\iota)\epsilon_\ell$ **then**
    $\mathcal{A}^e \leftarrow \mathcal{A}^e \cup \{a\}$
  end
end
$\hat{\theta} = \arg\min \sum_{a \in \mathcal{A}^e} \left(\mu^* - 2(1+\iota)\epsilon_\ell - \langle a, \hat{\theta} \rangle\right)^2$
return $\hat{\theta}$

to ensure that the mean reward of each arm in $\mathcal{A}^d$ can be estimated with error as low as possible. Second, we select arms with as large as possible pairwise angles between each other in order to appropriately control the condition number of the design matrix in Equation (1). We will pick d arms in d evenly spaced planes to ensure the latter property. In particular, we will select the i-th arm to be in the subspace spanned by the optimal arm a^* and the i-th vertex of a $d-1$-regular simplex.

Formally, consider any $d-1$-regular simplex $\mathcal{S}_i$ in $\mathbb{R}^d$ formed by the unit vectors $\{s_1, \dots, s_d\}$ such that $s_i \neq \alpha a^*$ for any $i \in [d], \alpha \neq 0$. To form the ith arm in $\mathcal{A}^e$, we will iterate through each arm a in the action set $\mathcal{A}$ and calculate two relevant metrics. The first is the distance between an arm a and its projection in the subspace $\text{span}(a^*, s_i)$. Formally, let $\text{proj}(a,i)$ denote the projection operation from an arm in $\mathcal{A}$ to the two dimensional subspace $\text{span}(a^*, s_i)$, i.e. $\text{proj}(a,i) := \underset{a' \in \text{span}(a^*, s_i)}{\arg\min} \|a - a'\|_2$. Then, the first metric is the distance between an arm a and the plane $\text{span}(a^*, s_i)$, i.e., $\text{dist}(a,i) := \|\text{proj}(a,i) - a\|_2$. The second metric we will calculate is the angle formed between the projection $\text{proj}(a,i)$ and the optimal arm a^*, which we will denote as $\tau(a,i) := \cos^{-1}\left(\frac{\langle \text{proj}(a,i), a^* \rangle}{\|\text{proj}(a,i)\|_2 \|a^*\|_2}\right)$. Thus, our goal is to find a subset of d arms $\mathcal{A}^e = \{a^1, \dots, a^d\}$ such that for the i-th arm a^i, a) $\text{dist}(a^i, i)$ is small, b) $\tau(a^i, i)$ is large, and c) $a^i \in \mathcal{A}_L \setminus \mathcal{A}_{L-1}$, i.e. a^i was eliminated in phase L.

It is worth noting that this specific subset of arms, $\mathcal{A}^e$, may not exist for an arbitrary action set $\mathcal{A}$ if the action set is not sufficiently dense or is very "sharp" around the optimal arm. Below, we state our assumptions on the action set to rule out these possibilities.

Assumption 4.1 *We assume that there exists a value $\bar{L}$ such that for all $i \in [d]$, for all $\ell \in [\bar{L}]$, and some $\omega > 1$, there exists an arm a^i with the properties:*

1. $\tau(a^i, i) \geq \beta$ *where* $\beta := (3(1-\iota)\epsilon_\ell)^{\frac{1}{\omega}}$
2. $\text{dist}(a^i, i) \leq \gamma \leq \frac{2\epsilon_{\bar{L}}}{\|\theta^*\|_2}$.
3. $\mu^* - 4(1-\iota)\epsilon_\ell \leq \langle \theta^*, a^* - a^i \rangle \leq \mu^* - 2(1-\iota)\epsilon_\ell$

A few comments on this assumption are in order. Part 1 of the assumption ensures that the angle between $\text{proj}(a^i, i)$ and the optimal arm a^* is sufficiently large; Part 2 ensures that a^i is close to

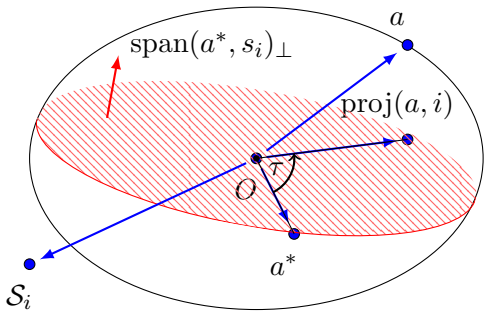


Figure 1: A visualization of the formation of an arm in $\mathcal{A}^e$. Here, we project an arm a onto the subspace $\text{span}(a^*, s_i)$ such that $\tau(a, i)$, the angle between the projection and a^*, is large and $\text{dist}(a, i)$ is small.

its projection onto the plane given by $\text{span}(a^*, i)$; and Part 3 ensures that the arm a^i is sufficiently suboptimal to be eliminated in phase L with high probability, but also sufficiently high in reward to stay active until phase L with high probability (Lemma 4.1, stated below). It is easy to find arms satisfying Part 2 as long as the action set is sufficiently "dense" (in the sense of satisfying a γ-covering of some continuous set in $\mathbb{R}^d$), and it is easy to find arms satisfying Parts 1 and 3 as long as the action set is sufficiently "smooth" around a^*, meaning that arms with a reward bounded away from the optimal reward *and* with a sufficiently large angular distance from a^* can be found.

Lemma 4.1 *Any arm a satisfying*

$$2(1-\iota)\epsilon_\ell < \langle a^* - a, \theta^* \rangle \leq 4(1-\iota)\epsilon_\ell$$

will be in $\mathcal{A}_\ell \setminus \mathcal{A}_{\ell-1}$ with probability at least $1-|\mathcal{A}|L\delta$. Therefore, with probability at least $1-|\mathcal{A}|L\delta$, the mean reward of any arm $a \in \mathcal{A}_L \setminus \mathcal{A}_{L-1}$ is bounded as

$$\mu^* - 4(1+\iota)\epsilon_\ell \leq \langle a, \theta^* \rangle \leq \mu^*.$$

This statement is proved in Appendix C. As long as $\mathcal{A}^e$ can be selected in this way, its condition number is upper bounded according to Lemma 4.2, stated below.

Lemma 4.2 (Condition Number of $\mathcal{A}^e$) *Let χ_2 and χ_1 be defined as $\chi_2 = \max_{a \in \mathcal{A}} \|a\|_2, \chi_1 = \min_{a \in \mathcal{A}} \|a\|_2$. Suppose that Assumption 4.1 holds, and we can select the action subset $\mathcal{A}^e$ according to Steps 4-6 of Algorithm 2. Then, with probability at least $1 - |\mathcal{A}|L\delta$, the condition number of the matrix whose rows are elements of $\mathcal{A}^e$ satisfies*

$$\text{cond}(\mathcal{A}^e) \leq \frac{\chi_1 + \gamma\sqrt{d}}{\chi_2 \left[(2d)^{-\frac{1}{2}} \beta^{\frac{1}{\omega}}\right] - \gamma\sqrt{d}}.$$

This statement is proved in Appendix D. Armed with these lemmas, we now provide upper bounds on the estimation error of the rewards for each arm in $\mathcal{A}^e$, which we will subsequently use to bound the estimation error of the reward parameter θ^*.

4.2. Estimating the rewards of actions in $\mathcal{A}^e$

We next estimate the mean reward for the arms from $\mathcal{A}^e$, i.e. $G_{\theta^*}(a) := \langle a, \theta^* \rangle$ for all $a \in \mathcal{A}^e$. Since each arm belongs to $\mathcal{A}_L \setminus \mathcal{A}_{L-1}$, it will have a mean reward less than the optimal reward μ^* and greater than $\mu^* - 4(1+\iota)\epsilon_\ell$ from Lemma 4.1. Consequently, the simple estimate $\hat{r} := \mu^* - 2(1+\iota)\epsilon_\ell$ satisfies the following upper bound on the estimation error.

Lemma 4.3 *Let r denote the vector of true rewards $\left\{R_{\theta^*}(a^i)\right\}_{i=1}^d$ and $\hat{r}$ denote a vector of our estimated rewards given by $\{\mu^* - 2(1+\iota)\epsilon_\ell\}_{i=1}^d$. Then, we have $\frac{\|r-\hat{r}\|_2}{\|r\|_2} \leq \frac{4\epsilon_L}{\mu^* - 8\epsilon_L} = \mathcal{O}\left(2^{-L}\right)$ with probability at least $1 - |\mathcal{A}|L\delta$.*

4.3. Main result: estimation error guarantee on θ^*

This statement is proved in Appendix D. The final step (Step 8 of Algorithm 2) constructs $\hat{\theta}$ as the LS estimate (Equation (1) using the action set $\mathcal{A}^e := \{a^1, \ldots, a^d\}$ as covariates and estimated rewards $\{\hat{r}\}_{i=1}^d$ as responses. We now present our main result, which is the error guarantee of the estimator from Algorithm 2.

Theorem 2 *Let χ_1, χ_2 be defined as $\chi_2 = \max_{a \in \mathcal{A}} \|a\|_2, \chi_1 = \min_{a \in \mathcal{A}} \|a\|_2$, and $J = \log\left(\frac{|\mathcal{A}|L(L+1)}{\delta}\right)$. Then, we have*

$$\frac{\left\|\hat{\theta} - \theta^*\right\|_2}{\|\theta^*\|_2} = \mathcal{O}\left(\frac{\chi_1 d^{\frac{2\omega-1}{2\omega}} J^{\frac{\omega-1}{\omega}}}{\chi_2 T^{\frac{\omega-1}{2\omega}}}\right)$$

with probability at least $1 - |\mathcal{A}|L\delta$. Note that $\omega > 1$ is the constant from Assumption 4.1.

This statement is proved in Appendix D. Since we have assumed $\omega > 1$, Theorem 2 implies consistent estimation of θ^* as $T \to \infty$. Moreover, as the smoothness parameter $\omega \to \infty$, the dependence on d becomes linear, and the dependence on T becomes $T^{-1/2}$; the latter is optimal in its dependence on T as shown in Theorem 3.

4.4. Discussion on Viability of Assumptions

A natural question is whether the assumptions taken on the action set are reasonable and whether the value of ω can be characterized for arbitrary action sets. The lemma below is a proof-of-concept that there exist valid action sets that satisfy our assumptions from Assumption 4.1 for any value $\omega \in [1, \infty)$. This statement is proved in Appendix F.

Lemma 4.4 *Let $G = \cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L$ for notational ease. Given any value $\omega \in [1, \infty)$, we can construct a bandit instance that satisfies Assumption 4.1. Specifically, Assumption 4.1 is satisfied by two-dimensional bandit instances that are rotationally isomorphic to the bandit instance where*

1. *θ^* forms an angle κ with the vector $(1, 0)$ where*

$$\kappa \in \Bigg[\max\left(-\cos^{-1}\left(\frac{3(1-\iota)\epsilon_L}{\|\theta^*\|_2}\right), \cos^{-1}(0) + \beta - \pi\right), \\ \min\left(\cos^{-1}\left(\frac{3(1-\iota)\epsilon_L}{\|\theta^*\|_2}\right), \cos^{-1}(0) - \beta\right)\Bigg].$$

2. $\forall (x,y) \in \mathcal{A}$ *s.t.* $(x,y) \neq (1,0)$, $\cos(\kappa + \tan^{-1}(y,x))\|\theta^*\|_2\sqrt{x^2+y^2} < \cos(\kappa)\|\theta^*\|_2$.

3. *The two points* $\left(\frac{G\cos(\beta)}{\cos(\kappa+\beta)\|\theta^*\|_2}, \frac{G\sin(\beta)}{\cos(k+\beta)\|\theta^*\|_2}\right), \left(\frac{G\cos(-\beta)}{\cos(\kappa-\beta)\|\theta^*\|_2}, \frac{G\sin(-\beta)}{\cos(k-\beta)\|\theta^*\|_2}\right) \in \mathcal{A}$.

We have defined two instances $\mathcal{M}_1 = (\theta_1^*, \mathcal{A}_1)$ *and* $\mathcal{M}_2 = (\theta_2^*, \mathcal{A}_2)$ *as rotationally isomorphic if there exists a rotation operation* $\mathcal{R}$ *such that* $\mathcal{R}(\theta_1^*) = \theta_2^*$ *and* $\mathcal{R}$ *is a bijective function from* $\mathcal{A}_1$ *to* $\mathcal{A}_2$.

5. Information-Theoretic Lower Bound

We now provide an information-theoretic lower bound on the accuracy achievable by any inverse estimator via the classical Le Cam binary testing approach (LeCam, 1973). Essentially, this approach creates two different bandit instances $\mathcal{M}_1 = (\theta_1^*, \mathcal{A}_1)$ and $\mathcal{M}_2 = (\theta_2^*, \mathcal{A}_2)$ and has the forward algorithm work with one of these bandit instances. Then, we show that the inverse algorithm will be unable to distinguish which of the bandit instances the forward algorithm interacted with given a single demonstration of *any* forward algorithm that incurs regret at least $\widetilde{\mathcal{O}}(\sqrt{dT})$ and sufficiently explores each direction. Since the fundamental limit on regret for stochastic linear bandits for finite action sets is known to be $\widetilde{\mathcal{O}}(\sqrt{dT})$ (Lattimore and Szepesvári, 2020), this implies a fundamental limit on inverse estimation on the order of $\sqrt{\frac{d}{T}}$, as stated below. Theorem 3 is proved in Appendix E.

Theorem 3 *For a bandit instance* $\mathcal{M}$ *characterized by reward parameter* θ_1^* *and action set* $\mathcal{A}$*, there exists a bandit instance* $\mathcal{M}'$ *with parameter* θ_2^* *and the same action set* $\mathcal{A}$ *such that any inverse estimator incurs error*

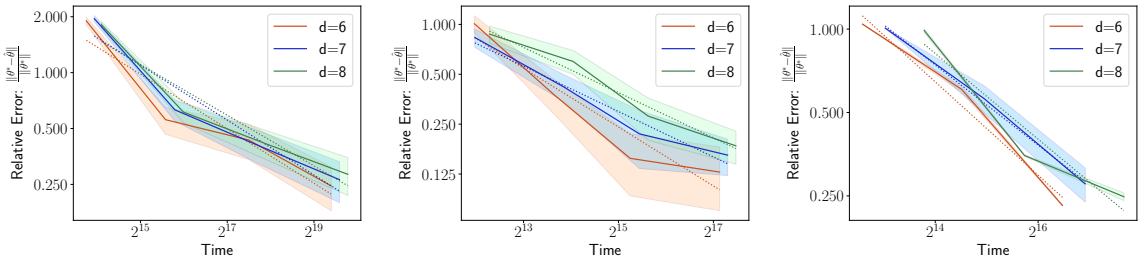
$$\max\{\|\hat{\theta} - \theta_2^*\|_2, \|\hat{\theta} - \theta_1^*\|_2\} = \widetilde{\Omega}\left(\sqrt{\frac{d}{T}}\right).$$

6. Experiments

To validate our results empirically, we implement our inverse estimator on both simulated and semi-synthetic environments, measuring the error in the estimate of θ^*. Throughout the experiments, the forward algorithm is Algorithm 3, and the noise in the observed reward is sampled from a Gaussian $\mathcal{N}(0, 0.2)$ distribution.

6.1. Simulations

To construct our action sets, we sample 4000 vectors from the surface of the unit ℓ_1, ℓ_2, and ℓ_5 balls and use this finite set as $\mathcal{A}$. This is done by independently sampling each entry from a generalized Gaussian distribution $\left(\text{having density proportional to } e^{-|x|^\beta}\right)$ with a β parameter of 1, 2, and 5 respectively, and then normalizing the resulting vector by its respective ℓ_p norm (Barthe et al., 2005). To run Phased Elimination and our estimator on these action sets most naturally, we choose to fix the number of phases instead of the number of rounds; see Appendix G for a formal description. Using this implementation, we run 100 bandit instances with maximum phase values from 3 to 6 and dimensions ranging from 3 to 8. Afterward, we run the inverse estimator on each instance and measure the metric of relative error of $\hat{\theta}$, defined as $\frac{\|\hat{\theta}-\theta^*\|_2}{\|\theta^*\|_2}$.



(*a*) ℓ_1 Ball. Slopes of $-0.487, -0.491, -0.453$ for orange, blue, and green best-fit lines.

(*b*) ℓ_2 Ball. Slopes of $-0.490, -0.435, -0.433$ for orange, blue, and green best-fit lines.

(*c*) ℓ_5 Ball. Slopes of $-0.455, -0.427, -0.413$ for orange, blue and green best-fit lines.

Figure 2: The inverse estimator's performance (averaged over 100 trials) over the ℓ_1, ℓ_2, and ℓ_5 balls across dimensions $d = 6, 7, 8$. The shaded region represents the standard deviation corresponding to each phase. Each graph is a log-log scale with orange, blue, and green dotted lines denoting a log-log linear fit for each dimension.

	ℓ_1 BALL		ℓ_2 BALL		ℓ_5 BALL	
d	INVERSE	FORWARD	INVERSE	FORWARD	INVERSE	FORWARD
3	0.247	0.053	0.011	0.002	0.054	0.083
4	0.352	0.097	0.071	0.002	0.172	0.124
5	0.464	0.230	0.108	0.004	0.247	0.178
6	0.499	0.401	0.138	0.122	0.338	0.249
7	0.551	0.586	0.247	0.391	0.324	0.451
8	0.587	1.392	0.281	1.136	0.379	0.722

Table 1: Inverse and Forward algorithms relative error on each action set. Our method achieves good accuracy over many action sets.

On the one hand, given the $T^{-\left(\frac{\omega-1}{2\omega}\right)}$ term from Theorem 2, we expect relative error to decay with time. From the log-log plots in Figure 2, we observe that this trend holds for each action set by examining the trend of each best-fit line. The lines in a Figure 2(*a*), Figure 2(*b*), and Figure 2(*c*) each contain slopes in the range $[-0.487, -0.413]$, indicating decay in T. The varying slopes between each plot may result from ω varying with the choice of action set. On the other hand, the $d^{\frac{2\omega-1}{2\omega}}$ term in Theorem 2 predicts that relative error should increase in d. In Figure 3, we plot the relative error of our inverse estimator on each unit ball for each dimension

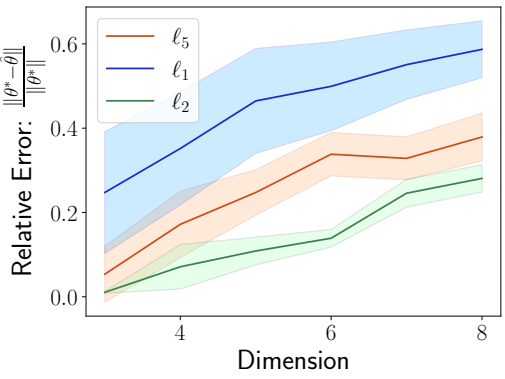


Figure 3: Error vs. d on each action set where shading denotes standard deviation. This corroborates our theoretical dependence of error on dimension.

from 3 to 8, verifying that higher dimensional action sets indeed incur higher relative error. Furthermore, from the results in Table 1, we observe that at dimensions of 6 or higher, the inverse algorithm performs comparably to the forward algorithm, occasionally incurring less relative error.

6.2. Semi-synthetic Experiments

To validate the performance of our estimator with more realistic data, we simulate the task of recommending movies to users on the MovieLens 25M dataset Lam and Herlocker (2006); Harper and Konstan (2015), as set up by Zhu and Kveton (2022). The MovieLens 25M dataset contains over 160,000 users, 60,000 movies, and 25 million ratings. To create an action set and θ^*, we randomly generate sample $u = 6,000$ users, $m = 4,000$ movies, and their corresponding ratings. We then perform a matrix factorization on $R \in \mathbb{R}^{u \times m}$, the matrix of ratings for each user and movie, using Alternating Least Squares. This yields matrices U and M where $UM^\top = R$, $U \in \mathbb{R}^{u \times d}$, and $M \in \mathbb{R}^{m \times d}$. Therefore, each row in M is a d dimensional embedding corresponding to a movie, while each row in U corresponds to the reward parameter for a given user. We then simulate a user's choices and ratings by randomly sampling a reward parameter $\theta^* = U_i$, and then running Algorithm 1 with M as the set of arms for 6 phases. Afterward, we estimate the user's reward parameter via Algorithm 2. We repeat this for ten randomly selected users and average the relative error of $\hat{\theta}$. We repeat this for four different values of d. Our numerical results can be found in Table 2. The results in table Table 2 contain the same trend of increased error with higher dimensional action sets.

7. Discussion

We have presented an inverse reinforcement learning algorithm for the setting of linear stochastic bandits and guarantees its convergence behavior as a function of the length T of the demonstration. We empirically verified the efficacy of our algorithm in both simulation and semi-synthetic settings. Moreover, we showed a lower bound on the best achievable error by any inverse learner. An interesting future direction would be extending a similar framework to nonlinear reward functions and general bandit settings.

MovieLens Relative Error		
d	Inverse Error	Forward Error
2	0.2859	0.0037
4	0.3666	0.0356
6	0.3641	0.1401
8	0.5030	0.4632

Table 2: Relative error of the inverse estimator on MovieLens 25M.

A fundamental limitation, even in the linear bandit settings, is that we limit our demonstrator to being the canonical Phased Elimination algorithm. In contrast, we could have studied other popular linear bandit algorithms. Moreover, we place assumptions on the density and geometry of the action set for our analysis, and weakening these assumptions is an important future direction.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

Naoki Abe and Philip M. Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 3–11, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

Kareem Amin, Nan Jiang, and Satinder Singh. Repeated inverse reinforcement learning. *CoRR*, abs/1705.05427, 2017. URL http://arxiv.org/abs/1705.05427.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016. URL http://arxiv.org/abs/1606.06565.

Christopher Madden Anderson. *Behavioral models of strategies in multi-armed bandit problems*. California Institute of Technology, 2001.

Debangshu Banerjee, Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Exploration in linear bandits with rich action sets and its implications for inference, 2022. URL https://arxiv.org/abs/2207.11597.

Franck Barthe, Olivier Guédon, Shahar Mendelson, and Assaf Naor. A probabilistic approach to the geometry of the ℓ pn-ball. 2005.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/chu11a.html.

Varsha Dani, 7 9, Thomas Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland*, pages 355–366, 2008. URL http://colt2008.cs.helsinki.fi/.

Hossein Esfandiari, Amin Karbasi, Abbas Mehrabian, and Vahab S. Mirrokni. Batched multi-armed bandits with optimal regret. *CoRR*, abs/1910.04959, 2019. URL http://arxiv.org/abs/1910.04959.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *CoRR*, abs/1710.11248, 2017. URL http://arxiv.org/abs/1710.11248.

Yang Gao, Huazhe Xu, Ji Lin, Fisher Yu, Sergey Levine, and Trevor Darrell. Reinforcement learning from imperfect demonstrations. *CoRR*, abs/1802.05313, 2018. URL http://arxiv.org/abs/1802.05313.

Sinong Geng, Houssam Nassif, Carlos Manzanares, Max Reppen, and Ronnie Sircar. Deep PQR: Solving inverse reinforcement learning using anchor actions. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3431–3441. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/geng20a.html.

Samuel J Gershman. Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, 71:1–6, 2016.

Wenshuo Guo, Kumar Krishna Agrawal, Aditya Grover, Vidya Muthukumar, and Ashwin Pananjady. Learning from an exploring demonstrator: Optimal reward estimation for bandits. *arXiv preprint arXiv:2106.14866*, 2021.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), dec 2015. ISSN 2160-6455. doi: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Alihan Hüyük, Daniel Jarrett, and Mihaela van der Schaar. Inverse contextual bandits: Learning how behavior evolves over time. In *International Conference on Machine Learning*, pages 9506–9524. PMLR, 2022.

Alexis Jacq, Matthieu Geist, Ana Paiva, and Olivier Pietquin. Learning from a learner. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2990–2999. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/jacq19a.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. 2014. doi: 10.48550/ARXIV.1407.4443. URL https://arxiv.org/abs/1407.4443.

R Krasnodębski. Dihedral angle of the regular a-simplex. *Commentationes Mathematicae*, 15(1), 1971.

Shyong Lam and Jon Herlocker. Movielens data sets. *Department of Computer Science and Engineering at the University of Minnesota*, 2006.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

Sergey Loyka. On singular value inequalities for the sum of two matrices, 2015.

James MacGlashan and Michael L Littman. Between imitation and intention learning. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pages 2586–2591, 2007.

Michal Valko, Remi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 46–54, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/valko14.html.

Rong Zhu and Branislav Kveton. Robust contextual linear bandits. *arXiv preprint arXiv:2210.14483*, 2022.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

Appendix A. Notation table

Symbol	Meaning
d	Dimension of environment
T	Time horizon
L	Number of phases
θ^*	True Reward Function Parameter
θ	Demonstrator's Reward Function Parameter
$\hat{\theta}$	Inverse Estimator's Estimated Reward Parameter
γ	Closeness parameter of action set
a_t	Action taken by demonstrator at time t
x_t	Reward seen by demonstrator at time t
η_t	Noise in reward function seen at time t
μ^*	Reward of optimal arm
a^*	Optimal action with the highest reward
$\mathcal{A}_\ell$	Set of remaining arms at phase ℓ
$\mathcal{A}_\ell \setminus \mathcal{A}_{\ell-1}$	Set of eliminated arms before phase ℓ
ϵ_ℓ	$2^{-\ell}$ used as criteria for elimination
ν_ℓ	Error parameter for G-Optimal Design
δ	Probability Parameter for G-Optimal Design

Appendix B. Technical Lemmas

B.1. Proof of Lemma B.1

Lemma B.1 *Given two arms a, b that are γ-close, i.e. $\|a-b\|_2 \leq \gamma$, the difference in their rewards is bounded by*

$$\langle a, \theta^* \rangle - \langle b, \theta^* \rangle \leq \gamma \|\theta^*\|_2.$$

Proof Simply,

$$\begin{aligned} \langle a, \theta^* \rangle - \langle b, \theta^* \rangle &= \langle a - b, \theta^* \rangle \\ &\leq \|a-b\|_2 \|\theta^*\|_2 \\ &\leq \gamma \|\theta^*\|_2 \end{aligned}$$

■

Appendix C. Phased Elimination Proofs

We first prove that the estimate of the reward parameter for the forward algorithm is an accurate estimate of θ^*. The central intuition behind this is that the G-Optimal design is chosen to ensure the forward algorithm explores each dimension in $\mathbb{R}^d$. This exploration helps ensure that the demonstrator's estimate of θ accurately predicts the sample mean rewards for any arm in the active

set, not just ones that point in specific favorable directions. Formally, it ensures that the demonstrator's estimate of the reward of any arm in the remaining active set of any phase ℓ is bounded by a ν_ℓ. This lemma is similar to that of Lemma 6.1 in Esfandiari et al. (2019).

Lemma C.1 (Demonstrator's Estimation Error) *From Esfandiari et al. (2019), given arms pulled in phase ℓ according to Algorithm 1, with probability at least $1 - |\mathcal{A}|L\delta$, for every $a \in \mathcal{A}_\ell$, we have*

$$|\langle a, \theta_\ell - \theta^* \rangle| \leq \nu_\ell.$$

Here, θ_ℓ estimates the forward algorithm reward parameter after the lth phase.

Proof From Lemma 6.1 of Esfandiari et al. (2019), for any $\delta, \nu_\ell \geq 0$, we know that we can find a multiset where after playing the multiset in batched bandits fashion, the least-squares estimate error for any arm a is $|\langle a, \theta_\ell - \theta^* \rangle| \leq \nu_\ell$ with probability $1 - \delta$. Therefore, we know that we can form a multiset such that for every arm $a \in \mathcal{A}$ and all phases l,

$$|\langle a, \theta_\ell - \theta^* \rangle| \leq \nu_\ell.$$

To get a lower bound on the probability that this event occurs, we need to find the probability of the union of all these events not happening. We can upper bound this by taking the union bound of all events. For all $|\mathcal{A}|$ arms and L phases, we get that the probability of any of these events not happening is upper bounded by $|\mathcal{A}|L\delta$. ■

This accuracy of the forward algorithm's θ_ℓ helps maintain its low regret properties. Given the accuracy of its reward parameter, it is intuitive that with high probability, the forward algorithm knows which arms are suboptimal and which are not. This intuition should include that of the optimal arm A^*, which is not suboptimal by definition. Therefore, with high probability, the forward algorithm does not eliminate the optimal arm.

Corollary C.1 *With probability $1 - |\mathcal{A}|L\delta$, for every phase l, $a^* \in \mathcal{A}_\ell$.*

Proof From Lemma C.1, for any suboptimal arm a,

$$\langle a, \theta_\ell \rangle - \langle a^*, \theta_\ell \rangle \leq (\langle a, \theta^* \rangle + \nu_\ell) - (\langle a^*, \theta^* \rangle - \nu_\ell) \leq 2\iota\epsilon_\ell \leq 2\epsilon_\ell.$$

The event from Lemma C.1 occurs with probability $1 - \delta$, so this result also happens with probability $1 - \delta$. ■

Given the event that the optimal arm remains in the active set, we can state with a high probability that suboptimal arms will be eliminated. This is clear from the elimination criteria; if an arm's reward is much worse than the best-estimated reward for any arm in the active set, it will be eliminated. Given that the optimal arm is still in the active set and the reward estimate is accurate, arms with a true reward much worse than the optimal arm will most likely also have an estimated reward worse than the optimal arm. This will lead to the elimination of that arm. We formalize this in Lemma 4.1.

Lemma 4.1 *Any arm a satisfying*

$$2(1 - \iota)\epsilon_\ell < \langle a^* - a, \theta^* \rangle \leq 4(1 - \iota)\epsilon_\ell$$

will be in $\mathcal{A}_\ell \setminus \mathcal{A}_{\ell-1}$ with probability at least $1 - |\mathcal{A}|L\delta$. Therefore, with probability at least $1 - |\mathcal{A}|L\delta$, the mean reward of any arm $a \in \mathcal{A}_L \setminus \mathcal{A}_{L-1}$ is bounded as

$$\mu^* - 4(1 + \iota)\epsilon_\ell \leq \langle a, \theta^* \rangle \leq \mu^*.$$

Proof Let $b_{\ell-1}$ be the arm that maximizes the reward $b_{\ell-1} = \arg\max_{b \in \mathcal{A}_{\ell-1}} \langle b, \theta_{\ell-1} \rangle$.

$$\begin{aligned}
\langle b_{\ell-1} - a, \theta_{\ell_1} \rangle &\leq \langle b_{\ell-1} - a, \theta^* \rangle + 2\nu_{\ell-1} \qquad (2) \\
&\leq \langle a^* - a, \theta^* \rangle + 2\nu_{\ell-1} \\
&\leq 4(1-\iota)\epsilon_\ell + 2\iota\epsilon_{\ell-1} \\
&\leq 2(1-\iota)\epsilon_{\ell-1} + 2\iota\epsilon_{\ell-1} \\
&= 2\epsilon_{\ell-1}
\end{aligned}$$

Here, Equation (2) comes from Lemma C.1 which happens with probability $1 - |\mathcal{A}|L\delta$. Therefore, arm a will not be deleted in phase $\ell - 1$. Moreover, let b_ℓ be the arm that maximizes the reward $b_\ell = \arg\max_{b \in \mathcal{A}_\ell} \langle b, \theta_\ell \rangle$.

$$\begin{aligned}
\langle b_\ell - a, \theta_\ell \rangle &= \langle b_\ell, \theta_\ell \rangle - \langle a, \theta_\ell \rangle \\
&\geq \langle a^*, \theta_\ell \rangle - \langle a, \theta_\ell \rangle \\
&\geq \langle a^* - a, \theta^* \rangle - 2\nu_\ell \qquad (3) \\
&= \langle a^* - a, \theta^* \rangle - 2\iota\epsilon_\ell \\
&\geq 2(1-\iota)\epsilon_\ell - 2\iota\epsilon_\ell \\
&= 2\epsilon_\ell
\end{aligned}$$

Here, Equation (3) comes from Lemma C.1, which again happens with the same probability. Therefore, arm a will be deleted in phase ℓ with probability $1 - |\mathcal{A}|L\delta$.

By the definition of μ^*,

$$\langle a, \theta^* \rangle \leq \mu^*.$$

Given arm a_i is in $\mathcal{A}_\ell \setminus \mathcal{A}_{\ell-1}$, it was not eliminated in the previous phase. For notational ease, let $b = \arg\max_{b \in \mathcal{A}_{\ell-1}} \langle b, \theta_{\ell-1} \rangle$. Therefore,

$$\begin{aligned}
2\epsilon_{l-1} &\geq \langle b - a, \theta_{\ell-1} \rangle \\
&= \langle b, \theta_{\ell-1} \rangle - \langle a, \theta_{\ell-1} \rangle \\
&= \langle b, \theta_{\ell-1} \rangle - \langle a, \theta_{\ell-1} - \theta^* \rangle - \langle a, \theta^* \rangle \\
&\geq \langle b, \theta_{\ell-1} \rangle - \nu_{\ell-1} - \langle a, \theta^* \rangle \qquad (4) \\
&\geq \langle a^*, \theta_{\ell-1} \rangle - \nu_{\ell-1} - \langle a, \theta^* \rangle \qquad (5) \\
&= \langle a^*, \theta_{\ell-1} - \theta^* \rangle + \langle a^*, \theta^* \rangle - \nu_{\ell-1} - \langle a, \theta^* \rangle \\
&\geq \langle a^*, \theta^* \rangle - 2\nu_{\ell-1} - \langle a, \theta^* \rangle \qquad (6)
\end{aligned}$$

Here, Equation (4) comes from Lemma C.1, which happens with probability at least $1 - |\mathcal{A}|L\delta$. Equation (5) comes from the fact that b achieves the maximum reward in $\mathcal{A}_{l-1}$ and $a^* \in \mathcal{A}_{l-1}$ with the same probability according to Corollary C.1. Also, Equation (6) comes from applying Lemma C.1 again. Therefore, we have

$$\begin{aligned}
\langle a, \theta^* \rangle &\geq \mu^* - 2\epsilon_{\ell-1} - 2\nu_{\ell-1} \\
&= \mu^* - 4\epsilon_\ell - 4\iota\epsilon_\ell \\
&= \mu^* - 4(1+\iota)\epsilon_\ell
\end{aligned}$$

■

Corollary C.2 *Given an arm a that is γ-close to arm b that has suboptimality*

$$\mu^* - 4(1-\iota)\epsilon_\ell + \gamma\|\theta^*\|_2^2 \leq \langle a^* - b, \theta^* \rangle \leq \mu^* - 2(1-\iota)\epsilon_\ell - \gamma\|\theta^*\|_2^2,$$

arm a will be eliminated before phase ℓ, i.e. $a \in \mathcal{A}_L \setminus \mathcal{A}_{L-1}$ with probability at least $1 - |\mathcal{A}|L\delta$.

Proof We have that $|\langle b - a, \theta^* \rangle| \leq \gamma\|\theta^*\|_2$ according to Lemma B.1. Therefore,

$$\begin{aligned}\langle a^* - b, \theta^* \rangle &\leq \langle a^* - a, \theta^* \rangle + \gamma\|\theta^*\|_2 \\ &\leq \mu^* - 4(1-\iota)\epsilon_\ell\end{aligned}$$

Moreover,

$$\begin{aligned}\langle a^* - b, \theta^* \rangle &\geq \langle a^* - a, \theta^* \rangle - \gamma\|\theta^*\|_2 \\ &\geq \mu^* - 2(1-\iota)\epsilon_\ell\end{aligned}$$

According to Lemma 4.1, which happens with probability at least $1 - |\mathcal{A}|L\delta$, arm a will be deleted. ■

Moreover, for simplicity, throughout this paper, we will do most of our calculations based on phase numbers, including L, the last phase number. However, given that the last phase is technically a random variable based on the G-optimal design, we provide a lower bound on the phase L in terms of T. Here, we see that L is lower bounded by the logarithm of T up to constants.

Lemma C.2 *The number of rounds that Phased Elimination takes and the total number of phases L exhibit the relationship*

$$\log(T) \leq \log(2\iota^{-2}dJ) + 2\log(2^L) + \log(2).$$

Here, J is a constant defined as $J := \left(\frac{|\mathcal{A}|L(L+1)}{\delta}\right)$.

Proof Let N_ℓ be the number of arms played in phase ℓ. From Lattimore and Szepesvári (2020), we have that any

$$\begin{aligned}N_\ell - \frac{d(d+1)}{2} &\leq \frac{2d}{\nu_\ell^2}\log\left(\frac{|\mathcal{A}|l(l+1)}{\delta}\right) \\ &\leq 2\iota^{-2}d \cdot 2^{2l}\left(\frac{|\mathcal{A}|l(l+1)}{\delta}\right)\end{aligned} \tag{7}$$

where the first equality comes from Lattimore and Szepesvári (2020). We will call $J := \left(\frac{|\mathcal{A}|L(L+1)}{\delta}\right)$ for notational ease.

$$
\begin{aligned}
\log\left(\sum_{\ell}^{L-1} N_\ell\right) &\leq \log\left(\sum_{\ell}^{L-1} 2\iota^{-2} d \cdot 2^{2l} \cdot (J) + \frac{d(d+1)}{2}\right) \\
&= \log\left(2\iota^{-2} d\,(J) \sum_{\ell}^{L-1} 2^{2l} + \sum_{\ell}^{L-1} \frac{d(d+1)}{2}\right) \\
&= \log\left(2\iota^{-2} d\,(J) \sum_{\ell}^{L-1} 2^{2l} + \sum_{\ell}^{L-1} \frac{d(d+1)}{2}\right) \\
&= \log\left(2\iota^{-2} d\,(J) \sum_{\ell}^{L-1} 2^{2l}\right) + \log\left(\frac{\sum_{\ell}^{L-1} \frac{d(d+1)}{2}}{2\iota^{-2} d\,(J) \sum_{\ell}^{L-1} 2^{2l})}\right) \\
&= \log\left(2\iota^{-2} d\,(J) \sum_{\ell}^{L-1} 2^{2l}\right) + \log\left(1 + \frac{\sum_{\ell}^{L-1} \frac{d+1}{4}}{2\iota^{-2} d\,(J) \sum_{\ell}^{L-1} 2^{2l})}\right) \\
&= \log\left(2\iota^{-2} d\,(J) \sum_{\ell}^{L-1} 2^{2l}\right) + \log\left(2\right) \\
&= \log\left(2\iota^{-2} d\,(J)\,(4^L - 4)\right) + \log\left(2\right) \\
&\leq \log(2\iota^{-2} dJ) + \log(4^L) + \log\left(2\right) \\
&\leq \log(2\iota^{-2} dJ) + 2\log(2^L) + \log\left(2\right)
\end{aligned}
$$

We have arrived at our final claim.

■

Appendix D. Inverse Estimator Properties

We restate a lemma connecting the error of our inverse estimate with the condition of matrix $\mathbf{A}$ and the reward estimates $\hat{b}$.

Lemma D.1 *Suppose r and $\hat{r}$ are vectors of the true rewards and estimated rewards for $\mathcal{A}^e$. The solution to $\hat{\theta} = \arg\min \sum_{a^i \in \mathcal{A}^e} (\hat{r}_i - \langle \theta, a^i \rangle)^2)$ where $\hat{r}_i$ is the estimate reward of a^i satisfies the bound the error in estimation of θ via*

$$
\frac{\left\|\hat{\theta} - \theta^*\right\|_2}{\|\theta\|_2} \leq \mathrm{cond}(\mathcal{A}^e) \frac{\|\hat{r} - r\|_2}{\|r\|_2}.
$$

Lemma 4.3 *Let r denote the vector of true rewards $\left\{R_{\theta^*}(a^i)\right\}_{i=1}^d$ and $\hat{r}$ denote a vector of our estimated rewards given by $\{\mu^* - 2(1+\iota)\epsilon_\ell\}_{i=1}^d$. Then, we have $\frac{\|r-\hat{r}\|_2}{\|r\|_2} \leq \frac{4\epsilon_L}{\mu^* - 8\epsilon_L} = \mathcal{O}\left(2^{-L}\right)$ with probability at least $1 - |\mathcal{A}| L\delta$.*

Proof r is a vector of rewards of arms in $\mathcal{A}_L \setminus \mathcal{A}_{L-1}$. Therefore, for an element r_a associated with an arm $a \in \mathcal{A}_L \setminus \mathcal{A}_{L-1}$, we know $a \notin \mathcal{A}_{L-1} \setminus \mathcal{A}_{L-2}$. Via Lemma 4.1, for any element r_i in r,

$$
\mu^* - 4(1+\iota)\epsilon_L \leq r_i \leq \mu^*.
$$

We remind the reader that $\hat{r}$ is the vector of all $\mu^* - 2(1+\iota)\epsilon_L$ from Algorithm 2. Therefore, the worst case error is when the true reward is exactly $r_i = \mu^*, \mu^* - 4(1+\iota)\epsilon_L$. In this, the error in the estimation of r is upper bounded by $|r_i - \hat{r}_i| \leq 2(1+\iota)\epsilon_L$. Therefore, the maximum of the ℓ_2 norm of the difference vector is

$$\|\hat{r} - r\|_2 \leq 2(1+\iota)\epsilon_L\sqrt{d}.$$

For calculating $\|r\|_2$, we acknowledge that the smallest r_a for any i can be is $\mu^* - 4(1+\iota)\epsilon_L$. Thus, ℓ_2 norm of the reward of true vectors is lower bounded by $\|r\|_2 \geq \sqrt{d}(\mu^* - 4(1+\iota)\epsilon_L)$. We have our final result with

$$\frac{\|r - \hat{r}\|_2}{\|r\|_2} \leq \frac{2(1+\iota)\epsilon_L}{\mu^* - 4(1+\iota)\epsilon_L}.$$

Since $\iota \leq 1$ from Assumption 4.1, we have that

$$\frac{2(1+\iota)\epsilon_L}{\mu^* - 4(1+\iota)\epsilon_L} \leq \frac{4\epsilon_L}{\mu^* - 8\epsilon_L} = \mathcal{O}\left(2^{-L}\right).$$

■

Lemma 4.2 (Condition Number of $\mathcal{A}^e$) *Let χ_2 and χ_1 be defined as $\chi_2 = \max_{a\in\mathcal{A}}\|a\|_2, \chi_1 = \min_{a\in\mathcal{A}}\|a\|_2$. Suppose that Assumption 4.1 holds, and we can select the action subset $\mathcal{A}^e$ according to Steps 4-6 of Algorithm 2. Then, with probability at least $1 - |\mathcal{A}|L\delta$, the condition number of the matrix whose rows are elements of $\mathcal{A}^e$ satisfies*

$$\text{cond}(\mathcal{A}^e) \leq \frac{\chi_1 + \gamma\sqrt{d}}{\chi_2\left[(2d)^{-\frac{1}{2}}\beta^{\frac{1}{\omega}}\right] - \gamma\sqrt{d}}.$$

Proof

We can now prove the original claim. For the help of this proof, we will denote $\mathbf{A}$ as the matrix version of $\mathcal{A}^e$, i.e. $\mathbf{A} = \begin{bmatrix} a^1 \\ a^2 \\ \vdots \\ a^d \end{bmatrix}$ where $a^1, \ldots, a^d \in \mathcal{A}^e$. We will break down the proof of the bound of the condition number into two parts. Decomposing $\mathbf{A}$ yields

$$\mathbf{A} = \mathbf{D}\tilde{\mathbf{A}} + \mathbf{N}.$$

Here, $\mathbf{D}$ is a diagonal matrix where the value of $\mathbf{D}_{i,i}$ is the ℓ_2 norm of the ith row of $\mathbf{A}$. Also, $\tilde{\mathbf{A}}$ is a matrix where the ith row of $\mathbf{A}$, call it v_i, is $v_i = \frac{\text{proj}(a^i, i)}{\|\text{proj}(a^i, i)\|_2}$. $\mathbf{N}$ is a matrix where the ith row is the vector $a^i - \text{proj}(a^i, i)$. Now, we need to lower bound $\sigma_{\min}(\mathbf{A})$ and upper bound $\sigma_{\max}(\mathbf{A})$. We begin with lower bounding $\sigma_{\min}(\mathbf{A}_\ell)$.

$$\begin{aligned} \sigma_{\min}(\mathbf{A}) &= \sigma_{\min}(\mathbf{D}\tilde{\mathbf{A}} + \mathbf{N}) \\ &\geq \sigma_{\min}(\mathbf{D}\tilde{\mathbf{A}}) - \sigma_{\max}(\mathbf{N}) \end{aligned} \tag{8}$$

Here, Equation (8) comes from Loyka (2015). We upper bound the $\sigma_{\max}(\mathbf{N})$ term via the following

$$\begin{aligned}\sigma_{\max}(\mathbf{N}) &= \sqrt{\|\mathbf{N}^\top \mathbf{N}\|_2} \\ &= \sqrt{\max_{x \text{ s.t. } \|x\|_2=1} x^\top \mathbf{N}^\top \mathbf{N} x} \\ &\leq \sqrt{d\gamma^2} \\ &= \gamma\sqrt{d}\end{aligned} \tag{9}$$

Here, Equation (9) comes from noticing that the rows of $\mathbf{N}$ have ℓ_2 norm at most γ. We can now move on to bounding $\sigma_{\min}(\mathbf{D}\tilde{\mathbf{A}}) \geq \sigma_{\min}(\mathbf{D})\sigma_{\min}(\tilde{\mathbf{A}})$.

By design, $\mathbf{D}$ is a diagonal matrix where the ith entry is ℓ_2 norm of the ith row. Therefore, the minimum singular value of $\mathbf{D}$ is lower bounded by the shortest arm in the action set, defined as constant χ_1. Therefore, we have

$$\sigma_{\min}(\mathbf{D}) \geq \min_{a \in \mathcal{A}} \|a\|_2 \to \chi_1.$$

We now have that

$$\sigma_{\min}(\mathbf{A}) \geq \chi_1 \sigma_{\min}(\tilde{\mathbf{A}}) - \gamma\sqrt{d}.$$

We now do the upper bound for the maximum singular value.

$$\begin{aligned}\sigma_{\max}(\mathbf{A}) &= \sigma_{\max}(\mathbf{D}\tilde{\mathbf{A}} + \mathbf{N}) \\ &\leq \sigma_{\max}(\mathbf{D}\tilde{\mathbf{A}}) + \sigma_{\max}(\mathbf{N}) \\ &\leq \sigma_{\max}(\mathbf{D}\tilde{\mathbf{A}}) + \gamma\sqrt{d}\end{aligned}$$

where the inequality comes from the Courant-Fischer min-max theorem, and the second inequality comes from the above analysis. Similarly, the maximum singular value of $\mathbf{D}$ is upper bounded by the length of the longest arm in the action set, defined as constant χ_2. Therefore, we have

$$\sigma_{\max}(\mathbf{D}) \leq \max_{a \in \mathcal{A}} \|a\|_2 \to \chi_2.$$

We now have that

$$\sigma_{\max}(\mathbf{A}) \leq \chi_2 \sigma_{\max}(\tilde{\mathbf{A}}) + \gamma\sqrt{d}.$$

We now need only analyze the minimum and maximum singular values of $\tilde{\mathbf{A}}$. We remind the reader that the rows of $\tilde{\mathbf{A}}$ are defined as $\frac{\text{proj}(a^i, i)}{\|\text{proj}(a^i, i)\|_2}$. First, we list three properties of our $\tilde{\mathbf{A}}$ matrix. We know that each row of $\tilde{\mathbf{A}}$ forms an angle of $\tau(a^i, i) \geq \beta$ with the optimal arm a^* from Assumption 4.1. We wish to find the condition number for the matrix $\tilde{\mathbf{A}}$. The smallest possible condition number is achieved when $\tau(a^i, i)$ is the smallest for each row v_i, i.e. $\tau(a^i, i) = \beta$. This is when the rows are the most colinear, leading to poor conditioning. To analyze the condition number of $\tilde{\mathbf{A}}$, we will first analyze the condition number of $\mathbf{B}$. We define the matrix $\mathbf{B} = \frac{1}{\sqrt{d}}\tilde{\mathbf{A}}^*$. We state the $\text{cond}(\tilde{\mathbf{A}}) = \text{cond}(\mathbf{B})$, so we need only find $\text{cond}(\mathbf{B})$. Moreover, we will do this by finding $\text{cond}(\mathbf{B}^*\mathbf{B})$. The condition number of this matrix is linked to that of $\mathbf{B}$ via

$$\sqrt{\text{cond}(\mathbf{B}^*\mathbf{B})} = \text{cond}(\mathbf{B}).$$

We note that $[\mathbf{B}^*\mathbf{B}]_{ij} = \frac{1}{d}\langle v_i, v_j\rangle$ where v_i and v_j are the ith and jth rows of $\tilde{\mathbf{A}}$. Note then that $[\mathbf{B}^*\mathbf{B}]_{ii} = \frac{1}{d}$. For $i \neq j$, then $\langle v_i, v_j\rangle$ is the following. We will assume the worst case, where the angle $\tau(a^i, i)$ is as small as possible, i.e. $\tau(a^i, i) = \beta$. We wish to first find the angle between our α vectors. We remind the reader that our α vectors form a $d-1$-dimensional simplex centered at the unit vector $u = \frac{a^*}{\|a^*\|_2}$. We will first find the radius of this simplex, i.e., $\|u - v_i\|_2$. The vectors u, v_i, and the origin form an isosceles triangle where u and v_i are unit-norm by definition. Therefore, by the Law of Sines

$$\begin{aligned}\|u - v_i\|_2 &= \frac{\sin(\tau(a^i, i))}{\sin\left(\frac{\pi - \tau(a^i,i)}{2}\right)} \\ &= 2\sin\left(\frac{\tau(a^i, i)}{2}\right)\end{aligned}$$

Therefore, we have that the radius of the simplex is $2\sin\left(\frac{\tau(a^i,i)}{2}\right)$, which we will call ρ for now. From Krasnodębski (1971), the angles formed between $u - v_i$ and $u - v_j$ is $\arccos\left(-\frac{1}{d-1}\right)$. Therefore, we have the distance between v_j and v_i satisfies

$$\begin{aligned}\|v_j - v_i\|_2^2 &= \|u - v_i\|_2^2 + \|u - v_j\|_2^2 - 2\|u - v_i\|_2\|u - v_j\|_2 \cos\left(\arccos\left(-\frac{1}{d-1}\right)\right) \\ &= 2\rho^2\left(1 - 2\cos\left(\arccos\left(-\frac{1}{d-1}\right)\right)\right) \\ &= 2\rho^2\frac{2d-1}{d-1}\end{aligned}$$

We also have that the angle we are looking for β, which is the angle between v_i and v_j, satisfies

$$\|v_j - v_i\|_2^2 = 2 - 2\cos(\beta).$$

Therefore, we have

$$\cos(\beta) = 1 - \frac{\rho^2 d}{d-1}$$

Next, we consider the structure of matrix $\mathbf{B}^*\mathbf{B}$. Its diagonal elements are $\frac{1}{d}$, and its nondiagonal elements are $\frac{1}{d}\cos(\beta)$, leading to an explicit unitary diagonalization. This matrix has singular values:

$$\sigma_1, \ldots, \sigma_{d-1} = \frac{1}{d} - \frac{1}{d}\cos(\beta)$$

$$\sigma_d = \frac{d-1}{d}\cos(\beta) + \frac{1}{d}.$$

We will upper bound the maximum singular value.

$$\begin{aligned}\sigma_d \leq& \frac{d-1}{d}\cos(\beta) + \frac{1}{d} \\ \leq& \frac{d-1}{d} + \frac{1}{d} \\ =& 1\end{aligned}$$

where the first inequality comes from the fact that $\cos(\beta) \leq 1$. For lower bounding the minimum singular value, we have

$$\begin{aligned}\sigma_1 &= \frac{1}{d} - \frac{1}{d}\cos(\beta) \\ &\geq \frac{\rho^2}{d-1}\end{aligned}$$

We can lower bound ρ^2 on the interval $\tau(a^i, i) \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ via its Taylor expansion as

$$\rho^2 \geq \frac{\tau(a^i, i)^2}{2}.$$

Therefore, we get that the minimum singular value is lower bounded by

$$\begin{aligned}\sigma_1 &\geq \frac{\tau(a^i, i)^2}{2d} \\ &\geq \frac{1}{2d}\beta^{\frac{2}{\omega}} \qquad (10)\end{aligned}$$

Here, Equation (10) comes from our assumption Assumption 4.1. Therefore, the maximum singular value for $\tilde{\mathbf{A}}$ is upper bounded by 1 and the minimum singular value for $\tilde{\mathbf{A}}$ is lower bounded by $(2d)^{-\frac{1}{2}}\beta^{\frac{1}{\omega}}$.

We have proved the condition number of $\tilde{\mathbf{A}}$. Now, we can find the total condition number for $\mathbf{A}$.

$$\begin{aligned}\text{cond}(\mathbf{A}) &= \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})} \\ &\leq \frac{\chi_1\sigma_{\max}(\tilde{\mathbf{A}}) + \gamma\sqrt{d}}{\chi_2\sigma_{\min}(\tilde{\mathbf{A}}) - \gamma\sqrt{d}} \\ &\leq \frac{\chi_1 + \gamma\sqrt{d}}{\chi_2\left[(2d)^{-\frac{1}{2}}\beta^{\frac{1}{\omega}}\right] - \gamma\sqrt{d}}\end{aligned}$$

■

Theorem 2 *Let* χ_1, χ_2 *be defined as* $\chi_2 = \max_{a\in\mathcal{A}}\|a\|_2, \chi_1 = \min_{a\in\mathcal{A}}\|a\|_2$, *and* $J = \log\left(\frac{|\mathcal{A}|L(L+1)}{\delta}\right)$. *Then, we have*

$$\frac{\left\|\hat{\theta} - \theta^*\right\|_2}{\|\theta^*\|_2} = \mathcal{O}\left(\frac{\chi_1 d^{\frac{2\omega-1}{2\omega}} J^{\frac{\omega-1}{\omega}}}{\chi_2 T^{\frac{\omega-1}{2\omega}}}\right)$$

with probability at least $1 - |\mathcal{A}|L\delta$. *Note that* $\omega > 1$ *is the constant from Assumption 4.1.*

Proof We remember that from Lemma D.1, we have that

$$\frac{\left\|\hat{\theta} - \theta^*\right\|_2}{\|\theta^*\|_2} \leq \text{cond}(\mathcal{A}^e)\frac{\|\hat{r} - r\|_2}{\|r\|_2}.$$

From Lemma 4.3, we know that

$$\text{cond}(\mathcal{A}^e) \leq \frac{\chi_1 + \gamma\sqrt{d}}{\chi_2 \left[(2d)^{-\frac{1}{2}} [\beta]^{\frac{1}{\omega}}\right] - \gamma\sqrt{d}}.$$

Here, from Assumption 4.1, $\beta = (3(1-\iota)\epsilon_L)^{\frac{1}{\omega}}$. Moreover, from Lemma 4.3, we have that the error in r is bounded by

$$\frac{\|r - \hat{r}\|_2}{\|r\|_2} \leq \frac{4\epsilon_L}{\mu^* - 8\epsilon_L}.$$

Combining these in Lemma D.1, we have that

$$\begin{aligned}
\frac{\left\|\hat{\theta} - \theta^*\right\|_2}{\|\theta\|_2} &\leq \frac{\chi_1 + \gamma\sqrt{d}}{\chi_2 \left[(2d)^{-\frac{1}{2}} [3(1-\iota)\epsilon_L]^{\frac{1}{\omega}}\right] - \gamma\sqrt{d}} \cdot \frac{4\epsilon_L}{\mu^* - 8\epsilon_L} \\
&\leq \frac{\chi_1 + \gamma\sqrt{d}}{2^{\frac{L(\omega-1)}{\omega}} \chi_2 \left[(2d)^{-\frac{1}{2}} [3(1-\iota)]^{\frac{1}{\omega}}\right] - 2^L\gamma\sqrt{d}} \cdot \frac{4}{\mu^* - 8\epsilon_L}
\end{aligned}$$

Now, for the last phase number, we wish to express this in terms of T instead of our dependence on L. We will use the result from Lemma C.2 that

$$\log(T) \leq \log(2\iota^{-2}dJ) + 2\log(2^L) + \log(2)$$

Using this, we have

$$\left[\frac{T}{4\iota^{-2}dJ}\right]^{\frac{1}{2}} \leq 2^L.$$

Since $\frac{1-\omega}{\omega}$ is negative, we have

$$2^{\frac{L(1-\omega)}{\omega}} \leq \left[\frac{T}{4\iota^{-2}dJ}\right]^{\frac{1-\omega}{2\omega}}.$$

Using this for our bound, we have that

$$\begin{aligned}
\frac{\left\|\hat{\theta} - \theta^*\right\|_2}{\|\theta\|_2} &\leq \frac{\chi_1 + \gamma\sqrt{d}}{2^{\frac{L(\omega-1)}{\omega}} \chi_2 \left[(2d)^{-\frac{1}{2}} [3]^{\frac{1}{\omega}}\right] - 2^l\gamma\sqrt{d}} \cdot \frac{4}{\mu^* - 8\epsilon_L} \\
&\leq \frac{\chi_1 + \gamma\sqrt{d}}{\left[\frac{T}{4\iota^{-2}dJ}\right]^{\frac{\omega-1}{2\omega}} \chi_2 \left[(2d)^{-\frac{1}{2}} [3]^{\frac{1}{\omega}}\right] - 2^L\gamma\sqrt{d}} \cdot \frac{4}{\mu^* - 8\epsilon_L}
\end{aligned}$$

Given $\gamma \leq \frac{2^L}{\|\theta\|_2}$ from Assumption 4.1, we can remove these small constants yielding

$$\frac{\left\|\hat{\theta} - \theta^*\right\|_2}{\|\theta\|_2} = \mathcal{O}\left(\frac{\chi_1 d^{\frac{2\omega-1}{2\omega}} J^{\frac{\omega-1}{2\omega}}}{\chi_2 T^{\frac{\omega-1}{\omega}}}\right)$$

■

Appendix E. Proof of Theorem 3

Lemma E.1 *Given Assumption 4.1, Banerjee et al. (2022) shows that the maximum eigenvalue λ_d of the gram matrix $\sum^T a_t a_t^\top = \mathcal{O}(T)$ and for all other eigenvalues λ_i for all $i \in [d-1]$ satisfies $\lambda_i = \mathcal{O}\left(\frac{T}{d}\right)$.*

Theorem 3 *For a bandit instance $\mathcal{M}$ characterized by reward parameter θ_1^* and action set $\mathcal{A}$, there exists a bandit instance $\mathcal{M}'$ with parameter θ_2^* and the same action set $\mathcal{A}$ such that any inverse estimator incurs error*

$$\max\{\|\hat{\theta} - \theta_2^*\|_2, \|\hat{\theta} - \theta_1^*\|_2\} = \widetilde{\Omega}\left(\sqrt{\frac{d}{T}}\right).$$

Proof This proof will follow the proof of Theorem 1 from Guo et al. (2021). We will establish two bandit instances. The first instance $\mathcal{M}$ is parameterized by the true θ_1^*. The second instance is $\mathcal{M}'$ which is parameterized by θ_2^* where $\theta_2^* := \theta_1^* - \epsilon v$ where $\epsilon \in \mathbb{R}$. We will choose $v \in \mathbb{R}^d$ as a random vector on the unit ball according to a uniform distribution. Suppose one of instances $\mathcal{M}$ and $\mathcal{M}'$ are chosen and we observe the sequence $\mathcal{E}_T := \{a_1, a_2, \ldots, a_T\}$. We denote the reward distribution for an arm a_t under bandit instances $\mathcal{M}$ and $\mathcal{M}'$ as $\mathcal{V}(a_t)$ and $\mathcal{V}'(a_t)$ respectively. Furthermore, we state that the rewards of these bandit instances are a sample from Normal Distributions with variance Σ^2. Formally, we state that $\mathcal{V}(a_t) \sim N(\langle \theta_1^*, a_T \rangle, \Sigma^2)$ and $\mathcal{V}'(a_t) \sim N(\langle \theta_2^*, a_T \rangle, \Sigma^2)$. We reduce the reward estimation error to that of binary testing between these two instances, as in the Le-Cam approach.

Given some series of actions $\mathcal{E} := \{a_1, a_2, \ldots, a_T\}$ generated by our demonstrator where $\mathcal{E} \in \mathcal{F}$ and $\mathcal{F}$ is the sigma-algebra of possible events, i.e. $\mathcal{F}_T = \sigma(\{a_1, a_2, \ldots, a_T\})$. Our bandit instances $\mathcal{M}$ and $\mathcal{M}'$ have the probability distributions over all possible series of actions $\mathbb{P}$ and $\mathbb{P}'$, acting over $\mathcal{F}_T$. Given LeCam (1973), any algorithm choosing between the two bandit instances with a decision $\hat{\theta}$, it must at least suffer an error

$$\begin{aligned}
\mathop{\mathbb{E}}_{v}\left[\max\{\mathbb{E}_1\left(\left\|\hat{\theta} - \theta_2^*\right\|_2\right), \mathbb{E}_2\left(\left\|\hat{\theta} - \theta_1^*\right\|_2\right)\}\right] &\geq \mathop{\mathbb{E}}_{v}\left[\frac{1}{2}\|\epsilon v\|(1 - \|\mathbb{P}' - \mathbb{P}\|_{\text{TV}})\right] \\
&\geq \mathop{\mathbb{E}}_{v}\left[\frac{1}{2}\|\epsilon v\|\left(1 - \sup_{\mathcal{E}\in\mathcal{F}_T}|\mathbb{P}(\mathcal{E}) - \mathbb{P}'(\mathcal{E})|\right)\right] \quad (11)
\end{aligned}$$

where Equation (11) comes from the definition of the total variation. Here, we rely on the result of Lemma 19 from Kaufmann et al. (2014) stating that

$$\sup_{\mathcal{E}\in\mathcal{F}_T}|\mathbb{P}(\mathcal{E}) - \mathbb{P}'(\mathcal{E})| \leq \sum_{t=1}^{T} \text{KL}(\mathcal{V}(a_t), \mathcal{V}'(a_t)).$$

However, remembering that the reward distributions are normally distributed with well-defined means and variances, we get

$$\text{KL}(\mathcal{V}(a_t), \mathcal{V}'(a_t)) = \frac{\epsilon^2(\langle a_t, v\rangle)^2}{2\Sigma^2}.$$

Here, we introduce the term $\alpha_{t,d} = \langle a_t, v \rangle$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{v}\left(\sum_{t=1}^{T}\left(\langle a_t, v\rangle\right)^2\right) &= \mathop{\mathbb{E}}_{v}\left(\sum_{t=1}^{T} v^\top a_t a_t^\top v\right) \\
&= \mathop{\mathbb{E}}_{v}\left(v^\top\left(\sum_{t=1}^{T} a_t a_t^\top\right) v\right) \\
&= \mathop{\mathbb{E}}_{v}\left(\sum_{i}^{d} \alpha_i^2 e_i^\top\left(\sum_{t=1}^{T} a_t a_t^\top\right) e_i^\top\right) \\
&= \mathop{\mathbb{E}}_{v}\left(\sum_{i}^{d} \alpha_i^2 e_i^\top\left(\sum_{t=1}^{T} a_t a_t^\top\right) e_i^\top\right) \\
&= \sum_{i}^{d} \frac{1}{d}\|e_i\|_2^2 \lambda_i \qquad (12) \\
&\leq \|e_i\|_2^2 \frac{T}{d} + \sum_{i}^{d-1} \frac{1}{d}\|e_i\|_2^2 \lambda_i \\
&\leq \|e_i\|_2^2 \frac{T}{d} + \max_{i\in[d-1]}\left(\frac{1}{d}\lambda_i \|e_i\|_2^2\right) \\
&\leq \frac{T}{d}\max_{i\in[d]}\left(\|e_i\|_2^2\right) \qquad (13)
\end{aligned}$$

where Equation (13) comes from Banerjee et al. (2022) saying $\lambda_i \leq \mathcal{O}\left(\frac{T}{d}\right)$ for $i \leq d-1$ and $\lambda_d \leq \mathcal{O}(T)$. We will call the quantity from Equation (13) as $\Lambda = \frac{T}{d}\max_{i\in[d-1]}\left(\|e_i\|_2^2\right)$. We finally have

$$\sup_{\mathcal{E}\in\mathcal{F}_T} |\mathbb{P}(\mathcal{E}) - \mathbb{P}'(\mathcal{E})| \leq \frac{\epsilon^2\Lambda}{\Sigma^2}$$

Therefore, we arrive at the final

$$\begin{aligned}
\mathop{\mathbb{E}}_{v}\left(\max\{\mathbb{E}_1\left(\left\|\hat{\theta} - \theta_2^*\right\|_2\right), \mathbb{E}_2\left(\left\|\hat{\theta} - \theta_1^*\right\|_2\right)\}\right) &\geq \frac{1}{2}\|\epsilon v\|\left(1 - \frac{\epsilon^2\Lambda}{\Sigma^2}\right) \\
&\geq \frac{\epsilon\|v\|}{2} - \frac{\epsilon^3\|v\|\Lambda}{2\Sigma^2}
\end{aligned}$$

To maximize the lower bound, we set $\epsilon = \frac{\Sigma}{\sqrt{2\Lambda}}$ to get ,

$$\mathop{\mathbb{E}}_{v}\left(\max\{\mathbb{E}_1\left(\left\|\hat{\theta} - \theta_2^*\right\|_2\right), \mathbb{E}_2\left(\left\|\hat{\theta} - \theta_1^*\right\|_2\right)\}\right) \leq \frac{\Sigma\|v\|}{3\sqrt{3\Lambda}}.$$

Substituting in Λ, we get

$$\begin{aligned}\mathop{\mathbb{E}}_{v}\left(\max\{\ \mathbb{E}_1\left(\left\|\hat{\theta}-\theta_2^*\right\|_2\right),\mathbb{E}_2\left(\left\|\hat{\theta}-\theta_1^*\right\|_2\right)\}\right) &\leq \frac{\Sigma\|v\|}{3\sqrt{3\Lambda}}\\ &\leq \frac{\Sigma\|v\|\sqrt{d}}{3\sqrt{3T}\max\limits_{i\in[d-1]}\left(\|e_i\|_2\right)}\\ &\leq \frac{\Sigma\sqrt{d}}{3\sqrt{3T}\max\limits_{i\in[d-1]}\left(\|e_i\|_2\right)}\end{aligned}$$

Therefore, we get our final claim

$$\mathop{\mathbb{E}}_{v}\left(\max\{\mathbb{E}_1\left(\left\|\hat{\theta}-\theta_2^*\right\|_2\right),\mathbb{E}_2\left(\left\|\hat{\theta}-\theta_1^*\right\|_2\right)\}\right)\leq\mathcal{O}\left(\sqrt{\frac{d\Sigma^2}{T}}\right).$$

Therefore, in expectation of v, we have the desired quantity.

■

Appendix F. Proof of Lemma 4.4

Lemma 4.4 *Let $G=\cos(\kappa)\|\theta^*\|_2-3(1-\iota)\epsilon_L$ for notational ease. Given any value $\omega\in[1,\infty)$, we can construct a bandit instance that satisfies Assumption 4.1. Specifically, Assumption 4.1 is satisfied by two-dimensional bandit instances that are rotationally isomorphic to the bandit instance where*

1. *θ^* forms an angle κ with the vector $(1,0)$ where*

$$\kappa\in\left[\max\left(-\cos^{-1}\left(\frac{3(1-\iota)\epsilon_L}{\|\theta^*\|_2}\right),\cos^{-1}(0)+\beta-\pi\right),\right.$$
$$\left.\min\left(\cos^{-1}\left(\frac{3(1-\iota)\epsilon_L}{\|\theta^*\|_2}\right),\cos^{-1}(0)-\beta\right)\right].$$

2. *$\forall(x,y)\in\mathcal{A}$ s.t. $(x,y)\neq(1,0),\ \cos(\kappa+\tan^{-1}(y,x))\|\theta^*\|_2\sqrt{x^2+y^2}<\cos(\kappa)\|\theta^*\|_2$.*
3. *The two points $\left(\frac{G\cos(\beta)}{\cos(\kappa+\beta)\|\theta^*\|_2},\frac{G\sin(\beta)}{\cos(k+\beta)\|\theta^*\|_2}\right),\left(\frac{G\cos(-\beta)}{\cos(\kappa-\beta)\|\theta^*\|_2},\frac{G\sin(-\beta)}{\cos(k-\beta)\|\theta^*\|_2}\right)\in\mathcal{A}$.*

We have defined two instances $\mathcal{M}_1=(\theta_1^*,\mathcal{A}_1)$ and $\mathcal{M}_2=(\theta_2^*,\mathcal{A}_2)$ as rotationally isomorphic if there exists a rotation operation $\mathcal{R}$ such that $\mathcal{R}(\theta_1^*)=\theta_2^*$ and $\mathcal{R}$ is a bijective function from $\mathcal{A}_1$ to $\mathcal{A}_2$.

Proof For visualization purposes, we will demonstrate the existence of an action set $\mathcal{A}\subset\mathbb{R}^2$, which satisfies our assumptions for a prechosen value ω. We provide an example visualization in Figure 4. Without loss of generality, set the optimal arm a^* to be the vector $(1,0)$. Let

$$\kappa\in\left[\max\left(-\arccos\left(\frac{3(1-\iota)\epsilon_L}{\|\theta^*\|_2}\right),\arccos(0)+\beta-\pi\right),\right.$$
$$\left.\min\left(\arccos\left(\frac{3(1-\iota)\epsilon_L}{\|\theta^*\|_2}\right),\arccos(0)-\beta\right)\right]$$

be the angle formed between θ^* and a^* where a^* is the reference point and $\theta^* \in \mathbb{R}^d$. In this setting, $\mu^* = \cos(\kappa)\|\theta^*\|_2$. We remind the reader that we set $\beta = (3(1-\iota)\epsilon_L)^{\frac{1}{\omega}}$.

The claim is that the following conditions are sufficient for an action set to satisfy Assumption 4.1 for a given ω.

1. $\forall (x, y) \in \mathcal{A}$ s.t. $(x, y) \neq a^*, \cos(\kappa + \text{atan2}(y, x))\|\theta^*\|_2\sqrt{x^2+y^2} < \cos(\kappa)\|\theta^*\|_2$
2. The points $\left(\frac{(\cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L)\cos(\beta)}{\cos(\kappa+\beta)\|\theta^*\|_2}, \frac{(\cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L)\sin(\beta)}{\cos(k+\beta)\|\theta^*\|_2}\right)$ and $\left(\frac{(\cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L)\cos(-\beta)}{\cos(\kappa-\beta)\|\theta^*\|_2}, \frac{(\cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L)\sin(-\beta)\|\theta^*\|_2}{\cos(k-\beta)\|\theta^*\|_2}\right)$ are both in $\mathcal{A}$.

In the visualization (Figure 4), the orange line denotes the first constraint so that all points to the left of the orange line satisfy the first constraint. Moreover, the points from the second constraint are Points 1 and 3 in Figure 4.

We will evaluate the reward of Point 1. Point 1 forms an angle of β with the optimal arm a^* and, thus, forms an angle of $\beta + \kappa$ with θ^*. Moreover, the ℓ_2 norm of Point 1 is

$$\left|\frac{(\cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L)}{\cos(\kappa+\beta)\|\theta^*\|_2}\right|.$$

Given the restriction on κ, we have that $\frac{(\cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L)}{\cos(\kappa+\beta)\|\theta^*\|_2}$ is strictly positive. Since

$$-\arccos\left(\frac{3(1-\iota)\epsilon_L}{\|\theta^*\|_2}\right) \leq \kappa \leq \arccos\left(\frac{3(1-\iota)\epsilon_L}{\|\theta^*\|_2}\right),$$

the numerator is positive. Moreover, since $\arccos(0) - \beta - \pi \leq \arccos(0) - \beta$ the denominator is positive. Therefore, its reward is

$$\begin{aligned}\frac{(\cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L)}{\cos(\kappa+\beta)\|\theta^*\|_2}\|\theta^*\|_2\cos(\beta+\kappa) &= \cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L \\ &= \mu^* - 3(1-\iota)\epsilon_L\end{aligned}$$

We now do this similarly for Point 3. Point 3 forms an angle of $-\beta$ with the optimal arm a^* and, thus, forms an angle of $\kappa - \beta$ with θ^*. Moreover, the ℓ_2 norm of Point 1 is

$$\left|\frac{(\cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L)}{\cos(\kappa-\beta)\|\theta^*\|_2}\right|.$$

Given the restrictions on κ, the value $\frac{(\cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L)}{\cos(\kappa-\beta)}$ is strictly positive. Since

$$-\arccos\left(\frac{3(1-\iota)\epsilon_L}{\|\theta^*\|_2}\right) \leq \kappa \leq \arccos\left(\frac{3(1-\iota)\epsilon_L}{\|\theta^*\|_2}\right),$$

the numerator is positive. Moreover, since $\arccos(0) + \beta - \pi \leq \arccos(0) + \beta$, the denominator is positive. Therefore, its reward is Therefore, its reward is

$$\begin{aligned}\frac{(\cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L)}{\cos(\kappa-\beta)\|\theta^*\|_2}\|\theta^*\|_2\cos(\kappa-\beta) &= \cos(\kappa)\|\theta^*\|_2 - 3(1-\iota)\epsilon_L \\ &= \mu^* - 3(1-\iota)\epsilon_L\end{aligned}$$

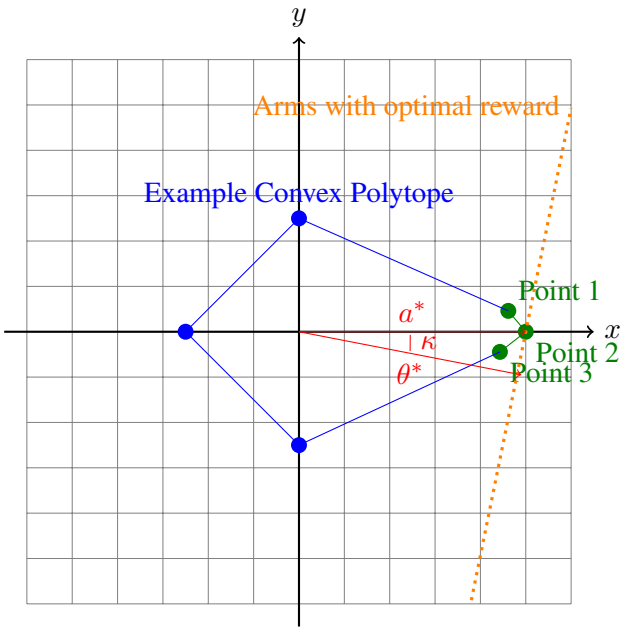


Figure 4: Example Configuration of action set detailed by the proof for Lemma 4.4. The green points are the three points referenced by the proof, the orange line is the line of vectors with the same optimal reward as the optimal Point 2, and the blue lines are example continuations of drawing the convex hull of the action set that satisfy Assumption 3.1. These are done when $\kappa = .2, L = 5,$ and $\beta = .1$.

Moreover, a^*, or Point 2 in Figure 4 has a reward of $\mu^* = \cos(\kappa)\|\theta^*\|_2$. The ℓ_2 norm of Point 2 is 1. It forms an angle of κ with θ^*. Therefore, its reward is $\cos(\kappa)\|\theta^*\|_2$. Given that all points in the action set obey constraint 1 except for a^*, by definition, they have a reward less than $\cos(\kappa)\|\theta^*\|_2$, which is the reward of a^*. Therefore, all points in $\mathcal{A}$ will be rewarded less than a^*. Also, Points 1 and 3 satisfy the first constaint as well. Therefore, these conditions are sufficient to satisfy Assumption 4.1.

■

Appendix G. Implementation details for Phased Elimination used in experiments

Algorithm 3: Phased Elimination

Input : δ (probability parameters), L (number of phases), $\{\nu_1, \ldots, \nu_L\}$ (error parameters)

Result: $a_1, \ldots, a_T$

1 $\ell \leftarrow 0$
2 $\mathcal{A}_1 \leftarrow \mathcal{A}$
3 $t_\ell \leftarrow 0$
4 **while** $\ell < L$ **do**
5 $\quad \varepsilon_\ell \leftarrow 2^{-\ell}$
6 $\quad \pi_\ell \leftarrow$ G-Optimal design of $\mathcal{A}_\ell$ with δ and ν_ℓ
7 $\quad N_\ell \leftarrow 0$
8 $\quad$ **for** $a \in \mathcal{A}_\ell$ **do**
9 $\quad\quad N_\ell(a) \leftarrow \left\lceil \frac{2d\pi_\ell(a)}{\nu_\ell^2} \log\left(\frac{k\ell(\ell+1)}{\delta}\right)\right\rceil$
10 $\quad\quad$ Play action a for $N_\ell(a)$ rounds
11 $\quad\quad N_\ell \leftarrow N_\ell + N_\ell(a)$
12 $\quad$ **end**
13 $\quad V_\ell \leftarrow \sum_{a \in \mathcal{A}_\ell} \pi_\ell(a) a a^\top$
14 $\quad \theta_\ell \leftarrow V_l^{-1} \sum_{t=t_\ell}^{t_\ell + N_\ell} a_t x_t$
15 $\quad \mathcal{A}_{\ell+1} \leftarrow \{a \in \mathcal{A}_\ell \text{ s.t. } \max_{b \in \mathcal{A}_\ell}(\langle \theta_\ell, b - a\rangle) \leq 2\varepsilon_l\}$
16 $\quad t_\ell \leftarrow t_\ell + T_\ell$
17 $\quad \ell \leftarrow \ell + 1$
18 **end**

Algorithm 3 formally describes the implementation of Phased Elimination used in our experiments. The behavior of this implementation only differs from Algorithm 1 in the choice of stopping criteria; here, we stop after a maximum number of phases, while Algorithm 1 fixes T and allows L to vary. Line 6 is computed via a convex program.