Inverse Reinforcement Learning for Phased Elimination of Linear Stochastic Bandits

Etash Guha
Georgia Institute of Technology
Atlanta, GA, 30309
etash@gatech.edu

Jim James
Georgia Institute of Technology
Atlanta, GA, 30309
jimjames@gatech.edu

Krishna Acharya
Georgia Institute of Technology
Atlanta, GA, 30309
krishna.acharya@gatech.edu

Vidya Muthukumar
Georgia Institute of Technology
Atlanta, GA, 30309
vmuthukumar8@gatech.edu

Ashwin Pananjady
Georgia Institute of Technology
Atlanta, GA, 30309
ashwinpm@gatech.edu

Abstract

The "Inverse Bandit" problem entails estimating the rewards seen by a low-regret demonstrator. Existing approaches mainly look at the Multi-Armed Bandit setting, where the arms' rewards are independent. However, in this paper, we turn our eyes to the Linear Stochastic Bandit setting, where the arms' rewards are linked together by some parameterization unknown to the inverse learner. Specifically, we analyze a demonstrator performing the Phased Elimination algorithm, where arms in the action set are sequentially eliminated from consideration. For a demonstrator performing the Phased Elimination algorithm, we provide a low-error inverse estimator that predicts the true rewards of the arms in the action set of the demonstrator given only the demonstrator's actions, including the eliminations. Furthermore, this estimator enjoys an error bound on the order of $\mathcal{O}\left(\sqrt{\frac{d}{T}}\right)$ where d is the dimension of the arms, and T is the number of actions chosen by the demonstrator. Providing empirical verification of these theoretical improvements, we provide experiments demonstrating the error of our estimator as compared to a random baseline.

1 Introduction

The task of Reward Specification is important for building Machine Learning systems that operate safely and effectively [Amodei et al., 2016]. While important, this task often results in rewards misspecified, even in seemingly simple structures where there is more access to information about the rewards [Amin et al., 2017, Gershman, 2016]. A recent paradigm known as **Inverse Reinforcement Learning** (IRL) focuses on the actions taken by a demonstrator or forward algorithm to estimate the reward function. IRL is a well-studied phenomenon with extensions to Apprenticeship Learning [Abbeel and Ng, 2004], imitation learning [Ho and Ermon, 2016], and explicitly learning the reward function [Ng et al., 2000].

Preprint. Under review.

A more difficult problem within the Inverse Reinforcement Learning paradigm is Reward Identifiability. This problem occurs when any inverse learner needs more than the final state of the demonstrator to approximate the reward function. In such a case, an inverse learner must learn from the demonstrator as it navigates its environment and focuses on lower-regret paths. This evolving nature of the demonstrator makes the problem of Reward Identifiability especially tricky. Specifically, with an *optimal* demonstrator, especially traditional IRL algorithms have been shown to struggle [Ng et al., 2000]. Therefore, the goal of Inverse Reinforcement Learning in this setting is to use the actions of evolving demonstrators to estimate the reward function as accurately as possible. Luckily, in the policy evolution process, the demonstrator leaks information about the reward function, namely, which actions are suboptimal and to what degree.

To make this intuition about the information leakage formal, we study the case of IRL for Bandit algorithms. This context is challenging as an optimal demonstrator does not leak any information. However, accurate IRL algorithms have been developed for low-regret demonstrators, such as Successive Arm Elimination [Guo et al., 2021], Upper Confidence Bound algorithm [Guo et al., 2021], Multi-Armed Bandit algorithms in general [Chan et al., 2019]. While these bandit algorithms provide efficient estimators, these estimators are restricted to the Multi-Arm Bandit setting, where the rewards of each arm are independent of each other.

However, for our setting, we analyze the Linear Stochastic Bandit setting, originating from [Abe and Long, 1999]. In this setting, the arms' rewards are linked to a true parameter θ, and a forward demonstrator iteratively takes actions that minimize the regret regarding this true θ. While this setting is natural and practical for modeling many ML tasks, regrettably, few IRL algorithms have been proposed for the linear bandit setting. One example is an IRL algorithm for the variant Linear Contextual Bandits [Hüyük et al., 2022]. To tackle IRL for Linear Stochastic Bandits more efficiently, we analyze the low-regret demonstrator of Phased Elimination. [Esfandiari et al., 2019]. This forward algorithm exhibits strong reward structures as suboptimal arms get eliminated throughout the policy evolution. We provide an IRL algorithm for estimating the reward function given the actions taken by a Phased Elimination demonstration utilizing this structure.

Specifically, our estimator takes advantage of the structure of the elimination criteria of Phased Elimination, including the connection between the rewards of each arm. Looking at a linearly independent set of eliminated arms is enough to generate an accurate estimate of the reward parameterization. With this intuition, we form an IRL algorithm that has an upper bound of error in terms of $\sqrt{\frac{d}{T}}$ where d is the dimension of the arms, and T is the number of actions taken by the demonstrator.

Moreover, we prove that any inverse estimator is information-theoretically bound to this $\sqrt{\frac{d}{T}}$ error rate, in that no other estimator can have a better dependence in d or T. This lower bound proves the optimality of our inverse estimator up to constant values.

Contributions Specifically, our contributions are as follows. We firstly provide some critical background information for understanding the problem setting in Section 2. Given this background, we provide some valuable lemmas explaining some behavior from the Phased Elimination algorithm in Section 3. Utilizing this behavior, we formally define our estimator and prove an upper bound in the error of this estimator in Section 4. In order to provide context for our error bounds for the inverse estimator, we provide an information-theoretic lower bound of the error rate achievable by any inverse estimator. To empirically verify the performance of our estimator, we provide several experiments in Section 6.

2 Preliminary

We formally define the problem of reward estimation formulation from the lens of Stochastic Linear Bandits. We add some formal technical definitions useful for analysis in Section 2.1 as an addendum. We begin with the motivation of the main objective of the Linear Stochastic Bandits in Section 2.2. We discuss several essential details about the shape of the action space that are useful for analysis in Section 2.5. Given the motivation of the setting, we discuss the background of the Phased Elimination algorithm in Section 2.3. Finally, we formally present the problem of Inverse Reinforcement Learning in the setting of Linear Stochastic Bandits in Section 2.5.

2.1 Technical Terminology

We introduce several terms here for technical use, which we will use later. We first define the term γ-close to determine a metric of geometric closeness.

Definition 1. *Two vectors $X, Y \in \mathbb{R}^d$ are γ-close if*

$$\|X - Y\|_2 \leq \gamma.$$

2.2 Linear Stochastic Bandits

The key detail in the Linear Stochastic Bandit setting is that the reward function is some linear function parameterized by θ. The reward of an action from the action set $A \in \mathbb{A}$ is defined as

$$X = \langle \theta, A \rangle + \eta.$$

The noise η, in this case, is sampled from some subgaussian distribution. Therefore, any forward algorithm or demonstrator will continually take actions A_t and experience rewards $X_t = \langle A_t, \theta \rangle + \eta_t$, where $t \in [T]$ and T is the time horizon of the problem, i.e., how many actions the demonstrator takes in total. The main goal of any demonstrator is to maximize the value $\sum_t^T X_t$. The performance of any demonstrator is usually compared to a strategy of pulling the optimal arm $A^* = \arg\max_{A \in \mathbb{A}} \langle \theta, A \rangle$. This difference in values, known as *regret*, is denoted as

Definition 2. *Expected Regret of a Demonstrator The Expected Regret of a demonstrator at time T is denoted as*

$$\mathbb{E}(R_T) = T\mu^* - \mathbb{E}(\sum_t^T \langle \theta, A_t \rangle)$$

where $\mu^* = \langle \theta, A^* \rangle$. For simplicity, we define $\nabla_A = \mu^* - \langle \theta, A \rangle$ as the suboptimality gap of any arm A.

Moreover, any two arms that are γ-close are also close in rewards.

Lemma 1. *Given two arms X, Y that are γ-close, the difference in their rewards is bounded by*

$$\mu_Y - \mu_X \leq \gamma^2 \|\theta\|_2^2.$$

Given the problem setting, Phased Elimination is one such algorithm used to solve this setting that achieves low expected regret.

2.3 Phased Elimination

There have been several algorithms have been analyzed within this linear setting. One such example is the demonstrator known as the Phased Elimination Algorithm. This algorithm sequentially selects arms and then eliminates arms in phases [Valko et al., 2014]. Notably, during a phase l, it picks an experimental design over the set of non-eliminated arms, plays the arms according to the said experimental design, updates its estimates of θ, and eliminates arms that it deems suboptimal. This algorithm is known to achieve the regret bound on the order of $\mathbb{O}(\sqrt{dT})$, the best regret bound known for linear bandit settings Valko et al. [2014]. To understand the algorithm, we must first establish some definitions.

Definition 3. *A* G-optimal design *for an action set $\hat{\mathbb{A}}$ is a function $\pi : \hat{\mathbb{A}} \to \mathbb{R}$ that maximizes $f(\pi) = \log(\det(V(\pi)))$ such that $\sum_{A \in \hat{\mathbb{A}}} \pi(A) = 1$ where $V(\pi) = \sum_{A \in \hat{\mathbb{A}}} \pi(A) A A^T$*

This G-Optimal design is an example of an experimental design that balances the trade between exploitation and exploration necessary to achieve its low-regret bound. For clarity, we provide further details of Phased Elimination in Algorithm 1.

Algorithm 1 Phased Elimination

Data: δ, T
Result: $A_1, \ldots, A_T$
$l \leftarrow 1$
$\mathbb{A}_0 \leftarrow \mathbb{A}$ $N_1 \leftarrow 0$
while $\sum_{i=1}^{l} N_{l-1} \leq T$ **do**
  $\varepsilon_l \leftarrow 2^{-l}$
  $\pi_l \leftarrow$ *G-Optimal design of* $\mathbb{A}_l$
  $N_l \leftarrow 0$
  for $A \in \mathbb{A}_l$ **do**
    $n_l(A) \leftarrow \pi_l(A) \cdot \frac{g(\pi_l)}{\varepsilon_l^2} \log \frac{1}{\delta}$
    $N_l \leftarrow N_l + n_l(A)$
  end
  Play each action $A \in \mathbb{A}_l$ *each* $n_l(A)$ *times*
  $V_l \leftarrow \sum_{A \in \mathbb{A}_l} n_l(A) A A^T$
  $\hat{\theta}_l \leftarrow V_l^{-1} \sum_{t=t_l}^{t_l+T_l} A_t X_t$
  $A_{l+1} \leftarrow \{A \in \mathbb{A}_l \text{ s.t. } \max_{B \in \mathbb{A}_l} (\langle \hat{\theta}_l, B - A \rangle) \leq 2\varepsilon_l\}$
  $l \leftarrow l + 1$
end

While Phased Elimination is an effective low-regret forward algorithm for Linear Stochastic Bandits, the inverse estimator still needs to be better defined.

2.4 Inverse Reinforcement Learning for Linear Stochastic Bandits

The process of Inverse Reinforcement Learning is that of learning the reward function given only the actions taken by a demonstrator. Specifically, given only knowledge of the actions taken within each phase $\pi_1, \ldots, \pi_L$ and the actions available at each phase $\mathbb{A}_1 \ldots \mathbb{A}_L$ where L is the last phase run, we want to estimate θ, the true parameter of the problem. This true parameter is also unknown to both the forward and inverse algorithms. Therefore, an estimator creates a prediction $\hat{\theta}$ minimizing $\frac{\|\hat{\theta}-\theta\|_2}{\|\theta\|_2}$.

2.5 Action Set Details

The set of actions from which a demonstrator can choose is some finite set of vectors $\mathbb{A}$. There are many possible shapes such a set can take. For example, if $\mathbb{A}$ consists only of orthogonal arms, this can be seen as the traditional Multi-Armed Bandit setting. However, a more interesting case is when many arms in $\mathbb{A}$ are not orthogonal, where their rewards are linked by how close they are to each other. We will define several characteristics of $\mathbb{A}$ that are useful for our analysis. Furthermore, we wish to study *nondegenerate* action sets for our purposes. An example of a degenerate action set could be the ℓ_2 unit ball but with a large spike representing A^*. In this case, any demonstrator could exploit such degeneracy, relinquishing its requirement to explore the action set. This degeneracy would make it difficult, intuitively, for any inverse estimator to learn about the rewards associated with directions orthogonal from the optimal arm. In this manner, any inverse estimator would suffer a significant error in its estimation of $\hat{\theta}$. In order to alleviate these issues, we will make several assumptions on our action set $\mathbb{A}$.

Assumption 1. *We assume that there exists a continuous function* $g : \mathbb{R}^d \to 0, 1$ *such that for in any point in the set* $S_g \coloneqq \{v \text{ s.t. } g(v) = 1\}$ *is* γ*-close to at least one point in* $\mathbb{A}$*, where* $\gamma \leq \frac{1}{T\|\theta\|^2}$*. Moreover, any point in the boundary of this set* S_g*,* i.e. $B(\mathbb{A}) \coloneqq \partial S_g$ *is also* γ*-close to any point in* $\mathbb{A}$*.*

Intuitively, this set $B(\mathbb{A})$ is the boundary of our action set or the points in the action set with the largest ℓ_2 norm. Now, a degenerate action set would mean two points in $B(\mathbb{A})$ that are geometrically

close to each other with very different rewards. We place a continuousness assumption to analyze nondegenerate action sets. To ease the analysis, we will consider the function $r(\mathrm{B}) \coloneqq \langle f(\mathrm{B}), \theta \rangle$ where $f(\mathrm{B}) : [0, 2\pi]^d \to \partial S_g$ takes a set of d angles in $\mathrm{B} \coloneqq [\mathrm{B}_1, \ldots, \mathrm{B}_d]$ and finds the vector in ∂S_g that most aligns with the set of angles with the largest norm. For notational ease, we define $\mathrm{B}^* \coloneqq [0]^d$ where $f(\mathrm{B}^*)$ is γ-close to A^*. A vector $f(\mathrm{B})$ forms an angle of B_i with the optimal arm A^* in the hyperplane defined by the ith and $i+1$th axes.

Assumption 2. *Furthermore, we assume that with any ray $\mathbf{v}_\mathrm{B}$, defined as the line that forms an angle B with the optimal arm A^* and goes through the origin. There is only one point in the intersection between $B(\mathbb{A})$ and $\mathbf{v}_\mathrm{B}$,* i.e.

$$|\mathbf{v}_\mathrm{B} \cap B(\mathbb{A})| = 1$$

and furthermore

$$\mathbf{v}_\mathrm{B} \cap B(\mathbb{A}) = \{f(\mathrm{B})\}.$$

This assumption is equivalent to the statements that the boundary $B(\mathbb{A})$ is a connected set and S_g contains no empty space. By definition, we state that B^* is associated with $\mu^* - \gamma^2\|\theta\|_2^2 \leq r(\mathrm{B}^*) \leq \mu^* + \gamma^2\|\theta\|_2^2$ and $f(\mathrm{B}^*)$ is γ-close to A^*.

Assumption 3. *We assume the following properties of $r(\mathrm{B})$. We assume for some neighborhood around B^*,* i.e. $\mathcal{N}_{\beta^*} \coloneqq \{\mathrm{B}$ *s.t.* $\|\mathrm{B} - \mathrm{B}^*\|_2 \leq \frac{1}{d}\}$*, the function r is concave. Moreover, we assume that the function $r(\mathrm{B})$ is 2-Holder Continuous with coefficient H with respect to the ℓ_2 norm. We denote the minimum eigenvalue of H as L. Formally, for any two points $\mathrm{B}, \mathrm{B}' \in \mathcal{N}_{\beta^*}$,*

$$r(\mathrm{B}) - r(\mathrm{B}') \geq H\|\mathrm{B} - \mathrm{B}'\|_2^2.$$

Furthermore, if B and B' only differ in one dimension i, then we can further write that

$$r(\mathrm{B}) - r(\mathrm{B}') \geq L(\mathrm{B}_i - \mathrm{B}'_i)^2.$$

To analyze this problem, we will slightly abuse notation. We state that we will say that β_i is the vector of d angles containing all values for 0 except at the ith index, where it contains the value β. This vector looks like

$$\beta_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \beta \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} 0 \\ \\ i-1 \\ i \\ i+1 \\ \\ d-1 \end{matrix}.$$

In this way, $f(\beta_i)$ is the arm formed by rotating the optimal arm in the hyperplane defined by ith and $i+1$th axes by angle β.

3 Properties of Phased Elimination

The inverse estimator utilizes the information leakage from Phased Elimination to establish its estimate of θ. The Phased Elimination demonstrator leaks information when it decides which arms are eliminated and when. Intuitively, arms eliminated in earlier phases will have larger suboptimality gaps, whereas arms eliminated later will have smaller suboptimality gaps.

This G-Optimal Design naturally helps select arms so that 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 ε_l

Lemma 2. ***Demonstrator's Estimation Error*** *From [Esfandiari et al., 2019], given a G-Optimal design with parameter δ is chosen at each phase, for any $A \in \mathbb{A}_l$, with probability $1 - \delta$, we have*

$$|\langle A, \hat{\theta}_l - \theta^* \rangle| \leq \varepsilon_l.$$

This accuracy of the forward algorithm's θ helps maintain its low regret properties. Furthermore, this property states that the forward algorithm will keep arms with small suboptimality gaps till later phases. More specifically, this property is explicit for the optimal arm A^*, as with a high probability, the forward algorithm will not eliminate the optimal arm.

Corollary 1. *With probability $1-\delta$, for any phase l, $A^* \in \mathbb{A}_l$.*

Since the active arm remains in the active set with high probability, we can make connections between the elimination criteria and the suboptimality gap of an eliminated arm. The elimination criteria solely depend on the estimated reward difference between the arms in the action set. Therefore, the forward algorithm will eliminate an arm when its suboptimality gap is double $2\varepsilon_l$. We present this formally in Lemma 3.

Lemma 3. ***Elimination of Suboptimal Arms*** *Let l_A be the first phase such that the suboptimality gap drops below double the elimination criteria $l_A = \min\{l \text{ s.t. } 4\varepsilon_l \leq \nabla_A\}$. With probability $1-\delta$, arm A will be deleted before phase l_A.*

Given this lemma, we can estimate an arm's true reward based on when the forward algorithm eliminates it. We now look at what set of arms is eliminated at each round.

This shape allows us to define the structure of the set of eliminated arms. We claim that within this shape of an action set, at every phase l, *a linearly independent set of arms are eliminated.* Intuitively, given the true parameter θ, we can find a set of actions where the inner product between any of the actions and θ is between the two bounds for suboptimality gaps as formed by the elimination criteria for phases $l-1$ and l. This claim is proved in Lemma 4.

Lemma 4. ***Linearly Independent Set of Eliminated Arms*** *Given a set of eliminated arms $\mathbb{E}_l = \mathbb{A}_l \setminus \mathbb{A}_{l-1}$, we prove that we can select a subset of arms $\mathbf{A}_l$ from $\mathbb{E}_l$ such that it is linearly independent and spans $\mathbb{R}^d$ with probability $1-d\delta$.*

4 The Inverse Estimator

We take advantage of two ideas to form our Inverse Estimator.

1. Given the connection between each arm's rewards, we only need an accurate estimate of the rewards of a linearly independent subset of the action space to form an estimate $\hat{\theta}$ of θ.
2. At each phase, the demonstrator eliminates a linearly independent subset of the arms.

Given these two ideas, we can intuitively form an estimator. We propose one central idea.

Use the arms eliminated by the last phase to form an estimate of θ.

Since the forward algorithm eliminates a linearly independent set of arms at every phase and we have a high probability bound for the suboptimality gaps for those arms, we can form an estimate of the true rewards for each arm. Our goal is simple. Given some independent set of arms where each arm makes up a row of $\mathbf{A}$ and knowledge of the true rewards for those arms making some vector b, we know the relation $\mathbf{A}\theta = b$ must hold. Intuitively, we have access to both at the last phase to a high level of accuracy. These properties yield the intuition for our inverse estimator. In order to formally construct our estimator, we assume knowledge of the mean of the optimal arm.

Assumption 4. *We assume our inverse learner knows μ^*.*

Assumption 4 is a reasonable assumption in most cases Guo et al. [2021]. Finally, we have all the intuition and assumptions needed to form our inverse learner.

Definition 4. *Our Inverse Estimator Given a demonstrator that has taken steps generating $\pi_1, \ldots, \pi_L$ and $\mathbb{A}_1 \ldots \mathbb{A}_L$, our inverse estimator firstly generates a matrix $\mathbf{A}_L \in \mathbb{R}^d$, where each row is a vector from an independent set of arms from $\mathbb{A}_L$. Our estimate is then $\hat{\theta} = \mathbf{A}_l^{-1}\hat{b}$ where $\hat{b}$ is the d-dimensional vector of all values $\mu^* - 6 * 2^{-L}$.*

We formally state our inverse estimator in Definition 4 and algorithmically in Algorithm 2. Given our knowledge of the optimal arm's true reward, any arm eliminated in the last phase has a reward probably between the elimination criteria of the last phase and the penultimate phase. Therefore,

we can solve for the true parameter θ given the eliminated arms and this estimate of the rewards. However, how to select $\mathbf{A}_l$ has yet to be stated, and it is not immediately clear which $\mathbf{A}_l$ will create the most accurate estimate of $\hat{\theta}$. We present the following technical lemma to provide motivation for how we pick our $\mathbf{A}_l$.

Lemma 5. *If we have the bound $\mathbf{A}\hat{x} = \hat{y}$ and $\mathbf{A}x = y$, we have*

$$\frac{\|\hat{x} - x\|_2}{\|x\|_2} \leq \text{cond}(\mathbf{A}) \frac{\|\hat{y} - y\|_2}{\|y\|_2}.$$

As seen in Lemma 5, to achieve the best error bound, we have the following requirements on our $\mathbf{A}_l$.

1. The rows of $\mathbf{A}_l$ should be linearly independent
2. The rows of $\mathbf{A}_l$ should be in $\mathbb{E}_l$.
3. The condition number of matrix $\mathbf{A}_l$ should be as small as possible.

To provide some motivation, let us first prove that such a well-behaved $\mathbf{A}_l$ exists. For example, given knowledge of the best arm A^*, we can perform the following method to find a well-behaved $\mathbf{A}_l$ To this end, we choose our $\mathbf{A}_l$ in the following manner.

Remark 1. *To form the ith row in $\mathbf{A}_l$, we search for an vector β_i where $\beta \geq \min(\sqrt{\frac{6*2^{-l}}{|L|}}, \frac{1}{d})$ and $\mu^* - 4 \cdot 2^{1-l} \leq r(\beta_i) \leq \mu^* - 4 \cdot 2^{-l}$, and returning the arm in $\mathbb{E}_l$ that is γ-close to the $f(\beta_i)$.*

A $\mathbf{A}_l$ chosen in such a manner satisfies each of our previous conditions. This set trivially satisfies the linear independence requirement given $f(\beta_i)$ used to generate the $\mathbf{A}_l$ are formed by rotating in different orthogonal hyperplanes. Given Lemma 4, there exists some angle at which rotating the optimal arm yields an arm that is close to an arm in $\mathbb{A}_0$. According to Lemma 3, the arms generated will be in $\mathbb{E}_l$ with high probability, satisfying our second property. We demonstrate that a $\mathbf{A}_l$ chosen by rotating the optimal arm by the angle β as in according to Remark 1 satisfies the third property in Lemma 7.

The condition number of $\mathbf{A}_l$ quantifies how linearly independent the arms in $\mathbf{A}_l$ are. Visually, they form some cone around the optimal arm. The cone's width directly correlates with how much these arms are codependent. We want to prove that there is some minimum radius of that cone. This minimum radius would help us prove an upper bound on the condition number of $\mathbf{A}_l$, limiting our error. We prove such a bound in Lemma 6. This lemma states the minimum angle of rotation β in any of the d hyperplanes of rotation needed such that $r(\beta_i)$ is between the elimination criteria of phases l and $l-1$ can be bounded using the Assumption 3.

Lemma 6. *For every $i \in [d]$, the solution β that solves $r(\beta_i) = \mu^* - 6 \cdot 2^{-l}$ obeys the lower bound $\beta \geq \min(\sqrt{\frac{6*2^{-l}}{|L|}}, \frac{1}{d})$.*

Due to the construction of $\mathbf{A}_l$ with angles according to Lemma 6, we can cleanly get an upper bound of the $\mathbf{A}_l$'s condition number as a function of d and l.

Lemma 7. ***Condition Number of $\mathbf{A}_l$*** *We state that the condition number of $\mathbf{A}_l$ generated according to Remark 1 satisfies*

$$\text{cond}(\mathbf{A}_l) = \mathcal{O}\left(\sqrt{2^l d}\right).$$

We have proven that there exists a subset of $\mathbb{E}_l$ arranged in a matrix $\mathbf{A}_l$ with a bounded condition number with high probability. While this matrix was constructed with θ, we need not know θ to find such a matrix in practice. Our inverse estimator has access to $\mathbb{E}_l$. Therefore, our inverse estimator can search through d sized subsets of $\mathbb{E}_l$ until it generates a $\mathbf{A}_l$ with condition number on the same order of Lemma 7. Again, with high probability, such a matrix exists.

A remaining question is from which phase l should we draw our $\mathbf{A}_l$ for our inverse estimator. As is evident in Lemma 7, the larger the phase l, the worse the condition number is. However, intuitively, the later the phase l is, the smaller the gap between the elimination criteria of phases l and $l-1$ is. This property means our estimate of $\hat{b}$ will be closer to b. This intuition is formalized in Lemma 8.

Lemma 8. *We state that $\frac{\|b-\hat{b}\|_2}{\|b\|_2} \leq \frac{2*2^{-l}}{\mu^* - 8*2^{-l}} = \mathcal{O}(2^{-l})$.*

Combining these two lemmas provide an interesting insight: despite the ill-conditioning of the matrix, the error is bounded on the order of $\sqrt{\frac{d}{2^l}}$. Therefore, it is clear that we should take arms from the last set of eliminated arms, where $l = L$. Noting the property $L \approx \log(T)$ where T is the number of arms selected, we get an error bound as in Theorem 1.

Theorem 1. ***Accuracy in terms of*** $\hat{\theta}$ *We claim that with probability at least* $1 - d\delta$*,*

$$\frac{\|\hat{\theta} - \theta\|_2}{\|\theta\|_2} \leq \mathcal{O}\left(\sqrt{\frac{d}{T}}\right).$$

Given this term, we see our intuition turns out to be true. Our estimator only gets more accurate with time, exhibiting an inverse root relationship with T. Unfortunately, our error also grows with the root of the dimension. However, as our lower bound proves, we cannot do any better than such an inverse estimator.

Algorithm 2 Our Inverse Estimator

Data: $[\pi_1, \ldots, \pi_L], [\mathbb{A}_1 \ldots \mathbb{A}_L]$
Result: $\hat{\theta}$
$\mathbb{E}_L \leftarrow \mathbb{A}_L \setminus \mathbb{A}_{L-1}$
for $\mathbf{A}_L \in [\mathbb{E}_L]^d$ **do**
  if $\mathrm{cond}(\mathbf{A}_L) = \mathcal{O}(\sqrt{dT})$ **then**
    $\hat{b} \leftarrow [\mu^* - 6 * 2^{-L}]^d$
    $\hat{\theta} \leftarrow \mathbf{A}_L^{-1}\hat{b}$
    return $\hat{\theta}$
  end
end

5 Information-Theoretic Lower Bound

In order to provide context for how accurate our inverse estimator is, we provide an information-theoretic lower bound on the accuracy achievable by any inverse estimator.

To generate a lower bound, we need an assumption on the structure of the action space from Banerjee et al. [2022].

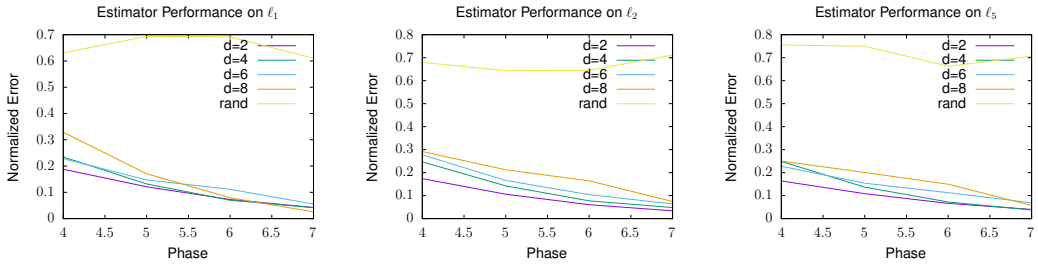
Assumption 5. *We assume our action space is a Locally Constant Hessian Action Space from Banerjee et al. [2022]. This yields the assumption that the minimum eigenvalue* λ_{min} *of* $\sum_1^T A_t A_t^\top$ *obeys* $\lambda_{min} \geq \eta\sqrt{T}$ *where* η *is a constant depending on the condition number of the Hessian of the space and the size of the parameter* θ*.*

Such an assumption entails an action space that is relatively smooth in a manner that each direction or axis is well-represented in the action space. In this manner, learners with low regret guarantees must explore every direction to a certain magnitude. This intuition is formalized by the inequality $\lambda_{\min} \geq \eta\sqrt{T}$ where the actions taken by a learner must not ignore any direction. This property, however, makes it difficult for an inverse learner as an inverse learner can only learn from nonuniform behavior among directions. In contrast, this minimum eigenvalue property enforces some uniform behavior. With such an assumption, we can establish the lower bound appropriately.

Theorem 2. *Firstly, we denote* Σ^2 *as the noise distribution's variance. For any bandit instance* $\mathcal{M}$ *characterized by parameter* θ*, there exists a bandit instance* $\mathcal{M}'$ *with parameter* θ' *such that any inverse estimator incurs error*

$$\max\{\|\hat{\theta} - \theta'\|, \|\hat{\theta} - \theta\|\} = \Omega\left(\sqrt{\frac{d}{T}}\right)$$

Such a bound applies to any reward estimation procedure. Theorem 2 relies on both Assumption 5, the work of Kaufmann et al. [2014], and Banerjee et al. [2022] in order to demonstrate that any



(a) ℓ_1 Ball as Action Set (b) ℓ_2 Ball as Action Set (c) ℓ_5 Ball as Action Set

Figure 1: Estimator performance over the ℓ_1, ℓ_2 and ℓ_5 balls across several dimensions. Our graphs are in terms of phase, not time; however, phase is approximately the logarithm of time. We plot our algorithm's error over phase number for several dimensions for each action set. We additionally provide a random inverse estimator to demonstrate that our inverse estimator is accurate.

inverse estimator needs at least a certain amount of pulls in any direction to establish a reasonable estimate of θ. Therefore, any estimate of the error is fundamentally limited to $\sqrt{\frac{d}{T}}$. This lower bound recovers our upper bound in error, demonstrating that our inverse estimator is optimal up to constants.

6 Experiments

In order to provide empirical verification for our results, we run our inverse estimator on several different settings, measuring the error of the estimate of θ.

6.1 Experiment Setup

We briefly explain the setup of our experiments. We will conduct experiments for three different datasets. The three environments are the ℓ_1, ℓ_2, and ℓ_5 balls. For each environment, we randomly sample a θ from the respective ball, and the action space is a densely sampled set from the respective ball, where each θ and action space makes a bandit instance. For each environment, we sample bandit instances with vector dimensions ranging from 2 to 8. For each dimension and ball pair, we sample 10 bandit instances. We measure the metric of relative error of $\hat{\theta}$, more specifically $\frac{\|\hat{\theta}-\theta\|_2}{\|\theta\|_2}$. As a baseline to compare our error estimates, we use an inverse estimator that randomly chooses θ and the estimate of θ from the demonstrator. We report our results in fig. 1a, fig. 1b, and fig. 1c.

6.2 Results

As indicated by the figures, our algorithm's error demonstrates an inverse root relationship with time. This relationship is coherent with the theoretical bounds promised by Theorem 1. Furthermore, our inverse estimator's estimator should demonstrate a root relationship with dimension according to theorem 1. This relationship is most apparent in fig. 1b. In fig. 1a and fig. 1c, this dependence in d is less clear but still visible. For example, in fig. 1c, the root d dependence is visible for all but the final and first phases. This trend demonstrates empirical verification of theorem 1. Furthermore, across all fig. 1a, fig. 1b, and fig. 1c, our algorithm vastly outperforms the random inverse estimator, as expected. Therefore, our inverse estimator is far more accurate than this random estimator across all action set settings.

References

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.

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.

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.

Lawrence Chan, Dylan Hadfield-Menell, Siddhartha Srinivasa, and Anca Dragan. The assistive multi-armed bandit. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 354–363. IEEE, 2019.

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.

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.

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.

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.

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

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

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.

A Notation Help

Here, we provide a table of notation to help the understanding of this paper.

Symbol	Meaning
d	dimension of environment
θ	True Reward Function Parameter
$\hat{\theta}$	Inverse Estimator's Estimated Reward Parameter
T	Number of Actions taken by Demonstrator
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 highest reward
$\mathbb{A}_l$	Set of remaining arms at phase l
$\mathbb{E}_l$	Set of eliminated arms before phase l
B	Set of d angles
B^*	Set of 0 angles such that $f(\mathrm{B}^*) = A^*$
β_i	Equivalent to B^* except at the ith index where it equals β
$B(\mathbb{A})$	the boundary of a set $\mathbb{A}$
$f(\mathrm{B})$	the arm γ-close to a rotation of B of the optimal arm B^*
$r(\mathrm{B})$	true reward of arm $f(\mathrm{B})$
H	Hölder Continuous Coefficient of reward function r near optimal arm
L	Minimum Eigenvalue of H
ϵ_l	2^{-l} used as criteria for elimination
δ	Parameter for G-Optimal Design, used for high-probability bounds
$\lambda_{\min}$	the minimum eigenvalue of $\sum_1^T A_t A_t^\top$
η	Constant from Banerjee et al. [2022] used to lower bound the exploration of a demonstrator

B Technical Lemmas

B.1 Proof of Lemma 1

Proof. Given two arms X, Y that are γ-close, the difference in their rewards is bounded by

$$\mu_Y - \mu_X \leq \gamma^2 \|\theta\|_2^2.$$

Let us say without loss of generality that the reward of arm X is $\mu_X = \langle X, \theta \rangle$. Therefore,

$$\begin{aligned} \mu_Y - \mu_X &= \langle Y - X, \theta \rangle && (1) \\ &\leq \|Y - X\|_2^2 \|\theta\|_2^2 && (2) \\ &\leq \gamma^2 \|\theta\|_2^2 && (3) \end{aligned}$$

□

C Phased Elimination Properties

C.1 Proof of Corollary Corollary 1

Proof. From Lemma 2, for any suboptimal arm A,

$$\langle A, \hat{\theta}_l \rangle - \langle A^*, \hat{\theta}_l \rangle \leq (\langle A, \theta^* \rangle + \varepsilon_l) - (\langle A^*, \theta^* \rangle - \varepsilon_l) \leq 2\varepsilon_l.$$

The event from $Lemma\ 2$ occurs with probability $1 - \delta$, so this result also happens with probability $1 - \delta$. □

C.2 Proof of Lemma 3

Proof. The statement arm A will be deleted before phase l_A is equivalent to the event $P(A \notin \mathbb{A}_{l_A} \cap A \in \mathbb{A}_{l_A-1})$. We consider the good event where $A^* \in \mathbb{A}_{l_A}$. Note that this trivially implies that A^* is active for all the previous phases. This event happens with $1-\delta$ probability according to Corollary 1. Given that this is true, we will first prove that this event implies $A \notin \mathbb{A}_{l_A}$.

The elimination criteria for arm A in phase l_A is $\max_{B \in \mathbb{A}_{l_A-1}} (\langle B - A, \hat{\theta}_{l_A} \rangle)$. Trivially,

$$\max_{B \in \mathbb{A}_{l_A-1}} (\langle B - A, \hat{\theta}_{l_A} \rangle) \geq \langle A^* - A, \hat{\theta}_{l_A} \rangle.$$

Given Lemma 2, $\langle A^* - A, \hat{\theta}_{l_A} \rangle \geq (\langle A^*, \theta^* \rangle - \epsilon_{l_A}) - (\langle A, \theta^* \rangle + \epsilon_{l_A})$. From the definition of the suboptimality gap, it holds

$$(\langle A^*, \theta^* \rangle - \varepsilon_{l_A}) - (\langle A, \theta^* \rangle + \varepsilon_{l_A}) \geq \delta_A - 2\varepsilon_{l_A}) \geq 2\varepsilon_{l_A}).$$

This connection implies that arm A will be eliminated before phase l_A. □

C.3 Proof of Lemma 4

Proof. In order to prove that such a linearly independent set of arms exists, we will build this set via construction. Our goal for this proof is to find some d sized set of arms $\mathcal{X}$ such that each element $X_i \in \mathbb{E}_l$ and the set is linearly independent. We will follow these steps to generate the ith element in $\mathcal{X}$.

1. Take the optimal arm A^* and rotate it in the hyperplane defined by the i and $i+1$th axis, where $i+1 = 0$ if $i = d-1$.
2. Stop rotating when the suboptimality gap of the resulting vector is at the midpoint between the elimination criteria of phases l and $l-1$
3. Find the arm in $\mathbb{A}_0$ closest to the generated vector and add it to $\mathcal{X}$ as X_i

For such a procedure to be valid, we need only prove that a resulting vector exists whose reward is between the elimination criteria of the consecutive phases and the closest arm in the action set has the same property. Given that these arms are generated by rotating in the orthogonal hyperplanes, the linear independence of these arms is natural.
The reward of such an arm would be

$$r(\beta_i) = \|f(\beta_i)\| \|\theta\| \cos(\rho)$$

where ρ is the angle between the resulting vector and θ. We note that we recover the optimal arm when $\mathrm{B} = 0$, as in $A^* = f(\mathbf{0})$, where $\mathbf{0}$ is the vector of all 0's. Therefore, at this point, $r(\beta)$ achieves its maximum at the point $\beta_i = 0$ for any $i \in [d]$. We are trying to show that for any i, there exists some β_i' such that $r(\beta_i') = \mu^* - 6 \cdot 2^{-l}$. Clearly, the minimum value that r takes is less than that of $\mu^* - 6 \cdot 2^{-l}$. This is formally shown as $\min_{\beta_i} r(\beta_i) < \mu^* - 6 \cdot 2^{-l}$. Therefore, given that our function g is continuous, r must be continuous. Moreover, r can go both higher and lower than $\mu^* - 6 \cdot 2^{-l}$, there must exist a value β' such that $r(\beta_i') = \mu^* - 6 \cdot 2^{-l}$. The vector that achieves this is $\tilde{X}_i \coloneqq f(\beta_i')$. We show an arm $X_i \in \mathbb{A}_0$ that is close to $\tilde{X}_i$, maintaining the property that its reward is between the elimination criteria of phases l and $l-1$.

Given Assumption 1, there must exist an arm $X_i \in \mathbb{A}_0$ where $\|X_i - \tilde{X}_i\|_2^2 \leq \gamma$. Therefore,

$$\begin{aligned} \langle X_i, \theta \rangle &= \langle X_i - \tilde{X}_i, \theta \rangle + \langle \tilde{X}_i, \theta \rangle && (4) \\ &\leq \gamma^2 \|\theta\|_2^2 + \mu^* - 6\varepsilon_l && (5) \\ &\leq \mu^* - 4\varepsilon_l && (6) \end{aligned}$$

where the first inequality comes from Cauchy-Schwarz and the second inequality comes from the fact that $\gamma^2 \|\theta\|_2^2 << 2\varepsilon_l$. We can similarly show that $\langle X_i, \theta \rangle \geq \mu^* - 8\epsilon_l$. Therefore, given these conditions, according to Lemma 3, arm X will be in set $\mathbb{E}_l$ with prob $1-\delta$. Therefore, with probability greater than $1 - d\delta$ via the union bound. We can generate a set $\mathcal{X}$ with such X's. They are a linearly independent set of arms of size d as a subset of $\mathbb{E}_l$.

□

D Inverse Estimator Properties

D.1 Proof of Lemma 8

Proof. Obviously, we remind that $\hat{b}$ is the vector of all $\mu^* - 6 * 2^{-l}$. Additionally, each individual value of b must be between $\mu^* - 4 * 2^{-l}$ and $\mu^* - 8 * 2^{-l}$. Therefore, the $\max(\hat{b_i} - b_i) = 2 * 2^{-l}$. Therefore, the maximum of the norm is

$$\|\hat{b} - b\|_2 \leq \sqrt{d} * 2 * 2^{-l}.$$

For $\|b\|$, we acknowledge that the smallest b_i can be is $\mu^* - 8 * 2^{-l}$. Thus, $\|b\| \geq \sqrt{d}(\mu^* - 8 * 2^{-l})$. We have our final result with

$$\frac{\|b - \hat{b}\|_2}{\|b\|_2} \leq \frac{2 * 2^{-l}}{\mu^* - 8 * 2^{-l}}.$$

□

D.2 Proof of Lemma 6

Proof. We begin the proof for this claim here.

Either our desired β satisfies the property $\beta \geq \frac{1}{d}$ or not. If the former is true, our desired property is true. If not, we can use Assumption 3. Therefore, we can say that

$$r(\beta_i') - r(\mathrm{B}^*) \geq L(\beta' - 0)^2 \tag{7}$$

$$r(\beta_i') \geq r(\mathrm{B}^*) + L(\beta')^2 \tag{8}$$

Given this, we wish to find a lower bound on how small β' is when the arm generated by β', *i.e.* $f(\beta_i)$, has mean reward $\mu^* - 6 * 2^{-l}$. Therefore, we can solve for such a lower bound by setting

$$r(\mathrm{B}^*) + L\beta'^2 \leq \mu^* - 6 * 2^{-l}.$$

Remembering that $r(\mathrm{B}^*) = \mu^*$ yields $\beta' \geq \sqrt{\frac{6*2^{-l}}{|L|}}$. □

D.3 Proof of Lemma 7

Proof. We will need a lower bound on how small β can be to prove this. If β is small, the arms in $\mathbf{A}_l$ are closer to each other in direction, which will significantly increase the condition number of $\mathbf{A}_l$ intuitively. We provide such a lower bound in Lemma 6

We can now prove the original claim. We will break down the proof of the bound of the condition number into two parts. Decomposing $\mathbf{A}_l$ yields $\mathbf{A}_l = \mathbf{D}(\mathbf{A} + \mathbf{N})$ where $\mathbf{A}$ is the matrix formed by rotation from a central arm, $\mathbf{N}$ is some matrix that takes the rotated vector to the γ-close arm in the action set, and $\mathbf{D}$ is a diagonal matrix where each element is the norm of the ith row of $\mathbf{A}_l$. In this manner, the rows of $\mathbf{A}$ are normalized by the ℓ_2 norm. We use the simple fact that

$$\mathrm{cond}(\mathbf{A}_l) \leq \mathrm{cond}(\mathbf{D})\,\mathrm{cond}(\mathbf{A} + \mathbf{N}).$$

We will first compute the condition number of $\mathbf{D}$. We will secondly compute the condition number $\mathbf{A}$ and then use it to compute the condition number of $\mathbf{A} + \mathbf{N}$.

By design, $\mathbf{D}$ is a diagonal matrix where the ith entry is ℓ_2 norm of the ith row. Therefore, the condition number of $\mathbf{D}$ is directly upper bounded by the ratio of length between the longest arm and the shortest arm in the action set, defined as constant χ. Therefore, we have

$$\mathrm{cond}(\mathbf{D}) \leq \frac{\max_{A \in \mathbb{A}_0} \|A\|_2}{\min_{A \in \mathbb{A}_0} \|A\|_2} = \chi.$$

We first list out three properties of our $\mathbf{A}$ matrix. We know that the rows are linearly independent, and each row is generated by rotating some central vector Z by some angle $\kappa \geq \sqrt{\frac{6*2^{-l}}{|L|}}$ from Lemma 6. We define the matrix $\mathbf{B} = \frac{1}{\sqrt{d}}\mathbf{A}^*$. We state the $\mathrm{cond}(\mathbf{A}) = \mathrm{cond}(\mathbf{B})$, so we need only find $\mathrm{cond}(\mathbf{B})$.

We note that $B^*B_{ij} = \frac{1}{d}\langle \alpha_i, \alpha_j \rangle$ where α_i and α_j are the ith and jth rows of A. Note then that $B^*B_{ii} = \frac{1}{d}$. For $i \neq j$, then $\langle \alpha_i, \alpha_j \rangle$ is the following. Therefore, since the vectors $Z - \alpha_i$ and $Z - \alpha_j$ exist in purely orthogonal planes, the plane of rotation, they form a right angle. They are of the same magnitude $\|Z - \alpha_i\| = \|Z - \alpha_j\| = 2 - 2\cos\kappa$ by the law of cosines. Given this, we have $\|\alpha_i - \alpha_j\| = 2\sqrt{2}(1 - \cos(\beta))$. Therefore, the angle between α_i and α_j by the Law of Cosines is γ where

$$2 - 2\cos\gamma = \|\alpha_i - \alpha_j\| = 2\sqrt{2}(1 - \cos(\kappa)).$$

Therefore, $\cos(\gamma) = 1 - \sqrt{2}(1 - \cos(\kappa))$. Therefore,

$$\langle \alpha_i, \alpha_j \rangle = \cos\gamma = 1 - \sqrt{2}(1 - \cos(\kappa)) \tag{9}$$

for any i, j where $i \neq j$. Consider matrix B^*B. Its diagonal elements are $\frac{1}{d}$ and its nondiagonal elements are $\frac{1}{d}\cos(\gamma)$. This matrix has eigenvalues:

$$\lambda_1, \ldots, \lambda_{d-1} = \frac{1}{d} - \frac{1}{d}\cos(\gamma)$$

$$\lambda_d = \frac{d-1}{d}\cos(\gamma) + \frac{1}{d}$$

Thus, the condition number of B, and therefore A is:

$$\text{cond}(\mathbf{A}) = \sqrt{\frac{(d-1)\cos(\gamma) + 1}{1 - \cos(\gamma)}}$$

Plug in $\cos(\gamma) = 1 - \sqrt{2}(1 - \cos(\kappa))$ to get:

$$\text{cond}(\mathbf{A}) = \sqrt{\frac{(d-1)(1 - \sqrt{2}(1 - \cos(\kappa))) + 1}{1 - (1 - \sqrt{2}(1 - \cos(\kappa)))}} \tag{10}$$

$$= \sqrt{\frac{d - (d-1)\sqrt{2}(1 - \cos(\kappa))}{\sqrt{2}(1 - \cos(\kappa))}} \tag{11}$$

$$= \sqrt{\frac{d - (d-1)\sqrt{2} + (d-1)\sqrt{2}\cos(\kappa)}{\sqrt{2}(1 - \cos(\kappa))}}. \tag{12}$$

We do some simplifications to this bound. Additionally,

$$\sqrt{\frac{d - (d-1)\sqrt{2} + (d-1)\sqrt{2}\cos(\kappa)}{\sqrt{2}(1 - \cos(\kappa))}} \leq \sqrt{\frac{(1-d)\sqrt{2} + d\sqrt{2}\cos(\kappa)}{\sqrt{2}(1 - \cos(\kappa))}} \tag{13}$$

$$= \sqrt{\frac{\mathcal{O}(d)}{\mathcal{O}(1 - \cos(\kappa))}} \tag{14}$$

We can upper bound $\cos(\kappa)$ via its Taylor expansion as

$$\cos(\kappa) \geq 1 - \frac{\kappa^2}{2!} + \frac{\kappa^4}{4!}$$

This in turn lower bounds

$$1 - \cos(\kappa) \leq \frac{\kappa^2}{2!} = \frac{2^{-l}}{2!}$$

Therefore, we finally arrive at the final condition number

$$\text{cond}(\mathbf{A}) \leq \mathcal{O}(\sqrt{2^l d})$$

The probability that this statement holds is $1 - \bigcup_{i=1}^{d} \mathbb{P}(\mathbf{A}_l^i \notin \mathbb{E}_l)$, where $\mathbf{A}_l^i$ is the ith row of $\mathbf{A}^l$, which is lower bounded, given the union bound, by $1 - d\delta$. We have proved the condition number of

$\mathbf{A}$. We will compute the condition number of $\mathbf{A}+\mathbf{N}$. Firstly, we bound the maximum eigenvalue of this matrix.

$$\max_{\mathbf{N}} \max_{x} \left(\frac{\|(\mathbf{A}+\mathbf{N})x\|_2}{\|x\|_2} \right) = \max_{\mathbf{N}} \max_{x} \left(\frac{\|\mathbf{A}x+\mathbf{N}x\|_2}{\|x\|_2} \right) \tag{15}$$

$$\leq \max_{\mathbf{N}} \max_{x} \left(\frac{\|\mathbf{A}x\|_2+\|\mathbf{N}x\|_2}{\|x\|_2} \right) \tag{16}$$

$$\leq \lambda_{\max} + \gamma\sqrt{d} \tag{17}$$

where the second inequality comes from the Triangle inequality, and the third inequality comes from setting x to the maximum eigenvector of A. In this case, remember that the max value of any entry in $\mathbf{N}$ must be less than γ. Therefore, the maximum value of $\frac{\|\mathbf{N}x\|_2}{\|x\|_2}$ is $\gamma\sqrt{d}$. In a similar manner, we can prove that $\min_{\mathbf{N}} \min_{x} \left(\frac{\|(\mathbf{A}+\mathbf{N})x\|_2}{\|x\|_2} \right) \geq \lambda_{\min} - \gamma\sqrt{d}$. Therefore, the condition number of $\mathbf{A}+\mathbf{N}$ is still $\mathrm{cond}(\mathbf{A}+\mathbf{N}) = \mathcal{O}(\sqrt{2^l d})$. Noting that $\mathrm{cond}(\mathbf{D})$ is bounded by some constant χ, we can say that in terms of dependence on l and d, we have

$$\mathrm{cond}(\mathbf{A}_l) = \mathcal{O}(\sqrt{2^l d}).$$

□

E Proof of Theorem 2

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. The second instance is $\mathcal{M}'$ which is parameterized by θ' where $\theta' \coloneqq \theta - \epsilon e_d$ where $\epsilon \in \mathbb{R}$ and e_d is the eigenvector corresponding the maximum eigenvalue λ_d of $V_T = \sum A_t A_t^T$. Suppose one of instances $\mathcal{M}$ and $\mathcal{M}'$ are chosen and we observe the sequence $\mathcal{E}_T \coloneqq \{A_1, A_2, \ldots, A_T\}$. 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 $\mathcal{V}(A_t) \sim N(\langle \theta, A_T \rangle, \Sigma^2)$ and $\mathcal{V}'(A_t) \sim N(\langle \theta', A_T \rangle, \Sigma^2)$. 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 $\coloneqq \{A_1, A_2, \ldots, A_T\}$ generated by our demonstrator where $\in \mathcal{F}$ 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 $\mathcal{P}$ and $\mathcal{P}'$, 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

$$\max\{\mathbb{E}_0(\|\hat{\theta}' - \theta'\|, \mathbb{E}_1(\|\hat{\theta} - \theta\|)\} \geq \frac{1}{2}\|\epsilon e_d\|(1 - \|\mathbb{P}' - \mathbb{P}\|_{\mathrm{TV}}) \tag{18}$$

$$\geq \frac{1}{2}\|\epsilon e_d\| \left(1 - \sup_{\mathcal{E}\in\mathcal{F}_T} |\mathbb{P}(\mathcal{E}) - \mathbb{P}'(\mathcal{E})|\right) \tag{19}$$

where the last inequality comes from the definition of the total variation.

Here, we rely on the result of Lemma 19 from Kaufmann et al. [2014] stating that

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

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

$$\mathrm{KL}(\mathcal{V}(A_t), \mathcal{V}'(A_t)) = \frac{\epsilon^2(\langle A_t, e_d \rangle)^2}{2\Sigma^2}.$$

Here, we introduce the term $\alpha_{t,d} = \langle A_t, e_d \rangle$. By definition, $\lambda_d^2 = \sum_{t=1}^{T} \alpha_{t,d}^2$ is the square of dth eigenvalue corresponding to e_d. Naturally, we know that $\sum_i^d \lambda_i^2 = T$. Therefore,

$$\sum_{t=1}^{T} \alpha_{t,d}^2 = \lambda_d^2 = T - \sum_{i}^{d-1} \lambda_i^2.$$

However, from Banerjee et al. [2022], we have the lower bound of $\lambda_i \geq \eta\sqrt{T}$. Therefore, we have an upper bound of

$$\sum_{t=1}^{T} \alpha_{t,d}^2 \leq (1-(d-1)\eta^2)T \tag{20}$$

We finally have

$$\sup_{\mathcal{E}\in\mathcal{F}_T} |\mathbb{P}(\mathcal{E}) - \mathbb{P}'(\mathcal{E})| \leq \frac{\epsilon^2(1-(d-1)\eta^2)T}{\Sigma^2} \tag{21}$$

Therefore, we arrive at the final

$$\max\{\mathbb{E}_0(\|\hat{\theta}' - \theta'\|, \mathbb{E}_1(\|\hat{\theta} - \theta\|)\} \geq \frac{1}{2}\|\epsilon e_d\| \left(1 - \frac{\epsilon^2(1-(d-1)\eta^2)T}{\Sigma^2}\right) \tag{22}$$

$$\geq \frac{\epsilon}{2} - \frac{\epsilon^3(1-(d-1)\eta^2)T}{2\Sigma^2} \tag{23}$$

Choosing $\epsilon = \sqrt{\frac{\Sigma^2}{3(1-(d-1)\eta^2)T}}$ in order to maximize the right-hand-side of the above yields,

$$\max\{\mathbb{E}_0(\|\hat{\theta}' - \theta'\|, \mathbb{E}_1(\|\hat{\theta} - \theta\|)\} \geq \frac{\|e_d\|}{3^{1.5}}\sqrt{\frac{\Sigma^2}{3(1-(d-1)\eta^2)T}} \tag{24}$$

Furthermore, from Banerjee et al. [2022], we have $\eta = \Omega(\frac{1}{\sqrt{d}})$. Therefore, we can say that for some constant C, we have $\eta^2 \geq \frac{C}{d}$. Using this yields

$$\frac{\|e_d\|}{3^{1.5}}\sqrt{\frac{\Sigma^2}{3(1-(d-1)\eta^2)T}} = \Omega\left(\sqrt{\frac{\Sigma^2}{1-\frac{d-1}{d}T}}\right) = \Omega\left(\sqrt{\frac{d\Sigma^2}{T}}\right)$$

□