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

A NOTATION TABLE

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,

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

□

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 $b_{\ell-1} = \arg\max_{b \in \mathcal{A}_{\ell-1}} \langle b, \theta_{\ell-1} \rangle$.

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

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 $b_\ell = \arg\max_{b \in \mathcal{A}_\ell} \langle b, \theta_\ell \rangle$.

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

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 $\mathcal{A}_\ell \backslash \mathcal{A}_{\ell-1}$, 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 $J \coloneqq \left(\frac{|\mathcal{A}|L(L+1)}{\delta}\right)$.

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 $J \coloneqq \left(\frac{|\mathcal{A}|L(L+1)}{\delta}\right)$ 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.

□

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*

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

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*

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

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 $\sigma_{\min}(\mathbf{A}_\ell)$.

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

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

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

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 $\mathbf{A}$.

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

□

Theorem 4.1. *Let* $\chi_2 = \max_{a\in\mathcal{A}}\|a\|_2$, $\chi_1 = \min_{a\in\mathcal{A}}\|a\|_2$ *and define* $J = \log\left(\frac{|\mathcal{A}|L(L+1)}{\delta}\right)$ *as shorthand. Then, we have*

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

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

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

From Lemma 4.3, we know that

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

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

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

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

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

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

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

□

E LOWER BOUND PROOFS

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 5.1. *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 $\theta_2^* \coloneqq \theta_1^* - \epsilon v$ 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 $\mathcal{E}_T \coloneqq \{a_1, a_2, \ldots, a_T\}$. 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 $\mathcal{V}(a_t) \sim N(\langle \theta_1^*, a_T \rangle, \Sigma^2)$ 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 $\mathcal{E} \coloneqq \{a_1, a_2, \ldots, a_T\}$ 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

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

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 $\lambda_d \leq \mathcal{O}\left(T\right)$. 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

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

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.

□

F PROOF OF LEMMA 4.4

Lemma F.1. *Given any value $\omega \in [1, \infty)$, there exists a linear bandit instance (i.e., a set of arms and a reward function) that satisfies Assumption 4.1.*

To prove the above claim, we will prove the following lemma which implies Lemma 4.4.

Lemma F.1. *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. *All arms (x, y) in the action set $\mathcal{A}$ that aren't $(1, 0)$ satisfyu*

$$\cos(\kappa + \tan^{-1}(y, x))\|\theta^*\|_2\sqrt{x^2+y^2} < \cos(\kappa)\|\theta^*\|_2.$$

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

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

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

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

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.

□

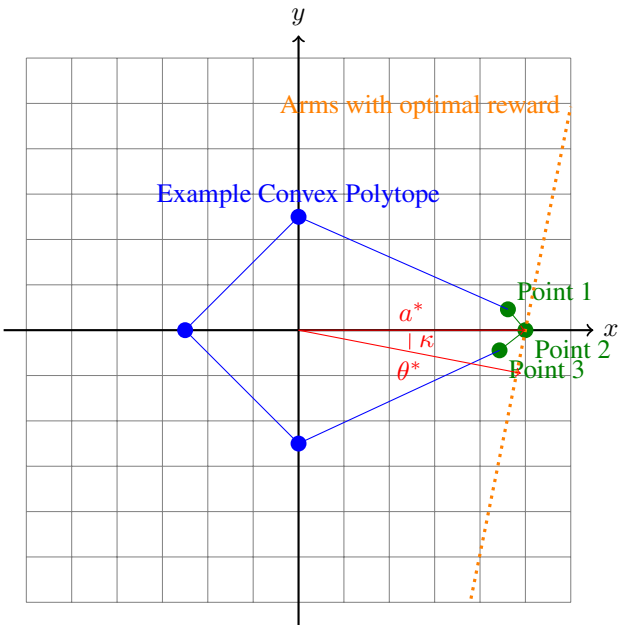

Figure 4: Example Configuration of action set detailed by the proof for Lemma 4.4. The green points are the three points referenced by the proof, the orange line is the line of vectors with the same optimal reward as the optimal Point 2, and the blue lines are example continuations of drawing the convex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

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: $a_1, \ldots, a_T$

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