Estimating Optimal Policy Value in Linear Contextual Bandits beyond Gaussianity

Jonathan N. Lee [1] **Weihao Kong** [2] **Aldo Pacchiano** [3] **Vidya Muthukumar** [4] **Emma Brunskill** [1]

Abstract

While much of bandit learning focuses on minimizing regret while learning an optimal policy, it is often of interest to estimate the maximum value achievable before learning the optimal policy, which can be of use as an input to downstream tasks like model selection. Prior work in contextual bandits has considered this in the Gaussian setting. Here, we study the problem of approximating the optimal policy value in the more general linear contextual bandit problem, and we focus on whether it is possible to do so with less data than what is needed to learn the optimal policy. We consider two objectives: (1) estimating upper bounds on the value and (2) estimating the value directly. For the first, we present an adaptive upper bound that is at most logarithmic factor larger than the value and tight when the data is Gaussian and show that it is possible to estimate this upper bound in $\widetilde{\mathcal{O}}(\sqrt{d})$ samples where d is the number of parameters. As a consequence of this bound, we show improved regret bounds for model selection. For the second objective, we present a moment-based algorithm for estimating the optimal policy value with sample complexity $\widetilde{\mathcal{O}}(\sqrt{d})$ for sub-Gaussian context distributions whose low order moments are known.

1. Introduction

Classic paradigms in multi-armed bandits (MAB), contextual bandits (CB), and reinforcement learning (RL) consider a plethora of objectives from best-policy identification to regret minimization. The meta-objective is typically to learn an explicit, near-optimal *policy* from samples. The actual performance of an optimal policy, typically denoted as *optimal value* V^* is often unknown ahead of time. This quantity may depend in complex ways on the nature of the context space, as well as the function approximators used to represent the policy class or value functions. In many applications, many learner's choices such as the richness of the policy class are often unclear a priori. In such learning situations it would be useful if it was possible to quickly estimate V^* and assess the target performance value in order to decide whether to adjust the learner's choices before spending valuable resources solving the task. For example, prior work that used online policy search to optimize educational activity selection has sometimes found that some of the educational activities contribute little to student success (Antonova et al., 2016). In such settings, if the resulting performance is inadequate, knowing this early could enable a system designer to halt and then explore improvements, like to introduce new actions (Mandel et al., 2017), refine the state representation to enable additional customization (Keramati and Brunskill, 2019) or explore alternate policy classes, in an effort to change the system in order to more effectively support student learning. Indeed a number of recent papers on *on-the-fly* automated online model selection for bandit and reinforcement learning settings (Agarwal et al., 2017; Foster et al., 2019; Chatterji et al., 2020; Pacchiano et al., 2020; Lee et al., 2021) leverage a V^*-estimation (or closely related gap estimation) subroutine. Lee et al. (2021) in particular provides a clean characterization of the additional advantages afforded to online model selection when V^* can be estimated faster than the optimal policy.

Despite its importance, the amount of data needed to estimate V^* is poorly understood in real-world settings. A naive and expensive way to do so would be to plug in the estimate of the value of an approximately optimal policy learned from samples; however, this necessitates fully deploying an algorithm before knowing what it is capable of. In this work, we pursue a more ambitious agenda, and ask *if it is possible to estimate the optimal value V^* faster than learning an optimal policy*. Prior work suggests that this is possible but has only considered a quite restricted setting: (Kong et al., 2020) show that in the setting of disjoint linear contextual bandits with Gaussian contexts with known covariances it is possible to estimate V^* accurately with only $\mathcal{O}_{\epsilon,K}(\sqrt{d})$ samples, a quantity significantly smaller than the $\mathcal{O}_{\epsilon,K}(d)$ samples required to learn a good policy. Here, d is

*Equal contribution [1]Stanford University [2]Google [3]University of California, Berkeley [4]Georgia Institute of Technology. Correspondence to: Jonathan Lee <jnl@stanford.edu>.

the number of parameters and $\mathcal{O}_{\epsilon,K}$ hides dependence on the target accuracy ϵ and number of actions K.

Our work shows that this fast rate is attainable under much more general distributional assumptions on the contexts. We provide a range of procedures to provably estimate V^* and upper-bound surrogates on V^* at a faster rate than estimating the optimal policy itself. Our strongest guarantees are on the surrogates on V^* that are sufficient for testing model misspecification, thereby providing improved model selection guarantees. Our results show that strong estimation-theoretic guarantees are possible even in large-action settings.

1.1. Our contributions

We consider the problem of estimating the optimal value V^* in a d-dimensional stochastic linear contextual bandit problem where we provide several new estimators of V^* and show that they achieve the fast rate $\mathcal{O}_{\epsilon,K}(\sqrt{d})$ in a variety of settings beyond Gaussianity. In particular: (1) We provide information-theoretic lower bounds on the rate of estimation of V^*. We show that if the action set is unrestricted, the rate of estimation for the optimal value V^* scales linearly with the input dimension d in the worst case. (2) When the contexts are sufficiently well-conditioned and second order moments are known, a related task of estimating an informative upper bound on V^* can be done with sample complexity that is sublinear in the problem dimension. Our upper bound is especially useful for approximating gaps between linear model classes and can be used to improve model selsection. (3) When the distribution of the context vectors is sub-Gaussian and moments are known up to an arbitrary order (i.e., known mean, covariance, etc.), we obtain estimates on V^* directly with a large dependence on K and the accuracy ϵ^{-1} but with sublinear sample complexity in the dimension d. This illustrates the surprising ability of estimating V^* with $\mathcal{O}_{\epsilon,K}(\sqrt{d})$ sample size to hold much more generally than the Gaussian setting studied in (Kong et al., 2020).

1.2. Related Work

One can show (see Proposition 1) that in the MAB setting, because there is a lack of shared information between the different arms, estimating the optimal arm's value is no easier than solving the best arm identification problem (Bubeck et al., 2009; Audibert et al., 2010; Gabillon et al., 2012; Karnin et al., 2013; Jun et al., 2016) (equivalently, minimizing the number of samples needed to identify the best arm with high confidence (Even-Dar et al., 2006; Maron and Moore, 1994; Mnih et al., 2008; Jamieson et al., 2014; Katz-Samuels and Jamieson, 2020)). Similarly, best-arm identification has been studied in the non-contextual linear bandit problem (Hoffman et al., 2014; Soare et al., 2014; Karnin, 2016; Tao et al., 2018; Xu et al., 2018; Fiez et al., 2019; Jedra and Proutiere, 2020) where each arm is associated with a fixed feature across rounds.

In this linear CB setting, as well as in the non-disjoint setting (Chu et al., 2011), there is significantly more shared structure across actions. Surprisingly little work has been spent on trying to directly address the problem of V^* estimation even in this case of shared structure. This problem was first proposed by Kong et al. (2020). In the Gaussian context setting they develop a information theoretically optimal and efficient algorithm which can estimate V^* up to ϵ error within $\widetilde{\mathcal{O}}(\frac{\sqrt{d}K}{\epsilon^2})$ in the *disjoint* contextual bandits setting. In this work, we consider the standard non-disjoint linear contextual bandits setting and show that V^*-estimation is possible under significantly broader distribution models for the contexts. A particularly critical application of V^*-estimation arises in online model selection in CB (Agarwal et al., 2017; Foster et al., 2019; Chatterji et al., 2020; Pacchiano et al., 2020; Lee et al., 2021). We leverage our faster estimators of V^* to improve the model selection results of (Foster et al., 2019) in the linear CB setting.

2. Preliminaries

We consider the stochastic contextual bandit problem with a set of contexts $\mathcal{X}$ and a finite set of actions $\mathcal{A} = [K]$ (with $K = |\mathcal{A}|$). At each timestep, a context-reward pair (X_t, Y_t) is sampled i.i.d from a fixed distribution $\mathcal{D}$, where $X_t \in \mathcal{X}$ and $Y_t \in \mathbb{R}^K$ is a reward vector indexable by actions from $\mathcal{A}$. Upon seeing the context X_t, the learner chooses an action A_t and collects reward $Y_t(A_t)$.

Let $f^*(x,a) = \mathbb{E}[Y(a) \mid x]$ and let π^* be the optimal policy such that $\pi^*(x) \in \arg\max_{a\in\mathcal{A}} f^*(x,a)$. The quantity of interest throughout this paper is the average value of the optimal policy, defined as

$$V^* := \mathbb{E}\, Y(\pi^*(X)) = \mathbb{E}_X \max_{a\in\mathcal{A}} f(X,a) \tag{1}$$

For an arbitrary policy $\pi : \mathcal{X} \to \mathcal{A}$, we define $V^\pi = \mathbb{E}\, Y(\pi(X))$. In contextual bandits the learner's goal is typically to either propose a sequence of policies $\{\pi_t\}_{t\in[T]}$ that minimizes cumulative regret

$$\mathrm{Reg}_T(\pi_{1:T}) = \sum_{t\in[T]} V^* - V^{\pi_t} \tag{2}$$

or to sample-efficiently find a policy $\hat{\pi}$ such that $V^* - V^\pi \le \epsilon$ for some target $\epsilon > 0$. We restrict our attention to the *linear* contextual bandit. We assume that there is a known feature map $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ and a vector $\theta \in \mathbb{R}^d$ such that $f(x,a) = \langle \phi(x,a), \theta \rangle$ for all $x \in \mathcal{X}$, $a \in \mathcal{A}$.

2.1. Distribution Assumptions

Throughout this work we make certain distributional assumptions whose definitions we introduce in this section. A random variable Z is sub-Gaussian if there exists $\sigma > 0$ such that $\mathbb{E}\left[|Z|^p\right]^{1/p} \leq \sigma\sqrt{p}$ for all $p \geq 1$ and we define $\|Z\|_\psi$ as the smallest such σ: $\|Z\|_{\psi_2} := \sup_{p\geq 1} p^{-1/2}\mathbb{E}\left[|Z|^p\right]^{1/p}$. A random vector $\overline{Z}$ is sub-Gaussian if there exists σ such that $\|\overline{Z}\|_{\psi_2} := \sup_{v\in\mathcal{S}^{d-1}} \|\langle \overline{Z}, v\rangle\|_{\psi_2} \leq \sigma$.

We assume that the mean reward for any fixed action is zero. That is, $\mathbb{E}f^*(X,a) = 0$ for all $a \in \mathcal{A}$ and $Y(a) - \mathbb{E}[Y(a)|X] \sim \mathrm{subG}(\sigma^2)$ for all $a \in \mathcal{A}$. Now, we state two distributional assumptions on the context distributions that will be common throughout the rest of the paper. Some additional assumptions will be required later for certain individual results.

Assumption 1. *The covariance matrices* $\Sigma_a = \Sigma_a := \mathbb{E}_X\left[\phi(X,a)\phi(X,a)^\top\right]$ *and* $\Sigma_{a,a'} = \mathbb{E}_X\left[(\phi(X,a)-\phi(X,a'))(\phi(X,a)-\phi(X,a'))^\top\right]$ *are known. Furthermore, The average covariance matrix of all actions* $a \in [K]$ *is the identity:* $\Sigma := \frac{1}{K}\sum_{a\in[K]}\Sigma_a = \mathbb{I}_d$.

The identity requirement is essentially requiring that the average covariance matrix Σ is well-conditioned as we can simply apply a linear transformation $\Sigma^{-1/2}\phi(X,A)$ where $A \sim \mathrm{Unif}[K]$ to whiten the data. That these covariance matrices are known[1] is realistic in settings when a large amount of non-interaction data is available, i.e., unlabeled data from $\mathcal{D}$ only containing the context X, in which case Σ_a and $\Sigma_{a,a'}$ can be estimated with high accuracy. This known covariance setting was also previously considered by (Kong et al., 2020).

Assumption 2. *Let* $(X,Y) \sim \mathcal{D}$ *be an independent context-reward pair. There are constants* $\sigma, \tau > 0$ *such that the following is satisfied. For all* $a \in [K]$, *(1)* $\mathbb{E}\phi(X,a) = 0$. *(2) the noise* $\eta(a) := Y(a) - f(X,a)$ *is independent of* X *and sub-Gaussian with* $\|\eta\|_{\psi_2} \leq \sigma$. *(3)* $\phi(X,a)$ *is sub-Gaussian with* $\|\phi(X,a)\|_{\psi_2} \leq \tau$.

The above is a major departure from (Kong et al., 2020) who instead made the restrictive assumption that η and $\phi(X,\cdot)$ be Gaussian. The weaker conditions of Assumption 2 allow for far more general distributions. Finally, we note that there is a careful balance between requiring that the average covariance matrix be identity and requiring that the sub-Gaussianity parameter τ be constant, which necessarily precludes ill-conditioned distributions. In the next section, we show that both known and well-conditioned covariance matrix Σ is in fact critical to obtain the sublinear estimation results of this work.

[1] In settings where the covariance matrix is unknown, we can frequently plug in an estimate of the covariance matrix from a larger number of *unlabeled* samples and obtain similar guarantees as in the known-covariance case; this is particularly useful in the model selection problem for contextual bandits (Foster et al., 2019). In the absence of a plethora of unlabeled data, impossibility results on estimating V^* are well-known even in the single-action setting (Verzelen et al., 2018).

3. Hardness Results

We begin the discussion of V^* estimation with several negative results, showing that, in certain problem settings, it is not possible to do significantly better than simply finding the optimal policy. A natural starting point is the classical K-armed bandit problem where $f^*(x,a)$ is independent of x (and for this part only we assume that the means are non-zero), equivalently represented as a mean vector $\mu \in \mathbb{R}^K$. The feedback is the same: $Y(a) = \mu_a + \eta(a)$. In this case, V^* is defined as $V^* = \max_a \mu_a$. The following proposition asserts that it is not possible to get significantly better dependence on K or ϵ in estimating V^*.

Proposition 1. *There exists a class of* K*-armed bandit problems satisfying* $\|\mu\|_1 = O(1)$ *such that any algorithm that returns an* ϵ*-optimal estimate of* V^* *with probability at least* $2/3$ *must use* $\Omega(K/\epsilon^2)$ *samples.*

The lower bound for the multi-armed bandit problem can be readily converted to a lower bound for certain linear contextual bandits as well, showing that one can expect at least linear dependence on d in the large action regime.

Proposition 2. *There exists a class of linear contextual bandit problems satisfying Assumption 2, with* $\phi : \mathcal{X}\times\mathcal{A} \to \mathbb{R}^d$ *and* $K \geq d$, *such that any algorithm that returns an* ϵ*-optimal estimate of* V^* *with probability at least* $2/3$ *must use* $\Omega(d/\epsilon^2)$ *samples. Under the same assumption, there exists a class of linear contextual bandit problem with* $K = 2$, *and an absolute constant* c, *such that any algorithm that returns an* c*-optimal estimate of* V^* *with probability at least* $2/3$ *must use* $\Omega(d)$ *samples.*

The first lower bound (for large action-settings) holds even when the algorithm has full knowledge of the covariance matrix Σ of all the actions. The second lower bound holds even in the Gaussian setting, but only when the covariance structure is unknown to the algorithm. Even with knowledge of the covariance, certain non-Gaussian distributions over covariates, such as the construction above, can force a lack of sharing of information that prevents the desired fast estimation even under the conditions of Assumption 2.

These lower bounds suggest that, in general, optimal policy value estimation is no easier than learning the optimal policy itself. It is natural then to ask if fast estimation is possible at all outside of the Gaussian case. In the following sections, we answer this affirmatively and present several positive results made possible by leveraging some additional a structure. In particular a known and well-conditioned covariance by virtue of Assumption 1 proves critical for

such positive results. Note that one of the key aspects of the lower bound construction in Proposition 2 is that the covariate distribution has fairly small singular values: $\mathbb{E}\phi(X,a)\phi(X,a)^\top = \frac{1}{d}\mathbb{I}_d$.

4. Estimating Upper Bounds on V^*

We now present several positive results made possible by leveraging some additional structure. The first will concern a variant of the V^* estimation problem: here, our goal will be to instead approximate V^* with an upper bound and then estimate this upper bound efficiently. Our main insight is to upper-bound the stochastic process given by $\{\langle\theta,\phi(X,a)\rangle\}_{a\in[K]}$ with a Gaussian process matched in covariance, and then estimate the covariance matrix of the Gaussian process. For this to work, we make the following *sub-Gaussian process* assumption.

Assumption 3. *There exists an absolute constant L_0 such that, for all $v \in \mathbf{S}^{d-1}$ and $a, a' \in [K]$,*

$$\| \langle\phi(X,a),v\rangle - \langle\phi(X,a'),v\rangle \|_{\psi_2} \tag{3}$$
$$\leq L_0 \| \langle\phi(X,a),v\rangle - \langle\phi(X,a'),v\rangle \|_{L^2} \tag{4}$$

Under Assumption 3, we now provide an explicit upper bound on V^* and estimate it at a fast rate.

Theorem 1. *Under Assumptions 1, 2,and 3, there are absolute constants $C_1, C_2 > 0$ and a Gaussian process $(Z_a)_{a\in[K]}$ with $U = \mathbb{E}\max_a Z_a$ such that*

$$V^* \leq C_1 \cdot U \leq C_2 \cdot V^* \sqrt{\log K} \tag{5}$$

Furthermore, for $\delta \leq 1/e$, there exists an estimator $\hat{U}$ such that with probability at least $1-\delta$,

$$|U - \hat{U}| \leq \mathcal{O}\left(\frac{\sqrt{\|\theta\|}\log(K/\delta)}{n^{1/4}} + \frac{d^{1/4}\log^{3/2}(dK/\delta)}{\sqrt{n}} \right) \tag{6}$$

When the process $\phi(X,\cdot)$ is Gaussian, we have $V^* = U$ and $\hat{U}$ estimates V^* exactly.

Perhaps surprisingly, the result can viewed as a direct combination of Talagrand's comparison inequality (which arises as a consequence of Talagrand's fundamental "generic chaining" approach in empirical process theory (Talagrand, 2006)), and similar techniques as those developed by (Kong et al., 2020) for estimating V^* in the Gaussian case, but for non-disjoint arms.

4.1. Application to Model Selection

Following a similar setup to that of (Foster et al., 2019), we consider two nested linear function classes $\mathcal{F}_1$ and $\mathcal{F}_2$ where $\mathcal{F}_i = \{(s,a) \mapsto \langle\phi_i,\theta\rangle \ : \ \theta \in \mathbb{R}^{d_i}\}$. Here, ϕ_i maps to $\mathbb{R}^{d_i}$ where $d_1 < d_2$, and the first d_i components of ϕ_1 are identical to ϕ_2. In other words, the function classes are *nested*, i.e. $\mathcal{F}_1 \subseteq \mathcal{F}_2$. The objective is to minimize regret, as defined in Equation (2).

Throughout, we assume that $\mathcal{F}_2$ realizes f^*; however, if $\mathcal{F}_1$ also realizes f^*, we would ideally like the regret to scale with $d_* := d_1$ instead of d_2, as $d_1 \ll d_2$ potentially. If $\mathcal{F}_1$ does not realize f^*, then we accept regret scaling with $d_* = d_2$. The ultimate goal of online model selection is to achieve a regret bound that is $\text{poly}(d_*, K, T)$. Our improved estimators for V^* imply improved bounds on online model selection, as stated below.

Theorem 2. *There exists a model selection algorithm such that, with probability at least $1-\delta$,*

$$\mathit{Reg}_T = \mathcal{O}\left(d_*^{1/4} T^{2/3} \log^{3/2}(d_* TK/\delta) \log^{1/2}(K) \right) \tag{7}$$
$$+ \mathcal{O}\left(\sqrt{d_* KT \log(d_*)} \cdot \log(TK/\delta) \right) \tag{8}$$

5. A Moment-Based Estimator of V^*

Thus far, we have shown an upper bound for V^* and a method of estimating it at a fast rate; however, the question of whether it is possible achieve sublinear complexity for estimating V^* itself in the general case remains open. In this section, we present our main result, which is an estimator that achieves this task in $\widetilde{\mathcal{O}}(\sqrt{d})$ samples. The full algorithm is presented in Algorithm 3 in Appendix C along with intuition for the estimator. The main idea is to first consider a t^{th}-order K-variate polynomial approximation of the K-variate max function. The problem of V^* estimation is then reduced to the problem of estimating the multivariate moments between the $\{\langle\theta,\phi(X,a)\rangle\}_{a\in\mathcal{A}}$ random variables.

Assumption 4. *For all $a \in \mathcal{A}$, the covariance matrix is identity: $\Sigma_a := \mathbb{E}_X\left[\phi(X,a)\phi(X,a)^\top\right] = \mathbb{I}_d$ and there exists a constant $L > 0$ such that for any $v, u \in \mathbb{R}^d$, $\mathbb{E}\left[(\phi(X,a)^\top v)^2(\phi(X,a)^\top u)^2\right] \leq L \cdot \mathbb{E}\left[(\phi(X,a)^\top v)^2\right]\mathbb{E}\left[(\phi(X,a)^\top u)^2\right]$.*

Assumption 5. *For all $a \in \mathcal{A}$, the expected reward conditioned on X satisfies $\langle\phi(X,a),\theta\rangle \in [-1,1]$.*

We furthermore assume that all moments up to degree t of $\langle\phi(X,\cdot),v\rangle$ for any $v \in \mathbf{S}^{d-1}$ are known or can be computed.

5.1. Main Result

Our main result, stated below, shows that it is indeed possible to estimate V^* to high accuracy with $\mathcal{O}_{\epsilon,K}(\sqrt{d})$ samples under these assumptions using Algorithm 3 to estimate the expected value of a particular degree t polynomial, $p_t : [-1,1]^K \to \mathbb{R}$.

Theorem 3. *Let assumptions 1, 2, 4, and 5 hold. Let $\hat{S}_n = \sum_{\alpha \,:\, |\alpha|\leq t} c_\alpha \hat{S}_{n,\alpha}$ as defined in Algorithm 3 be*

the estimator of $\mathbb{E}_X p_t$ up to degree t. There is an absolute constant $C > 0$ such that with probability at least $1 - t(et/K + e)^K\delta$,

$$|V^* - \hat{S}_n| \leq \frac{C_K}{t} \tag{9}$$

$$+ t\,(et/K + e)^K\, c_{\max} \cdot C^{t/2} t^t \cdot \sum_{s=1}^{t} \left(\frac{\sqrt{d}}{n} \cdot \log(1/\delta) \right)^{s/2} \tag{10}$$

where C_K is a constant that depends only on K.

The bound shows that it is indeed possible to estimate V^* with sublinear sample complexity in d beyond only the Gaussian case. In particular, it is possible to estimate V^* even when $d \gg n$, i.e. for high-dimensional problems. Note that the degree of the polynomial p_t controls the approximation error. With an appropriate choice of t, we have the following corollary, making this trade-off explicit.

Corollary 1. *Under the same assumptions as Theorem 3, estimator $\hat{S}_n$ generated by Algorithm 3 satisfies $|V^* - \hat{S}_n| \leq \epsilon$ for $\epsilon < 1$ with probability at least $1 - \delta$ and sample complexity*

$$\mathcal{O}\left(K \left(\frac{C_K}{\epsilon} \right)^{K + C_K/\epsilon} \cdot \frac{\sqrt{d}}{\epsilon^2} \cdot \log\left(\frac{C_K}{\epsilon\delta} \right) \right) \tag{11}$$

where C_K is a constant that depends only on K.

The above corollary explicitly exhibits the $\sqrt{d}$ dependence in the sample complexity of estimation task. However, the major drawback of this estimator is exponential dependence on both ϵ^{-1} and K, which arise as a result of exponentially large variance for estimators of the moments. It remains an open question whether a $\sqrt{d} \cdot \text{poly}(K, \epsilon^{-1})$ algorithm exists for V^* estimation in this general setting.

6. Discussion

In this paper, we studied the problem of estimating the optimal policy value in a linear contextual bandit problem before learning the optimal policy itself. We considered this problem beyond the Gaussian case, and presented estimators for both V^* and informative upper bounds on V^*. In particular, we showed that a fast $\mathcal{O}_{\epsilon,K}(\sqrt{d})$ is possible for estimating V^* directly, given that moments of the context distribution are known up to an arbitrary degree. However, there are several remaining open questions. While Theorem 3 shows that a $\mathcal{O}_{\epsilon,K}(\sqrt{d})$ sample complexity is possible, the algorithm is not very practical as it requires a large polynomial approximation and, as a result, has exponential dependence on ϵ^{-1} and K. We ask whether there is a practical algorithm with a similar guarantee or, even better, a practical algorithm with only $\sqrt{d} \cdot \text{poly}(K, \epsilon^{-1})$ sample complexity. It would also be interesting to explore how the sample complexity degrades if the higher moments of the context distribution are not known exactly or must be estimated from the same data that is collected.

References

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

Rika Antonova, Joe Runde, Min Hyung Lee, and Emma Brunskill. Automatically learning to teach to the learning objectives. In *Proceedings of the third (2016) ACM conference on learning@ scale*, pages 317–320, 2016.

Jean-Yves Audibert, Sebastien Bubeck, and Remi Munos. Best arm identification in multi-armed bandits. 2010.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International Conference on Algorithmic Learning Theory*, 2009.

Sourav Chatterjee. An error bound in the sudakov-fernique inequality. *arXiv preprint math/0510424*, 2005.

Niladri Chatterji, Vidya Muthukumar, and Peter Bartlett. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1844–1854, 2020.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. 2006.

Tanner Fiez, Lalit Jain, Kevin Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *arXiv preprint arXiv:1906.08399*, 2019.

Dylan Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. *arXiv preprint arXiv:1906.00531*, 2019.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. pages 3212–3220, 2012.

David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random

variables. *The Annals of Mathematical Statistics*, 42(3): 1079–1083, 1971.

Matthew Hoffman, Bobak Shahriari, and Nando Freitas. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*, pages 365–374, 2014.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439, 2014.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Kwang-Sung Jun, Kevin Jamieson, Robert Nowak, and Xiaojin Zhu. Top arm identification in multi-armed bandits with batch arm pulls. In *Artificial Intelligence and Statistics*, pages 139–148. PMLR, 2016.

Zohar Karnin. Verification based solution for structured mab problems. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 145–153, 2016.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. 2013.

Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, pages 1781–1791. PMLR, 2020.

Ramtin Keramati and Emma Brunskill. Value driven representation for human-in-the-loop reinforcement learning. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pages 176–180. ACM, 2019.

Weihao Kong and Gregory Valiant. Estimating learnability in the sublinear data regime. *Advances in Neural Information Processing Systems*, 31:5455–5464, 2018.

Weihao Kong, Emma Brunskill, and Gregory Valiant. Sublinear optimal policy value estimation in contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4377–4387. PMLR, 2020.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.

Jonathan Lee, Aldo Pacchiano, Vidya Muthukumar, Weihao Kong, and Emma Brunskill. Online model selection for reinforcement learning with function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3340–3348. PMLR, 2021.

Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. Where to add actions in human-in-the-loop reinforcement learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Oded Maron and Andrew W Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. pages 59–66, 1994.

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical bernstein stopping. pages 672–679. ACM, 2008.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *arXiv preprint arXiv:1506.03271*, 2015.

Aldo Pacchiano, My Phan, Yasin Abbasi-Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. *arXiv preprint arXiv:2003.01704*, 2020.

Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *arXiv preprint arXiv:1409.6110*, 2014.

Michel Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media, 2006.

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

Kevin Tian, Weihao Kong, and Gregory Valiant. Learning populations of parameters. *arXiv preprint arXiv:1709.02707*, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Nicolas Verzelen, Elisabeth Gassiat, et al. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 24(4B):3683–3710, 2018.

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

Krzysztof Zajkowski. Bounds on tail probabilities for quadratic forms in dependent sub-gaussian random variables. *Statistics & Probability Letters*, 167:108898, 2020.

A. Proofs of Results in Section 3

A.1. Additional Notation

We use $[n] = \{1, \dots, n\}$ for $n \in \mathbb{N}$. For any vector $v \in \mathbb{R}^d$, $\|v\| = \|v\|_2$. For any matrix $M \in \mathbb{R}^{d\times d}$, $\|M\|$ denotes the operator norm and $\|M\|_F$ the Frobenius norm. For a random variable Z, $\|Z\|_{L^2} = (\mathbb{E}Z^2)^{1/2}$. We denote the d-dimensional unit sphere $\mathbf{S}^{d-1} = \{v \in \mathbb{R}^d : \|v\| = 1\}$. We call $\binom{[n]}{s}$ the set of s-combinations of $[n]$ and use the symbol $\mathbb{I}_d$ to denote the $d \times d$ identity matrix. We use $C_1, C_2, \dots$ to refer to absolute constants independent of problem parameters. Throughout, we use δ to represent a desired failure probability and assume $\delta \le 1/e$. The symbol $\lesssim$ means $\le$ up to a constant.

A.2. Proposition 1

The proof of the lower bound for the K-armed bandit problem follows a standard argument via Le Cam's method. Let $\hat{V}_n$ denote the output of a fixed algorithm $\mathcal{A}$ after n interactions with the bandit that achieves $|\hat{V}_n - V^*| \le \epsilon$ with probability at least $2/3$. We let ν and ν' denote two separate bandit instances, determined by their distributions.

For shorthand, P_ν and $P_{\nu'}$ denote measures under these instances for the fixed, arbitrary algorithm (and similarly expectations $\mathbb{E}_\nu$ and $\mathbb{E}_{\nu'}$). N_a denotes the (random) number of times the fixed algorithm sampled arm a.

We let ν be distributed as $\mathcal{N}(\mu_a, 1)$ for all a where $\mu = (\epsilon, 0, \dots, 0)$. Then, define $a' \in \arg\min_{a\neq 1} \mathbb{E}_\nu[T_a]$ and let ν' be distributed as $\mathcal{N}(\mu'_a, 1)$ where $\mu'_a = \mu_a$ for all $a \neq a'$ and $\mu'_{a'} = 4\epsilon$. We define the successful events $E_\nu = \{\hat{V}_n \in [0, 2\epsilon]\}$ and $E_{\nu'} = \{\hat{V}_n \in [3\epsilon, 5\epsilon]\}$.

By Le Cam's lemma and Pinsker's inequality,

$$P_\nu(E_\nu^c) + P_{\nu'}(E_\nu) \gtrsim 1 - \sqrt{D_{KL}(P_\nu, P_{\nu'})} \tag{12}$$

where $D_{KL}(P_\nu, P_{\nu'}) \lesssim \mathbb{E}_\nu[N_{a'}]\,\epsilon^2 \le \frac{n\epsilon^2}{K-1}$ (Lattimore and Szepesvári, 2018, Lemma 15.1). It then follows that the probability of the successful event is bounded as

$$P_\nu(E_\nu) \le P_{\nu'}(E_\nu) + C\sqrt{\frac{n\epsilon^2}{K-1}} \tag{13}$$

$$\le P_{\nu'}(E_{\nu'}^c) + C\sqrt{\frac{n\epsilon^2}{K-1}} \tag{14}$$

$$\le \frac{1}{3} + C\sqrt{\frac{n\epsilon^2}{K-1}} \tag{15}$$

for some constant $C > 0$. Thus, in order for $P_\nu(E_\nu) \ge 2/3$ it must be that $n \ge \frac{(K-1)}{9C^2\epsilon^2}$. It follows that any algorithm that achieves such a condition must incur sample complexity $\Omega(K/\epsilon^2)$.

A.3. Proposition 2

Proof. **Proof of the first lower bound.** Fix algorithm $\mathcal{A}$ for the linear contextual bandit problem. Then consider the class of $\frac{d}{2}$-armed bandit problem with means vectors satisfying $\|\mu\| = \mathcal{O}(1)$. From this class, we construct the following class of linear contextual bandits. Let $\theta = \begin{bmatrix}\mu \\ -\mu\end{bmatrix} \in \mathbb{R}^d$. The set of contexts is $\mathcal{X} = \{1, 2\}$ and the feature map is defined as

$$\phi(x, a) = \begin{cases} e_a & x = 1 \\ -e_a & x = 2 \end{cases} \tag{16}$$

where $\{e_1, \dots, e_d\} \subseteq \mathbb{R}^d$ denotes the set of standard basis vectors. Then, $X = 1$ and $X = 2$ each with probability $\frac{1}{2}$. This ensures that $\mathbb{E}\phi(X, a) = 0$ for any fixed action a. Furthermore, $\|X_a\|_{\psi_2} = \Theta(1)$. Note that $V^* = \mathbb{E}\max_a \langle \phi(X, a), \theta\rangle = \max_a \mu_a$.

Now we construct the reduction by specifying an algorithm $\mathcal{B}$ for the $\frac{d}{2}$-armed bandit. At each round, $\mathcal{B}$ samples $X \sim \mathrm{Unif}\{1, 2\}$ and queries $\mathcal{A}$ for an arm A. Upon observing feedback $Y(A) = \mu_A + \eta(A)$, $\mathcal{B}$ feeds $Y(A)$ back to $\mathcal{A}$ if $X = 1$ and $-Y(A)$ if $X = 2$. This process is repeated for n rounds and $\mathcal{A}$ outputs an estimate $\hat{V}_n$, which $\mathcal{B}$ also outputs. If $\mathcal{A}$ outputs $\hat{V}_n$ such that $|\hat{V}_n - V^*| \le \epsilon$ for any given instance in the linear contextual bandit then $\hat{V}_n$ is also an ϵ-optimal estimate of $\max_a \mu_a$. Therefore, to satisfy $|\hat{V}_n - V^*| \le \epsilon$, it follows that $n = \Omega(d/\epsilon^2)$.

Proof of the second lower bound. Here we prove the second statement of the proposition that even for $K = 2$, it takes $\Omega(d)$ samples to estimate V^* up to small constant additive error c. The proof simply follows from the hard instance for signal-noise-ratio (SNR) estimation problem in Theorem 3 of (Kong and Valiant, 2018).

Let $\mathcal{Q}_n(\mathcal{P})$ be the distribution of $(x_1, y_1, \dots, x_n, y_n)$ such that $(\theta, \sigma, \Sigma) \sim P$, $x_i \sim N(0, \Sigma)$, $y_i = x_i + \eta_i$, $\eta_i \sim N(0, \sigma^2)$. Let the "pure noise" distribution $\mathcal{P}_0$ satisfies $\theta = 0, \Sigma = \mathbb{I}_d, \sigma^2 = 1$ almost surely. Theorem 3 of (Kong and Valiant, 2018) relies on the fact that there exists a "pure signal" distribution $\mathcal{P}_1$ over (θ, σ, Σ) which is constructed by randomly rotating a d-dimensional isotropic Gaussian distribution in the $d + 1$ dimensional space. The covariance Σ drawn from distribution $\mathcal{P}_1$ is bad conditioned and has smallest eigenvalue being $O(1/d)$ with large probability. In addition, the distribution $\mathcal{P}_1$ satisfies $\mathbb{P}(\|\theta\| \ge 1/2) \ge 1/2$, and $\|\Sigma\| \le 1, \sigma = 0$ almost surely, and it holds that

$$d_{TV}(\mathcal{Q}_n(\mathcal{P}_0), \mathcal{Q}_n(\mathcal{P}_1)) \le 1/3, \tag{17}$$

with $n = c \cdot d$ for some constant c. The hard case for the bandits problem follows immediately. We construct the "pure signal" bandit instance using $(\theta, \sigma, \Sigma) \sim \mathcal{P}_1$, and for each arm $a \in [2]$, define $\phi(X, a) \sim N(0, \Sigma), Y(a) =$

Algorithm 1 Estimator of Upper Bound on V^*

Input: Number of interactions n, failure probability $\delta \leq 1/e$.
Set $m = \frac{n}{2}$.
Initialize empty dataset D
for $i = 1, \ldots, n$ **do**
  Sample independently $x_i \sim \mathcal{D}$ and $a_i \sim \text{Unif}[K]$. Receive reward y_i
  Add tuple (x_i, a_i, y_i) to D
end for
Split dataset D evenly into $\{x_i, a_i, y_i\}_{i\in[m]}$ and $\{x'_i, a'_i, y'_i\}_{i\in[m]}$.
Compute estimators $\hat{\theta} = \frac{1}{m}\sum_{i\in[m]} y_i\phi(x_i, a_i)$ and $\hat{\theta}' = \frac{1}{m}\sum_{i\in[m]} y'_i\phi(x'_i, a'_i)$
for $a, a' \in [K]$ such that $a \neq a'$ **do**
  Set $\hat{\beta}_{a,a'} := \hat{\theta}^\top \Sigma_{a,a'}\hat{\theta}'$
end for
$\tilde{\Lambda} = \arg\min_{\lambda\in\mathbb{S}^K_+} \max_{a\neq a'} |\lambda_{a,a} + \lambda_{a',a'} - 2\lambda_{a,a'} - \hat{\beta}_{a,a'}|$
Return $\mathbb{E}\max_{a\in[K]} \tilde{Z}$ where $\tilde{Z} \sim \mathcal{N}(0, \tilde{\Lambda})$

$\theta^\top\phi(X, a)$. It is easy to see that in this case, $\mathbb{E}V^* = \Omega(1)$. The other bandit instance is a simple "pure noise" example where $\phi(X, a) \sim N(0, \mathbb{I}_d), Y(a) \sim N(0, 1)$, and clearly $\mathbb{E}V^* = 0$ since $\theta = 0$. For any bandit algorithm for estimating V^*, after $cn/2$ rounds, even if all the rewards (regardless of which arm gets pulled) are shown to the algorithm the total variation distance is between the two example is still bounded by $1/3$ through Equation 17. Therefore, we conclude any bandit algorithm must incur $\Omega(1)$ error for estimating V^* with probability at least $2/3$ when $n = c \cdot d$.

□

B. Proofs of Results in Section 4

B.1. Proof of Theorem 1

The proof relies on Talagrand's comparison inequality for sub-Gaussian processes. Here, we state a version that appears in (Vershynin, 2018).

Lemma 1. *Let $(W_a)_{a\in[K]}$ be a mean zero sub-Gaussian process and $(Z_a)_{a\in[K]}$ a mean zero Gaussian process satisfying $\|W_a - W_{a'}\|_{\psi_2} \lesssim \|Z_a - Z_{a'}\|_{L^2}$. Then,*

$$\mathbb{E}\max_{a\in[K]} W_a \lesssim \mathbb{E}\max_{a\in[K]} Z_a \tag{18}$$

By Assumption 3, note that

$$\| \langle\phi(X, a) - \phi(X, a'), \theta\rangle \|^2_{\psi_2} \leq L_0^2 \| \langle\phi(X, a) - \phi(X, a'), \theta\rangle \|^2_{L^2} \tag{19}$$

Thus, we can define a Gaussian process $Z \sim \mathcal{N}(0, \Lambda)$ that satisfies the condition in Talagrand's inequality by choosing its mean to be zero and its covariance matrix to match the increment of the original sub-Gaussian process $\phi(X, \cdot)$. Note that such a process trivially exists since we can let Λ satisfy:

$$\Lambda_{a,a'} = \text{cov}(Z_a, Z_{a'}) = \mathbb{E}\left[\langle\phi(X, a), \theta\rangle \langle\phi(X, a'), \theta\rangle\right] \tag{20}$$

Then, the first inequality in the theorem is satisfied with $U = \mathbb{E}\max_{a\in[K]} Z_a$. The proof of the second inequality is deferred to Section B.1.1.

Since θ is unknown, our goal now is to estimate the increment $\| \langle\phi(X, a) - \phi(X, a'), \theta\rangle \|^2_{L^2}$ from samples. Specifically, we aim to estimate the following quantities:

- For all $a, a' \in [K]$ such that $a \neq a'$, $\beta_{a,a'} := \mathbb{E}\left[\langle\phi(X, a) - \phi(X, a'), \theta\rangle^2\right] = \theta^\top\Sigma_{a,a'}\theta$ where $\Sigma_{a,a'} = \mathbb{E}\left[(\phi(X, a) - \phi(X, a')(\phi(X, a) - \phi(X, a'))^\top\right]$.

We can construct fast estimators for these quantities using similar techniques as those developed in (Kong and Valiant, 2018). While a similar final result is obtained in that paper by Chebyshev's inequality and counting, here we present a version that is carried out with a couple simple applications of Bernstein's inequality. Algorithm 1 specifies the form of the estimator and the data collection procedure.

Lemma 2. *Fix $a, a' \in [K]$ such that $a \neq a'$ and define $\xi^2 = \tau^2(\tau^2\|\theta\|^2 + \sigma^2)$. Let $\delta \leq 1/e$. Given the dataset $D_n = \{x_i, a_i, y_i\}$, with probability at least $1 - 3\delta$,*

$$|\hat{\beta}_{a,a'} - \beta_{a,a'}| \leq \sqrt{\frac{\xi^2\|\Sigma\|^2\|\theta\|^2}{C_1 m}} \cdot \log(2/\delta) \tag{21}$$

$$+ \sqrt{\frac{\xi^4\|\Sigma\|^2 d}{C_2 m^2}} \cdot \log^2(2d/\delta) \tag{22}$$

for absolute constants $C_1, C_2 > 0$.

Proof. Consider an arbitrary pair a, a' and covariance matrix $\Sigma_{a,a'}$. For convenience, we drop the subscript notation and just write Σ. The argument will be the same for all pairs, including when $a = a'$. The dataset D_n is split into two independent datasets D_m and D'_m of size $m = \frac{n}{2}$. Let $\phi_i := \phi(x_i, a_i)$ as shorthand and the same for ϕ'_i.

First, we verify that $\hat{\beta}_{a,a'}$ is indeed an unbiased estimator of $\beta_{a,a'}$:

$$\mathbb{E}\left[\hat{\theta}^\top\Sigma\hat{\theta}'\right] = \mathbb{E}\left[y_i y'_j \phi_i^\top\Sigma\phi'_i\right] = \theta^\top\Sigma\theta \tag{23}$$

which follows by independence of the datasets D_m and D'_m and the fact that the covariance matrix under the uniform data collection policy is the identity. By adding and subtracting and then applying the triangle inequality, we have

$$|\hat{\theta}^\top \Sigma \hat{\theta}' - \theta^\top \Sigma \theta| = \underbrace{|\theta^\top \Sigma \hat{\theta}' - \theta^\top \Sigma \theta|}_{\text{Term I}} + \underbrace{|\hat{\theta}^\top \Sigma \hat{\theta}' - \theta^\top \Sigma \hat{\theta}'|}_{\text{Term II}} \tag{24}$$

and we focus on bounding each term individually. We start with the first. Note that $\|\theta^\top \Sigma \phi'_i\|_{\psi_2} \leq \|\Sigma\theta\|\tau$ and $\|y'_i\|_{\psi_2} \lesssim \sqrt{\tau^2\|\theta\|^2 + \sigma^2}$. Therefore, we have that the term $\phi'_{i,k} y'_i$ is sub-exponential with parameter $\|\phi'_{i,k} y'_i\|_{\psi_1} \lesssim \tau\|\Sigma\theta\|\sqrt{\tau^2\|\theta\|^2 + \sigma^2} = \xi\|\Sigma\theta\|$, where recall that we have defined $\xi^2 = \tau^2(\tau^2\|\theta\|^2 + \sigma^2)$. Then, by Bernstein's inequality,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i\in[n]} \theta^\top \Sigma \phi'_{i,k} y'_i - \theta^\top \Sigma \theta \geq t\right) \tag{25}$$

$$\leq \exp\left(-C \min\left\{\frac{nt^2}{\|\Sigma\theta\|^2\xi^2}, \frac{nt}{\|\Sigma\theta\|\xi}\right\}\right) \tag{26}$$

for some absolute constant $C > 0$, and the negative event occurs with the same upper bound on the probability. This implies

$$|\frac{1}{m}\sum_{i\in[m]} \theta^\top \Sigma \phi'_i y'_i - \theta^\top \Sigma \theta| \leq \sqrt{\frac{\xi^2\|\Sigma\theta\|^2}{Cm}} \cdot \log(2/\delta) \tag{27}$$

For the second term, we condition on the data in D' and then apply the same calculations. The difference is that $\|\phi_i \Sigma \hat{\theta}'\|_{\psi_2} \leq \tau\|\Sigma\hat{\theta}'\|$ and so the bound becomes

$$|\frac{1}{m}\sum_{i\in[m]} y_i \phi_i^\top \Sigma \hat{\theta}' - \theta^\top \Sigma \hat{\theta}'| \leq \sqrt{\frac{\xi^2\|\Sigma\hat{\theta}'\|^2}{Cm}} \cdot \log(2/\delta) \tag{28}$$

with probability at least $1 - \delta$.

It suffices now to obtain a high probability bound on $\|\hat{\theta}'\|$, showing that it is close in value to $\|\theta\|$. Let $\phi'_{i,k}$ and θ_k denote the kth elements of ϕ'_i and θ_k, respectively. Similar to the previous proof, we have that

$$\|\phi'_{i,k} y'_i\|_{\psi_1} \lesssim \xi \tag{29}$$

by multiplication of the sub-Gaussian random variables. By Bernstein's inequality, with probability $1-\delta$, for all $k \in [d]$,

$$|\frac{1}{m}\sum_{i\in[m]} \phi'_{i,k} y'_i - \theta_k| \leq \sqrt{\frac{\xi^2}{Cm}} \cdot \log(2d/\delta) \tag{30}$$

for some constant $C > 0$. Under the same event,

$$\|\hat{\theta}' - \theta\| \leq \sqrt{\frac{d\xi^2}{Cm}} \cdot \log(2d/\delta) \tag{31}$$

by standard norm inequalities. The triangle inequality then yields

$$\|\hat{\theta}'\| \leq \|\theta\| + \sqrt{\frac{d\xi^2}{Cm}} \cdot \log(2d/\delta) \tag{32}$$

Finally, we are able to put these three events together:

$$|\hat{\theta}^\top \Sigma \hat{\theta}' - \theta^\top \Sigma \theta| \leq \sqrt{\frac{\xi^2\|\Sigma\theta\|^2}{Cm}} \cdot \log(2/\delta) + \sqrt{\frac{\xi^2\|\Sigma\hat{\theta}'\|^2}{Cm}} \cdot \log(2/\delta) \tag{33}$$

$$\leq \sqrt{\frac{\xi^2\|\Sigma\theta\|^2}{Cm}} \cdot \log(2/\delta) + \sqrt{\frac{2\xi^2\|\Sigma\|^2\|\theta\|^2}{Cm}} \cdot \log(2/\delta) \tag{34}$$

$$+ \sqrt{\frac{2\xi^4\|\Sigma\|^2 d}{C_1 m^2}} \cdot \log^2(2d/\delta) \tag{35}$$

$$\leq \sqrt{\frac{8\xi^2\|\Sigma\|^2\|\theta\|^2}{C_2 m}} \cdot \log(2/\delta) \tag{36}$$

$$+ \sqrt{\frac{2\xi^4\|\Sigma\|^2 d}{C_2 m^2}} \cdot \log^2(2d/\delta) \tag{37}$$

with probability at least $1 - 3\delta$ by the union bound over the three events. □

Define $\tilde{\beta}_{a,a'} = \tilde{\Lambda}_{a,a} + \tilde{\Lambda}_{a',a'} - 2\tilde{\Lambda}_{a,a'}$, and $\tilde{Z} \sim N(0, \tilde{\Lambda})$ where $\tilde{\Lambda}$ is the result of the projection onto $\mathbb{S}_+^K$ using $\hat{\beta}$ as defined in Line 13 of Algorithm 1. Since Λ is positive semidefinite, the fact that

$$|\beta_{a,a'} - \hat{\beta}_{a,a'}| \leq \mathcal{O}\left(\frac{\|\theta\|\log(K/\delta)}{\sqrt{n}} + \frac{\sqrt{d}\cdot\log^2(dK/\delta)}{n}\right), \tag{38}$$

and the optimality of $\tilde{\Lambda}$ in Algorithm 1, we have

$$|\hat{\beta}_{a,a'} - \tilde{\beta}_{a,a'}| \leq \mathcal{O}\left(\frac{\|\theta\|\log(K/\delta)}{\sqrt{n}} + \frac{\sqrt{d}\cdot\log^2(dK/\delta)}{n}\right). \tag{39}$$

Triangle inequality then immediately implies the following element-wise error bound on the increment

$$|\beta_{a,a'} - \tilde{\beta}_{a,a'}| \leq \mathcal{O}\left(\frac{\|\theta\|\log(K/\delta)}{\sqrt{n}} + \frac{\sqrt{d}\cdot\log^2(dK/\delta)}{n}\right) \tag{40}$$

with probability at least $1 - \delta$.

Now we apply the following error bound due to (Chatterjee, 2005).

Lemma 3 (Theorem 1.2, (Chatterjee, 2005)). *Let W and $\tilde{W}$ be two Gaussian random vectors with $\mathbb{E}W_a = \mathbb{E}\tilde{W}_a$ for all $a \in [K]$. Define $\gamma_{a,a'} = \|W_a - W_{a'}\|_{L_2}^2$ and $\tilde{\gamma}_{a,a'} = \|\tilde{W}_a - \tilde{W}_{a'}\|_{L_2}^2$ and $\Gamma = \max_{a,a'} |\tilde{\gamma}_{a,a'} - \gamma_{a,a'}|$. Then,*

$$|\mathbb{E}\max_{a\in[K]} W_a - \mathbb{E}\max_{a\in[K]} \tilde{W}_a| \le \sqrt{\Gamma \log K} \tag{41}$$

Therefore, by the union bound over at most K^2 terms $\beta_{a,a'}$, the final bound becomes

$$|U - \mathbb{E}\max_{a\in[K]} \tilde{Z}_a| \le \mathcal{O}\left(\frac{\sqrt{\|\theta\|}\log(K/\delta)}{n^{1/4}} + \frac{d^{1/4}\log^{3/2}(dK/\delta)}{\sqrt{n}}\right) \tag{42}$$

with probability at least $1-\delta$.

B.1.1. Proof of the second inequality

Here we prove the second inequality in the theorem statement that $\sqrt{\log K}\cdot V^* \gtrsim U$

Lemma 4. *Let $(W_a)_{a\in[K]}$ be a mean zero sub-Gaussian process such that $\|W_a - W_{a'}\|_{\psi^2} \lesssim \|W_a - W_{a'}\|_{L^2}$, then*

$$\mathbb{E}\max_{a\in[K]} W_a \gtrsim \max_{a,a'\in[K]} \|W_a - W_{a'}\|_{L^2} \tag{43}$$

Proof. Let random variable $W_b, W_{b'}$ achieve the maximum for $\max_{a,a'\in[K]} \|W_a - W_{a'}\|_{L^2}$.

$\mathbb{E}\max_{a\in[K]} W_a \ge \mathbb{E}\max(W_b, W_{b'})$ Define $Z = W_{b'} - W_b$, then

$$\begin{aligned}
&\mathbb{E}\max(W_b, W_{b'})\\
=&\mathbb{E}[W_b|Z\le 0]\mathbb{P}[Z\le 0] + \mathbb{E}[W_b + Z|Z>0]\mathbb{P}[Z>0]\\
=&\mathbb{E}[W_b|Z\le 0]\mathbb{P}[Z\le 0] + \mathbb{E}[W_b|Z>0]\mathbb{P}[Z>0] + \mathbb{E}[Z|Z>0]\mathbb{P}[Z>0]\\
=&\mathbb{E}[W_b] + \mathbb{E}[Z|Z>0]\mathbb{P}[Z>0]\\
=&\mathbb{E}[Z|Z>0]\mathbb{P}[Z>0]
\end{aligned}$$

Since $\mathbb{E}[Z|Z>0]\mathbb{P}[Z>0] + \mathbb{E}[Z|Z<0]\mathbb{P}[Z<0] = 0$, we have

$$\mathbb{E}[Z|Z>0]\mathbb{P}[Z>0] = \mathbb{E}[|Z|]/2 \tag{44}$$

Thus, we just need to lower bound $\mathbb{E}[|Z|]$. Due to the sub-Gaussian assumption on Z, it holds that for a constant K_0,

$$\mathbb{P}(|Z|>t) \le \exp(-\frac{t^2}{K_0\|Z\|_{L^2}^2})$$

Let C be a constant such that

$$\begin{aligned}
&\int_{C\|Z\|_{L^2}}^{\infty} t\exp(-\frac{t^2}{K_0\|Z\|_{L^2}^2})dt\\
=&K_0\|Z\|_{L^2}^2\exp(-\frac{C^2}{K_0})\\
=&\|Z\|_{L^2}^2/20.
\end{aligned}$$

Then,

$$\begin{aligned}
\|Z\|_{L^2}^2 &= 2\int_0^\infty t\mathbb{P}(|Z|>t)dt\\
&= 2\int_0^{C\|Z\|_{L^2}} t\mathbb{P}(|Z|>t)dt + 2\int_{C\|Z\|_{L^2}}^\infty t\mathbb{P}(|Z|>t)dt\\
&le 2\int_0^{C\|Z\|_{L^2}} t\mathbb{P}(|Z|>t)dt + 2\int_{C\|Z\|_{L^2}}^\infty t\exp(-\frac{t^2}{K_0\|Z\|_{L^2}^2})dt\\
&\le 2C\|Z\|_{L^2}\int_0^{C\|Z\|_{L^2}} \mathbb{P}(|Z|>t)dt + \|Z\|_{L^2}^2/10\\
&\le 2C\|Z\|_{L^2}\mathbb{E}[|Z|] + \|Z\|_{L^2}^2/10.
\end{aligned}$$

This implies that $\mathbb{E}[|Z|] \ge \frac{9}{20C}\|Z\|_{L^2}$. Combining with Equation 44 yields

$$\mathbb{E}\max_{a\in[K]} W_a \gtrsim \|W_{b'} - W_b\|_{L^2}$$

□

Proposition 3. *Let $(Z_a)_{a\in[K]}$ be a mean zero Gaussian process, then*

$$\mathbb{E}\max_{a\in[K]} Z_a \lesssim \sqrt{\log K}\max_{a,a'\in[K]} \|Z_a - Z_{a'}\|_{L^2} \tag{45}$$

Proof. This is a simple corollary of Sudakov-Fernique's inequality (see, e.g. Theorem 7.2.11 in (Vershynin, 2018)). Define mean zero Gaussian process $Y_a, a \in [K]$ such that each Y_a is sampled independently from $N(0, \max_{a,a'\in[K]}\|Z_a - Z_{a'}\|_{L^2}^2)$. By Sudakov-Fernique's inequality, it holds that

$$\mathbb{E}\max_{a\in[K]} Z_a \le \mathbb{E}\max_{a\in[K]} Y_a.$$

We conclude the proof by combining with classical fact that

$$\max_{a\in[K]} Y_a \lesssim \sqrt{\log K}\max_{a,a'\in[K]} \|Z_a - Z_{a'}\|_{L^2}$$

□

Applying Lemma 4 on V^* yields

$$V^* \gtrsim \max_{a,a'\in[K]} \|\langle\phi(X,a) - \phi(X,a'), \theta\rangle\|_{L^2}.$$

By the definition of the Gaussian process Z, its increment is bounded by $\max_{a,a'\in[K]}\|\langle\phi(X,a) - \phi(X,a'),\theta\rangle\|_{L^2}$, therefore applying Proposition 3 for U yields

$$U \lesssim \sqrt{\log K}\max_{a,a'\in[K]} \|\langle\phi(X,a) - \phi(X,a'),\theta\rangle\|_{L^2}.$$

This concludes the proof.

Algorithm 2 Model Selection with Gaussian Process Upper Bound

Input: Rounds T, failure probability $\delta \leq 1/e$, constants C_0, C_1
Set $t_{\min} = C_0 \log^{3/2}(T \log T/\delta)$
Set $\alpha_t = C_1 \cdot \frac{d_2^{1/4} \log^{3/2}(d_2 KT/\delta)}{t^{1/3}}$
Initialize exploration dataset $S_0 = \{\}$
Initialize algorithm $\text{Alg}_1 \leftarrow$ Exp4-IX($\mathcal{F}_1$).
Sampler Bernoulli $Z_t \sim \text{ber}(t^{-1/3})$ for all $t \in [T]$
for $t = 1, \ldots, T$ **do**
  Sample independently $x_t \sim \mathcal{D}$ and
  if $Z_t = 1$ **then**
    Sample $a_t \sim \text{Unif}[K]$, observe y_t
    Add to dataset: $S_t = (x_t, a_t, y_t) \cup S_{t-1}$
  else
    Sample a_t from Alg_t, observe y_t
    Update Alg_t with (x_t, a_t, y_t)
    $S_t = S_{t-1}$
    $\text{Alg}_{t+1} \leftarrow \text{Alg}_t$
  end if
  Estimate $\hat{U}_t$ from S_t.
  if $t \geq t_{\min}$ and $\hat{U}_t > 2\alpha_t$ **then**
    Set algorithm $\text{Alg}_{t+1} \leftarrow$ Exp4-IX($\mathcal{F}_2$)
  end if
end for

B.2. Proof of Theorem 2

The algorithm that achieves the regret bound in Theorem 2 is presented in Algorithm 2. The main idea is that the algorithm starts with model class $\mathcal{F}_1$, the simpler one, and runs an Exp4-like algorithm under $\mathcal{F}_1$. However, it will randomly allocate some timesteps for exploratory actions where the uniform random policy is applied. From the exploration data, if it is detected that the gap is non-zero with high confidence, then the algorithm switches to $\mathcal{F}_2$. The critical component of the algorithm is in detecting the non-zero gap and then bounding the worst-case performance when the gap is non-zero but it has not been detected yet.

We require several intermediate results in order to prove the regret bound. The first is a generic high probability regret bound for a variant of Exp4-IX as given by Algorithm 4 of (Foster et al., 2019), which is a modification of the algorithm proposed by (Neu, 2015). Let π_{θ_i} be the argmax policy induced by θ_i where θ_i is defined is defined as

$$\theta_i = \arg\min_{\theta} \frac{1}{K} \sum_{a \in [K]} \mathbb{E}\left(\langle \phi_i(X, a), \theta \rangle - Y\right)^2 \tag{46}$$

Note that the policy π_{θ_1} may not be the same as the policy that maximizes value.

Lemma 5 ((Foster et al., 2019), Lemma 23). *With probability at least $1 - \delta$, for any $t \in [T]$, Exp4-IX for model class $\mathcal{F}_i$ satisfies*

$$\sum_{s=1}^{t} V^{\pi_{\theta_i}} - V^{\pi_s} \leq \mathcal{O}\left(\sqrt{d_i t K \log(d_i)} \cdot \log(TK/\delta)\right) \tag{47}$$

The second result we require is high probability upper and lower bounds on the number of exploration samples we should expect to have at any time $t \in [T]$. We appeal to Lemma 2 of (Lee et al., 2021), as the exploration schedules are identical.

Lemma 6 ((Lee et al., 2021), Lemma 2). *There are constants $C_1, C_2 > 0$ such that, with probability $1 - \delta$, $C_1 t^{2/3} \leq |S_t| \leq C_2 t^{2/3}$ for $t \geq C_0 \log^{3/2}(T \log T/\delta)$.*

The last intermediate result leverages the upper bound estimator from Theorem 1. We will define a Gaussian process, which we prove will act as an upper bound on the gap in value between the model classes. Let $Z \sim \mathcal{N}(0, \Lambda)$ where

$$\Lambda_{a,a'} = \mathbb{E}\left[\langle \phi(X, a), \theta_{\text{diff}} \rangle \langle \phi(X, a'), \theta_{\text{diff}} \rangle\right] \tag{48}$$

for all $a, a' \in [K]$ and $\theta_{\text{diff}} = \theta_2 - \begin{bmatrix} \theta_1 \\ 0 \end{bmatrix}$. The following lemma establishes these upper bounds and shows that we can estimate $\mathbb{E} \max_{a \in [K]} Z_a$ at a fast rate. The critical property of this upper bound is that it is 0 when $\mathcal{F}_1$ satisfies realizability.

A simple transformation of the feature vectors allows us to apply the results from before. For datapoints (x_i, a_i, y_i) collected by the uniform random policy, the following is an unbiased estimator of $\theta_2 - \begin{bmatrix} \theta_1 \\ 0 \end{bmatrix}$:

$$y_i \left(\phi_2(x_i, a_i) - \begin{bmatrix} \phi_1(x_i, a_i) \\ 0 \end{bmatrix}\right) = y_i \left(\phi_2(x_i, a_i) - \begin{bmatrix} \phi_1(x_i, a_i) \\ 0 \end{bmatrix}\right) \tag{49}$$

$$= y_i \begin{bmatrix} 0 \\ \phi_{d_1:d_2}(x_i, a_i) \end{bmatrix} \tag{50}$$

where $\phi_{d_1:d_2}$ denotes the bottom $d_2 - d_1$ coordinates of the feature map ϕ. As shorthand, we define $\tilde{\phi}_i(x, a) = \begin{bmatrix} 0 \\ \phi_{d_1:d_2}(x, a) \end{bmatrix}$. Note that $\|\tilde{\phi}_i\|_{\psi_2} \leq \tau$ and this feature vector still satisfies the conditions of Assumption 3 as we can simply zero the top coordinates. Furthermore, define $\tilde{\Sigma}_{a,a'} = \mathbb{E}\left(\tilde{\phi}(X, a) - \tilde{\phi}(X, a')\right)\left(\tilde{\phi}(X, a) - \tilde{\phi}(X, a')\right)^\top$ for $a \neq a'$. The estimators for this transformed problem are

then

$$\hat{\theta} = \frac{1}{m}\sum_i y_i \tilde{\phi}_i \tag{51}$$

$$\hat{\theta}' = \frac{1}{m}\sum_i y'_i \tilde{\phi}'_i \tag{52}$$

And, as before, the quadratic form estimators are analogously

$$\hat{\beta}_{a,a'} = \hat{\theta}^\top \tilde{\Sigma}_{a,a'} \hat{\theta}' \tag{53}$$

Lemma 7. *There is a constant C such that the Gaussian process Z*

$$V^* - V^{\pi_{\theta_1}} \le 2C \cdot \mathbb{E} \max_{a\in[K]} Z_a \tag{54}$$

and, with probability at least $1-\delta$, for all $n \in [T]$, the estimator $\hat{U}$ defined in Algorithm 2 with n independent samples satisfies

$$|\mathbb{E} \max_{a\in[K]} Z_a - \hat{U}| \tag{55}$$

$$\le \mathcal{O}\left(\frac{\sqrt{\|\theta_{\mathit{diff}}\|}\log(TK/\delta)}{n^{1/4}} + \frac{d_2^{1/4}\log^{3/2}(d_2KT/\delta)}{\sqrt{n}}\right) \tag{56}$$

Proof. It is immediate that

$$V^* - \max_{\pi\in\Pi_1} V^\pi \le V^* - V^{\pi_{\theta_1}} \tag{57}$$

since $\theta_1 \in \mathcal{F}_1$ by definition and π_{θ_1} is an argmax policy. This gap can then be bounded as

$$V^* - V^{\pi_{\theta_1}} = V^{\pi_{\theta_2}} - V^{\pi_{\theta_1}} \tag{58}$$

$$= \mathbb{E}\langle \phi_2(X, \pi_{\theta_2}(X)), \theta_2\rangle - \mathbb{E}\langle \phi_2(X, \pi_{\theta_1}(X)), \theta_2\rangle \tag{59}$$

$$= \mathbb{E}\langle \phi_2(X, \pi_{\theta_2}(X)), \theta_2\rangle - \mathbb{E}\langle \phi_2(X, \pi_{\theta_1}(X)), \theta_2\rangle \tag{60}$$

$$+ \mathbb{E}\langle \phi_1(X, \pi_{\theta_1}(X)), \theta_1\rangle - \mathbb{E}\langle \phi_1(X, \pi_{\theta_1}(X)), \theta_1\rangle \tag{61}$$

$$\le \mathbb{E}\left\langle \phi_2(X, \pi_{\theta_2}(X)), \theta_2 - \begin{bmatrix}\theta_1\\0\end{bmatrix}\right\rangle \tag{62}$$

$$+ \mathbb{E}\left\langle \phi_2(X, \pi_{\theta_1}(X)), \begin{bmatrix}\theta_1\\0\end{bmatrix} - \theta_2\right\rangle \tag{63}$$

$$\le \mathbb{E}\max_{a\in[K]}\left\langle \phi_2(X, a), \theta_2 - \begin{bmatrix}\theta_1\\0\end{bmatrix}\right\rangle \tag{64}$$

$$+ \mathbb{E}\max_{a\in[K]}\left\langle \phi_2(X, a), \begin{bmatrix}\theta_1\\0\end{bmatrix} - \theta_2\right\rangle \tag{65}$$

The Gaussian process $Z \sim \mathcal{N}(0, \Lambda)$ satisfies the conditions of Lemma 1, which implies the Gaussian process upper bound on both of the above terms and, thus, the first claim.

Now we prove the estimation error bound. We apply Algorithm 1 with the constructed fast estimators for quadratic forms $\theta_{\text{diff}}^\top \tilde{\Sigma}_{a,a'}\theta_{\text{diff}}$ for all $a, a' \in [K]$. Let $\tilde{Z} \sim N(0, \tilde{\Lambda})$. We can apply Theorem 1 and get

$$|\mathbb{E}\max_{a\in[K]} Z_a - \mathbb{E}\max_{a\in[K]} \tilde{Z}_a| \le \mathcal{O}\left(\frac{\sqrt{\|\theta\|}\log(K/\delta)}{n^{1/4}} + \frac{d^{1/4}\log^{3/2}(dK/\delta)}{\sqrt{n}}\right) \tag{66}$$

Setting $\hat{U} = \mathbb{E}\max_{a\in[K]} \tilde{Z}_a$ gives the result.

□

Lemma 8. *Let $\hat{U}$ be the estimate of $\mathbb{E}\max_{a\in[K]} Z_a$ from Lemma 7 using the same method. Then, with probability $1-\delta$,*

$$\mathbb{E}\max_a Z_a \le C\hat{U}\log^{1/2}(K) \tag{67}$$

$$+ \mathcal{O}\left(\frac{(\|\theta_{\mathit{diff}}\|^{1/2} + d^{1/4})\log(d_2K/\delta)\log^{1/2}(K)}{\sqrt{n}}\right) \tag{68}$$

for some constant $C > 0$.

Proof. Here we let C represent an absolute constant, which may change from line to line. For this, we require a multiplicative error bound, which is stated formally in Theorem 5. It is similar to the additive one developed in the proof of Theorem 1. From Theorem 5, and applying the union bound over all pairs of actions in $[K]$, we have with probability at least $1 - K^2\delta$, for all $a' \ne a$,

$$|\beta_{a,a'} - \hat{\beta}_{a,a'}| \le \frac{\beta_{a,a'}}{2c} + \mathcal{O}\left(\frac{c(\|\Sigma_{a,a'}^{1/2}\theta\| + \sqrt{d})\log^2(d/\delta)}{n}\right) \tag{69}$$

where we simply prepend $\Sigma^{1/2}$ to θ and the estimators and $c \ge 1$ is to be chosen later.

With this concentration, we now show that if $\hat{U} = \mathbb{E}\max_{a\in[K]} \tilde{Z}_a$ is small, then this must mean that $\max_{a,a'} \beta_{a,a'}$ is also small.

$$\left(\hat{U}\right)^2 \ge \left(\mathbb{E}\max_{a\in[K]} \tilde{Z}_a\right)^2 \tag{70}$$

$$\ge C \max_{a,a'\in[K]\,:\,a\ne a'} \|\tilde{Z}_a - \tilde{Z}_{a'}\|_{L^2}^2 \tag{71}$$

$$= \mathcal{O}\left(\max_{a,a'} \beta_{a,a'} - \frac{\beta_{a,a'}}{2c} - \frac{c(\|\theta_{\text{diff}}\| + \sqrt{d})\log^2(d/\delta)}{n}\right) \tag{72}$$

for an absolute constant C. The second line uses Lemma 4. The third line uses the concentration above. Choosing c

large enough (dependent only on absolute constants), we get

$$\max_{a\neq a'} \beta_{a,a'} \leq 2\hat{U}^2 + \mathcal{O}\left(\frac{(\|\theta_{\text{diff}}\| + \sqrt{d})\log^2(d/\delta)}{n}\right) \tag{73}$$

Then, from Proposition 3, we get the statement:

$$U \leq C\sqrt{\log K} \cdot \sqrt{\max_{a\neq a'} \beta_{a,a'}} \tag{74}$$

$$\leq C\hat{U}\sqrt{\log K} \tag{75}$$

$$+ \mathcal{O}\left(\frac{(\|\theta_{\text{diff}}\|^{1/2} + d^{1/4})\log(d/\delta)\log^{1/2}(K)}{\sqrt{n}}\right) \tag{76}$$

Changing the variable $\delta' = \delta/K^2$ gives the result.

□

Armed with these facts, we can prove the regret bound. Let $E = E_1 \cap E_2 \cap E_3 \cap E_4$ denote the good event that satisfies the conditions laid out in the intermediate results where

1. E_1 is the event that $\sum_{s\in\mathcal{I}} V^{\pi_{\theta_i}} - V^{\pi_s} \leq \mathcal{O}\left(\sqrt{d_i|\mathcal{I}|K\log(d_i)} \cdot \log(TK/\delta)\right)$ for any interval of times up to $|\mathcal{I}| \leq T$.
2. E_2 is the event that $C_1 t^{2/3} \leq |S_t| \leq C_2 t^{2/3}$ for $t \geq t_{\min}$
3. E_3 is the event that the following inequality is satisfied for all $t_{\min} \leq t \leq T$:

$$|\mathbb{E}\max_{a\in[K]} Z_a - \hat{U}_t| \tag{77}$$

$$\leq \mathcal{O}\left(\frac{\sqrt{\|\theta_{\text{diff}}\|}\log(TK/\delta)}{t^{1/6}} + \frac{d_2^{1/4}\log^{3/2}(d_2KT/\delta)}{t^{1/3}}\right) \tag{78}$$

4. E_4 is the event that the following is satisfied for all $t_{\min} \leq t \leq T$:

$$V^* - V^{\pi_1} \leq C\hat{U}_t \log^{1/2}(K) \tag{79}$$

$$+ \mathcal{O}\left(\frac{(\|\theta_{\text{diff}}\|^{1/2} + d^{1/4})\log(d_2KT/\delta)\log^{1/2}(K)}{t^{1/3}}\right) \tag{80}$$

Proof of Theorem 2. First note that event E holds with probability at least $1 - 4\delta$ via an application of the union bound (over T) and the intermediate results. We now work under the assumption that E holds. The proof is divided into cases when $\mathcal{F}_1$ does and does not satisfy realizability.

First, we bound the instantaneous regret incurred during the exploration rounds. Note that the average value of the uniform policy is zero and $V^* \leq \mathcal{O}\left(\|\theta\|\sqrt{\log K}\right)$ by standard maximal inequalities. This establishes the bound on the instantaneous regret for these rounds.

1. When $\mathcal{F}_1$ satisfies realizability, the algorithm is already running Exp4-IX with model class $\mathcal{F}_1$ from the beginning, so we are left with verifying that a switch to $\mathcal{F}_2$ never occurs in this setting. This can be shown by realizing that $\mathbb{E}\max_{a\in[K]} Z_a = 0$ whenever $\mathcal{F}_1$ satisfies realizability. Therefore $\theta_{\text{diff}} = 0$ and, under the good event, we have that

$$\hat{U}_t \leq C\frac{d_2^{1/4}\log^{3/2}(d_2KT/\delta)}{t^{1/3}} \tag{81}$$

for a some constant $C > 0$. Therefore, for C_1 chosen large enough, $\hat{U}_t \leq 2\alpha_t$ for all $t \geq t_{\min}$ and thus a switch never occurs. In this case, the regret incurred is

$$\text{Reg}_T \leq \widetilde{\mathcal{O}}\left(T^{2/3} \cdot \log^{1/2}(K)\right) \tag{82}$$

$$+ \widetilde{\mathcal{O}}\left(\sqrt{d_1 TK\log(d_1)} \cdot \log(TK/\delta) + t_{\min}\right) \tag{83}$$

where the first term is due to the upper bound on the number of exploration rounds in E_2 and the second term is due to the regret bound for Exp4-IX under model $\mathcal{F}_1$.

2. In the second case when $\mathcal{F}_1$ does not satisfy realizability we must bound the regret when the algorithm is still using $\mathcal{F}_1$. The regret may therefore be decomposed as

$$\text{Reg}_T \leq (V^* - V^{\pi_{\theta_1}}) \cdot t_* + \sum_{t\in[t_*]} V^{\pi_{\theta_1}} - V^{\pi_t} + \sum_{t=t_*+1}^{T} V^* - V^{\pi_t} \tag{84}$$

where t^* is the timestep that the switch is detected. From t_* onward, the algorithm runs Exp4-IX with $\mathcal{F}_2$, so this last term is simply bounded by $\widetilde{\mathcal{O}}(\sqrt{d_2 KT})$ under event E. The same is true for the middle term.

Note that before the switch occurs it must be that $\hat{U}_{t_*-1} \leq \alpha_{t_*-1}$. Therefore, from event E,

$$V^* - V^{\pi_{\theta_1}} \tag{85}$$

$$\leq C\hat{U}_t \log^{1/2}(K) \tag{86}$$

$$+ \mathcal{O}\left(\frac{(\|\theta_{\text{diff}}\|^{1/2} + d_2^{1/4})\log(d_2KT/\delta)\log^{1/2}(K)}{t^{1/3}}\right) \tag{87}$$

$$\leq \mathcal{O}\left(\frac{d_2^{1/4}\log^{3/2}(d_2KT/\delta) \cdot \log^{1/2}(K)}{t^{1/3}}\right) \tag{88}$$

The final regret bound for this case is then

$$\mathrm{Reg}_T \le \mathcal{O}\left(d_2^{1/4}T^{2/3}\cdot\log^{3/2}(d_2KT/\delta)\cdot\log^{1/2}(K)\right) \tag{89}$$

$$+\mathcal{O}\left(\sqrt{d_1TK\log(d_1)}\cdot\log(TK/\delta)\right) \tag{90}$$

$$+\mathcal{O}\left(\sqrt{d_2TK\log(d_2)}\cdot\log(TK/\delta)+t_{\min}\right) \tag{91}$$

□

B.3. A bandit instance that satisfies Assumption 2 but not Assumption 3

Given a constant C, let us define a bandit instance with $K=2$ as follows:

$$\phi(X,1)\sim N(0,1) \tag{92}$$

$$\phi(X,2)=\begin{cases}\phi(X,1) & \text{if } |\phi(X,1)|\le C;\\ -\phi(X,1) & \text{if } |\phi(X,1)|>C.\end{cases} \tag{93}$$

Since the marginal distribution of $\phi(X,1)$ and $\phi(X,2)$ are both $N(0,1)$, it is easy to see that Assumption 2 is satisfied. To see how Assumption 3 fails to hold, we compute the sub-Gaussian norm and L_2 norm of $Z:=\phi(X,1)-\phi(X,2)$. Notice that Z has Gaussian tail when $|Z|>2C$, and thus $\|Z\|_{\psi_2}=\Theta(1)$. For the L_2 norm, note that $\|Z\|_{L_2}=O(\int_C^\infty t^2\exp(-t^2)dt)=O(C^2\exp(-C^2))$ which can be made arbitrarily small by choosing large C. Therefor for any constant L, there exist a bandit instance such that Assumption 2 holds but Assumption 3 does not.

C. Moment-Based Estimator Details

We now describe the construction of the estimator in Algorithm 3 in more detail. As mentioned, the estimator that achieves this sublinear in d rate relies on approximating the $\max$ function with a degree-t multivariate polynomial approximation $p_t:[-1,1]^K\to\mathbb{R}$ of the form

$$p_t(z_1,\dots,z_K)=\sum_{|\alpha|\le t}c_\alpha\prod_{a\in[K]}z_a^{\alpha_a}. \tag{95}$$

Here, $z\in[-1,1]^K$, α is a multiset given by $\alpha=\{\alpha_1,\dots,\alpha_K\}$ for $\alpha_a\in\mathbb{N}$, and we denote $|\alpha|=\sum_{a\in[K]}\alpha_a$. The task then becomes estimating the associated moments up to degree t. The following lemma bounds the approximation error in terms of the degree.

Lemma 9. *Let $f:[-1,1]^K\to\mathbb{R}$ be defined as $f(z)=\max_a z_a$. There exists a degree-t polynomial $p_t:[-1,1]^K\to\mathbb{R}$ of the form (95) such that*

$$\sup_{z\in[0,1]^K}|f(z_1,\dots,z_K)-p_t(z_1,\dots,z_K)|\le\frac{C_K}{t} \tag{96}$$

Algorithm 3 Moment-Based Estimator

Input: Number of samples n, degree t, failure probability δ, coefficients $\{c_\alpha\}_{|\alpha|\le t}$.
Define $q=48\log(1/\delta)$, $m=\frac{n}{q}$.
Initialize empty datasets $D^1,\dots,D^q$
for $k=1,\dots,q$ **do**
  for $i=1,\dots,m$ **do**
    Sample independently $x_i^k\sim\mathcal{D}$ and $a_i^k\sim\mathrm{Unif}[K]$. Receive reward y_i^k.
    Set $\phi_i^k=\phi(x_i^k,a_i^k)$
    Add tuple (ϕ_i^k,y_i^k) to D^k
  end for
end for
for α such that $s:=|\alpha|\le t$ **do**
  Compute independent moment estimators:

$$\hat{S}^k_{m,\alpha}:=\frac{1}{\binom{m}{s}}\sum_{\ell\in\binom{[m]}{s}}\mathbb{E}_X\prod_{j\in[s]}\left\langle y^k_{\ell_j}\phi^k_{\ell_j},\phi(X,a_{(j)})\right\rangle\quad\forall k=1,\dots,q \tag{94}$$

  Set $\hat{S}_{n,\alpha}\leftarrow\mathrm{median}\{\hat{S}^k_{m,\alpha}\}_{k=1}^q$
end for
Return $\hat{S}_n:=\sum_{\alpha\,:\,|\alpha|\le t}c_\alpha\hat{S}_{n,\alpha}$

for some constant C_K. Furthermore, $|c_\alpha|\le\frac{(2et)^{2K+1}2^t}{K^K}=:c_{\max}$ for all α such that $|\alpha|\le t$.

Naturally, the approximation error can be decreased by increasing the degree of the polynomial. Then, we turn our attention obtaining a good estimate of

$$\mathbb{E}_X\left[p_t\left(\langle\phi(X,1),\theta\rangle,\dots,\langle\phi(X,K),\theta\rangle\right)\right] \tag{97}$$

$$=\sum_{|\alpha|\le t}c_\alpha\mathbb{E}_X\prod_{a\in[K]}\langle\phi(X,a),\theta\rangle^{\alpha_a}. \tag{98}$$

We achieve this by estimating the individual moments $\mathbb{E}_X\prod_{a\in[K]}\langle\phi(X,a),\theta\rangle^{\alpha_a}$ for all α up to degree t using the estimators $\hat{S}^k_{m,\alpha}$ specified in (94). Note that in (94), for the multiset α of size $s:=|\alpha|$ and $j\in[s]$, we use the notation $a_{(j)}$ to mean the action $a_{(j)}=\max\{a':\sum_{b<a'}\alpha_b\le j\}$. That is to say, if we considered the tuple $(\phi(X,1),\dots,\phi(X,1),\phi(X,2),\dots,\phi(X,2),\dots,\phi(X,K))$ where $\phi(X,a)$ is repeated α_a times, $\phi(X,a_{(j)})$ refers to the jth element of this tuple.

The algorithm proceeds as follows. As before, the input specifies the number of samples $n\in\mathbb{N}$ and target confidence $\delta>0$. For ease of calculations, we assume that $q:=48\log(1/\delta)$ is in $\mathbb{N}$ and that n is evenly divisible by q for reasons that will become clear momentarily. The algorithm then collects data from the environment by sampling actions from $\mathcal{A}$ uniformly at random resulting in a

total of p datasets $D_m^1, \ldots, D_m^q$ each of size $m = \frac{n}{p}$ where $D_m^k = \{x_i^k, a_i^k, y_i^k\}_{i\in[m]}$. Note that the datasets are independent. For each $k \in [q]$ and $i \in [m]$, x_i^k and a_i^k are i.i.d copies from the distribution $\mathcal{D}$ and Unif $\mathcal{A}$, respectively, and $y_i^k = \langle \phi(x_i^k, a_i^k), \theta\rangle + \eta_i^k(a_i^k)$ as defined by the environment.

To estimate a particular moment $\mathbb{E}_X \prod_{a\in[K]} \langle \phi(X, a), \theta\rangle^{\alpha_a}$, the algorithm constructs q independent estimators $\hat{S}_{m,\alpha}^1, \ldots, \hat{S}_{m,\alpha}^q$ of the form of (94). The intuition is that there are $\binom{m}{s}$ ways to construct a product of unbiased estimators of θ. Then the median is taken across the results of all q datasets $\hat{S}_{n,\alpha}$. Finally, to estimate $\mathbb{E}_X p_t$, it simply constructs the weighted sum of these according to the coefficients of the polynomial and the returns the result, $\hat{S}_n$. Our main result, Theorem 3, states an error bound on the estimate $\hat{S}_n$, combining both the approximation error of the polynomial p_t and the estimation error of $\hat{S}_n$.

D. Proofs of Results in Section 5

D.1. Proof of Lemma 9

Lemma 9. *Let $f : [-1,1]^K \to \mathbb{R}$ be defined as $f(z) = \max_a z_a$. There exists a degree-t polynomial $p_t : [-1,1]^K \to \mathbb{R}$ of the form (95) such that*

$$\sup_{z\in[0,1]^K} |f(z_1, \ldots, z_K) - p_t(z_1, \ldots, z_K)| \le \frac{C_K}{t} \quad (96)$$

for some constant C_K. Furthermore, $|c_\alpha| \le \frac{(2et)^{2K+1}2^t}{K^K} =: c_{\max}$ for all α such that $|\alpha| \le t$.

Proof. It follows from Lemma 2 of (Tian et al., 2017) that, for any 1-Lipschitz g supported on $[0,1]^K$, a polynomial $q(\hat{z}) = \sum_{\alpha\,:\,|\alpha|\le t} \hat{c}_\alpha \prod_{u\in\alpha} \hat{z}^u$ exists satisfying (96) with $|c_\alpha| \le (2t)^K 2^t := \hat{c}_{\max}$ and constant $\frac{C_K}{2}$. The max function g is 1-Lipschitz and thus satisfies this condition. Let $g(\hat{z}) = \max_a \hat{z}_a$ and $\hat{z}_a = \frac{z_a+1}{2}$. Note that $\hat{z} \in [0,1]^K$ by this definition and $f(z) = 2g(\hat{z}) - 1$. Furthermore $p(z) = 2q(\hat{z}) - 1$ degree t polynomial such that $p(z) = \sum_{|\alpha|\le t} c_\alpha \prod_{u\in\alpha} z^u$. Therefore, for any z, $|f(z) - p(z)| \le C_K/t$.

The coefficients c_α are different from $\hat{c}_\alpha$ as a result of the change of variables. Note that there are $\sum_{s=0}^t s + K - 1K - 1 \le (t+1)(et/K + e)^K \le 2t(2et/K)^K$ terms. Therefore $|c_\alpha| \le \frac{(2et)^{2K+1}2^t}{K^K}$. □

D.2. Proof of Theorem 3

First, we verify that $\hat{S}_{m,\alpha}^k$ for $k = 1, \ldots, \lceil 48\log(1/\delta)\rceil$ are unbiased estimators of the moments of interest.

Lemma 10. *Given $\hat{S}_{m,\alpha}^k$ defined in (94), it holds that $\mathbb{E}_{D^k}\left[\hat{S}_{m,\alpha}^k\right] = \mathbb{E}_X \prod_{a\in[K]} \langle \theta, X_a\rangle^{\alpha_a}$.*

Proof. We drop the superscript k notation denoting which of the datasets is being used as the argument is identical. Fix $\ell \in \binom{[m]}{s}$ as an s-combination of the indices $[n]$. Since the data in D is i.i.d, we have that

$$\mathbb{E}_D\mathbb{E}_X \prod_{j\in[s]} \left\langle y_{\ell_j} x_{\ell_j}, \phi(X, a_{(j)})\right\rangle = \mathbb{E}_X \prod_{j\in[s]} \left\langle \mathbb{E}_D\left[y_{\ell_j} x_{\ell_j}\right], \phi(X, a_{(j)})\right\rangle \quad (99)$$

$$= \mathbb{E}_X \prod_{j\in[s]} \left\langle \theta, \phi(X, a_{(j)})\right\rangle \quad (100)$$

$$= \mathbb{E}_X \prod_{a\in[K]} \langle \theta, \phi(X, a)\rangle^{\alpha_a} \quad (101)$$

where the second equality uses the fact that $\mathbb{E}_D x_i x_i^\top = \mathbb{I}_d$. for all $i \in [m]$. □

Next, we establish a bound on the variance in preparation to apply Chebyshev's inequality.

Lemma 11. *There exists a constant C such that*

$$\mathrm{var}(\hat{S}_{m,\alpha}^k) = C^s s^{2s} \cdot \sum_{u=1}^s \left(\frac{\sqrt{d}}{m}\right)^u \quad (102)$$

where $s = |\alpha|$.

Proof. As before, we will drop the superscript k notation as the argument is identical for each independent estimator. Let $s = |\alpha|$.

By definition, the variance is given by

$$\operatorname*{var}_D\left(\hat{S}_{m,\alpha}\right) = \mathbb{E}_D\left[\hat{S}_{m,\alpha}^2\right] - \mathbb{E}_D\left[\hat{S}_{m,\alpha}\right]^2 \quad (103)$$

where

$$\hat{S}_{m,\alpha}^2 = \frac{1}{\binom{m}{s}^2} \sum_{\ell,\ell'} \mathbb{E}_{X,X'} \prod_{i\in[s]} \left\langle y_{\ell_i} x_{\ell_i}, \phi(X, a_{(j)})\right\rangle \quad (104)$$

$$\cdot \prod_{i\in[s]} \left\langle y_{\ell'_i} x_{\ell'_i}, \phi(X', a_{(i)})\right\rangle \quad (105)$$

$$\mathbb{E}_D\left[\hat{S}_{n,\alpha}\right]^2 = \mathbb{E}_{X,X'} \prod_a \langle \theta, \phi(X, a)\rangle^{\alpha_a} \cdot \prod_a \langle \theta, \phi(X', a)\rangle^{\alpha_a} \quad (106)$$

where again ℓ and ℓ' are s-combinations $[n]$. Similar to (Kong and Valiant, 2018), we can analyze the variance as

individual terms in the sum over ℓ and ℓ':

$$\mathbb{E}_D\mathbb{E}_{X,X'}\prod_{i\in[s]}\left\langle y_{\ell_i}x_{\ell_i},\phi(X,a_{(i)})\right\rangle\cdot\prod_{j\in[s]}\left\langle y_{\ell'_i}x_{\ell'_i},\phi(X',a_{(i)})\right\rangle \tag{107}$$

$$-\mathbb{E}_{X,X'}\prod_a\langle\theta,\phi(X,a)\rangle^{\alpha_a}\cdot\prod_a\langle\theta,\phi(X',a)\rangle^{\alpha_a} \tag{108}$$

There are two important cases to consider: (1) when ℓ and ℓ' do not share any indices and (2) when there is partial or complete overlap of indices.

1. **No intersection of ℓ and ℓ'** In this case, we may see that there is no contribution to the variance for this term due to independence:

$$\mathbb{E}_D\mathbb{E}_{X,X'}\prod_{i\in[s]}\left\langle y_{\ell_i}x_{\ell_i},\phi(X,a_{(i)})\right\rangle\cdot\prod_{i\in[s]}\left\langle y_{\ell'_i}x_{\ell'_i},\phi(X',a_{(i)})\right\rangle \tag{109}$$

$$=\mathbb{E}_{X,X'}\prod_{i\in[s]}\left\langle\theta,\phi(X,a_{(i)})\right\rangle\cdot\prod_{i\in[s]}\left\langle\theta,\phi(X',a_{(i)})\right\rangle \tag{110}$$

$$=\mathbb{E}_{X,X'}\prod_a\langle\theta,\phi(X,a)\rangle^{\alpha_a}\cdot\prod_a\langle\theta,\phi(X',a)\rangle^{\alpha_a} \tag{111}$$

   This term simply cancels with $-\mathbb{E}\left[\hat{S}_{m,\alpha}\right]^2$.

2. **Partial or complete intersection of ℓ and ℓ'**

   In this case, there are some samples that appear twice. Let $\beta=\{(i,j)\ :\ \ell_i=\ell'_j\}$ be the set of indices that refer to the same sample in D. Also define $\gamma,\gamma'\subseteq[s]$ as the subsets of indices of ℓ and ℓ' respectively that are not shared.

   The left-hand side of this term can be then be written as

$$\mathbb{E}_D\mathbb{E}_{X,X'}\prod_{i\in[s]}\left\langle y_{\ell_i}\phi_{\ell_i},\phi(X,a_{(j)})\right\rangle\cdot\prod_{j\in[s]}\left\langle y_{\ell'_j}\phi_{\ell'_i},\phi(X',a_{(i)})\right\rangle \tag{112}$$

$$=\mathbb{E}_{X,X'}\prod_{(i,i')\in\beta}\mathbb{E}_{D_n}\left[y^2_{\ell_i}\left\langle\phi_{\ell_i},\phi(X,a_{(i)})\right\rangle\left\langle\phi_{\ell_i},\phi(X',a_{(i')})\right\rangle\right] \tag{113}$$

$$\times\prod_{i\in\gamma}\left\langle\theta,X_{(i)}\right\rangle\prod_{i'\in\gamma'}\left\langle\theta,X'_{(i')}\right\rangle \tag{114}$$

   To proceed, we require the following lemma which bounds separate moments in the factors that come from shared indices. The proof given in Section E.

Lemma 12. *Let $p\geq1$ be an integer and define $M=\mathbb{E}_{D_n}\left[y^2_{\ell_i}\phi_{\ell_i}\phi^\top_{\ell_i}\right]$. There is a constant C such that*

$$\left(\mathbb{E}_{X,X'}|\phi(X,a_{(i)})^\top M\phi(X',a_{(i)})|^p\right)^{1/p} \tag{115}$$

$$\leq C\cdot p\tau^2(\sigma^2+L\|\theta\|^2)\sqrt{d} \tag{116}$$

Through standard sub-Gaussian arguments, we also have that, for $p\geq1$, $\left(\mathbb{E}|\left\langle\theta,\phi(X,a_{(i)})\right\rangle|^p\right)^{1/p}\leq C\tau\|\theta\|\sqrt{p}$ for some constant $C>0$. And the same holds for the X' factors.

For convenience, let $\zeta=(\sigma^2+L\|\theta\|^2)$ and let $m=|\beta|\leq s$ be the size of the overlap. By the generalized Holder inequality, the term in (113) is upper bounded by

$$\left(\prod_{(i,i')\in\beta}\mathbb{E}_{X,X'}|\phi(X,a_{(i)})^\top M\phi(X',a_{(i)})|^{2s}\prod_{i\in\gamma}\mathbb{E}_{X,X'}|\left\langle\theta,\phi(X,a_{(i)})\right.\right. \tag{117}$$

$$\left.\left.\times\left(\prod_{i'\in\gamma'}\mathbb{E}_{X,X'}|\left\langle\theta,\phi(X',a_{(i')})\right\rangle|^{2s}\right)^{1/2s}\right. \tag{118}$$

$$\leq(C_0\cdot(2s)\tau^2\zeta\sqrt{d})^m\cdot\left(C_1\cdot\tau\|\theta\|\sqrt{2s}\right)^{2s-2m} \tag{119}$$

$$\leq C_2^{2s}\cdot\tau^{2s}\zeta^m\|\theta\|^{2s-2m}\cdot s^s\cdot d^{m/2} \tag{120}$$

where $C_0,C_1,C_2>0$ are problem-independent constants.

In summary, we have shown that there is no contribution to the variance when no indices are shared between ℓ and ℓ' and the contribution to the variance when m indices are shared is bounded by $\mathcal{O}(d^{m/2})$. It suffices now to count the terms to see the total contribution for each $u=1,\ldots,s$.

It can be checked that the number of terms where the size of the intersection $|\beta|=u$ is

$$\binom{m}{k}\binom{k}{u}\binom{m-k}{k-u}\leq\left(\frac{me}{s}\right)^s\left(\frac{se}{u}\right)^u\left(\frac{(m-s)e}{s-u}\right)^{s-u} \tag{121}$$

$$\leq\frac{m^{2s-u}e^{2s}}{s^su^u(s-u)^{s-u}} \tag{122}$$

since there are s elements ℓ, u of which may have an intersection, and a remaining $s-u$ elements to be chosen for ℓ' that are not shared with ℓ. Similarly, we have $\binom{m}{s}^2\geq\left(\frac{m}{s}\right)^{2s}$ which implies that the variance can be

bounded as

$$\operatorname*{var}_D\left(\hat{S}_{m,\alpha}\right) \leq \frac{1}{\binom{m}{s}^2}\sum_{u=1}^{s}\frac{m^{2s-u}e^{2s}}{s^s u^u (s-u)^{s-u}}\cdot C_0^{2s} \tag{123}$$

$$\cdot\,\tau^{2s}\zeta^u\|\theta\|^{2s-2u}\cdot s^s\cdot d^{u/2} \tag{124}$$

$$\leq \sum_{u=1}^{s} C_1^{2s}\cdot\left(\frac{\sqrt{d}}{m}\right)^u\cdot s^{2s}\tau^{2s}\zeta^u\|\theta\|^{2s-2u} \tag{125}$$

for absolute constants $C_0, C_1, C_2 > 0$. Since it was assumed that τ, σ^2, L and $\|\theta\|$ are $\mathcal{O}(1)$, the final claim follows. □

The error bound result on the median of the estimators follows almost immediately.

Theorem 4. *There exists a constant $C = \mathcal{O}(1)$ for all k such that, with probability at least $2/3$,*

$$|\hat{S}^k_{m,\alpha} - \mathbb{E}\hat{S}^k_{m,\alpha}| \leq \epsilon(m,d,s) \tag{126}$$

where

$$\epsilon(m,d,s) := C^{s/2}s^s\sum_{u=1}^{s}\left(\frac{\sqrt{d}}{m}\right)^{u/2} \tag{127}$$

Furthermore, defining $\hat{S}_{n,\alpha} = \operatorname{median}\{\hat{S}^k_{m,\alpha}\}_{k=1}^q$, with probability $1-\delta$,

$$|\hat{S}_{n,\alpha} - \mathbb{E}\hat{S}^k_{n,\alpha}| \leq \epsilon(m,d,s) \tag{128}$$

Proof. The first statement follows immediately from Chebyshev's inequality and the second applies the median of means trick for the independent estimators $\{\hat{S}^k_{m,\alpha}\}_{k=1}^q$ given the choice of q (Kong et al., 2020, Fact 11). □

D.2.1. Final Bound

We now combine the estimation and approximation error bounds to derive the final result, which is reproduced here.

Theorem 3. *Let assumptions 1, 2, 4, and 5 hold. Let $\hat{S}_n = \sum_{\alpha\,:\,|\alpha|\leq t} c_\alpha \hat{S}_{n,\alpha}$ as defined in Algorithm 3 be the estimator of $\mathbb{E}_X p_t$ up to degree t. There is an absolute constant $C > 0$ such that with probability at least $1 - t(et/K+e)^K\delta$,*

$$|V^* - \hat{S}_n| \leq \frac{C_K}{t} \tag{9}$$

$$+\,t\,(et/K+e)^K\,c_{\max}\cdot C^{t/2}t^t\cdot\sum_{s=1}^{t}\left(\frac{\sqrt{d}}{n}\cdot\log(1/\delta)\right)^{s/2} \tag{10}$$

where C_K is a constant that depends only on K.

Proof. We first start by bounding the full estimation error $|\mathbb{E}_X p(X) - \hat{S}_n|$. The degree 0 and degree 1 moments are already known exactly; thus we may consider $2 \leq s \leq t$. By the union bound and triangle inequality combined with the result of Theorem 4, with probability at least $1 - t(et/K+e)^K\delta$,

$$|\mathbb{E}_X p(X) - \hat{S}_n| \leq \sum_{\substack{s\in[2,t]\\ \alpha\,:\,|\alpha|\leq t}} c_\alpha|\hat{S}_{n,\alpha} - \mathbb{E}\hat{S}^k_{n,\alpha}| \tag{129}$$

$$\leq \sum_{\substack{s\in[2,t]\\ \alpha\,:\,|\alpha|\leq t}} c_{\max}\epsilon(n,d,s) \tag{130}$$

For each s, there are $\binom{s+K-1}{K-1} \leq (es/K+e)^K$ monomials for possible choices of α. Therefore, the good event implies that

$$|\mathbb{E}_X p(X) - \hat{S}_n| \leq t\,(et/K+e)^K\,c_{\max}\cdot\epsilon(n,d,t) \tag{131}$$

Next, we may apply the approximation error. By the triangle inequality

$$|\mathbb{E}\max_a\langle\theta, X_a\rangle - \hat{S}_n| \leq \frac{C_K}{t} + (es/K+e)^K\,c_{\max}\cdot\epsilon(n,d,t) \tag{132}$$

□

Corollary 1. *Under the same assumptions as Theorem 3, estimator $\hat{S}_n$ generated by Algorithm 3 satisfies $|V^* - \hat{S}_n| \leq \epsilon$ for $\epsilon < 1$ with probability at least $1-\delta$ and sample complexity*

$$\mathcal{O}\left(K\left(\frac{C_K}{\epsilon}\right)^{K+C_K/\epsilon}\cdot\frac{\sqrt{d}}{\epsilon^2}\cdot\log\left(\frac{C_K}{\epsilon\delta}\right)\right) \tag{11}$$

where C_K is a constant that depends only on K.

Proof. To ensure that each term in the sum is at most $\frac{\epsilon}{t}$, it suffices to take

$$n = c_{\max}^2 C^t\cdot t^{2t+4}\,(et/K+e)^{2K}\cdot\sqrt{d}\epsilon^{-2}\cdot\log(1/\delta) \tag{133}$$

Then, choose $t = 2C_K/\epsilon$. Therefore, by the definition of $c_{\max}$,

$$n = \mathcal{O}\left(\left(\frac{C\cdot C_K}{\epsilon}+1\right)^{K+C_K/\epsilon}\cdot\frac{\sqrt{d}}{\epsilon^2}\right), \tag{134}$$

for some constant $C \geq 1$. Since we require $48\log(1/\delta)$ estimators to ensure the good event occurs with probability at least , the total sample complexity is

$$\mathcal{O}\left(\left(\frac{C\cdot C_K}{\epsilon}+1\right)^{K+C_K/\epsilon}\cdot\frac{\sqrt{d}}{\epsilon^2}\cdot\log\left(\frac{C_K\,(C_K/\epsilon+1)^K}{\epsilon\delta'}\right)\right) \tag{135}$$

□

with probability $1-\delta'$ where $\delta' = \frac{\delta}{t(et/K+e)^K}$

E. Supporting Lemmas

The following is a restatement of the moment bound in Lemma 13.

Lemma 13. *Let $p \geq 1$ be an integer and define $M = \mathbb{E}_{D_n}\left[y_{\ell_i}^2 \phi_{\ell_i}\phi_{\ell_i}^\top\right]$. There is a constant C such that*

$$\left(\mathbb{E}_{X,X'}|\phi(X,a_{(i)})^\top M \phi(X',a_{(i)})|^p\right)^{1/p} \tag{136}$$
$$\leq C \cdot p\tau^2(\sigma^2 + L\|\theta\|^2)\sqrt{d} \tag{137}$$

Proof. For convenience, define $X_{(i)} = \phi(X, a_{(i)})$ and the same for $X'_{(i)}$. Define $A = \begin{bmatrix} \mathbf{0}_d & M \\ \mathbf{0}_d & \mathbf{0}_d \end{bmatrix}$ and $Z = \begin{bmatrix} X_{(i)} \\ X'_{(i)} \end{bmatrix}$. Note that $Z^\top A Z = X_{(i)}^\top M X'_{(i)}$ and $A^\top A = \begin{bmatrix} M^\top M & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d \end{bmatrix}$. By Lemma 16, $Z \sim \mathrm{subG}(C_0\tau^2)$. Furthermore, $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ^\top = \mathbb{I}_d$. The remaining proof utilizes a variation of the Hanson-Wright inequality due to (Zajkowski, 2020), stated in Lemma 14[2]. By this inequality, there exists a constant $C > 0$ such that

$$\mathbb{P}\left(|Z^\top AZ - \mathbb{E}\left[Z^\top AZ\right]| \geq \xi\right) \tag{138}$$
$$\leq \exp\left(-C\min\left\{\frac{\xi^2}{\tau^4\|A\|_F^2}, \frac{\xi}{\tau^2\|A\|_F}\right\}\right) \tag{139}$$

By direct calculation, we have that $\mathbb{E}\left[Z^\top AZ\right] = \mathrm{tr}\,\mathbb{E}\left[ZZ^\top A\right] = 0$ and by Lemma 15, $\|A\|_F \leq \sqrt{d}(\sigma^2 + L\|\theta\|^2)$. To bound the moment, we use the tail-sum-expectation for non-negative random variables. For convenience, define $\sigma_1 = \tau^2\|A\|_F$.

$$\mathbb{E}|Z^\top AZ|^p = \int_0^\infty \mathbb{P}\left(|Z^\top AZ|^p \geq u\right) du \tag{140}$$
$$= \int_0^\infty pv^{p-1}\mathbb{P}\left(|Z^\top AZ| \geq v\right) dv \tag{141}$$
$$\leq \int_0^\infty pv^{p-1}\max\left\{e^{\frac{Cv^2}{\sigma_1^2}}, e^{\frac{Cv}{\sigma_1}}\right\} dv \tag{142}$$
$$\leq \int_0^\infty pv^{p-1}e^{\frac{Cv^2}{\sigma_1^2}} dv + \int_0^\infty pv^{p-1}e^{\frac{Cv}{\sigma_1}} dv \tag{143}$$

[2]Critically, Lemma 14 applies to quadratic forms of sub-Gaussian, dependent random variables, rather than requiring the coordinates of Z to be independent as in the traditional Hanson-Wright inequality (Rudelson et al., 2013; Hanson and Wright, 1971). As a consequence, the second term in the minimum of the above tail bound depends on $\|A\|_F$ as opposed to the operator norm $\|A\|$. Further discussion may be found in (Zajkowski, 2020).

The first inequality used Lemma 14. Consider the second term first. Let $r = Cv/\sigma_1$. Then, by a change of variables,

$$\int_0^\infty pv^{p-1}e^{\frac{Cv}{\sigma_1}} dv = p(\sigma_1/C)^p \int_0^\infty r^{p-1}e^{-r} dr \leq 3p(\sigma_1/C)^p \cdot p^p \tag{144}$$

where we have used the Gamma function inequality $\int_0^\infty r^{p-1}e^{-r}dr \leq 3p^p$ (Vershynin, 2018). Consider the first term. Let $r = Cv^2/\sigma_1^2$. Like the previous part, we may apply a change of variables.

$$\int_0^\infty pv^{p-1}e^{-Cv^2/\sigma_1^2} dv = \frac{1}{2}\int_0^\infty p\left(\frac{\sigma_1^2 r}{C}\right)^{\frac{p-1}{2}} e^{-r} \cdot \sqrt{\frac{\sigma_1^2}{rC}} \cdot dr \tag{145}$$
$$= \frac{p}{2}\left(\frac{\sigma_1^2}{C}\right)^{\frac{p}{2}} \int_0^\infty r^{\frac{p}{2}-1}e^{-r} dr \tag{146}$$
$$\leq \frac{3p}{2}\left(\frac{\sigma_1^2}{C}\right)^{\frac{p}{2}} \cdot (p/2)^{(p/2)}. \tag{147}$$

Taking these two together,

$$\left(\mathbb{E}|Z^\top AZ|^p\right)^{1/p} \leq \left(3p(\sigma_1/C)^p \cdot p^p + \frac{3p}{2}\left(\frac{\sigma_1^2}{C}\right)^{\frac{p}{2}} \cdot (p/2)^{(p/2)}\right)^{1/p} \tag{148}$$
$$\leq C' \cdot \sigma_1(p + \sqrt{p}), \tag{149}$$

for some other constant $C' > 0$ since $p^{1/p}$ is bounded by a constant. Since we only consider $p \geq 1$, the claim follows. □

Lemma 14 (Restatement of Corollary 2.8 of (Zajkowski, 2020))**.** *Let $X \sim \mathrm{subG}(\tau^2)$ be a centered random vector in $\mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$. Then, there exists a constant $C > 0$ such that*

$$\mathbb{P}\left(|X^\top AX - \mathbb{E}\left[X^\top AX\right]|\right) \leq \exp\left(-C\min\left\{\frac{\xi^2}{\tau^4\|A\|_F^2}, \frac{\xi}{\tau^2\|A\|_F}\right\}\right) \tag{150}$$

where $\|\cdot\|_F$ is the Frobenius norm.

Lemma 15. *Let (ϕ, y) be generated under the uniform-random policy. Define $M = \mathbb{E}\left[y^2\phi\phi^\top\right]$ and $A = \begin{bmatrix} \mathbf{0}_d & M \\ \mathbf{0}_d & \mathbf{0}_d \end{bmatrix}$. Under Assumption 4, $\|A\| \leq L\|\theta\|^2 + \sigma^2$ and $\|A\|_F \leq \sqrt{d}(L\|\theta\|^2 + \sigma^2)$.*

Proof. By definition $\|A\|^2 = \sup_{v\,:\,\|v\|=1} v^\top A^\top Av = \sup_{v\,:\,\|v\|=1} v_1^\top M^\top M v_1 = \|M\|^2$ where v_1 denotes the

first d coordinates of v. The first equality follows since $A^\top A = \begin{bmatrix} M^\top M & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d \end{bmatrix}$. Since M is positive semi-definite,

$$\|M\| = \sup_{v \in \mathbb{R}^d \,:\, \|v\|=1} v^\top M v \tag{151}$$

$$= \sup_{v \in \mathbb{R}^d \,:\, \|v\|=1} \mathbb{E}\left[y^2 (\phi^\top v)^2\right] \tag{152}$$

$$= \sup_{v \in \mathbb{R}^d \,:\, \|v\|=1} \left\{\mathbb{E}\left[(\phi^\top v)^2 (\phi^\top \theta)^2\right] + \mathbb{E}\left[(\phi^\top v)^2 \eta^2\right]\right\} \tag{153}$$

$$\tag{154}$$

The second term is simply $\mathbb{E}\eta^2 = \sigma^2$ since $\mathbb{E}\phi\phi^\top = \mathbb{I}_d$ and ϕ and η are independent. The first term may be bounded as $\mathbb{E}\left[(\phi^\top v)^2 (\phi^\top \theta)^2\right] \le L \cdot \mathbb{E}\left[(\phi^\top v)^2\right] \mathbb{E}\left[(\phi^\top \theta)^2\right] = L\|\theta\|^2$ by Assumption 4. This concludes the first claim. For the second, we note that $\|A\|_F^2 = \operatorname{tr} A^\top A = \operatorname{tr} M^\top M \le d\|M\|^2$ and the second claim follows by applying the first. □

Lemma 16. *Let X, Y $\mathrm{subG}(\tau^2)$ be two independent sub-Gaussian vectors in $\mathbb{R}^d$. Then, $Z = \begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathrm{subG}(C_0\tau^2)$ for some constant $C_0 > 0$.*

Proof. Let $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \in \mathbb{R}^{2d}$ where $v_1, v_2 \in \mathbb{R}^d$ and $\|v\|_2 = 1$. Then, $v^\top Z = v_1^\top X + v_2^\top Y$ is the sum of independent sub-Gaussian variables where $v_1^\top X \sim \mathrm{subG}(\|v_1\|_2^2 \tau^2)$ and $v_2^\top Y \sim \mathrm{subG}(\|v_2\|_2^2 \tau^2)$ where both $\|v_1\|_2 \le 1$ and $\|v_2\|_2 \le 1$. Therefore $v^\top Z \sim \mathrm{subG}(C_0\tau^2)$ for a constant $C_0 > 0$. Since v was arbitrary, the statement follows. □

E.1. Multiplicative Error Bound for Estimating Norms

In this section, we prove a multiplicative error bound for estimating $\|\theta\|^2$, which can potentially be faster. The key is an application of the AM-GM inequality, similar to the work of (Foster et al., 2019). As before, we will consider a dataset of n samples split evenly into $D = \{\phi_i, y_i\}$ and $D' = \{\phi_i', y_i'\}$ each of size $m = \frac{n}{2}$. Define

$$\hat{\theta} = \frac{1}{m} \sum_{i \in [m]} \phi_i y_i \tag{155}$$

$$\hat{\theta}' = \frac{1}{m} \sum_{i \in [m]} \phi_i' y_i' \tag{156}$$

Then, we estimate $\theta^\top \theta$ with $\hat{\theta}^\top \hat{\theta}'$.

Theorem 5. *Let $\delta \le 1/e$ and let $c > 1$ be a constant. With $\hat{\theta}$ and $\hat{\theta}'$ defined above with n total samples, the following error bound holds with probability at least $1 - \delta$:*

$$|\hat{\theta}^\top \hat{\theta}' - \theta^\top \theta| \le \frac{\theta^\top \theta}{2c} + \mathcal{O}\left(\frac{c(\|\theta\| + \sqrt{d}) \max\{\xi^2, \xi\} \log^2(d/\delta)}{n}\right) \tag{157}$$

Proof. Similar to the proof of Theorem 1, we apply the triangle inequality use Bernstein's inequality to bound two terms individually with high probability.

The decomposition becomes

$$|\hat{\theta}^\top \hat{\theta}' - \theta^\top \theta| \le |\hat{\theta}^\top \theta - \theta^\top \theta| + |\hat{\theta}^\top \theta' - \hat{\theta}^\top \theta| \tag{158}$$

We start with the first term. By Bernstein's inequality there is a constant $C > 0$ such that

$$\mathbb{P}\left(|\hat{\theta}^\top \theta - \theta^\top \theta| \ge \epsilon\right) \le \exp\left(-C \min\left\{\frac{m\epsilon^2}{\|\theta\|^2 \xi^2}, \frac{m\epsilon}{\|\theta\| \xi}\right\}\right) \tag{159}$$

since $y_i \theta^\top x_i$ is sub-exponential with $\|y_i \theta^\top x_i\|_{\psi_1} \le \xi \|\theta\|$, as before. Rearranng, we have that with probability at least $1 - \delta$,

$$|\hat{\theta}^\top \theta - \theta^\top \theta| \le \sqrt{\frac{\|\theta\|^2 \xi^2 \log(1/\delta)}{Cm}} + \frac{\|\theta\| \xi \log(1/\delta)}{Cm} \tag{160}$$

$$\le \frac{\|\theta\|^2}{4c} + \frac{c\xi^2 \log(1/\delta)}{Cm} + \frac{c\|\theta\| \xi \log(1/\delta)}{Cm} \tag{161}$$

where the second line follows from the AM-GM inequality. Similarly, conditioned on the dataset D, the second term in the triangle inequality may be bounded as

$$|\hat{\theta}^\top \theta' - \hat{\theta}^\top \theta| \le \sqrt{\frac{\|\hat{\theta}\|^2 \xi^2 \log(1/\delta)}{Cm}} + \frac{\|\hat{\theta}\| \xi \log(1/\delta)}{Cm} \tag{162}$$

$$\le \|\hat{\theta}\| \cdot \left(\sqrt{\frac{\xi^2 \log(1/\delta)}{Cm}} + \frac{\xi \log(1/\delta)}{Cm}\right) \tag{163}$$

with probability at least $1 - \delta$. Finally the proof Theorem 1 shows that, with probability $1 - \delta$,

$$\|\hat{\theta}\| \le \|\theta\| + \sqrt{\frac{d\xi^2}{Cm}} \cdot \log(2d/\delta) \tag{164}$$

Under both of these events, we have

$$|\hat{\theta}^\top \theta' - \hat{\theta}^\top \theta| \leq \sqrt{\frac{\|\theta\|^2 \xi^2 \log(1/\delta)}{Cm}} + \frac{\|\theta\| \xi \log(1/\delta)}{Cm} \tag{165}$$

$$+ \frac{\sqrt{d}\xi^2 \log^{3/2}(2d/\delta)}{Cm} + \frac{\sqrt{d}\xi^2 \log^2(2d/\delta)}{(Cm)^{3/2}} \tag{166}$$

$$\leq \frac{\|\theta\|^2}{4c} + \frac{c\xi^2 \log(1/\delta)}{Cm} + \frac{\|\theta\| \xi \log(1/\delta)}{Cm} \tag{167}$$

$$+ \frac{\sqrt{d}\xi^2 \log^{3/2}(2d/\delta)}{Cm} + \frac{\sqrt{d}\xi^2 \log^2(2d/\delta)}{(Cm)^{3/2}} \tag{168}$$

where the second line again uses the AM-GM inequality. Putting all three events together and applying the union bound, we have with probability $1 - 3\delta$,

$$|\hat{\theta}^\top \hat{\theta}' - \theta^\top \theta| \leq \frac{\|\theta\|^2}{2c} + \mathcal{O}\left(\frac{c\|\theta\| \max\{\xi^2, \xi\} \log(1/\delta)}{m} + \frac{c\sqrt{d}\xi^2 \log^2(2d/\delta)}{m} \right) \tag{169}$$

Simplifying the error term gives the result. □