SPEED: Experimental Design for Policy Evaluation in Linear Heteroscedastic Bandits

Subhojyoti Mukherjee* **Qiaomin Xie*** **Josiah P. Hanna*** **Robert Nowak***
*University of Wisconsin-Madison

Abstract

In this paper, we study the problem of optimal data collection for policy evaluation in linear bandits. In policy evaluation, we are given a *target* policy and asked to estimate the expected reward it will obtain when executed in a multi-armed bandit environment. Our work is the first work that focuses on such an optimal data collection strategy for policy evaluation involving heteroscedastic reward noise in the linear bandit setting. We first formulate an optimal design for weighted least squares estimates in the heteroscedastic linear bandit setting with the knowledge of noise variances. This design minimizes the mean squared error (MSE) of the estimated value of the target policy and is termed the oracle design. Since the noise variance is typically unknown, we then introduce a novel algorithm, SPEED (**S**tructured **P**olicy **E**valuation **E**xperimental **D**esign), that tracks the oracle design and derive its regret with respect to the oracle design. We show that regret scales as $\widetilde{O}(d^3 n^{-3/2})$ and prove a matching lower bound of $\Omega(d^2 n^{-3/2})$. Finally, we evaluate SPEED on a set of policy evaluation tasks and demonstrate that it achieves MSE comparable to an optimal oracle and much lower than simply running the target policy.

1 INTRODUCTION

Bandit policy optimization has been applied in various applications such as web marketing [Bottou et al., 2013], web search [Li et al., 2011], and healthcare recommendations [Zhou et al., 2017]. In practice, before widely deploying a learned policy, it is often necessary to have an accurate estimation of its performance (i.e., expected reward). To this effect, *policy evaluation* is often a critical step as it allows practitioners to determine if a learned policy truly represents improved task performance. While off-policy evaluation has been extensively studied as a potential solution [Dudík et al., 2014, Li et al., 2015, Swaminathan et al., 2017, Wang et al., 2017, Su et al., 2020, Kallus et al., 2021, Cai et al., 2021], in practice, some amount of online evaluation is often required before widescale deployment. For instance, in web-marketing it is common to run an A/B test with a subset of users before a potential new policy is deployed for all users [Kohavi and Longbotham, 2017]. When online policy evaluation is required, we desire methods that provide an accurate estimate of policy performance with a minimal amount of data collected. The default choice for online policy evaluation is to simply run the target policy and average the resulting rewards. However, this approach is sub-optimal when the action space is large or different actions have reward distributions with different variances.

In this paper, we formulate a new experimental design for allocating action samples so as to obtain minimal mean squared error (MSE) for policy evaluation. Specifically, we consider optimal policy evaluation under the following linear heteroscedastic bandit model.

Let $\mathcal{A}$ be the set of *actions* and each $a \in \mathcal{A}$ is associated with a feature vector $\mathbf{x}(a) \in \mathbb{R}^d$ and $|\mathcal{A}| = A$. The reward distribution for each action a has mean $\boldsymbol{\theta}_*^\top \mathbf{x}(a)$, for some $\boldsymbol{\theta}_* \in \mathbb{R}^d$. Often the variance of the reward distribution is assumed to be the same for all actions, but in this paper, we depart from this assumption. We consider the setting that the variance is governed by a quadratic function of the form $\mathbf{x}(a)^\top \boldsymbol{\Sigma}_* \mathbf{x}(a)$, for some symmetric positive definite matrix $\boldsymbol{\Sigma}_* \in \mathbb{R}^{d \times d}$. This assumption allows us to capture problems in which both the mean reward and the variance may depend on the action taken, but both vary smoothly in $\mathbf{x}(a)$.

We briefly contrast our studied setting with other work. In policy evaluation, the common metric of algorithm performance is regret with respect to the mean squared

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

error of an oracle algorithm that has knowledge of the variances of different reward distributions (i.e., knows Σ^*). There has been an increasing focus on studying data collection for policy evaluation in bandit settings [Zhu and Kveton, 2021, 2022, Wan et al., 2022] and there has been some theoretical progress [Chaudhuri et al., 2017, Fontaine et al., 2021]. Several works [Antos et al., 2008, Carpentier and Munos, 2012, Carpentier et al., 2015, Fontaine et al., 2021] have shown that in the classical bandit setting a regret of $\widetilde{O}(An^{-3/2})$ is possible where n is the total budget of actions that can be tried and $\widetilde{O}$ hides logarithmic factors. These works have also shown that simply running the target policy to take actions results in a slower decrease of regret at the rate of $\widetilde{O}(An^{-1})$. Note that collecting data through running the target policy is called on-policy sampling. The work of Zhu and Kveton [2022], Wan et al. [2022] studies the same setting under safety constraints and provides asymptotic error bounds. However, none of the above works provides a finite-time regret guarantee for data collection for policy evaluation in the heteroscedastic linear bandit setting.

The closest works to ours [Antos et al., 2008, Carpentier and Munos, 2012, Carpentier et al., 2015, Fontaine et al., 2021] either consider unstructured settings or consider the classical bandit setting. As many real-world bandit applications have $d \ll A$, a natural question arises as to how to build an algorithm for policy evaluation in the heteroscedastic linear bandit setting with unknown $\boldsymbol{\theta}_*$ and $\boldsymbol{\Sigma}_*$ that can leverage the structure. Further, we want the regret of such an algorithm to decrease at a rate faster than $\widetilde{O}(n^{-1})$ (the on-policy regret rate) and to scale with the dimension d instead of actions as $A \gg d$. Note that the regret should scale at least by d^2 because the learner needs to probe in d^2 dimensions to estimate $\boldsymbol{\Sigma}_* \in \mathbb{R}^{d\times d}$ [Wainwright, 2019]. Thus, the goal of our work is to answer the question:

Can we design an algorithm to collect data for policy evaluation that adapts to the variance of each action, and its regret decreases at a rate faster than $\widetilde{O}(d^2n^{-1})$?

In this paper, we answer this question affirmatively. We make the following novel contributions to the growing literature on online policy evaluation:

1. We are the first to formulate the policy evaluation problem for heteroscedastic linear bandit setting where the variance of each action $a \in \mathcal{A}$ depends on a lower dimensional co-variance matrix parameter $\boldsymbol{\Sigma}_* \in \mathbb{R}^{d\times d}$ such that variance $\sigma^2(a) = \mathbf{x}(a)^\top \boldsymbol{\Sigma}_* \mathbf{x}(a)$. This is a more general heteroscedastic linear bandit setting than studied in Chaudhuri et al. [2017], Kirschner and Krause [2018], Fontaine et al. [2021], and different than the time-dependent variance model of Zhang et al. [2021], Zhao et al. [2022].

2. We characterize the MSE in this setting and show that the optimal design, denoted as **P**olicy **E**valuation (PE) Optimal design that minimizes the MSE is different than A-, D-, E-, G-optimality [Pukelsheim, 2006]. We establish several key properties of this novel PE-Optimal design and discuss how we can solve for the design efficiently.

3. Finally, we propose the agnostic algorithm, SPEED, that does not know the underlying covariance matrix $\boldsymbol{\Sigma}_*$. SPEED tracks the oracle design and we analyze its MSE. We then bound the regret of SPEED compared to an oracle strategy that follows the optimal design with the knowledge of $\boldsymbol{\Sigma}_*$. We show that the regret scales as $O(\frac{d^3\log(n)}{n^{3/2}})$ which is an improvement over the regret for the stochastic non-structured bandit setting which scales as $O(\frac{A\log(n)}{n^{3/2}})$ [Carpentier and Munos, 2011, 2012, Carpentier et al., 2015, Fontaine et al., 2021]. Hence, we answer positively to our main query. We also prove the first lower bound for this setting that scales as $\Omega(\frac{d^2\log(n)}{n^{3/2}})$. Finally, we conduct experiments on synthetic and real-life data sets and show that SPEED lowers the MSE of policy evaluation compared to baseline methods. We discuss more related works and motivations in Appendix A.1.

2 PRELIMINARIES

We study the linear bandit setting where the expected reward for each action is assumed to be a linear function [Mason et al., 2021, Jamieson and Jain, 2022]. We define $[m] := [1, 2, \ldots, m]$. We denote the action space as $\mathcal{A}$ and $|\mathcal{A}| = A$. Actions are indexed by $a \in [A]$, and each action a is associated with a feature vector $\mathbf{x}(a) \in \mathbb{R}^d$ with dimension $d \ll A$. Denote by $\triangle(\mathcal{A})$ the probability simplex over the action space $\mathcal{A}$ and a policy $\pi \in \triangle(\mathcal{A})$ as a mapping $\pi : \mathcal{A} \to [0, 1]$ such that $\sum_a \pi(a) = 1$.

Data collection is performed over n rounds of action selection. Specifically, at each round $t \in [n]$, the selected action a_t yields a reward: $r_t = \mathbf{x}(a_t)^\top \boldsymbol{\theta}_* + \eta_t$, where $\boldsymbol{\theta}_* \in \mathbb{R}^d$ is the *unknown* reward parameter, and η_t is zero-mean noise with variance $\sigma^2(a_t)$ and we further assume that η_t is κ^2-subgaussian. We assume that for each action $a \in \mathcal{A}$ the variance $\sigma^2(a)$ has a lower-dimensional structure such that $\sigma^2(a) = \mathbf{x}(a)^\top \boldsymbol{\Sigma}_* \mathbf{x}(a)$ where $\boldsymbol{\Sigma}_* \in \mathbb{R}^{d\times d}$ is an *unknown* variance parameter. Observe that the variance depends on the action features, which is called the heteroscedastic noise model [Greene, 2002, Chaudhuri et al., 2017] which differs from the unknown time-dependent variance model of Zhang et al. [2021], Zhao et al. [2022]. Moreover,

Fontaine et al. [2021] do not consider structure in variances and Chaudhuri et al. [2017] only consider a special case of our setting where $\mathbf{\Sigma}_*$ is a rank-1 matrix. We also assume that the norms of the features are bounded such that $H_L^2 \leq \|\mathbf{x}(a)\|^2 \leq H_U^2$ for all $a \in \mathcal{A}$. In our heteroscedastic linear bandit setting selecting any action gives information about $\boldsymbol{\theta}_*$ and also gives information about the noise covariance matrix $\mathbf{\Sigma}_*$.

The value of a policy π is defined as $v(\pi) := \mathbb{E}[R_t]$ where the expectation is taken over $a_t \sim \pi, R_t \sim \mathbf{x}(a_t)^\top \boldsymbol{\theta}_* + \eta_t$. In the policy evaluation problem, we are given a fixed, target policy π and asked to estimate $v(\pi)$. Estimating $v(\pi)$ requires a dataset of actions and their associated rewards, $\mathcal{D} := \{(a_1, r_1, ..., a_n, r_n)\}$, which is collected by executing some policy. We refer to the policy that collects $\mathcal{D}$ as the *behavior policy*, denoted by $\mathbf{b} \in \triangle(\mathcal{A})$. We then define the value estimate of a policy π as Y_n, where n is the sample budget. The exact nature of the value estimate for the linear bandit setting will be made clear in Section 3.1. Our goal is to choose a behavior policy that minimizes the mean squared error (MSE) defined as $\mathbb{E}_{\mathcal{D}}[(Y_n - v(\pi))^2]$, where the expectation is over the collected data set $\mathcal{D}$.

We now state an assumption on the boundedness on the variance of each action $a \in [A]$. Let the singular value decomposition of $\mathbf{\Sigma}_*$ be $\mathbf{UDP}^\top$ with orthogonal matrices $\mathbf{U}, \mathbf{P}^\top$ and $\mathbf{D} = \text{diag}\,(\lambda_1, \ldots, \lambda_d)$ where $\{\lambda_i\}$ are singular values. It follows that $\sigma^2_{\min} \leq \sigma^2(a) \leq \sigma^2_{\max}$ where $\sigma^2_{\min} = \min_i |\lambda_i| H_L^2$ and $\sigma^2_{\max} = \max_i |\lambda_i| H_U^2$ (see Remark 1).

Assumption 1. *We assume that $\mathbf{\Sigma}_*$ has its minimum and maximum eigenvalues bounded such that for every action $a \in [A]$ the following holds $\sigma^2_{\min} \leq \sigma^2(a) \leq \sigma^2_{\max}$.*

3 OPTIMAL DESIGN FOR POLICY EVALUATION

In this section, we first discuss why following the target policy to take actions can lead to a poor estimation of the value of the policy. This discussion motivates how a different behavior policy can produce more accurate estimates of the target policy's value. After this motivation, we derive an expression for policy evaluation error in terms of the behavior sampling proportion $\mathbf{b} \in \triangle(\mathcal{A})$, target policy π, and action features $\mathbf{x}(a) \in \mathbb{R}^d$. We call the minimizer of this expression the "optimal design" [Pukelsheim, 2006] as it minimizes the mean squared error for policy evaluation. We then analyze the error incurred by an oracle that can compute and follow the optimal behavior policy through knowledge of problem-dependent parameters.

Motivating Example: Consider the linear bandit environment where $d = 2$ and $A = 100$ actions. Let one action be along the x-axis, one action along the y-axis, and 98 actions along the direction of $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Assume $\boldsymbol{\theta}_*$ is in the direction of x-axis (so action 1 is the optimal action). A similar canonical linear bandit setting has been studied by Fiez et al. [2019], Katz-Samuels et al. [2020]. Consider a target policy π such that $\pi(1) = 0.9$ and it distributes 0.1 probability equally on the remaining actions. In this case, just running the target policy π for n rounds leads to sampling uninformative actions for identifying $\boldsymbol{\theta}_*$. In fact, in our experiments, we show that the estimate $v(\pi)$ will be inaccurate compared to running the optimal behavior policy (called Oracle policy; see Figure 1 top-left).

Now suppose we divide the budget of n samples across the actions, and let $T_n(1), T_n(2), \ldots, T_n(A)$ be the number of samples allocated to actions $1, 2, \ldots, A$ at the end of n rounds. After observing n samples, let the *weighted* least square estimate (WLS) be:

$$\widehat{\boldsymbol{\theta}}_n := \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{n} \tfrac{1}{\sigma^2(a_t)} (r_t - \mathbf{x}(a_t)^\top \boldsymbol{\theta})^2 \qquad (1)$$

where a_t is the action sampled at round t and $\sigma^2(a_t)$ is the variance of action a_t. Also note that this is an unbiased estimator of $\boldsymbol{\theta}_*$ (see Remark 2). In a linear bandit, we can define the value estimate of a *target policy* as $Y_n := \sum_a \mathbf{w}(a)^\top \widehat{\boldsymbol{\theta}}_n$, where $\mathbf{w}(a) := \pi(a)\mathbf{x}(a)$ is the expected feature for each action $a \in \mathcal{A}$ under the target policy, and $\widehat{\boldsymbol{\theta}}_n$ is an unbiased estimate of $\boldsymbol{\theta}_*$ computed with n samples in $\mathcal{D}$. As $\widehat{\boldsymbol{\theta}}_n$ is an unbiased estimate, we have that $\mathbb{E}_{\mathcal{D}}[Y_n] = \sum_{a=1}^{A} \mathbf{w}(a)^\top \boldsymbol{\theta}_* = v(\pi)$. Since we have an unbiased estimator of $v(\pi)$, minimizing the MSE is equivalent to minimizing the variance, $\min \mathbb{E}_{\mathcal{D}}\big[\,(Y_n - \mathbb{E}[Y_n])^2\,\big] = \min \mathbb{E}_{\mathcal{D}}\big[\big(\sum_{a=1}^{A} \mathbf{w}(a)^\top (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*)\big)^2\big]$, where the minimization is with respect to the data distribution $\mathcal{D}$, which is determined by the behavior policy. In general, the behavior policy that minimizes the MSE may be different from the target policy. To identify this optimal behavior policy, following the optimal design literature [Pukelsheim, 2006, Fedorov, 2013] we define the design or information matrix $\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*} \in \mathbb{R}^{d \times d}$ w.r.t. each $\mathbf{b} \in \Delta(\mathcal{A})$ as

$$\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*} = \sum_{a \in \mathcal{A}} \mathbf{b}(a) \big(\tfrac{\mathbf{x}(a)}{\sigma(a)}\big) \big(\tfrac{\mathbf{x}(a)}{\sigma(a)}\big)^\top = \sum_{a \in \mathcal{A}} \mathbf{b}(a) \widetilde{\mathbf{x}}(a) \widetilde{\mathbf{x}}(a)^\top \quad (2)$$

where $\widetilde{\mathbf{x}}(a) = \mathbf{x}(a)/\sigma(a)$. Observe that our design matrix in (2) captures the information about the action features $\mathbf{x}(a)$, and variance $\sigma^2(a)$ and weights them by the sampling proportion $\mathbf{b}(a)$. Then in the following proposition, we exactly characterize the MSE with respect to the design matrix $\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}$, target policy π and action features $\mathbf{x}$. Moving forward, we will use the term *loss* interchangeably with MSE.

Proposition 1. *Let $\widehat{\boldsymbol{\theta}}_n$ be the Weighted Least Square (WLS) estimate (1) of $\boldsymbol{\theta}_*$ after observing n samples*

and define $\mathbf{w}(a) = \pi(a)\mathbf{x}(a)$. *Define the design matrix as* $\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}$ *(see (2)). Then the loss is given by*

$$\mathbb{E}_{\mathcal{D}}\Big[\Big(\sum_{a=1}^{A}\mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_*)\Big)^2\Big]=\underbrace{\frac{1}{n}\sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}\mathbf{w}(a')}_{:=\mathcal{L}_n(\pi,\mathbf{b},\mathbf{\Sigma}_*)}.$$

Proof (Overview) The key idea is to show that the linear model yields for each action $a \in [A], \widetilde{Y}_n(a) = \widetilde{\mathbf{x}}_n(a)^\top\boldsymbol{\theta}^\star + \widetilde{\eta}_n(a)$ where we define

$$\widetilde{Y}_n(a) = \sum_{i=1}^{T_n(a)}\frac{R_i(a)}{\sigma(a)\sqrt{T_n(a)}}, \widetilde{\mathbf{x}}_n(a) = \frac{\sqrt{T_n(a)}\mathbf{x}(a)}{\sigma(a)},$$

$$\widetilde{\eta}_n(a) = \sum_{i=1}^{T_n(a)}\frac{\eta_i(a)}{\sigma(a)\sqrt{T_n(a)}},$$

with $R_i(a)$ being the reward observed for action a taken for the i-th time, $\eta_i(a)$ being the corresponding noise, and $T_n(a)$ is the number of samples of action a. Next, observe that using the independent noise assumption, we have that $\mathbb{E}[\widetilde{\eta}_n(a)] = 0$ and $\mathbf{Var}[\widetilde{\eta}_n(a)] = 1$. Let $\mathbf{X} = \left(\widetilde{\mathbf{x}}_n(1)^\top, \cdots, \widetilde{\mathbf{x}}_n(A)^\top\right)^\top \in \mathbb{R}^{A\times d}$ be the induced feature matrix of the policy and $\mathbf{Y} = [\widetilde{Y}_n(1), \widetilde{Y}_n(2), \ldots, \widetilde{Y}_n(A)]^\top$. The above weighted least squares (WLS) problem has an optimal unbiased estimator $\widehat{\boldsymbol{\theta}}_n = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}$ [Fontaine et al., 2021]. Substituting the definition of $\widehat{\boldsymbol{\theta}}_n$ yields the desired expression of the loss as stated in the proposition. The detailed proof is given in Appendix A.3. ■

Observe that the loss in our setting depends on the inverse of the design matrix denoted by $\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}$, the target policy, as well as features of action pairs $(a, a') \in \mathcal{A}\times\mathcal{A}$. Hence, minimizing the loss is equivalent to minimizing the quantity $1/n(\sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}\mathbf{w}(a'))$. As this design is different than a number of existing notions of optimality such as D-, E-, T-, or G-optimality [Pukelsheim, 2006, Fedorov, 2013, Jamieson and Jain, 2022], we call this the *PE-Optimal design*. None of these previously proposed designs capture the objective of minimal MSE for policy evaluation. For example, G-optimality (as studied by [Katz-Samuels et al., 2020, Mason et al., 2021, Katz-Samuels et al., 2021, Mukherjee et al., 2023, 2024]) minimizes the worst-case error of $\max_{\mathbf{x}(a)}\mathbb{E}_{\mathcal{D}}[(\mathbf{x}(a)^\top(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_*))^2]$ by minimizing the quantity $\max_{\mathbf{x}(a)}\mathbf{x}(a)^\top\mathbf{A}_{\mathbf{b}}^{-1}\mathbf{x}(a)$ for homoscedastic noise. The E-optimal design minimizes $\max_{\|\mathbf{u}\|\le 1}\mathbb{E}_{\mathcal{D}}[(\mathbf{u}^\top(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_*))^2]$ by minimizing the minimum eigenvalue of the inverse of design matrix [Mukherjee et al., 2022b] and the A-optimal design minimizes $\mathbb{E}_{\mathcal{D}}[(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_*)^2]$ by minimizing the trace of the inverse of design matrix [Fontaine et al., 2021].

We now state a few more notations for ease of exposition. Using Proposition 1 we define the optimal behavior policy when the matrix $\mathbf{\Sigma}_*$ is known as:

$$\mathbf{b}_* := \arg\min_{\mathbf{b}}\mathcal{L}_n(\pi,\mathbf{b},\mathbf{\Sigma}_*), \tag{3}$$

where the loss $\mathcal{L}_n(\pi,\mathbf{b},\mathbf{\Sigma}_*)$ is defined in Proposition 1. We define the optimal loss (with knowledge of $\mathbf{\Sigma}_*$) as:

$$\mathcal{L}^*_n(\pi,\mathbf{b}_*,\mathbf{\Sigma}_*) = \min_{\mathbf{b}}\mathcal{L}_n(\pi,\mathbf{b},\mathbf{\Sigma}_*). \tag{4}$$

3.1 Computation of the optimal design $\mathbf{b}_*$

In this section, we digress a bit to discuss the computational aspect of $\mathcal{L}_n(\pi,\mathbf{b},\mathbf{\Sigma}_*)$. Since PE-Optimal design is a new type of design, the natural question to ask is *how to optimize this loss function w.r.t.* $\mathbf{b}$*?* We show in Proposition 2 that the loss $\mathcal{L}_n(\pi,\mathbf{b},\mathbf{\Sigma}_*)$ for any arbitrary design proportion $\mathbf{b} \in \triangle(\mathcal{A})$ is strictly convex with respect to the proportion $\mathbf{b}$. The proposition and its proof are given in Appendix A.4. Next in Proposition 3 we show that the gradient of the loss function is bounded. Due to space constraints, both propositions and their proofs are given in Appendix A.4 and Appendix A.5 respectively. We first state an assumption that the minimum eigenvalue satisfies $\lambda_{\min}\left(\sum_{a=1}^{A}\mathbf{w}(a)\mathbf{w}(a)^\top\right) > 0$, which is required for proving Proposition 3.

Assumption 2. ***(Distribution of*** π***)*** *We assume that the set of actions* a *such that* $\pi(a)>0$*, spans* $\mathbb{R}^d$ *and* $\mathbb{R}^{d\times d}$*.*

Note that this is a realistic and not a restrictive assumption, since if the target policy never takes an action that is needed to cover some dimension then we can avoid identifying $\boldsymbol{\theta}_*$ in that dimension. Using Proposition 2, 3 we can effectively solve the PE-Optimal design with gradient descent approaches [Lacoste-Julien and Jaggi, 2013, Berthet and Perchet, 2017]. We capture this convergence guarantee with the assumption of the existence of an approximation oracle.

Assumption 3. ***(Approximation Oracle)*** *We assume access to an approximation oracle. Given a convex loss function* $\mathcal{L}_n(\pi,\mathbf{b},\mathbf{\Sigma}_*)$ *with minimizer* $\mathbf{b}_*$*, the approximation oracle returns a proportion* $\widehat{\mathbf{b}}_* = \arg\min_{\mathbf{b}}\mathcal{L}_n(\pi,\mathbf{b},\mathbf{\Sigma}_*)$ *such that* $|\mathcal{L}_n(\pi,\widehat{\mathbf{b}}_*,\mathbf{\Sigma}_*) - \mathcal{L}_n(\pi,\mathbf{b}_*,\mathbf{\Sigma}_*)| \le \epsilon$*.*

Therefore from Proposition 2, and 3 and using Assumption 2, and 3 we can get a computationally efficient solution to $\min_{\mathbf{b}\in\Delta(\mathcal{A})}\mathcal{L}_n(\pi,\mathbf{b},\mathbf{\Sigma}_*)$.

3.2 Oracle Loss

Recall from Section 1, that our final goal is to control the regret (excess loss) of an agnostic algorithm that does not know $\mathbf{\Sigma}_*$, with respect to an oracle that already knows $\mathbf{\Sigma}_*$. Towards this goal, in this section,

we develop our theory for optimal data collection by considering an oracle for the heteroscedastic linear bandit setting. Specifically, we consider an oracle that has knowledge of $\mathbf{\Sigma}_*$ but does not know $\boldsymbol{\theta}_*$. With this knowledge, it can solve Equation (3) (Assumption 3) to determine the PE-Optimal design, $\mathbf{b}_*$, that minimizes the loss. The oracle takes actions in proportion $\mathbf{b}_*$ for n samples and then computes the WLS estimate $\widehat{\boldsymbol{\theta}}_n$ using $\mathbf{\Sigma}_*$. The following proposition then bounds the loss of the oracle after n samples.

Proposition 5. ***(Oracle Loss)*** *Let the oracle sample each action a for $\lceil n\mathbf{b}_*(a)\rceil$ times, where $\mathbf{b}_*$ is the solution to (3). Define $\lambda_1(\mathbf{V})$ as the maximum eigenvalue of $\mathbf{V} := \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top$. Then the loss satisfies* $\mathcal{L}_n^*(\pi, \mathbf{b}_*, \mathbf{\Sigma}_*) \leq O_{\kappa^2, H_U^2}{}^{1}\left(\frac{d\lambda_1(\mathbf{V})\log n}{n}\right) + O_{\kappa^2,H_U^2}\left(\frac{1}{n}\right).$

Proof (Overview) Note that the oracle knows the $\mathbf{\Sigma}_*$ and uses $\widehat{\boldsymbol{\theta}}_n$ in (1) to estimate $\boldsymbol{\theta}_*$. We use Corollary 1 to show that $\mathcal{L}_n(\pi, \mathbf{b}_*, \mathbf{\Sigma}_*) \leq \lambda_1(\mathbf{V})d$ where $\mathbf{V} = \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top$. The proof follows by showing that $(\sum_{a=1}^A \mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*))^2$ is a sub-exponential variable. Then using sub-exponential concentration inequality in Lemma 4 (Appendix A.2) and setting $\delta = O(1/n^2)$ we can bound the expected loss with high probability. The full proof is given in Appendix B.1. ■

Connection to prior work: Prior work has considered a similar oracle for the basic stochastic bandit setting, which is a special case of our setting with $\mathbf{x}(a)$ being a one-hot vector in $\mathbb{R}^A$. In this case, we can see that $\mathbf{b}_* = \arg\min_{\mathbf{b}} \sum_a \frac{\pi^2(a)\sigma^2(a)}{\lceil \mathbf{b}(a)n\rceil}$. This captures the optimal number of times the actions should be pulled weighted by the target policy and their variance. Solving for $\mathbf{b}_*$, we obtain $\mathbf{b}_*(a) \propto \pi^2(a)\sigma^2(a)$. This solution matches the optimal sampling proportion given by Antos et al. [2008], Carpentier and Munos [2011, 2012], Carpentier et al. [2015] for this special case. The loss in prior work decays at the rate of $\widetilde{O}^2(An^{-1})$ whereas the loss in Proposition 5 decreases at the rate of $\widetilde{O}(dn^{-1})$. Also note the loss in Proposition 5 scales with d instead of d^2 as the oracle knows the $\mathbf{\Sigma}_*$ and does not need to explore d^2 directions to estimate $\mathbf{\Sigma}_*$. So we obtain an equivalence between the PE-Optimal design and the solution from prior work in the basic bandit setting while considering a more general setting.

4 SPEED AND REGRET ANALYSIS

In this section, we first introduce an agnostic algorithm called SPEED for data collection that does not know $\mathbf{\Sigma}_*$, and then analyze its regret. Here, regret refers to the excess loss relative to the oracle that knows $\mathbf{\Sigma}_*$.

4.1 Details of Algorithm SPEED

In practice, $\mathbf{\Sigma}_*$ is unknown and so the oracle behavior policy cannot be directly computed. Instead, we first conduct a small amount of exploration to estimate $\mathbf{\Sigma}_*$ and then use the estimate in place of $\mathbf{\Sigma}_*$ in (2). Specifically, we define the forced exploration phase as the first Γ rounds in which the algorithm conducts exploration to estimate $\mathbf{\Sigma}_*$. To ensure adequate exploration, we first apply Principal Component Analysis (PCA) on the feature matrix $\mathbf{X}$ and choose the most significant d directions (directions having the highest variance). Then we choose one random action for each of these d significant directions and sample these actions uniform randomly for Γ rounds. Since the algorithm explores first and then uses the estimate to compute the PE-Optimal design, it can be viewed as an explore-then-commit algorithm [Rusmevichientong and Tsitsiklis, 2010, Lattimore and Szepesvári, 2020b]. As we consider a structured setting we call this algorithm **S**tructured **P**olicy **E**valuation **E**xperimental **D**esign (SPEED). After $\Gamma = \sqrt{n}$ rounds, SPEED estimates the covariance matrix $\widehat{\mathbf{\Sigma}}_\Gamma$ as follows:

$$\widehat{\mathbf{\Sigma}}_\Gamma = \min_{\mathbf{S}\in\mathbb{R}^{d\times d}} \sum_{t=1}^{\Gamma} \left[\langle \mathbf{x}(a_t)\mathbf{x}(a_t)^\top, \mathbf{S}\rangle - (r_t - \mathbf{x}(a_t)^\top \widehat{\boldsymbol{\theta}}_\Gamma)^2\right]^2 \quad (5)$$

where $\widehat{\boldsymbol{\theta}}_\Gamma$ is the ordinary least square (OLS) estimate of $\boldsymbol{\theta}_*$ using the data from the first Γ rounds. Note that the OLS estimate is given by $\widehat{\boldsymbol{\theta}}_\Gamma = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$, where $\mathbf{X} = \left(\mathbf{x}_1^\top, \cdots, \mathbf{x}_\Gamma^\top\right)^\top$ and $\mathbf{Y} = [r_1, \ldots, r_\Gamma]^\top$. A covariance estimation technique similar to (5) has been considered for the active regression setting though only for the case when $\mathbf{\Sigma}_*$ has rank 1 [Chaudhuri et al., 2017]. The estimate of the covariance matrix $\widehat{\mathbf{\Sigma}}_\Gamma$ is then fed to the oracle optimizer (Assumption 3) to compute the sampling proportion $\widehat{\mathbf{b}}$. Actions are chosen according to $\widehat{\mathbf{b}}$ for the remaining $n-\Gamma$ rounds and then the WLS estimate $\widehat{\boldsymbol{\theta}}_{n-\Gamma}$ is computed using $\widehat{\mathbf{\Sigma}}_\Gamma$ as the covariance matrix parameter (Equation (1)). Finally, SPEED outputs the dataset $\mathcal{D}$ to estimate the value of target policy π and $\widehat{\boldsymbol{\theta}}_{n-\Gamma}$. Full pseudocode is given in Algorithm 1.

4.2 Regret Analysis of SPEED

In this section, we first state our regret definition and then analyze the regret of the agnostic algorithm SPEED. As an agnostic algorithm, SPEED does not know the true covariance matrix $\mathbf{\Sigma}_*$ and must estimate the covariance matrix $\widehat{\mathbf{\Sigma}}_\Gamma$ after conducting exploration for Γ rounds. We define the loss of an algorithm after exploring for Γ rounds as the MSE of the resulting

[1]Here $O_{\kappa^2,H_U^2}()$ hides the sub-Gaussian factor κ^2 and upper bound H_U^2 on feature norm

[2]Here $\widetilde{O}$ hides logarithmic and problem dependent factors like $\sigma_{\min}^2, \kappa^2, H_U^2$.

Algorithm 1 Structured Policy Evaluation Experimental Design (SPEED)

1: **Input:** Action set $\mathcal{A}$, target policy π, budget n.
2: Conduct forced exploration for $\Gamma = \sqrt{n}$ rounds and estimate $\widehat{\mathbf{\Sigma}}_\Gamma$ using (5).
3: Let $\widehat{\mathbf{b}} \in \triangle(\mathcal{A})$ be the minimizer of $\mathcal{L}_n(\pi, \mathbf{b}, \widehat{\mathbf{\Sigma}}_\Gamma)$.
4: Pull each action a exactly $T_n(a) = \left\lfloor \widehat{\mathbf{b}}(a)(n-\Gamma) \right\rfloor$ times, and let $\mathcal{H}(a) := \{a, R_i(a)\}_{i=1}^{T_n(a)}$ be the corresponding data. Set $\mathcal{D} \leftarrow \cup_a \mathcal{H}(a)$.
5: Construct the weighted least squares estimator $\widehat{\boldsymbol{\theta}}_{n-\Gamma}$ using only the observations $\mathcal{D}$ from step 4.
6: **Output:** $\mathcal{D}$ and $\widehat{\boldsymbol{\theta}}_{n-\Gamma}$.

value estimate as follows:

$$\overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma) := \mathbb{E}_{\mathcal{D}}\big[\big(\sum_{a=1}^{A} \mathbf{w}(a)^\top (\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_*)\big)^2\big], \quad (6)$$

where $\widehat{\boldsymbol{\theta}}_{n-\Gamma}$ is the WLS estimate of $\boldsymbol{\theta}_*$ calculated from data of last $n-\Gamma$ rounds. We now define the regret for the agnostic algorithm with the estimated behavior policy $\widehat{\mathbf{b}}$ as

$$\mathcal{R}_n = \overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}_n^*(\pi, \mathbf{b}_*, \mathbf{\Sigma}_*). \quad (7)$$

where $\overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma)$ is the loss of the agnostic algorithm and $\mathcal{L}_n(\pi, \mathbf{b}_*, \mathbf{\Sigma}_*)$ is the oracle loss defined in (4). We now state the main theorem for the regret of SPEED.

Theorem 1. ***(Regret of Algorithm 1, informal)*** *Running Algorithm 1 with budget $n \geq O_{\kappa^2, H_U^2}(\frac{d^4 \sigma_{\max}^4 \log^2(A/\delta)}{\sigma_{\min}^4})$, the resulting regret satisfies $\mathcal{R}_n = O_{\kappa^2, H_U^2}\left(\frac{d^3 \sigma_{\max}^2 \log(n)}{\sigma_{\min}^2 n^{3/2}}\right)$.*

Discussion of Regret: Theorem 1 states that the regret of Algorithm 1 scales as $O_{\kappa^2, H_U^2}(d^3 \sigma_{\max}^2 \log(n)/n^{3/2})$ where d is the dimension of $\boldsymbol{\theta}_*$. Note that our regret bound depends on the underlying feature dimension d instead of actions A, and scales as $\widetilde{O}(d^3 n^{-3/2})$ which gives a positive answer to the main question of whether such a result is possible. In comparison to earlier work, when $d^3 < A$, we have a tighter bound than that given by Carpentier and Munos [2011]. Furthermore, the results of Carpentier and Munos [2011, 2012], Carpentier et al. [2015] are for the standard multi-armed bandit setting and cannot be easily extended to incorporate structure in the linear bandit setting. Our new bound also improves upon the A-optimal design method given by Fontaine et al. [2021], as their regret depends on the number of actions A and scales as $O(\frac{A \log n}{n^{3/2}})$.

Proof (Overview) of Theorem 1: We now outline the key steps for proving Theorem 1.

Step 1 (Regret Decomposition): We first decompose the regret $\mathcal{R}_n = \overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}_n^*(\pi, \mathbf{b}_*, \mathbf{\Sigma}_*)$. Recall that $\mathbf{b}_* \in \triangle(\mathcal{A})$ is the optimal design in (3) and $\widehat{\mathbf{b}} \in \triangle(\mathcal{A})$ is the design followed by SPEED. However, we cannot directly go after the loss $\overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma)$ as it does not admit a simple structure like $\mathcal{L}_n^*(\pi, \mathbf{b}_*, \mathbf{\Sigma}_*)$. Rather we establish an upper bound on the loss $\overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma)$, given by $\mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}_*, \widehat{\mathbf{\Sigma}}_\Gamma)$ (defined formally in (9)). Consequently, we can decompose the regret $\mathcal{R}_n$ into three parts as follows:

$$\begin{aligned}\mathcal{R}_n &\overset{(a)}{\leq} \underbrace{\mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}_*, \widehat{\mathbf{\Sigma}}_\Gamma)}_{\textbf{Approximation error}} \\ &+ \underbrace{\mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}_*, \widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}_n(\pi, \mathbf{b}_*, \widehat{\mathbf{\Sigma}}_\Gamma)}_{\textbf{Comparing two different loss}} \\ &+ \underbrace{\mathcal{L}_n(\pi, \mathbf{b}_*, \widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}_n^*(\pi, \mathbf{b}_*, \mathbf{\Sigma}_*)}_{\textbf{Estimation error of } \mathbf{\Sigma}_*}. \end{aligned} \quad (8)$$

where (a) follows as we show that

$$\begin{aligned}&\overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma) = \mathbb{E}\left[\big(\sum_{a=1}^{A} \mathbf{w}(a)^\top (\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_*)\big)^2\right] \\ &\leq \tfrac{1}{n-\Gamma}\left(1 + \tfrac{2Cd^2 \log(A/\delta)}{\sigma_{\min}^2 \Gamma}\right) \sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}^{-1}_{\widehat{\mathbf{b}}_*, \widehat{\mathbf{\Sigma}}_\Gamma} \mathbf{w}(a') \\ &:= \mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}_*, \widehat{\mathbf{\Sigma}}_\Gamma), \end{aligned} \quad (9)$$

where $C > 0$ is a constant. Note that the inequality above is shown in Proposition 6 which we discuss in depth in step 2. Finally note that $\widehat{\mathbf{b}}_*$ is the empirical PE-Optimal design returned by the approximator after it is supplied with $\widehat{\mathbf{\Sigma}}_\Gamma$.

Step 2 (Bounding the loss $\overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma)$): In this step we discuss how to upper bound the agnostic loss $\overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma)$ with $\mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}_*, \widehat{\mathbf{\Sigma}}_\Gamma)$ as defined in (9).

We first state a concentration lemma that is key to proving this upper bound. This lemma is novel for our proof because we estimate the underlying covariance matrix $\mathbf{\Sigma}_*$ using OLS estimator for Γ rounds. We then use the estimation $\widehat{\mathbf{\Sigma}}_\Gamma$ in the WLS estimator. For our lemma, we first define the variance concentration good event under Γ rounds of forced exploration as:

$$\xi_\delta^{var}(\Gamma) := \left\{\forall a, \left|\mathbf{x}(a)^\top (\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*)\mathbf{x}(a)\right| < \frac{2Cd^2 \sigma_{\max}^2 \log(A/\delta)}{\Gamma}\right\} \quad (10)$$

Lemma 1. ***(OLS-WLS Concentration Lemma)*** *After Γ samples of exploration, we can show that $\mathbb{P}(\xi_\delta^{var}(\Gamma)) \geq 1 - 8\delta$, where $C > 0$ is a constant.*

Proof (Overview) of Lemma 1: Note that we construct an initial estimate $\widehat{\boldsymbol{\theta}}_\Gamma$ of $\boldsymbol{\theta}_*$ using OLS estimate based on the first Γ rounds of data $\{a_t, r_t\}_{t=1}^{\Gamma}$.

Let the feature of a_t be $\mathbf{x}_t$ and the squared residual $y_t := (\mathbf{x}_t^\top \widehat{\boldsymbol{\theta}}_\Gamma - r_t)^2$. Recall that SPEED estimates $\boldsymbol{\Sigma}_*$ via $\min_{\mathbf{S}\in\mathbb{R}^{d\times d}} \sum_{t=1}^{\Gamma}(\langle \mathbf{x}_t\mathbf{x}_t^\top, \mathbf{S}\rangle - y_t)^2$. Let $\zeta_\Gamma := \widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*$. Then we can show that $y_t = \mathbf{x}_t^\top \boldsymbol{\Sigma}_* \mathbf{x}_t + \epsilon_t$ and the noise ϵ_t can be bounded by

$$\epsilon_t = \underbrace{\eta_t^2 - \mathbb{E}[\eta_t^2]}_{\textbf{Part A}} + \underbrace{2\eta_t \mathbf{x}_t^\top \zeta_\Gamma}_{\textbf{Part B}} + \underbrace{\left(\mathbf{x}_t^\top \zeta_\Gamma\right)^2}_{\textbf{Part C}}.$$

For the part A, observe that η_t^2 is a sub-exponential random variable as $\eta_t \sim \mathcal{SG}(0, \mathbf{x}_t^\top \boldsymbol{\Sigma}_* \mathbf{x}_t)$. Hence we can use sub-exponential concentration inequality from Lemma 4 (Appendix A.2) to bound it. For part C first recall that $\zeta_\Gamma := \widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*$ and we use Lemma 5 (Appendix A.2) to bound it. Finally, for part B, we can decompose $2\eta_t \mathbf{x}_t^\top \zeta_\Gamma \le 2\eta_t^2 + \frac{1}{2}(\mathbf{x}_t^\top \zeta_\Gamma)^2$. Then using the same technique for parts A and C we bound the total deviation for part B. Combining the three parts gives the desired concentration inequality. The proof is in Appendix B.2. ■

Lemma 1 directly leads to Corollary 3 (Appendix B.4) which shows that for $n \ge 16C^2d^4\log^2(A/\delta)/\sigma_{\min}^4$ we have that $\overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\Sigma}}_\Gamma) \le \mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\Sigma}}_\Gamma)$. Compared to earlier work, Fontaine et al. [2021] does not require this approach as the variances of each action lack a common structure. Similarly, this approach differs from the time-dependent variance model of Zhang et al. [2021], Zhao et al. [2022].

Step 3 (Bounding the approximation error and comparing two different losses): For the approximation error in (8) we need access to an optimization oracle that gives ϵ approximation error (Assumption 3). Then setting $\epsilon = \frac{1}{\sqrt{n}}$ we have that the estimation error is upper bounded by $n^{-3/2}$. For comparing the two different losses in (8), we use their definition of to bound it as $O_{\kappa^2, H_u^2}(\frac{d^2\log(A/\delta)}{n^{3/2}})$ as shown in (29) in Appendix B.3.

Step 4 (Bounding Estimation Error): Now observe that the third quantity in (8) (estimation error of $\boldsymbol{\Sigma}_*$) contains $\mathcal{L}_n(\pi, \mathbf{b}_*, \widehat{\boldsymbol{\Sigma}}_\Gamma)$ that depends on the design matrix $\mathbf{A}^{-1}_{\mathbf{b}_*, \widehat{\boldsymbol{\Sigma}}_\Gamma}$ which in turn depends on the estimation of $\widehat{\boldsymbol{\Sigma}}_\Gamma$. Similarly $\mathcal{L}_n(\pi, \mathbf{b}_*, \boldsymbol{\Sigma}_*)$ in the third quantity depends on the design matrix $\mathbf{A}^{-1}_{\mathbf{b}_*, \boldsymbol{\Sigma}_*}$ which in turn depends on the true $\boldsymbol{\Sigma}_*$. Hence, we now bound the concentration of the loss under $\mathbf{A}^{-1}_{\mathbf{b}_*, \widehat{\boldsymbol{\Sigma}}_\Gamma}$ against the design matrix $\mathbf{A}^{-1}_{\mathbf{b}_*, \boldsymbol{\Sigma}_*}$ in the following lemma.

Lemma 2. ***(Concentration of the design matrix)*** *Let $\widehat{\boldsymbol{\Sigma}}_\Gamma$ be the empirical estimate of $\boldsymbol{\Sigma}_*$, and $\mathbf{V} = \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top$. For any arbitrary proportion $\mathbf{b}$, with probability at least $(1-\delta)$, we have the following:*

$$\left|\sum_{a,a'} \mathbf{w}(a)^\top (\mathbf{A}^{-1}_{\mathbf{b}_*, \widehat{\boldsymbol{\Sigma}}_\Gamma} - \mathbf{A}^{-1}_{\mathbf{b}_*, \boldsymbol{\Sigma}_*}) \mathbf{w}(a')\right| \le \frac{2CB^* d^3 \sigma_{\max}^2 \log(A/\delta)}{\Gamma},$$

where B^* is a problem-dependent quantity and $C > 0$ is a universal constant.

Proof (Overview) of Lemma 2: We can upper bound $|\sum_{a,a'} \mathbf{w}(a)^\top (\mathbf{A}^{-1}_{\mathbf{b}_*, \widehat{\boldsymbol{\Sigma}}_\Gamma} - \mathbf{A}^{-1}_{\mathbf{b}_*, \boldsymbol{\Sigma}_*}) \mathbf{w}(a')| \le \|\mathbf{u}\| \underbrace{\left\|\mathbf{A}_{\mathbf{b}_*, \boldsymbol{\Sigma}_*} - \mathbf{A}_{\mathbf{b}_*, \widehat{\boldsymbol{\Sigma}}_\Gamma}\right\|}_{\Delta} \|\mathbf{v}\|$ where, $\|\mathbf{u}\| = \|\mathbf{A}^{-1}_{\mathbf{b}_*, \boldsymbol{\Sigma}_*}\mathbf{w}\|$ and $\|\mathbf{v}\| = \|\mathbf{A}^{-1}_{\mathbf{b}_*, \widehat{\boldsymbol{\Sigma}}_\Gamma}\mathbf{w}\|$. First, observe that $\|\mathbf{u}\|$ is a problem-dependent quantity. Then to bound Δ we use the Lemma 1 on the concentration of $\widehat{\sigma}_\Gamma^2(a)$. Finally to bound $\|\mathbf{v}\|$ we need to bound $\widehat{\sigma}_\Gamma^2(a) \le \sigma^2(a) + \frac{2Cd^2\sigma_{\max}^2\log(A/\delta)}{\Gamma}$ where $\widehat{\sigma}_\Gamma^2(a)$ is the empirical variance of $\sigma^2(a)$. Combining everything yields the desired result. The proof is in Appendix B.4 ■

One of our key technical contributions in Lemma 2 is to show that the difference between the two losses $\mathcal{L}_n(\pi, \mathbf{b}_*, \widehat{\boldsymbol{\Sigma}}_\Gamma)$, and $\mathcal{L}_n(\pi, \mathbf{b}_*, \boldsymbol{\Sigma}_*)$ scales with d^3 instead of the number of actions A. In contrast to prior work, a similar loss concentration in Fontaine et al. [2021] scales with A. Now using Lemma 2, setting the exploration factor $\Gamma = \sqrt{n}$, and $\delta = \frac{1}{n}$ we can show that the estimation error is upper bounded by $\frac{B^* C d^3 \sigma_{\max}^2 \log(n)}{\sigma_{\min}^2 n^{3/2}} + \frac{d^2}{n^2}\mathbf{Tr}(\sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top)$. Combining steps 1 – 4 we have the regret of SPEED as $O_{\kappa^2, H_U^2}(\frac{B^* d^3 \sigma_{\max}^2 \log(n)}{\sigma_{\min}^2 n^{3/2}})$. The full proof of Theorem 1 is in Appendix B.5. ■.

4.3 Lower Bound

Theorem 1 upper bounds the regret of our agnostic algorithm SPEED compared to an oracle algorithm with knowledge of $\boldsymbol{\Sigma}_*$. To quantify the tightness of our upper bound, we now turn to the question of whether we can lower bound the regret for any behavior policy learning algorithm. For our final theoretical result, we consider a slightly different notion of regret: $\mathcal{R}'_n := \mathcal{L}_n(\pi, \widehat{\mathbf{b}}, \boldsymbol{\Sigma}_*) - \mathcal{L}_n(\pi, \mathbf{b}_*, \boldsymbol{\Sigma}_*)$. This notion of regret captures how sub-optimal the estimated $\widehat{\mathbf{b}}$ is compared to $\mathbf{b}_*$, *without* additional error incurred by using an estimate of $\boldsymbol{\Sigma}_*$ in the WLS estimator. We conjecture that $\mathcal{R}'_n$ is indeed a lower bound to $\mathcal{R}_n$ as we have established in Proposition 1 that the minimum variance estimator is the WLS estimator using $\boldsymbol{\Sigma}_*$. Intuitively, $\mathcal{L}_n(\pi, \widehat{\mathbf{b}}, \boldsymbol{\Sigma}_*)$ is a lower bound to $\overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\Sigma}}_\Gamma)$ as estimation error will likely increase when using $\widehat{\boldsymbol{\Sigma}}_\Gamma$ in place of $\boldsymbol{\Sigma}_*$ in the WLS estimator. We leave proving that $\mathcal{R}'_n$ is a lower bound to $\mathcal{R}_n$ to future work.

Theorem 2. ***(Lower Bound)*** *Let* $|\mathbf{\Theta}|=2^d$, $\boldsymbol{\theta}_* \in \mathbf{\Theta}$. *Then any arbitrary* δ*-PAC policy following the design* $\mathbf{b} \in \triangle(\mathcal{A})$ *satisfies* $\mathcal{R}'_n = \mathcal{L}_n(\pi, \mathbf{b}, \mathbf{\Sigma}_*) - \mathcal{L}_n(\pi, \mathbf{b}_*, \mathbf{\Sigma}_*) \geq \Omega\left(\frac{d^2 \lambda_d(\mathbf{V}) \log(n)}{n^{3/2}}\right)$ *for the environment specified in* (33).

Proof (Overview:) The proof follows the change of measure argument [Lattimore and Szepesvári, 2020b] and we follow the proof technique of Huang et al. [2017], Mukherjee et al. [2022b]. We reduce the policy evaluation problem to the hypothesis testing setting and state a worst-case environment as in (33). We then show that the regret of any δ-PAC algorithm against an oracle in this environment must scale as $\Omega(\log n/n^{3/2})$. The proof is given in Appendix C. ■

From the above result, the upper bound of SPEED regret $\mathcal{R}_n$ matches the lower bound of regret $\mathcal{R}'_n$ in n but suffers an additional factor of d.

5 EXPERIMENTS

We now conduct numerical experiments to show that SPEED decreases MSE faster than other baselines. These experiments complement our theoretical analysis as they do not have the conditions on budget n required in Theorem 1. Thus, our experimental analysis will show that the theoretically motivated SPEED algorithm still provides benefit even outside of the sample regime considered in theory. As baselines, we compare against On-policy, Oracle, A-Optimal [Fontaine et al., 2021], and G-Optimal [Wan et al., 2022]. The On-policy algorithm simply runs the target policy to collect data, whereas the Oracle (as discussed in Section 3) samples according to the optimal $\mathbf{b}_*$. Of existing optimal design methods, A-Optimal , and G-Optimal are the closest in relation to our work. We experiment with A-Optimal design because this criterion minimizes the average variance of the estimates of the regression coefficients and is most closely aligned with our goal. The work of Wan et al. [2022] considers data collection under safety constraints using Inverse Propensity Weighting. In our unconstrained policy evaluation setting their approach boils down to just G-optimal design. Further experimental details are in Appendix D.

Unit Ball: We perform this experiment on a set of 5 actions that are arranged in a unit ball in $\mathbb{R}^2$ to show that SPEED allocates proportion to the most informative action (weighted by their variance). Figure 1 (Top Left) shows that SPEED reduces the MSE faster than On-policy, G-Optimal , and A-Optimal . We also include Oracle in this setting to show how quickly SPEED converges to it. However, for settings based on real-life data, we do not have such oracles.

Movielens Dataset: Consider a startup that wants

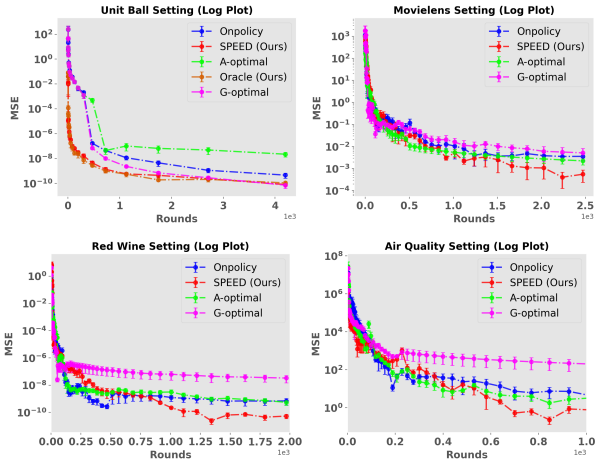

Figure 1: (Top-left) MSE plot for the Unit ball. (Top-right) MSE plot for the Movielens dataset. (Bottom-left) MSE plot for Red Wine Quality dataset. (Bottom-right) MSE plot for Air Quality dataset. The vertical axis gives MSE and the horizontal axis is the number of rounds. The vertical axis is log-scaled and confidence bars show one standard error.

to recommend movies to users based on their ratings. They have access to a target policy and want to evaluate it on a limited informative dataset before deploying it for full public use. We use real-world Movielens 1M dataset [Lam and Herlocker, 2016] datasets for this experiment. We apply low-rank factorization to the rating matrix to obtain 5-dimensional representations of users and movies. We then fit a weighted least square estimate of $\boldsymbol{\theta}_*$ and $\mathbf{\Sigma}_*$. We generate the reward using this $\boldsymbol{\theta}_*$ and $\mathbf{\Sigma}_*$. Then we use SPEED and other baselines to generate the small informative dataset to evaluate the target policy and this experiment is shown in Figure 1 (Top Right). SPEED initially conducts forced exploration to estimate $\boldsymbol{\theta}_*$, $\mathbf{\Sigma}_*$ and incurs slightly higher MSE but the MSE decreases faster than other baselines as the number of rounds increases.

Red Wine Quality: Consider an online wine company that wants to recommend wines to users and wants to evaluate a target policy before full deployment. We perform this experiment on real-world dataset *Red Wine Quality* from UCI datasets [Cortez et al., 2009]. The dataset consists of 1600 samples (actions) of red wine with each sample a having feature $\mathbf{x}(a) \in \mathbb{R}^{11}$ and their ratings. We fit a weighted least square estimate to the original dataset and get an estimate of $\boldsymbol{\theta}_*$ and $\mathbf{\Sigma}_*$. Then we use SPEED to generate the informative dataset to evaluate the target policy. Figure 2 (Bottom-left) shows SPEED outperforming other baselines as horizon increases.

Air Quality: We now consider a setting where a government agency wants to record air quality and

notify the public. However, it wants to evaluate a target policy on a limited informative dataset before full deployment. We perform this experiment on real-world dataset *Air-Quality* from UCI datasets [De Vito et al., 2008]. The dataset consists of 1500 samples (actions) with each sample a having feature $\mathbf{x}(a) \in \mathbb{R}^6$ and their air quality value. Similar to red wine dataset we estimate of $\boldsymbol{\theta}_*$ and $\boldsymbol{\Sigma}_*$. Then we use SPEED and other baselines (which do not know $\boldsymbol{\theta}_*$ and $\boldsymbol{\Sigma}_*$) to generate the informative dataset to evaluate the target policy and this experiment is shown in Figure 1 (Bottom-right). Observe that SPEED's MSE decreases faster than other baselines as the number of rounds increases.

6 CONCLUSIONS AND FUTURE DIRECTIONS

We proposed SPEED for optimal data collection for policy evaluation in linear bandits with heteroscedastic reward noise. We formulated a novel optimal design problem, PE-Optimal design, for which the optimal behavior policy is the solution that will produce minimal MSE policy evaluation when using a weighted least square estimate of the hidden reward parameters $\boldsymbol{\theta}_*$ and $\boldsymbol{\Sigma}_*$. We showed the regret of SPEED degrades at the rate of $\widetilde{O}(d^3 n^{-3/2})$ and matches the lower bound of $\widetilde{O}(d^2 n^{-3/2})$ except a factor of d. In contrast the On-policy suffers a regret of $\widetilde{O}(n^{-1})$ [Carpentier et al., 2015]. We showed empirically that our design outperforms other optimal designs. In future work, we intend to extend the result to a more general class of hard problems such as collecting data to minimize the MSE of multiple target policies.

Acknowledgements

We would like to thank all the reviewers for their helpful feedback. Q. Xie is supported in part by NSF grant CNS-1955997. J. Hanna is supported in part by American Family Insurance through a research partnership with the University of Wisconsin-Madison's Data Science Institute. We also thank Justin Weltz, Blake Mason, and Lalit Jain for pointing out errors in the previous version of this paper and helping to improve the draft.

References

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

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

András Antos, Varun Grover, and Csaba Szepesvári. Active learning in multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 287–302. Springer, 2008.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2):235–256, May 2002. ISSN 1573-0565. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Quentin Berthet and Vianney Perchet. Fast rates for bandit optimization with upper-confidence frank-wolfe. *Advances in Neural Information Processing Systems*, 30, 2017.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

Guillaume Bouchard, Théo Trouillon, Julien Perez, and Adrien Gaidon. Online learning to sample. *arXiv preprint arXiv:1506.09016*, 2016.

Hengrui Cai, Chengchun Shi, Rui Song, and Wenbin Lu. Deep jump learning for off-policy evaluation in continuous treatment settings. *Advances in Neural Information Processing Systems*, 34:15285–15300, 2021.

Alexandra Carpentier and Rémi Munos. Finite-time analysis of stratified sampling for monte carlo. In *NIPS-Twenty-Fifth Annual Conference on Neural Information Processing Systems*, 2011.

Alexandra Carpentier and Rémi Munos. Minimax number of strata for online stratified sampling given noisy samples. In *International Conference on Algorithmic Learning Theory*, pages 229–244. Springer, 2012.

Alexandra Carpentier, Remi Munos, and András Antos. Adaptive strategy for stratified monte carlo sampling. *J. Mach. Learn. Res.*, 16:2231–2271, 2015.

Kamalika Chaudhuri, Prateek Jain, and Nagarajan Natarajan. Active heteroscedastic regression. In *International Conference on Machine Learning*, pages 694–702. PMLR, 2017.

Kamil Ciosek and Shimon Whiteson. OFFER: Off-environment reinforcement learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

Nicholas Corrado and Josiah P. Hanna. On-policy policy gradient reinforcement learning without on-

policy sampling. In *Arxiv Pre-print*, September 2023. URL https://arxiv.org/abs/2311.08290.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.

Saverio De Vito, Ettore Massera, Marco Piga, Luca Martinotto, and Girolamo Di Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2):750–757, 2008.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. 2014.

Yuguang Fang, Kenneth A Loparo, and Xiangbo Feng. Inequalities for the trace of matrix product. *IEEE Transactions on Automatic Control*, 39(12):2489–2490, 1994.

Valerii Vadimovich Fedorov. *Theory of optimal experiments*. Elsevier, 2013.

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

Xavier Fontaine, Pierre Perrault, Michal Valko, and Vianney Perchet. Online a-optimal design and active linear regression. In *International Conference on Machine Learning*, pages 3374–3383. PMLR, 2021.

William H Greene. 000. econometric analysis, 2002.

Josiah P. Hanna, Philip S. Thomas, Peter Stone, and Scott Niekum. Data-Efficient Policy Evaluation Through Behavior Policy Search. *arXiv:1706.03469 [cs]*, June 2017. URL http://arxiv.org/abs/1706.03469. arXiv: 1706.03469.

Jean Honorio and Tommi Jaakkola. Tight bounds for the expected risk of linear classifiers and pac-bayes finite-sample guarantees. In *Artificial Intelligence and Statistics*, pages 384–392. PMLR, 2014.

Ruitong Huang, Mohammad M. Ajallooeian, Csaba Szepesvári, and Martin Müller. Structured best arm identification with fixed confidence. In Steve Hanneke and Lev Reyzin, editors, *International Conference on Algorithmic Learning Theory, ALT 2017, 15-17 October 2017, Kyoto University, Kyoto, Japan*, volume 76 of *Proceedings of Machine Learning Research*, pages 593–616. PMLR, 2017. URL http://proceedings.mlr.press/v76/huang17a.html.

Kevin Jamieson and Lalit Jain. Interactive machine learning. 2022.

Nathan Kallus, Yuta Saito, and Masatoshi Uehara. Optimal off-policy evaluation from multiple logging policies. In *International Conference on Machine Learning*, pages 5247–5256. PMLR, 2021.

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

Julian Katz-Samuels, Jifan Zhang, Lalit Jain, and Kevin Jamieson. Improved algorithms for agnostic pool-based active classification. In *International Conference on Machine Learning*, pages 5334–5344. PMLR, 2021.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR, 2018.

Ron Kohavi and Roger Longbotham. Online controlled experiments and a/b testing. *Encyclopedia of machine learning and data mining*, 7(8):922–929, 2017.

Simon Lacoste-Julien and Martin Jaggi. An affine invariant linear convergence analysis for frank-wolfe algorithms. *arXiv preprint arXiv:1312.7864*, 2013.

T. L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, March 1985. ISSN 0196-8858. doi: 10.1016/0196-8858(85)90002-8. URL https://www.sciencedirect.com/science/article/pii/0196885885900028.

Shyong Lam and Jon Herlocker. MovieLens Dataset. http://grouplens.org/datasets/movielens/, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020a.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020b.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011.

Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pages 608–616. PMLR, 2015.

Ting Li, Chengchun Shi, Jianing Wang, Fan Zhou, et al. Optimal treatment allocation for efficient policy evaluation in sequential decision making. *Advances in Neural Information Processing Systems*, 36, 2024.

Blake Mason, Romain Camilleri, Subhojyoti Mukherjee, Kevin Jamieson, Robert Nowak, and Lalit Jain. Nearly optimal algorithms for level set estimation. *arXiv preprint arXiv:2111.01768*, 2021.

Subhojyoti Mukherjee, Josiah P Hanna, and Robert D Nowak. Revar: Strengthening policy evaluation via reduced variance sampling. In *Uncertainty in Artificial Intelligence*, pages 1413–1422. PMLR, 2022a.

Subhojyoti Mukherjee, Ardhendu S Tripathy, and Robert Nowak. Chernoff sampling for active testing and extension to active regression. In *International Conference on Artificial Intelligence and Statistics*, pages 7384–7432. PMLR, 2022b.

Subhojyoti Mukherjee, Ruihao Zhu, and Branislav Kveton. Efficient and interpretable bandit algorithms. *arXiv preprint arXiv:2310.14751*, 2023.

Subhojyoti Mukherjee, Qiaomin Xie, Josiah Hanna, and Robert Nowak. Multi-task representation learning for pure exploration in bilinear bandits. *Advances in Neural Information Processing Systems*, 36, 2024.

Harrie Oosterhuis and Maarten de Rijke. Taking the Counterfactual Online: Efficient and Unbiased Online Evaluation for Ranking. *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 137–144, September 2020. doi: 10.1145/3409256.3409820. URL http://arxiv.org/abs/2007.12719. arXiv: 2007.12719.

Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 813(814):46, 2015.

Carlos Riquelme, Mohammad Ghavamzadeh, and Alessandro Lazaric. Active learning for accurate estimation of linear models. In *International Conference on Machine Learning*, pages 2931–2939. PMLR, 2017.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, December 1933. ISSN 0006-3444. doi: 10.1093/biomet/25.3-4.285. URL https://doi.org/10.1093/biomet/25.3-4.285.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

Aaron David Tucker and Thorsten Joachims. Variance-Optimal Augmentation Logging for Counterfactual Evaluation in Contextual Bandits. *arXiv:2202.01721 [cs]*, February 2022. URL http://arxiv.org/abs/2202.01721. arXiv: 2202.01721.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Runzhe Wan, Branislav Kveton, and Rui Song. Safe exploration for efficient policy evaluation and comparison. *arXiv preprint arXiv:2202.13234*, 2022.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

P Whittle. A multivariate generalization of tchebichev's inequality. *The Quarterly Journal of Mathematics*, 9(1):232–240, 1958.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems*, 34:4342–4355, 2021.

Heyang Zhao, Dongruo Zhou, Jiafan He, and Quanquan Gu. Bandit learning with general function classes: Heteroscedastic noise and variance-dependent regret bounds. *arXiv preprint arXiv:2202.13603*, 2022.

Rujie Zhong, Duohan Zhang, Lukas Schäfer, Stefano V. Albrecht, and Josiah P. Hanna. Robust on-policy sampling for data-efficient policy evaluation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, December 2022.

Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *arXiv preprint arXiv:2205.11507*, 2022.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, and Michael R Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal*

of the American Statistical Association, 112(517): 169–187, 2017.

Ruihao Zhu and Branislav Kveton. Safe data collection for offline and online policy learning. *arXiv preprint arXiv:2111.04835*, 2021.

Ruihao Zhu and Branislav Kveton. Safe optimal design with applications in off-policy learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2436–2447. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/zhu22a.html.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] We present the model and assumptions in Section 2. The algorithms are presented in Section 3 and Section 4.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] Results are presented in Section 3 and Section 4.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] In supplementary material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Section 2 for the details.
   (b) Complete proofs of all theoretical results. [Yes] Proofs outlines are provided in the main paper, with all detailed proofs are deferred to the appendix.
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] In the supplementary material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] In the supplementary material.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] In the supplementary material.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

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

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

A APPENDIX

A.1 Related Works and Motivations

Our work is most closely related to existing work on data collection for policy evaluation. Perhaps the most natural choice of behavior policy is to simply run the target policy, i.e., on-policy data collection [Sutton and Barto, 2018]. The works in adaptive Monte Carlo for bandits [Oosterhuis and de Rijke, 2020, Tucker and Joachims, 2022] and MDPs [Hanna et al., 2017, Ciosek and Whiteson, 2017, Bouchard et al., 2016, Zhong et al., 2022, Corrado and Hanna, 2023] have shown how to lower the variance of Monte Carlo estimation through the choice of behavior policy. In contrast to these works, we consider estimating $v(\pi)$ by estimating the reward distributions rather than using Monte Carlo estimation. Such *certainty-equivalence* estimators take advantage of the setting's structure and are thus typically of lower variance than Monte Carlo estimators [Sutton and Barto, 2018]. The work of Wan et al. [2022] studies a different estimator for reducing the variance of the importance sampling in constrained MDP setting whereas we study certainty equivalence estimator. Another set of work has studied sample allocation for stratified Monte Carlo estimators – a problem that is formally equivalent to behavior policy selection for policy evaluation in the bandit setting with linearly independent arms [Antos et al., 2008, Carpentier et al., 2015]. This line of work was recently extended to tabular, tree-structured MDPs by Mukherjee et al. [2022a]. In contrast, we consider the structured linear bandit setting which incorporates generalization across actions. Li et al. [2024] use A-optimal design to find an optimal behavior policy for the doubly robust estimator. Their focus is different though as they consider tabular MDPs rather than linear heteroscedastic bandits.

Our work is closely related to optimal experimental design and active learning literature. We formulate determining the optimal behavior policy in the bandit setting as an optimal design problem. In contrast to prior work, we introduce a new type of optimality that is tailored to the policy evaluation problem. We are also, to the best of our knowledge, the first to consider both heteroscedastic noise and weighted least squares estimators in formulating our design. The heteroscedastic noise model and weighted least squares estimator have been considered by Chaudhuri et al. [2017] in the active learning literature and in linear bandit setting by Kirschner and Krause [2018] using information directed sampling. In contrast to these works (and the active learning setting in general), we aim to minimize the weighted error $\sum_{a\in\mathbf{A}} \pi(a)\mathbf{x}(a)^\top(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})^2$ whereas in the active learning setting the goal is to minimize $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\|^2$ which results in A-optimal design [Fontaine et al., 2021, Pukelsheim, 2006]. Moreover the regret bounds in Fontaine et al. [2021] holds for $d = |\mathcal{A}|$. Riquelme et al. [2017] extends the results of Carpentier and Munos [2011] to a different linear regression setting than ours but under the homoscedastic noise model.

Data collection for policy evaluation is also related to the problem of exploration for policy learning in MDPs or best-arm identification in bandits. In those contexts, the aim of exploration is to find the optimal policy and the exploration-exploitation trade-off describes the tension between reducing uncertainty and focusing on known promising actions. In bandits, the exploration-exploitation trade-off is often navigated under the "**O**ptimism in the **F**ace of **U**ncertainty" principle using techniques such as UCB [Lai and Robbins, 1985, Auer et al., 2002, Abbasi-Yadkori et al., 2011] or Thompson Sampling [Thompson, 1933, Agrawal and Goyal, 2012]. In contrast to the standard exploration problem, we focus on evaluating a fixed policy. Instead of balancing exploration and exploitation, a behavior policy for policy evaluation should take actions that reduce uncertainty about $v(\pi)$ with emphasis on actions that have high probability under π. Also, note that heteroscedastic bandits have been studied from the perspective of policy improvement [Kirschner and Krause, 2018, Zhao et al., 2022] however, in this paper we focus on optimal data collection for policy evaluation.

We note that heteroscedasticity is also studied for the policy improvement setup [Kirschner and Krause, 2018, Zhou and Gu, 2022, Zhou et al., 2021, Zhang et al., 2021, Zhao et al., 2022]. In these prior works the reward variances are time-dependent as opposed to the quadratic structure studied in this paper. Note that policy improvement requires a different approach than policy evaluation. These works build tight confidence sets around the unknown model parameter $\boldsymbol{\theta}_*$ by employing weighted ridge regression involving an estimated upper bound to the time-dependent variances. However, in our setting, the variances of each action share the unknown low dimensional co-variance matrix $\boldsymbol{\Sigma}_*$. Hence we deviate from these approaches and employ an alternating OLS-WLS estimation to learn the underlying parameter $\boldsymbol{\Sigma}_*$.

A.2 Probability Tools

Lemma 3. *[Kiefer and Wolfowitz, 1960] Assume that $\mathcal{A} \subset \mathbb{R}^d$ is compact and $\text{span}(\mathcal{A}) = \mathbb{R}^d$. Let $\pi : \mathcal{A} \to [0,1]$ be a distribution on $\mathcal{A}$ so that $\sum_{a\in\mathcal{A}} \pi(a) = 1$ and $\mathbf{V}(\pi) \in \mathbb{R}^{d\times d}$ and $g(\pi) \in \mathbb{R}$ be given by*

$$\mathbf{V}(\pi) = \sum_{a\in\mathcal{A}} \pi(a) aa^\top, \quad g(\pi) = \max_{a\in\mathcal{A}} \|a\|^2_{\widetilde{\mathbf{X}}(\pi)^{-1}}$$

Then the following are equivalent:

(a) π^* is a minimizer of g.

(b) π^* is a maximizer of $f(\pi) = \log\det \mathbf{V}(\pi)$.

(c) $g(\pi^*) = d$.

Furthermore, there exists a minimizer π^* of g such that $|\text{Supp}(\pi^*)| \le d(d+1)/2$.

Lemma 4. ***(Sub-Exponential Concentration)*** *Suppose that X is sub-exponential with parameters (ν, α). Then*

$$\mathbb{P}[X \ge \mu + t] \le \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \text{if } 0 \le t \le \frac{\nu^2}{\alpha} \\ e^{-\frac{t}{2\alpha}} & \text{if } t > \frac{\nu^2}{\alpha} \end{cases}$$

which can be equivalently written as follows:

$$\mathbb{P}[X \ge \mu + t] \le \exp\left\{-\frac{1}{2}\min\left\{\frac{t}{\alpha}, \frac{t^2}{\nu^2}\right\}\right\}.$$

Lemma 5. ***(Restatement of Theorem 2.2 in Rigollet and Hütter [2015])*** *Assume that the linear model holds where the noise $\varepsilon \sim \text{subG}_n(\sigma^2)$. Then the least squares estimator $\widehat{\boldsymbol{\theta}}_\Gamma$ satisfies*

$$\mathbb{E}\left[\text{MSE}\left(\mathbf{X}\widehat{\boldsymbol{\theta}}_\Gamma\right)\right] = \frac{1}{n}\mathbb{E}\left|\mathbf{X}\widehat{\boldsymbol{\theta}}_\Gamma - \mathbf{X}\theta^*\right|_2^2 \lesssim \sigma^2\frac{r}{n}$$

where $r = \text{rank}\left(\mathbf{X}^\top\mathbf{X}\right)$. Moreover, for any $\delta > 0$, with probability at least $1-\delta$, it holds

$$\text{MSE}\left(\mathbf{X}\widehat{\boldsymbol{\theta}}_\Gamma\right) \lesssim \sigma^2\frac{r + \log(1/\delta)}{n}$$

A.3 Formulation for PE-Optimal Design to Reduce MSE

Proposition 1. *Let $\widehat{\boldsymbol{\theta}}_n$ be the Weighted Least Square (WLS) estimate (1) of $\boldsymbol{\theta}_*$ after observing n samples and define $\mathbf{w}(a) = \pi(a)\mathbf{x}(a)$. Define the design matrix as $\mathbf{A}_{\mathbf{b},\boldsymbol{\Sigma}_*}$ (see (2)). Then the loss is given by*

$$\mathbb{E}\left[\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*)\right)^2\right] = \frac{1}{n}\left(\sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\mathbf{b},\boldsymbol{\Sigma}_*}\mathbf{w}(a')\right).$$

Proof. Let $T_n(a) \ge 0$ be the number of samples of $\mathbf{x}(a)$, hence $n = \sum_{a=1}^A T_n(a)$. For each $a \in [A]$, the linear model yields:

$$\frac{1}{T_n(a)}\sum_{i=1}^{T_n(a)} R_i(a) = \mathbf{x}(a)^\top\boldsymbol{\theta}_* + \frac{1}{T_n(a)}\sum_{i=1}^{T_n(a)}\eta_i(a).$$

with $R_i(a)$ being the reward observed for action a taken for the i-th time, $\eta_i(a)$ being the corresponding noise, and $T_n(a)$ is the number of samples of action a. We define the following:

$$\widetilde{Y}_n(a) = \sum_{i=1}^{T_n(a)} \frac{R_i(a)}{\sigma(a)\sqrt{T_n(a)}}, \qquad \widetilde{\mathbf{x}}_n(a) = \frac{\sqrt{T_n(a)}\mathbf{x}(a)}{\sigma(a)}, \qquad \widetilde{\eta}_n(a) = \sum_{i=1}^{T_n(a)} \frac{\eta_i(a)}{\sigma(a)\sqrt{T_n(a)}}$$

so that for all $a \in [A], \widetilde{Y}_n(a) = \widetilde{\mathbf{x}}_n(a)^\top \boldsymbol{\theta}_* + \widetilde{\eta}_n(a)$ where we can show the following regarding the expectation of $\widetilde{\eta}_n(a)$ as

$$\mathbb{E}[\widetilde{\eta}_n(a)] = \mathbb{E}\left[\sum_{i=1}^{T_n(a)} \frac{\eta_i(a)}{\sigma(a)\sqrt{T_n(a)}}\right] = \sum_{i=1}^{T_n(a)} \frac{\mathbb{E}\left[\eta_i(a)\right]}{\sigma(a)\sqrt{T_n(a)}} = 0$$

and the variance as

$$\mathbf{Var}\left[\widetilde{\eta}_n(a)\right] = \mathbf{Var}\left[\sum_{i=1}^{T_n(a)} \frac{\eta_i(a)}{\sigma(a)\sqrt{T_n(a)}}\right] \overset{(a)}{=} \sum_{i=1}^{T_n(a)} \mathbf{Var}\left[\frac{\eta_i(a)}{\sigma(a)\sqrt{T_n(a)}}\right] = \sum_{i=1}^{T_n(a)} \frac{\mathbf{Var}\left[\eta_i(a)\right]}{\sigma^2(a)T_n(a)} = \frac{T_n(a)\sigma^2(a)}{\sigma^2(a)T_n(a)} = 1$$

where (a) follows as the noises are independent. We denote by $\mathbf{X} = \left(\widetilde{\mathbf{x}}_n(1)^\top, \cdots, \widetilde{\mathbf{x}}_n(A)^\top\right)^\top \in \mathbb{R}^{A\times d}$ the induced design matrix of the policy. Under the assumption that $\mathbf{X}$ has full rank, the above weighted least squares (WLS) problem has an optimal unbiased estimator $\widehat{\boldsymbol{\theta}}_n = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}$, where $\mathbf{Y} = [\widetilde{Y}_n(1), \widetilde{Y}_n(2), \ldots, \widetilde{Y}_n(A)]^\top$. Let $\eta = [\widetilde{\eta}_n(1), \widetilde{\eta}_n(2), \ldots, \widetilde{\eta}_n(A)]^\top$. Let $\mathbf{w}(a) = \pi(a)\mathbf{x}(a)$. Then the objective is to bound the loss as follows

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top\widehat{\boldsymbol{\theta}}_n - \sum_{a=1}^{A}\mathbf{w}(a)^\top\boldsymbol{\theta}_*\right)^2\right] = \mathbb{E}\left[\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*)\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y} - \boldsymbol{\theta}_*\right)\right)^2\right] = \mathbb{E}\left[\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}_* + \eta) - \boldsymbol{\theta}_*\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\eta\right)^2\right] \overset{(a)}{=} \mathbb{E}\left[\mathbf{Tr}\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\eta\eta^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\sum_{a=1}^{A}\mathbf{w}(a)\right)\right] \\
&= \mathbf{Tr}\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{E}\left[\eta\eta^\top\right]\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\sum_{a=1}^{A}\mathbf{w}(a)\right) \\
&\overset{(b)}{=} \mathbf{Tr}\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{I}\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\sum_{a=1}^{A}\mathbf{w}(a)\right) \\
&= \mathbf{Tr}\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\sum_{a=1}^{A}\mathbf{w}(a)\right) = \mathbf{Tr}\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top\left(\sum_{a=1}^{A}\widetilde{\mathbf{x}}_n(a)\widetilde{\mathbf{x}}_n(a)^\top\right)^{-1}\sum_{a=1}^{A}\mathbf{w}(a)\right) \\
&= \frac{1}{n}\mathbf{Tr}\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top\left(\sum_{a=1}^{A}\frac{\mathbf{b}(a)\mathbf{x}(a)\mathbf{x}(a)^\top}{\sigma(a)^2}\right)^{-1}\sum_{a=1}^{A}\mathbf{w}(a)\right) \\
&\overset{(c)}{=} \frac{1}{n}\mathbf{Tr}\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top\left(\sum_{a=1}^{A}\mathbf{b}(a)\widetilde{\mathbf{x}}(a)\widetilde{\mathbf{x}}(a)^\top\right)^{-1}\sum_{a=1}^{A}\mathbf{w}(a)\right) \\
&= \frac{1}{n}\mathbf{Tr}\left(\sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\mathbf{b},\boldsymbol{\Sigma}_*}\mathbf{w}(a')\right)
\end{aligned}$$

where, in (a) we can introduce the trace operator as for any vector $\mathbf{x}$ we have $\mathbf{Tr}(\mathbf{x}^\top\mathbf{x}) = \|\mathbf{x}\|^2$, (b) follows as the matrix $\mathbb{E}[\eta\eta^\top]$ has all the non-diagonal element as 0 (since noises are independent and $\mathbf{Cov}(\widetilde{\epsilon}_n(a), \widetilde{\epsilon}_n(a')) = 0$) and the diagonal element are the $\mathbf{Var}[\widetilde{\epsilon}_n(a)] = 1$, and (c) follows as we redefine $\widetilde{\mathbf{x}}(a) = \mathbf{x}(a)/\sigma(a)$. □

A.4 Loss is convex

Proposition 2. *The loss function*

$$\mathcal{L}_n(\pi, \mathbf{b}, \mathbf{\Sigma}_*) = \frac{1}{n}\left(\sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}^{-1} \mathbf{w}(a')\right)$$

for any arbitrary design proportion $\mathbf{b} \in \triangle(\mathcal{A})$ *and co-variance matrix* $\mathbf{\Sigma}_*$ *is strictly convex.*

Proof. Let $\mathbf{b}, \mathbf{b}' \in \triangle(\mathcal{A})$, so that $\mathbf{A}_\mathbf{b}$ and $\mathbf{A}_{\mathbf{b}'}$ are invertible. Recall that we have the loss for a design proportion $\mathbf{b}$ as

$$\mathcal{L}_n(\pi, \mathbf{b}, \mathbf{\Sigma}_*) = \frac{1}{n}\left(\sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}^{-1} \mathbf{w}(a')\right) \overset{(a)}{=} \frac{1}{n}\mathbf{Tr}\left(\sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}^{-1} \mathbf{w}(a')\right) = \frac{1}{n}\mathbf{Tr}\left(\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}^{-1} \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top\right)$$
$$= \frac{1}{n}\mathbf{Tr}\left(\mathbf{V}\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}^{-1}\right)$$

where, in (a) we can introduce the trace as the R.H.S. is a scalar quantity, $\mathbf{w}(a) = \pi(a)\mathbf{x}(a)$ and $\mathbf{V} = \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top$. Similarly for a $\lambda \in [0,1]$ we have

$$\mathcal{L}_n(\pi, \lambda\mathbf{b} + (1-\lambda)\mathbf{b}', \mathbf{\Sigma}_*) = \frac{1}{n}\mathbf{Tr}\left(\mathbf{A}_{\mathbf{b},\mathbf{b}',\mathbf{\Sigma}_*}^{-1} \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top\right) = \frac{1}{n}\mathbf{Tr}\left(\mathbf{V}\mathbf{A}_{\mathbf{b},\mathbf{b}',\mathbf{\Sigma}_*}^{-1}\right).$$

Let the matrix $\mathbf{A}_{\mathbf{b},\mathbf{b}',\mathbf{\Sigma}_*}$ be defined as

$$\mathbf{A}_{\mathbf{b},\mathbf{b}',\mathbf{\Sigma}_*} \coloneqq \lambda\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*} + (1-\lambda)\mathbf{A}_{\mathbf{b}',\mathbf{\Sigma}_*}.$$

Now observe that

$$\mathbf{A}_{\mathbf{b},\mathbf{b}',\mathbf{\Sigma}_*} = \lambda\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*} + (1-\lambda)\mathbf{A}_{\mathbf{b}',\mathbf{\Sigma}_*} = \sum_{a=1}^{A} \left(\lambda\mathbf{b}(a) + (1-\lambda)\mathbf{b}'(a)\right)\widetilde{\mathbf{x}}(a)\widetilde{\mathbf{x}}(a)^\top.$$

Also observe that this is a positive semi-definite matrix. Now using Lemma 1 from [Whittle, 1958] we can show that

$$\left(\lambda\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*} + (1-\lambda)\mathbf{A}_{\mathbf{b}',\mathbf{\Sigma}_*}\right)^{-1} \prec \lambda\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}^{-1} + (1-\lambda)\mathbf{A}_{\mathbf{b}',\mathbf{\Sigma}_*}^{-1}$$

for any positive semi-definite matrices $\mathbf{A}_\mathbf{b}, \mathbf{A}_{\mathbf{b}'}$, and $\lambda \in [0,1]$. Now taking the trace on both sides we get

$$\mathbf{Tr}\left(\lambda\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*} + (1-\lambda)\mathbf{A}_{\mathbf{b}',\mathbf{\Sigma}_*}\right)^{-1} \prec \mathbf{Tr}\lambda\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}^{-1} + \mathbf{Tr}(1-\lambda)\mathbf{A}_{\mathbf{b}',\mathbf{\Sigma}_*}^{-1}.$$

Now using Lemma 2 from Whittle [1958] we can show that

$$\mathbf{Tr}\left(\lambda\mathbf{V}\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*} + (1-\lambda)\mathbf{V}\mathbf{A}_{\mathbf{b}',\mathbf{\Sigma}_*}\right)^{-1} \prec \mathbf{Tr}\lambda\mathbf{V}\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}^{-1} + \mathbf{Tr}(1-\lambda)\mathbf{V}\mathbf{A}_{\mathbf{b}',\mathbf{\Sigma}_*}^{-1}.$$

for any positive semi-definite matrix $\mathbf{V}$. This implies that

$$\mathcal{L}_n(\pi, \lambda\mathbf{b} + (1-\lambda)\mathbf{b}', \mathbf{\Sigma}_*) < \lambda\mathcal{L}_n(\pi, \mathbf{b}, \mathbf{\Sigma}_*) + (1-\lambda)\mathcal{L}_n(\pi, \mathbf{b}', \mathbf{\Sigma}_*).$$

Hence, the loss function is convex. □

Remark 1. (Bound on variance) We can use singular value decomposition of $\mathbf{\Sigma}_*$ as $\mathbf{\Sigma}_* = \mathbf{U}\mathbf{D}\mathbf{P}^\top$ with orthogonal matrices $\mathbf{U}, \mathbf{P}^\top$ and $\mathbf{D} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_d)$ where λ_i denotes a singular value. Then we can bound $\mathbf{x}(a)^\top\mathbf{\Sigma}_*\mathbf{x}(a)$ as

$$\left\|\mathbf{x}(a)^\top\mathbf{\Sigma}_*\mathbf{x}(a)\right\| = \left\|\mathbf{x}(a)^\top\mathbf{U}\mathbf{D}\mathbf{P}^\top\mathbf{x}(a)\right\| \overset{(a)}{=} \left\|\mathbf{u}^\top\mathbf{D}\mathbf{p}\right\| \leq \left\|\mathbf{u}^\top\right\| \max_i |\lambda_i| \, \|\mathbf{p}\|$$
$$\overset{(b)}{=} \|\mathbf{x}(a)\| \max_i |\lambda_i| \, \|\mathbf{x}(a)\| = \max_i |\lambda_i| \, \|\mathbf{x}(a)\|^2$$

where in (a) we have $\mathbf{u} = \mathbf{U}^\top \mathbf{x}(a)$, $\mathbf{p} = \mathbf{P}^\top \mathbf{x}(a)$ and (b) uses the fact that $\left\|\mathbf{U}^\top \mathbf{x}(a)\right\| = \|\mathbf{x}(a)\|$ for any orthogonal matrix $\mathbf{U}^\top$. Similarly we can show that $\left\|\mathbf{x}(a)^\top \mathbf{\Sigma}_* \mathbf{x}(a)\right\| \geq \min_i |\lambda_i| \|\mathbf{x}(a)\|^2$. Let $H_L^2 \leq \|\mathbf{x}(a)\|^2 \leq H_U^2$ for any $a \in [A]$. This implies that

$$\underbrace{\min_i |\lambda_i| H_L^2}_{\sigma^2_{\min}} \leq \min_i |\lambda_i| \|\mathbf{x}(a)\|^2 \leq \underbrace{\mathbf{x}(a)^\top \mathbf{\Sigma}_* \mathbf{x}(a)}_{\sigma^2(a)} \leq \max_i |\lambda_i| \|\mathbf{x}(a)\|^2 \leq \underbrace{\max_i |\lambda_i| H_U^2}_{\sigma^2_{\max}}$$

A.5 Loss Gradient is Bounded

Proposition 3. *Let $\mathbf{b}, \mathbf{b}' \in \triangle(\mathcal{A})$, so that $\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}$ and $\mathbf{A}_{\mathbf{b}',\mathbf{\Sigma}_*}$ are invertible and define $\mathbf{V} = \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top$. Then the gradient of the loss function is bounded such that*

$$\|\nabla_{\mathbf{b}(a)} \mathcal{L}(\pi, \mathbf{b}, \mathbf{\Sigma}_*) - \nabla_{\mathbf{b}(a)} \mathcal{L}(\pi, \mathbf{b}', \mathbf{\Sigma}_*)\|_2 \leq C_\kappa$$

where, the

$$C_\kappa = \frac{\lambda_d(\mathbf{V}) H_U^2}{\sigma^2(a) \left(\min_{a' \in \mathcal{A}} \frac{\mathbf{b}(a')}{\sigma(a')^2} \lambda_{\min}\left(\sum_{a=1}^A \mathbf{w}(a)\mathbf{w}(a)^\top\right)\right)^2} + \frac{\lambda_1(\mathbf{V}) H_U^2}{\sigma^2(a) \left(\min_{a' \in \mathcal{A}} \frac{\mathbf{b}'(a')}{\sigma(a')^2} \lambda_{\min}\left(\sum_{a=1}^A \mathbf{w}(a)\mathbf{w}(a)^\top\right)\right)^2}.$$

Proof. Let $\mathbf{b}, \mathbf{b}' \in \triangle(\mathcal{A})$, so that $\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}$ and $\mathbf{A}_{\mathbf{b}',\mathbf{\Sigma}_*}$ are invertible. Observe that the gradient of the loss is given by

$$\begin{aligned}
\nabla_{\mathbf{b}(a)} \mathcal{L}(\pi, \mathbf{b}, \mathbf{\Sigma}_*) &= \nabla_{\mathbf{b}(a)} \mathbf{Tr}\left(\sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*} \mathbf{w}(a')\right) \\
&\overset{(a)}{\leq} \lambda_1(\mathbf{V}) \nabla_{\mathbf{b}(a)} \mathbf{Tr}(\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}) \\
&= -\lambda_1(\mathbf{V}) \mathbf{Tr}\left(\left(\frac{\mathbf{w}(a)\mathbf{w}(a)^\top}{\sigma^2(a)}\right) \mathbf{A}^{-2}_{\mathbf{b},\mathbf{\Sigma}_*}\right) \\
&= -\lambda_1(\mathbf{V}) \frac{1}{\sigma^2(a)} \left\|\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*} \mathbf{w}(a)\right\|_2^2
\end{aligned}$$

where, in (a) we denote $\mathbf{V} = \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top$. Similarly, the gradient of the loss is lower bounded by

$$\nabla_{\mathbf{b}(a)} \mathcal{L}(\pi, \mathbf{b}, \mathbf{\Sigma}_*) \geq -\lambda_d(\mathbf{V}) \frac{1}{\sigma^2(a)} \left\|\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*} \mathbf{w}(a)\right\|_2^2$$

which yields a bound on the gradient difference as

$$\begin{aligned}
\|\nabla_{\mathbf{b}(a)} \mathcal{L}(\pi, \mathbf{b}, \mathbf{\Sigma}_*) - \nabla_{\mathbf{b}'(a)} \mathcal{L}(\pi, \mathbf{b}', \mathbf{\Sigma}_*)\|_2 &\leq \left\|\lambda_d(\mathbf{V}) \frac{1}{\sigma^2(a)} \left\|\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*} \mathbf{w}(a)\right\|_2^2 - \lambda_1(\mathbf{V}) \frac{1}{\sigma^2(a)} \left\|\mathbf{A}^{-1}_{\mathbf{b}',\mathbf{\Sigma}_*} \mathbf{w}(a)\right\|_2^2\right\|_2 \\
&\leq \left|\lambda_d(\mathbf{V}) \frac{1}{\sigma^2(a)} \left\|\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*} \mathbf{w}(a)\right\|_2^2\right| + \left|\lambda_1(\mathbf{V}) \frac{1}{\sigma^2(a)} \left\|\mathbf{A}^{-1}_{\mathbf{b}',\mathbf{\Sigma}_*} \mathbf{w}(a)\right\|_2^2\right|.
\end{aligned}$$

So now we focus on the quantity

$$\left\|\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*} \mathbf{w}(a)\right\|_2^2 \leq \|\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}\|_2^2 \|\mathbf{w}(a)\|_2^2 \leq \|\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}\|_2^2 H_U^2.$$

Now observe that when $\mathbf{b}(a) \in \triangle(\mathcal{A})$ and initialized uniform randomly, then the optimization in (6) results in a non-singular $\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}$ if each action has been sampled at least once which is satisfied by SPEED. So now we need to bound the minimum eigenvalue of $\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}$ denoted as $\lambda_{\min}(\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*})$. Using Lemma 7 of Fontaine et al. [2021] we have that for all $\mathbf{b} \in \triangle(\mathcal{A})$,

$$\min_{a \in [A]} \frac{\mathbf{b}(a)}{\sigma(a)^2} \sum_{a=1}^A \mathbf{w}(a)\mathbf{w}(a)^\top \preccurlyeq \sum_{a=1}^A \frac{\mathbf{b}(a)}{\sigma(a)^2} \mathbf{w}(a)\mathbf{w}(a)^\top.$$

And finally

$$\min_{a\in[A]} \frac{\mathbf{b}(a)}{\sigma(a)^2}\lambda_{\min}\left(\sum_{a=1}^{A}\mathbf{w}(a)\mathbf{w}(a)^\top\right) \leq \lambda_{\min}(\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*})$$

This implies that

$$\lambda_{\min}(\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}) \leq \frac{1}{\min_{a\in[A]}\frac{\mathbf{b}(a)}{\sigma(a)^2}\lambda_{\min}\left(\sum_{a=1}^{A}\mathbf{w}(a)\mathbf{w}(a)^\top\right)}$$

Plugging everything back we get that

$$\begin{aligned}\|\nabla_{\mathbf{b}(a)}\mathcal{L}(\pi,\mathbf{b},\mathbf{\Sigma}_*) - \nabla_{\mathbf{b}'(a)}\mathcal{L}(\pi,\mathbf{b}',\mathbf{\Sigma}_*)\|_2 \leq & \frac{\lambda_d(\mathbf{V})H_U^2}{\sigma^2(a)\left(\min_{a'\in\mathcal{A}}\frac{\mathbf{b}(a')}{\sigma(a')^2}\lambda_{\min}\left(\sum_{a=1}^{A}\mathbf{w}(a)\mathbf{w}(a)^\top\right)\right)^2} \\ & + \frac{\lambda_1(\mathbf{V})H_U^2}{\sigma^2(a)\left(\min_{a'\in\mathcal{A}}\frac{\mathbf{b}'(a')}{\sigma(a')^2}\lambda_{\min}\left(\sum_{a=1}^{A}\mathbf{w}(a)\mathbf{w}(a)^\top\right)\right)^2}.\end{aligned}$$

The claim of the lemma follows. □

A.6 Kiefer-Wolfowitz Equivalence

We now introduce a Kiefer-Wolfowitz type equivalence [Kiefer and Wolfowitz, 1960] for the quantity $\mathbf{Tr}(\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*})$ for optimal $\mathbf{b}_* \in \Delta(\mathcal{A})$ and co-variance matrix $\mathbf{\Sigma}_*$ in Proposition 4.

Proposition 4. ***(Kiefer-Wolfowitz for PE-Optimal)*** *Define the heteroscedastic design matrix as* $\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*} = \sum_{a=1}^{A}\mathbf{b}(a)\widetilde{\mathbf{x}}(a)\widetilde{\mathbf{x}}(a)^\top$. *Assume that* $\mathcal{A}\subset\mathbb{R}^d$ *is compact and* $\text{span}(\mathcal{A}) = \mathbb{R}^d$. *Then the following are equivalent:*

(a) $\mathbf{b}_*$ *is a minimiser of* $\widetilde{g}(\mathbf{b},\mathbf{\Sigma}_*) = \mathbf{Tr}\left(\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}\right)$.

(b) $\mathbf{b}_*$ *is a maximiser of* $f(\mathbf{b},\mathbf{\Sigma}_*) = \log\det\left(\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}\right)$.

(c) $\widetilde{g}\left(\mathbf{b}_*,\mathbf{\Sigma}_*\right) = d$.

Furthermore, there exists a minimiser $\mathbf{b}_*$ *of* $\widetilde{g}(\mathbf{b},\mathbf{\Sigma}_*)$ *such that* $|\text{Supp}\left(\mathbf{b}_*\right)| \leq d(d+1)/2$.

Proof. We follow the proof technique of Lattimore and Szepesvári [2020b]. Let $\mathbf{b}:\mathcal{A}\to[0,1]$ be a distribution on $\mathcal{A}$ so that $\sum_{a\in\mathcal{A}}\mathbf{b}(a) = 1$ and $\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}\in\mathbb{R}^{d\times d}$ and $g(\mathbf{b})\in\mathbb{R}$ be given by

$$\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*} = \sum_{a=1}^{A}\mathbf{b}(a)\pi^2(a)\sigma^{-2}(a)\ \mathbf{x}(a)\mathbf{x}(a)^\top = \sum_{a=1}^{A}\mathbf{b}(a)\frac{\pi(a)\mathbf{x}(a)}{\sigma(a)}\left(\frac{\pi(a)\mathbf{x}(a)}{\sigma(a)}\right)^\top$$

where, (a) follows by setting $\widetilde{\mathbf{x}}(a) = \mathbf{x}(a)/\sigma(a)$. First recall that for a square matrix $\mathbf{A}$ let adj $(\mathbf{A})$ be the transpose of the cofactor matrix of $\mathbf{A}$. Use the facts that the inverse of a matrix $\mathbf{A}$ is $\mathbf{A}^{-1} = \text{adj}(\mathbf{A})^\top/\det(\mathbf{A})$ and that if $\mathbf{A}:\mathbb{R}\to\mathbb{R}^{d\times d}$, then

$$\frac{d}{dt}\det(\mathbf{A}(t)) = \mathbf{Tr}\left(\text{adj}(\mathbf{A})\frac{d}{dt}\mathbf{A}(t)\right).$$

It follows then that

$$\begin{aligned}\nabla f(\mathbf{b},\mathbf{\Sigma}_*)_{b(a)} &\overset{(a)}{=} \frac{\mathbf{Tr}\left(\text{adj}(\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*})\widetilde{\mathbf{x}}(a)\widetilde{\mathbf{x}}(a')^\top\right)}{\det(\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*})} \\ &= \frac{\widetilde{\mathbf{x}}(a)^\top\text{adj}(\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*})\widetilde{\mathbf{x}}(a')}{\det(\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*})} \overset{(b)}{=} \widetilde{\mathbf{x}}(a)^\top\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}\widetilde{\mathbf{x}}(a') = \widetilde{g}(\mathbf{b})\end{aligned}$$

where, in (a) we show the a-th component of $f(\mathbf{b})$ when we differentiate w.r.t to $\mathbf{b}(a)$, and (b) follows as $\frac{\text{adj}(\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*})}{\det(\mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*})} = \mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}$. Also observe that

$$\left(\sum_{a=1}^{A} \mathbf{b}(a)\|\widetilde{\mathbf{x}}(a)\|^2_{\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}}\right) = \mathbf{Tr}\left(\sum_{a=1}^{A} \mathbf{b}(a)\widetilde{\mathbf{x}}(a)\widetilde{\mathbf{x}}(a')^\top \mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}\right) = d. \tag{11}$$

Hence, $\max_{\mathbf{b}} \log\det \mathbf{A}_{\mathbf{b},\mathbf{\Sigma}_*}$ is lower bounded by d as in average we have that $\left(\sum_{a=1}^{A} \mathbf{b}(a)\|\widetilde{\mathbf{x}}(a)\|^2_{\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}}\right) = d$.

(b) $\Rightarrow$ (a): Suppose that $\mathbf{b}_*$ is a maximiser of f. By the first-order optimality criterion, for any $\mathbf{b}$ distribution on $\mathcal{A}$,

$$\begin{aligned}
0 &\geq \langle \nabla f\left(\mathbf{b}_*, \mathbf{\Sigma}_*\right), \mathbf{b} - \mathbf{b}_* \rangle \\
&\geq \left(\sum_{a=1}^{A} \mathbf{b}(a)\|\widetilde{\mathbf{x}}(a)\|^2_{\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*}} - \sum_{a=1}^{A} \mathbf{b}_*(a)\|\widetilde{\mathbf{x}}(a)\|^2_{\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*}}\right) \\
&\geq \left(\sum_{a=1}^{A} \mathbf{b}(a)\|\widetilde{\mathbf{x}}(a)\|^2_{\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*}} - d\right).
\end{aligned}$$

For an arbitrary $a \in \mathcal{A}$, choosing $\mathbf{b}$ to be the Dirac at $a \in \mathcal{A}$ proves that $\sum_{a=1}^{A} \|\widetilde{\mathbf{x}}(a)\|^2_{\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*}} \leq d$. Since $\widetilde{g}(\mathbf{b}) \geq d$ for all $\mathbf{b}$ by (11), it follows that $\mathbf{b}_*$ is a minimiser of $\widetilde{g}$ and that $\min_{\mathbf{b}} \widetilde{g}(\mathbf{b}) = d$.

(c) $\Longrightarrow$ (b): Suppose that $\widetilde{g}\left(\mathbf{b}_*\right) = d$. Then, for any $\mathbf{b}$,

$$\langle \nabla f\left(\mathbf{b}_*, \mathbf{\Sigma}_*\right), \mathbf{b} - \mathbf{b}_* \rangle = \left(\sum_{a=1}^{A} \mathbf{b}(a)\|\widetilde{\mathbf{x}}(a)\|^2_{\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*}} - d\right) \leq 0.$$

And it follows that $\mathbf{b}_*$ is a maximiser of f by the first-order optimality conditions and the concavity of f. This can be shown as follows:

Let $\mathbf{b}$ be a Dirac at a and $\mathbf{b}(t) = \mathbf{b}_* + t\left(\mathbf{b}_* - \mathbf{b}\right)$. Since $\mathbf{b}_*(a) > 0$ it follows for sufficiently small $t > 0$ that $\mathbf{b}(t)$ is a distribution over $\mathcal{A}$. Because $\mathbf{b}_*$ is a minimiser of f,

$$0 \geq \frac{d}{dt} f(\mathbf{b}(t), \mathbf{\Sigma}_*)\Big|_{t=0} = \langle \nabla f\left(\mathbf{b}_*, \mathbf{\Sigma}_*\right), \mathbf{b}_* - \mathbf{b} \rangle = d - \sum_{a=1}^{A} \|\widetilde{\mathbf{x}}(a)\|^2_{\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}}.$$

We now show (a) $\Longrightarrow$ (c). To prove the second part of the theorem, let $\mathbf{b}_*$ be a minimiser of $\widetilde{g}$, which by the previous part is a maximiser of f. Let $S = \text{Supp}\left(\mathbf{b}_*\right)$, and suppose that $|S| > d(d+1)/2$. Since the dimension of the subspace of $d \times d$ symmetric matrices is $d(d+1)/2$, there must be a non-zero function $v : \mathcal{A} \to \mathbb{R}$ with $\text{Supp}(v) \subseteq S$ such that

$$\sum_{a \in S} v(a)\widetilde{\mathbf{x}}(a)\widetilde{\mathbf{x}}(a)^\top = \mathbf{0}. \tag{12}$$

Notice that for any $\widetilde{\mathbf{x}}(a) \in S$, the first-order optimality conditions ensure that $\sum_{a=1}^{A} \|\widetilde{\mathbf{x}}(a)\|^2_{\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*}} = d$. Hence

$$d\sum_{a \in S} v(a) = \sum_{a \in S} v(a)\|\widetilde{\mathbf{x}}(a)\|^2_{\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*}} = 0,$$

where the last equality follows from (12). Let $\mathbf{b}(t) = \mathbf{b}_* + tv$ and let $\tau = \max\{t > 0 : \mathbf{b}(t) \in \mathcal{P}_{\mathcal{A}}\}$, which exists since $v \neq 0$ and $\sum_{a \in S} v(a) = 0$ and $\text{Supp}(v) \subseteq S$. By (12), $\mathbf{A}_{\mathbf{b}(t),\mathbf{\Sigma}_*} = \mathbf{A}_{\mathbf{b}_*,\mathbf{\Sigma}_*}$, and hence $f(\mathbf{b}(\tau), \mathbf{\Sigma}_*) = f\left(\mathbf{b}_*, \mathbf{\Sigma}_*\right)$, which means that $\mathbf{b}(\tau)$ also maximises f. The claim follows by checking that $|\text{Supp}(\mathbf{b}(T))| < |\text{Supp}\left(\mathbf{b}_*\right)|$ and then using induction. □

Corollary 1. *1 From Proposition 4 we know that $\mathbf{b}_*$ is a minimizer for $\mathbf{Tr}(\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*})$ and $\mathbf{Tr}(\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*}) = d$. This implies that the loss is bounded at $\mathbf{b}_*$ as $\frac{\lambda_d(\mathbf{V})d}{n} \leq \mathcal{L}_n(\pi, \mathbf{b}_*, \mathbf{\Sigma}_*) \leq \frac{\lambda_1(\mathbf{V})d}{n}$ where $\mathbf{V} = \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top$.*

Proof. First recall that we can rewrite the loss for any arbitrary proportion $\mathbf{b}$ and co-variance $\mathbf{\Sigma}_*$ as

$$\mathcal{L}_n(\pi, \mathbf{b}, \mathbf{\Sigma}_*) = \frac{1}{n}\left(\sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*} \mathbf{w}(a')\right) = \frac{1}{n}\left(\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*} \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top\right) = \frac{1}{n}\left(\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}\mathbf{V}\right).$$

From [Fang et al., 1994] we know that for any positive semi-definite matrices $\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}$ and $\mathbf{V}$ we have that

$$\lambda_d(\mathbf{V})\mathbf{Tr}(\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}) \le \mathbf{Tr}(\mathbf{V}\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*}) \le \lambda_1(\mathbf{V})\mathbf{Tr}(\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*})$$

where $\lambda_i(\mathbf{V})$ is the i th largest eigenvalue of $\mathbf{V}$. Now from Proposition 4 we know that for $\mathbf{b}_*$ is a minimizer for $\mathbf{Tr}(\mathbf{A}^{-1}_{\mathbf{b},\mathbf{\Sigma}_*})$ and $\mathbf{Tr}(\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*}) = d$. This implies that the loss is bounded at $\mathbf{b}_*$ as

$$\lambda_d(\mathbf{V})\mathbf{Tr}(\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*}) \le \mathbf{Tr}(\mathbf{V}\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*}) \le \lambda_1(\mathbf{V})\mathbf{Tr}(\mathbf{A}^{-1}_{\mathbf{b}_*,\mathbf{\Sigma}_*}) \implies \frac{\lambda_d(\mathbf{V})d}{n} \le \mathcal{L}_n(\pi, \mathbf{b}_*, \mathbf{\Sigma}_*) \le \frac{\lambda_1(\mathbf{V})d}{n}.$$

The claim of the corollary follows. □

Remark 2. Note that the estimator $\widehat{\boldsymbol{\theta}}_n$ is an unbiased estimator of $\boldsymbol{\theta}_*$. Recall that

$$\widehat{\boldsymbol{\theta}}_n \coloneqq \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{n} \tfrac{1}{\sigma^2(a_t)}(r_t - \mathbf{x}(a_t)^\top \boldsymbol{\theta})^2$$

where, a_t is the action sampled at timestep t. Define the $\mathbf{diag}(\mathbf{\Sigma}_n) = [\sigma^2(a_1), \sigma^2(a_2), \ldots, \sigma^2(a_n)]$, $\mathbf{R}_n = [r_1, r_2, \ldots, r_n]^\top \in \mathbb{R}^{n\times 1}$ be the n rewards observed and $\eta \in \mathbb{R}^{n\times 1}$ is the noise vector, where $a_1, a_2, \ldots, a_n$ are the actions pulled at time $t = 1, 2, \ldots, n$. Then it can be shown that

$$\begin{aligned}
\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_n\right] - \boldsymbol{\theta}_* &= \mathbb{E}\left[\left(\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n\right)^{-1}\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{R}_n\right] - \boldsymbol{\theta}_* \\
&= \mathbb{E}\left[\left(\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n\right)^{-1}\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\left(\mathbf{X}_n\boldsymbol{\theta}_* + \eta\right)\right] - \boldsymbol{\theta}_* \\
&= \mathbb{E}\left[\left(\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n\right)^{-1}\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n\boldsymbol{\theta}_*\right] + \mathbb{E}\left[\left(\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n\right)^{-1}\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\eta\right] - \boldsymbol{\theta}_* \\
&= \boldsymbol{\theta}_* + \left(\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n\right)^{-1}\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbb{E}\left[\eta\right] - \boldsymbol{\theta}_* \overset{(a)}{=} 0
\end{aligned}$$

where, (a) follows as noise is zero mean.

B Bandit Regret Proofs

B.1 Loss of Bandit Oracle

Proposition 5. ***(Bandit Oracle MSE)*** *Let the oracle sample each action a for $\lceil n\mathbf{b}_*(a)\rceil$ times, where $\mathbf{b}_*$ is the solution to (3). Define $\lambda_1(\mathbf{V})$ as the maximum eigenvalue of $\mathbf{V} \coloneqq \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top$. Then the loss satisfies*

$$\mathcal{L}^*_n(\pi, \mathbf{b}_*, \mathbf{\Sigma}_*) \le O_{\kappa^2, H_U^2}\left(\frac{d\lambda_1(V)\log n}{n}\right) + O_{\kappa^2, H_U^2}\left(\frac{1}{n}\right).$$

Proof. Recall the matrix $\mathbf{X}_n = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times d}$ are the observed features for the n samples taken. Let $\mathbf{R}_n = [r_1, r_2, \ldots, r_n]^\top \in \mathbb{R}^{n\times 1}$ be the n rewards observed and $\eta \in \mathbb{R}^{n\times 1}$ is the noise vector. Then using weighted least square estimates we have

$$\widehat{\boldsymbol{\theta}}_n \coloneqq \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{n} \frac{1}{\sigma^2(a_t)}(r_t - \mathbf{x}(a_t)^\top \boldsymbol{\theta})^2$$

where, in (a) we a_t is the action sampled at timestep t. Recall that the $\mathbf{diag}(\mathbf{\Sigma}_n) = [\sigma^2(a_1), \sigma^2(a_2), \ldots, \sigma^2(a_n)]$, where $a_1, a_2, \ldots, a_n$ are the actions pulled at time $t = 1, 2, \ldots, n$. We have that:

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_* = (\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n)^{-1}\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\eta$$

where the noise vector $\eta \sim \mathcal{SG}(0, \mathbf{\Sigma}_n)$ where $\mathbf{\Sigma}_n \in \mathbb{R}^{n\times n}$. For any $\mathbf{z} \in \mathbb{R}^d$ we have

$$\mathbf{z}^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) = \mathbf{z}^\top(\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n)^{-1}\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\eta.$$

Let $\mathbf{b}_*$ be the PE-Optimal design for $\mathcal{A}$ defined in (3). Then the oracle pulls action $a \in \mathcal{A}$ exactly $\lceil n\mathbf{b}_* \rceil$ times for some $n > d(d+1)/2$ and computes the least square estimator $\widehat{\boldsymbol{\theta}}_n$. Observe that

$$\sum_{a=1}^{A} \mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) \sim \mathcal{SG}\left(0, \sum_{a,a'} \mathbf{w}(a)^\top(\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n)^{-1}\mathbf{w}(a')\right).$$

So $\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*)\right)^2 \sim \mathcal{SE}\left(0, \sum_{a,a'} \mathbf{w}(a)^\top(\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n)^{-1}\mathbf{w}(a')\right)$ where $\mathcal{SE}$ denotes the sub-exponential distribution. Denote the quantity

$$t := \sqrt{2\sum_{a,a'} \mathbf{w}(a)^\top(\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n)^{-1}\mathbf{w}(a')\log(1/\delta)}.$$

Now using sub-exponential concentration inequality in Lemma 4, setting

$$\nu^2 = \sum_{a,a'} \mathbf{w}(a)^\top(\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n)^{-1}\mathbf{w}(a'),$$

and $\alpha = \nu$, we can show that

$$\mathbb{P}\left(\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*)\right)^2 > t\right) \le \delta, \qquad \text{if } t \in (0,1]$$
$$\mathbb{P}\left(\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*)\right)^2 > t^2\right) \le \delta, \qquad \text{if } t > 1.$$

Combining the above two we can show that

$$\mathbb{P}\left(\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*)\right)^2 > \min\{t, t^2\}\right) \le \delta, \forall t > 0.$$

Further define matrix $\overline{\mathbf{\Sigma}}_n \in \mathbb{R}^{d\times d}$ as $\overline{\mathbf{\Sigma}}_n^{-1} := (\mathbf{X}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{X}_n)^{-1}$. This means that we have with probability $(1-\delta)$ that

$$\begin{aligned}
\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*)\right)^2 &\le \min\left\{\sqrt{2\sum_{a,a'} \mathbf{w}(a)^\top \overline{\mathbf{\Sigma}}_n^{-1}\mathbf{w}(a')\log(1/\delta)}, 2\sum_{a,a'} \mathbf{w}(a)^\top \overline{\mathbf{\Sigma}}_n^{-1}\mathbf{w}(a')\log(1/\delta)\right\} \\
&\overset{(a)}{=} \min\left\{\sqrt{\frac{2}{n}\sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b}_*,\mathbf{\Sigma}_*}^{-1}\mathbf{w}(a')\log(1/\delta)}, \frac{2}{n}\sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b}_*,\mathbf{\Sigma}_*}^{-1}\mathbf{w}(a')\log(1/\delta)\right\} \\
&\overset{(b)}{\le} \min\left\{\sqrt{\frac{8d\lambda_1(\mathbf{V})\log(1/\delta)}{n}}, \frac{8d\lambda_1(\mathbf{V})\log(1/\delta)}{n}\right\}
\end{aligned}$$

and we have taken at most n pulls such that $n > \frac{d(d+1)}{2}$ pulls. Here (a) follows as $n\mathbf{A}_{\mathbf{b}_*,\mathbf{\Sigma}_*} = \overline{\mathbf{\Sigma}}_n$ and observing that oracle has access to $\mathbf{\Sigma}_*$, and optimal proportion $\mathbf{b}_*$. The (b) follows from applying Corollary 1 such that $\sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b}_*,\mathbf{\Sigma}_*}^{-1}\mathbf{w}(a') \le d\lambda_1(\mathbf{V})$ where $\mathbf{V} = \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top$. Thus, for any $\delta \in (0,1)$ we have

$$\mathbb{P}\left(\left\{\left(\sum_{a=1}^{A} \widetilde{\mathbf{x}}(a)^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*)\right)^2 > \min\left\{\sqrt{\frac{8d\lambda_1(\mathbf{V})\log(1/\delta)}{n}}, \frac{8d\lambda_1(\mathbf{V})\log(1/\delta)}{n}\right\}\right\}\right) \le \delta. \tag{13}$$

Define the good event $\xi_\delta(n)$ as follows:

$$\xi_\delta(n) := \left\{ \left(\sum_{a=1}^{A} \widetilde{\mathbf{x}}(a)^\top (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) \right)^2 \leq \min \left\{ \sqrt{\frac{8d\lambda_1(\mathbf{V})\log(1/\delta)}{n}}, \frac{8d\lambda_1(\mathbf{V})\log(1/\delta)}{n} \right\} \right\}.$$

Then the loss of the oracle following PE-Optimal $\mathbf{b}_*$ is given by

$$\begin{aligned}
\mathcal{L}_n^*(\pi, \mathbf{b}_*, \boldsymbol{\Sigma}_*) &= \mathbb{E}_{\mathcal{D}} \left[\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_* \right) \right)^2 \right] \\
&\leq \mathbb{E}_{\mathcal{D}} \left[\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_* \right) \right)^2 \xi_\delta(n) \right] + \mathbb{E}_{\mathcal{D}} \left[\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_* \right) \right)^2 \xi_\delta^c(n) \right] \\
&\overset{(a)}{\leq} \mathbb{E}_{\mathcal{D}} \left[\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_* \right) \right)^2 \xi_\delta(n) \right] + \sum_{t=1}^{n} AH_U^2 \kappa^2 \mathbb{P}(\xi_\delta^c(n)) \\
&\overset{(b)}{\leq} \min \left\{ \sqrt{\frac{8d\lambda_1(\mathbf{V})\log(1/\delta)}{n}}, \frac{8d\lambda_1(\mathbf{V})\log(1/\delta)}{n} \right\} + \sum_{t=1}^{n} AH_U^2 \kappa^2 \mathbb{P}(\xi_\delta^c(n)) \\
&\overset{(c)}{\leq} \min \left\{ \sqrt{\frac{16d\lambda_1(\mathbf{V})\log n}{n}}, \frac{16d\lambda_1(\mathbf{V})\log n}{n} \right\} + O_{\kappa^2, H_U^2}\left(\frac{1}{n}\right) \\
&\leq \frac{48d\lambda_1(\mathbf{V})\log n}{n} + O_{\kappa^2, H_U^2}\left(\frac{1}{n}\right)
\end{aligned}$$

where, (a) follows as the noise $\eta^2 \leq \kappa^2$ and $\sum_a \|\mathbf{x}(a)\|^2 \leq AH_U^2$ which implies

$$\mathbb{E}_{\mathcal{D}} \left[\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_* \right) \right)^2 \right] \leq nAH_U^2 \kappa^2.$$

The (b) follows from (13), and (c) follows by setting $\delta = 1/n^3$, and noting that $n > A$. □

B.2 OLS-WLS Concentration Lemma

Lemma 6. ***(Concentration Lemma)*** *After Γ samples of exploration, we can show that $\mathbb{P}\left(\xi_\delta^{var}(\Gamma)\right) \geq 1 - 8\delta$ where, $C > 0$ is a constant.*

Proof. We observed $(\mathbf{x}_t, r_t) \in \mathbb{R}^d \times \mathbb{R}, i = 1, \ldots, \Gamma$ from the model

$$r_t = \mathbf{x}_t^\top \boldsymbol{\theta}_* + \eta_t, \tag{14}$$

$$\eta_t \sim \mathcal{SG}(0, \mathbf{x}_t^\top \boldsymbol{\Sigma}_* \mathbf{x}_t), \tag{15}$$

where $\boldsymbol{\theta}_* \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_* \in \mathbb{R}^{d \times d}$ are unknown.

Given an initial estimate $\widehat{\boldsymbol{\theta}}_\Gamma$ of $\boldsymbol{\theta}_*$, we first compute the squared residual $y_t := \left(\mathbf{x}_t^\top \widehat{\boldsymbol{\theta}}_\Gamma - r_t \right)^2$, and then obtain an estimate of $\boldsymbol{\Sigma}_*$ via

$$\min_{\mathbf{S} \in \mathbb{R}^{d \times d}} \sum_{t=1}^{\Gamma} \left(\left\langle \mathbf{x}_t \mathbf{x}_t^\top, \mathbf{S} \right\rangle - y_t \right)^2. \tag{16}$$

Observe that if $\widehat{\boldsymbol{\theta}}_\Gamma = \boldsymbol{\theta}_*$, then the expectation of the squared residual y_t is

$$\mathbb{E}\left[y_t\right] = \mathbb{E}\left[\left(\mathbf{x}_t^\top \boldsymbol{\theta}_* - r_t \right)^2 \right] = \mathbb{E}\left[\eta_t^2\right] = \mathbf{x}_t^\top \boldsymbol{\Sigma}_* \mathbf{x}_t = \left\langle \mathbf{x}_t \mathbf{x}_t^\top, \boldsymbol{\Sigma}_* \right\rangle,$$

which is a linear function of $\boldsymbol{\Sigma}_*$. The program (16) is thus a least square formulation for estimating $\boldsymbol{\Sigma}_*$.

Let $\mathbf{X}_t := \mathbf{x}_t\mathbf{x}_t^\top$. Below we abuse notation and view $\mathbf{\Sigma}_*, \widehat{\mathbf{\Sigma}}_\Gamma, \mathbf{X}_t, \mathbf{S}$ as vectors in $\mathbb{R}^{d^2}$ endowed with the trace inner product $\langle\cdot,\cdot\rangle$. Let $\mathbf{X} \in \mathbb{R}^{\Gamma\times d^2}$ have rows $\{\mathbf{X}_t\}$, and $y = (y_1,\dots,y_\Gamma)^\top \in \mathbb{R}^\Gamma$. Suppose $\mathbf{x}_t$ can only take on M possible values from $\{\phi_1,\dots,\phi_M\}$, so $\mathbf{X}_t \in \{\Phi_1,\dots,\Phi_M\}$, where $\Phi_m := \phi_m\phi_m^\top$. Note that for the forced exploration setting we have $M = d < A$. Moreover, each value appears exactly Γ/M times. Then (16) can be rewritten as

$$\min_{\mathbf{S}\in\mathbb{R}^{d^2}} \sum_{m=1}^{M} \sum_{t:\mathbf{X}_t=\Phi_m} \left(\langle\Phi_m,\mathbf{S}\rangle - y_t\right)^2 = \min_{\mathbf{S}\in\mathbb{R}^{d^2}} \sum_{m=1}^{M} \left(\langle\Phi_m,\mathbf{S}\rangle - \frac{1}{\Gamma/M}\sum_{t:\mathbf{X}_t=\Phi_m} y_t\right)^2.$$

Let $z_m := \frac{1}{\Gamma/M}\sum_{t:\mathbf{X}_t=\Phi_m} y_t$. Then it becomes

$$\min_{\mathbf{S}\in\mathbb{R}^{d^2}} \sum_{m=1}^{M} \left(\langle\Phi_m,\mathbf{S}\rangle - z_m\right)^2 = \min_{\mathbf{S}\in\mathbb{R}^{d^2}} \|\Phi\mathbf{S} - z\|_2^2,$$

where $\Phi \in \mathbb{R}^{m\times d^2}$ has rows $\{\Phi_m\}$, and $z := (z_1,\dots,z_m)^\top \in \mathbb{R}/M$. Note that $\{\Phi_m\}$ may or may not span $\mathbb{R}^{d^2}$. Observe that $\widehat{\mathbf{\Sigma}}_\Gamma$ be an optimal solution to the above problem. Then

$$\begin{aligned}\left\|\Phi(\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*)\right\|_2^2 + \|\Phi\mathbf{\Sigma}_* - z\|_2^2 + 2\left\langle\Phi(\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*), \Phi\mathbf{\Sigma}_* - z\right\rangle &= \left\|\Phi\widehat{\mathbf{\Sigma}}_\Gamma - \Phi\mathbf{\Sigma}_* + \Phi\mathbf{\Sigma}_* - z\right\|_2^2 \\ &= \left\|\Phi\widehat{\mathbf{\Sigma}}_\Gamma - z\right\|_2^2 \le \|\Phi\mathbf{\Sigma}_* - z\|_2^2.\end{aligned}$$

Hence, we can show that

$$\begin{aligned}\left\|\Phi(\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*)\right\|_2^2 &\le -2\left\langle\Phi(\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*), \Phi\mathbf{\Sigma}_* - z\right\rangle \\ &\overset{(a)}{\le} 2\left\|\Phi(\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*)\right\|_2 \|\Phi\mathbf{\Sigma}_* - z\|_2.\end{aligned}$$

where, (a) follows from Cauchy Schwarz inequality. So

$$\left\|\Phi(\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*)\right\|_2 \le 2\|\Phi\mathbf{\Sigma}_* - z\|_2.$$

Observe that the RHS does not contain the $\widehat{\mathbf{\Sigma}}_\Gamma$ anymore. Note that the m-th entry of $\Phi\mathbf{\Sigma}_* - z$ is

$$\langle\Phi_m,\mathbf{\Sigma}_*\rangle - z_m = \phi_m^\top\mathbf{\Sigma}_*\phi_m - \frac{1}{\Gamma/M}\sum_{t:\mathbf{X}_t=\Phi_m} y_t.$$

Let $\zeta_\Gamma := \widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*$ where $\widehat{\boldsymbol{\theta}}_\Gamma$ is the estimation of $\boldsymbol{\theta}_*$ after $\Gamma = \sqrt{n}$ rounds of exploration. The noise η_t is σ_t^2 sub-Gaussian. Then

$$\begin{aligned}y_t &= \left(\mathbf{x}_t^\top\widehat{\boldsymbol{\theta}}_t - r_t\right)^2 \\ &= (\eta_t + \mathbf{x}_t^\top\zeta_\Gamma)^2 \\ &= \eta_t^2 + 2\eta_t\mathbf{x}_t^\top\zeta_\Gamma + \left(\mathbf{x}_t^\top\zeta_\Gamma\right)^2 \\ &= \mathbf{x}_t^\top\mathbf{\Sigma}_*\mathbf{x}_t + \epsilon_t = \langle\mathbf{\Sigma}_*, \mathbf{x}_t\mathbf{x}_t^\top\rangle + \epsilon_t \overset{(a)}{=} \langle\widetilde{\boldsymbol{\theta}}_*, \mathbf{z}_t\rangle + \epsilon_t\end{aligned}$$

where, in (a) we denote the $\widetilde{\boldsymbol{\theta}}_* \in \mathbb{R}^{d^2}$ as the vector reshaping $\mathbf{\Sigma}_*$ and $\mathbf{z}_t \in \mathbb{R}^{d^2}$ is the vector reshaping $\mathbf{x}_t\mathbf{x}_t^\top$. This shows that the feedback y_t is linear. Now we need to show that ϵ_t is sub-exponential. We proceed as follows: We have that

$$\epsilon_t := y_t - \mathbf{x}_t^\top\mathbf{\Sigma}_*\mathbf{x}_t = \underbrace{\eta_t^2 - \mathbb{E}[\eta_t^2]}_{\textbf{Part A}} + \underbrace{2\eta_t\mathbf{x}_t^\top\zeta_\Gamma}_{\textbf{Part B}} + \underbrace{\left(\mathbf{x}_t^\top\zeta_\Gamma\right)^2}_{\textbf{Part C}}$$

The goal is to prove that $\mathbb{P}(\epsilon_t > \mathbb{E}[\epsilon_t] + s) \le \exp(-s/2\sigma_{\max}^2)$ for some $s \in \mathbb{R}$.

For part A, we know that the η_t^2 is a sub-exponential random variable with $\eta_t^2 \sim \mathcal{SE}(\nu, \alpha)$ where $\nu = 4\sigma_t^2\sqrt{2}, \alpha = 4\sigma_t^2$, and $\sigma_t^2 = \mathbf{x}_t^\top \mathbf{\Sigma}_* \mathbf{x}_t$. This follows from Equation 37 in Appendix B of Honorio and Jaakkola [2014]. It shows that if X is a centered sub-Gaussian random variable with sub-Gaussian parameter σ^2 then X^2 is sub-exponential with parameters $\nu = 4\sigma^2\sqrt{2}, \alpha = 4\sigma^2$.

From Lemma 4 we know that

$$\mathbb{P}\left(\eta_t^2 \geq \mathbb{E}[\eta_t^2] + 8\sigma_t^2 \log(A/\delta)\right) \leq \exp\left(-\frac{1}{2}\min\left\{\frac{8\sigma_t^2\log(A/\delta)}{4\sigma_t^2}, \frac{64\sigma_t^4\log^2(A/\delta)}{32\sigma_t^4}\right\}\right) = \exp\left(-\log(A/\delta)\right).$$

Hence, $\eta_t^2 \leq 4\sigma_t^2\sqrt{2} + 8\sigma_t^2\log(A/\delta) \leq 16\sigma_{\max}^2\log(A/\delta)$ with probability $(1-\delta)$ as $\mathbb{E}[\eta_t^2] = \nu$. Equivalently we can write that

$$\mathbb{P}\left(\eta_t^2 \geq \mathbb{E}[\eta_t^2] + 16\sigma_t^2\log(A/\delta)\right) \leq \exp\left(-\frac{s_1^2}{16\sigma_{\max}^2}\right) = \exp\left(-\frac{s_1^2}{2c'\sigma_{\max}^2}\right). \tag{17}$$

where $s_1 = \sqrt{2c'\sigma_{\max}^2\log(A/\delta)}$ and $c' > 0$ is a constant.

For part C we proceed as follows:

$$\begin{aligned}(\mathbf{x}_t^\top(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*))^2 &\leq (\mathbf{x}^\top\mathbf{x})\|\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*\|^2 \leq (\mathbf{x}^\top\mathbf{x})\frac{MSE(\mathbf{X}(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*))}{\lambda_{\min}} \\ &\leq \frac{H_U^2}{\lambda_{\min}\Gamma}\left(8\log(6)\sigma_{\max}^2 r + 8\sigma_{\max}^2\log(1/\delta)\right) \leq \frac{2c''\sigma_{\max}^2 d^2\log(A/\delta)}{\Gamma}\end{aligned}$$

The first inequality follows by Cauchy Schwarz and the second by Remark 2.3 of Rigollet and Hütter [2015]. Therefore it follows that

$$\mathbb{P}\left((\mathbf{x}_t^\top(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*))^2 \geq \frac{2c''\sigma_{\max}^2 d^2\log(A/\delta)}{\Gamma}\right) \leq \delta.$$

Assuming $\Gamma > d^2$ we can also show that

$$\mathbb{P}\left((\mathbf{x}_t^\top(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*))^2 \geq 2c''\sigma_{\max}^2\log(A/\delta)\right) \leq \delta$$

which drops the dependence on Γ and d. Equivalently we can write that

$$\mathbb{P}\left((\mathbf{x}_t^\top(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*))^2 \geq 2c''\sigma_{\max}^2\log(A/\delta)\right) \leq \exp\left(-\frac{s_3^2}{2c''\sigma_{\max}^2}\right). \tag{18}$$

where $s_3 = \sqrt{2c''\sigma_{\max}^2\log(A/\delta)}$.

For part B we proceed as follows:

$$2\underbrace{\eta_t}_{a}\underbrace{\mathbf{x}_t^\top\left(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*\right)}_{b} \overset{(a)}{\leq} 2\eta_t^2 + \frac{1}{2}\left(\mathbf{x}_t^\top\left(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*\right)\right)^2$$

where, (a) follows as $2ab \leq 2a^2 + \frac{1}{2}b^2$. It follows then that

$$\begin{aligned}\mathbb{P}\left(2\eta_t\mathbf{x}_t^\top\left(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*\right) \geq s_1^2 + s_3^3\right) &\overset{(a)}{\leq} \mathbb{P}\left(2\eta_t^2 + \frac{1}{2}(\mathbf{x}_t^\top\left(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*\right))^2 > s_1^2 + s_3^2\right) \\ &\leq \mathbb{P}\left(2\eta_t^2 > s_1^2 + s_3^2\right) + \mathbb{P}\left(\frac{1}{2}(\mathbf{x}_t^\top\left(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*\right))^2 > s_1^2 + s_3^2\right) \\ &= \mathbb{P}\left(\eta_t^2 > \frac{s_1^2 + s_3^2}{2}\right) + \mathbb{P}\left((\mathbf{x}_t^\top\left(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*\right))^2 > 2(s_1^2 + s_3^2)\right) \\ &\overset{(b)}{\leq} \mathbb{P}\left(\eta_t^2 > \frac{s_1^2 + s_3^2}{2}\right) + \mathbb{P}\left((\mathbf{x}_t^\top\left(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*\right))^2 > \frac{s_1^2 + s_3^2}{2}\right) \\ &\overset{(c)}{\leq} \mathbb{P}\left(\eta_t^2 > \frac{s_1^2}{2}\right) + \mathbb{P}\left((\mathbf{x}_t^\top\left(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*\right))^2 > \frac{s_3^2}{2}\right) \\ &\overset{(d)}{\leq} \exp\left(-\frac{s_1^2}{c'\sigma_{\max}^2}\right) + \exp\left(-\frac{s_3^2}{c''\sigma_{\max}^2}\right)\end{aligned}$$

where, (a) follows as LHS $2\eta_t^2 + \frac{1}{2}(\mathbf{x}_t^\top \left(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*\right))^2 > 2\eta_t \mathbf{x}_t^\top \left(\widehat{\boldsymbol{\theta}}_\Gamma - \boldsymbol{\theta}_*\right)$ (that is LHS is larger). The (b) follows as RHS $\frac{s_1^2+s_3^2}{2} < 2(s_1^2+s_3^2)$ and (c) follows as RHS $\frac{s_1^2+s_3^2}{2} < \frac{s_1^2}{2}$ (that is RHS is smaller). The (d) follows from (17), and (18).

We now estimate the expectation of ϵ_t. Observe that for $\Gamma > d^2$ we have that

$$\mathbb{E}_{\eta,\zeta}[\epsilon_t] = \mathbb{E}_{\eta,\zeta}\left[\eta_t^2 - \mathbb{E}_\eta[\eta_t^2] + 2\eta_t \mathbf{x}_t^\top \zeta_\Gamma + \left(\mathbf{x}_t^\top \zeta_\Gamma\right)^2\right] = \mathbb{E}_\eta[\eta_t^2] - \mathbb{E}_{\eta,\zeta}[\mathbb{E}_\eta[\eta_t^2]] + 2\mathbb{E}_{\eta,\zeta}[\eta_t \mathbf{x}_t^\top \zeta_\Gamma] + \mathbb{E}_\zeta[\left(\mathbf{x}_t^\top \zeta_\Gamma\right)^2]$$
$$\geq 2\mathbb{E}_{\eta,\zeta}[\eta_t \mathbf{x}_t^\top \zeta_\Gamma] + \mathbb{E}_\zeta[\left(\mathbf{x}_t^\top \zeta_\Gamma\right)^2] \geq \frac{\sigma_{\max}^2 d}{\Gamma}.$$

Similarly we can get an upper bound to $\mathbb{E}[\epsilon_t]$ for $\Gamma > d^2$ as follows:

$$\mathbb{E}_{\eta,\zeta}[\epsilon_t] = \mathbb{E}_{\eta,\zeta}\left[\eta_t^2 - \mathbb{E}_\eta[\eta_t^2] + 2\eta_t \mathbf{x}_t^\top \zeta_\Gamma + \left(\mathbf{x}_t^\top \zeta_\Gamma\right)^2\right] \overset{(a)}{\leq} 2\mathbb{E}_\eta[\eta_t^2] + \frac{1}{2}\mathbb{E}_\zeta[\left(\mathbf{x}_t^\top \zeta_\Gamma\right)^2] + \mathbb{E}_\zeta[\left(\mathbf{x}_t^\top \zeta_\Gamma\right)^2]$$
$$\leq 2\mathbb{E}_\eta[\eta_t^2] + \frac{2c^{''}\sigma_{\max}^2 d^2 \log(A/\delta)}{\Gamma} \overset{(b)}{\leq} 16\sigma_{\max}^2 + \frac{2c^{''}\sigma_{\max}^2 d^2 \log(A/\delta)}{\Gamma}$$

where, (a) follows for $2ab \leq 2a^2 + \frac{1}{2}b^2$, (b) follows as $\mathbb{E}[\eta_t^2] = \nu \leq 8\sigma_{\max}^2$.

Define $s^2 = (s_1^2 + s_3^2)$. Then combining Part A, B and C it follows that

$$\mathbb{P}(\epsilon_t \geq s^2) = \mathbb{P}(\eta_t^2 + 2\eta_t \mathbf{x}_t^\top \zeta_\Gamma + \left(\mathbf{x}_t^\top \zeta_\Gamma\right)^2 \geq s^2) \leq \mathbb{P}(\eta_t^2 \geq s^2) + \mathbb{P}(2\eta_t \mathbf{x}_t^\top \zeta_\Gamma \geq s^2) + \mathbb{P}(\left(\mathbf{x}_t^\top \zeta_\Gamma\right)^2 \geq s^2)$$
$$\overset{(a)}{\leq} \mathbb{P}(\eta_t^2 \geq s_1^2) + \mathbb{P}(2\eta_t \mathbf{x}_t^\top \zeta_\Gamma \geq s_1^2 + s_3^2) + \mathbb{P}(\left(\mathbf{x}_t^\top \zeta_\Gamma\right)^2 \geq s_3^2)$$
$$\leq 2\exp\left(-\frac{s_1^2}{c'\sigma_{\max}^2}\right) + 2\exp\left(-\frac{s_3^2}{c^{''}\sigma_{\max}^2}\right) \tag{19}$$

where (a) follows as RHS $s_1^2 < s^2$, and $s_3^2 < s^2$ (that is the RHS is smaller). Let there be some constant $C > 0$ such that $s^2/C < \max\{s_1^2/c', s_3^2/c''\}$. Then it follows that

$$\mathbb{P}(\epsilon_t \geq \mathbb{E}[\epsilon_t] + s^2) \overset{(a)}{\leq} \mathbb{P}\left(\epsilon_t \geq \frac{\sigma_{\max}^2 d}{\Gamma} + s^2\right) \leq \mathbb{P}(\epsilon_t \geq s^2) \leq 4\exp\left(-\frac{s^2}{C\sigma_{\max}^2}\right).$$

where, (a) as the RHS $\mathbb{E}[\epsilon_t] \geq \frac{\sigma_{\max}^2 d}{\Gamma}$ is smaller. This shows that ϵ_t is a sub-exponential random variable using Lemma 4. Then using Lemma 4 and $\Gamma > d^2$ we can show that

$$\mathbb{P}\left(\epsilon_t \geq \mathbb{E}[\epsilon_t] + \frac{Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma}\right) \leq 4\exp\left(-\frac{Cd^2\sigma_{\max}^2 \log(A/\delta)}{C\Gamma\sigma_{\max}^2}\right) \leq 4\frac{\delta}{A}.$$

This implies that $\epsilon_t \leq \mathbb{E}[\epsilon_t] + \frac{Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma} \leq 16\sigma_{\max}^2 + \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma}$ with probability greater than $1 - 4\frac{\delta}{A}$.

Combining all of the steps above we can show that

$$\mathbb{P}\left(\langle \Phi_m, \boldsymbol{\Sigma}_* \rangle - z_m > \frac{d}{\sqrt{n}} \sum_{t:\mathbf{X}_t = \Phi_m} \left(16\sigma_{\max}^2 + \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma}\right)\right)$$
$$\overset{(a)}{=} \mathbb{P}\left(\langle \Phi_m, \boldsymbol{\Sigma}_* \rangle - z_m > \left(16\sigma_{\max}^2 + \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma}\right)\right)$$
$$\overset{(b)}{\leq} \mathbb{P}\left(\langle \Phi_m, \boldsymbol{\Sigma}_* \rangle - z_m > \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma}\right) \leq 4\delta/A,$$

where, (a) follows by setting $\Gamma = \sqrt{n}$ and $M = d < A$ and noting that the m-th row consist of $\sqrt{n}/d$ entries. The (b) follows as the Hence the above implies that

$$\mathbb{P}\left(\mathbf{x}(a)^\top \widehat{\boldsymbol{\Sigma}}_\Gamma \mathbf{x}(a) - \mathbf{x}(a)^\top \boldsymbol{\Sigma}_* \mathbf{x}(a) \geq \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma}\right) \leq 4\delta/A.$$

Similarly, we can bound the other tail inequality as

$$\mathbb{P}\left(\mathbf{x}(a)^\top \widehat{\mathbf{\Sigma}}_\Gamma \mathbf{x}(a) - \mathbf{x}(a)^\top \mathbf{\Sigma}_* \mathbf{x}(a) \leq -\frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma}\right) \leq 4\delta/A.$$

Hence we can show by union bounding over all actions $A > d$ that

$$\mathbb{P}\left(\forall a, \left|\mathbf{x}(a)^\top \left(\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*\right)\mathbf{x}(a)\right| \geq \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma}\right) \leq 2A\frac{4\delta}{A} = 8\delta.$$

The claim of the lemma follows. □

Lemma 7. ***(Operator Norm Concentration Lemma)*** *We have that*

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*\| \geq \frac{2Cd^2\sigma_{\max}^2 \lambda_{\min}^{-1}(\mathbf{Y}) \log(A/\delta)}{\Gamma}\right) \leq 8\delta$$

for a constant $C > 0$.

Proof. Define the set of actions $\mathcal{Z}$ such that it has a span over $\mathbf{X}$ and $\mathbf{X}\mathbf{X}^\top$. Define the vector $\mathbf{y}(a) = \mathbf{x}(a)\mathbf{x}(a)^\top \in \mathbb{R}^{d^2}$. Also observe that $|\mathcal{Z}| = d^2$. Without loss of generosity, we assume that $\mathcal{Z} = \{1, 2, \ldots, d^2\}$. Now define the matrix $\mathbf{Y} \in \mathbb{R}^{d^2 \times d^2}$ such that

$$\mathbf{Y} = [\mathbf{y}(1), \mathbf{y}(2), \ldots, \mathbf{y}(|\mathcal{Z}|)]$$

We further assume that the $\lambda_{\min}(\mathbf{Y}) > 0$. We already have from Lemma 1 that

$$\begin{aligned}
&\mathbb{P}\left(\forall a \in \mathcal{A}, \left|\mathbf{x}(a)^\top (\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*)\mathbf{x}(a)\right| \leq \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma}\right) \geq 1 - 8\delta \\
\overset{(a)}{\implies} &\mathbb{P}\left(\forall a \in \mathcal{Z}, \left|\langle \widehat{\mathbf{\Sigma}}_\Gamma, \mathbf{y}(a)\rangle - \langle \mathbf{\Sigma}_*, \mathbf{y}(a)\rangle\right| \leq \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma}\right) \geq 1 - 8\delta.
\end{aligned}$$

where, (a) follows by the fact that $\mathcal{Z} \subset \mathcal{A}$. Now take an arbitrary vector $\mathbf{x}$ in unit ball such that $\|\mathbf{x}\|_2 \leq 1$. Now we define the vector $\mathbf{y} = \mathbf{x}\mathbf{x}^\top$ such that $\mathbf{y} \in \mathbb{R}^{d^2}$. Then following Assumption 2 we have that

$$\mathbf{x}\mathbf{x}^\top = \mathbf{y} = \sum_{a \in \mathcal{Z}} \alpha(a)\mathbf{y}(a) = \alpha\mathbf{Y} \overset{(a)}{\implies} \alpha = \mathbf{Y}^{-1}\mathbf{y}$$

where, in (a) we can take the inverse because $\lambda_{\min}(\mathbf{Y}) > 0$. Now we want to bound

$$\begin{aligned}
\|\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*\| = \left|\mathbf{x}^\top \left(\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*\right)\mathbf{x}\right| = \left|\langle \widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*, \mathbf{y}\rangle\right| &\leq \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma} \left\|\underbrace{\sum_a \alpha(a)}_{\alpha}\right\| \\
&= \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma} \|\mathbf{Y}^{-1}\mathbf{y}\| \\
&\leq \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma} \|\mathbf{Y}^{-1}\| \|\mathbf{x}\|^2 \\
&\leq \frac{2Cd^2\sigma_{\max}^2 \lambda_{\min}^{-1}(\mathbf{Y}) \log(A/\delta)}{\Gamma}.
\end{aligned}$$

The claim of the lemma follows. □

Corollary 2. *For, $n \geq 4C^2d^2\sigma_{\max}^2 \log^2(A/\delta)/\sigma_{\min}^2$, we have that with probability at least $1 - 8\delta$, the following holds: for all action a, $\dfrac{\sigma^2(a)}{\widehat{\sigma}_\Gamma^2(a)} \leq 1 + \frac{4Cd^2\sigma_{\max}^2 \log(A/\delta)}{\sigma_{\min}^2 \Gamma}$.*

Proof. From the Lemma 1, we know that $\left|\mathbf{x}(a)^\top(\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*)\mathbf{x}(a)\right| \leq \frac{2Cd^2\log(A/\delta)}{\Gamma}$ with probability at least $1-8\delta$. Hence we can show that

$$\begin{aligned}
|\widehat{\sigma}^2_\Gamma(a) - \sigma^2(a)| \leq \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\Gamma} &\implies \sigma^2(a) - \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\Gamma} \leq \widehat{\sigma}^2_\Gamma(a) \leq \sigma^2(a) + \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\Gamma} \\
&\implies 1 - \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2(a)\Gamma} \leq \frac{\widehat{\sigma}^2_\Gamma(a)}{\sigma^2(a)} \leq 1 + \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2(a)\Gamma} \\
&\implies 1 - \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma} \leq \frac{\widehat{\sigma}^2_\Gamma(a)}{\sigma^2(a)} \leq 1 + \frac{2Cd^2\sigma^2_{\max}\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma} \\
&\implies \frac{1}{1 + \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma}} \leq \frac{\sigma^2(a)}{\widehat{\sigma}^2_\Gamma(a)} \leq \frac{1}{1 - \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma}}.
\end{aligned}$$

It follows then that

$$\frac{\sigma^2(a)}{\widehat{\sigma}^2_\Gamma(a)} \leq \frac{1}{1 - \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma}} \overset{(a)}{\leq} 1 + \frac{4Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma}$$

where, (a) follows for $x = \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma}$ and

$$\frac{1}{1-x} \leq 1 + 2x \implies 1 \leq 1 + x - 2x^2 \implies x(1-2x) \geq 0$$

which holds for $x = \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma} \leq \frac{1}{2}$. For $n \geq 4C^2d^2\sigma^2_{\max}\log^2(A/\delta)/\sigma^2_{\min}$ we can show that $x \leq \frac{1}{2}$. The claim of the corollary follows. □

B.3 Bounding the Loss of Algorithm 1

Proposition 6. ***(Loss of Algorithm 1, formal)*** *Let $\widehat{\mathbf{b}}$ be the empirical PE-Optimal design followed by Algorithm 1 and it samples each action a as $\lceil n\widehat{\mathbf{b}}(a)\rceil$ times. Then the MSE of Algorithm 1 for for $n \geq \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma}$ is given by*

$$\overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma) \leq \underbrace{O_{\kappa^2, H_U^2}\left(\frac{d^3\lambda_1(\mathbf{V})\log n}{\sigma^2_{\min} n}\right)}_{\substack{\textbf{PE-Optimal MSE} \\ \textbf{and exploration error}}} + \underbrace{O_{\kappa^2, H_U^2}\left(\frac{d^2\lambda_1(\mathbf{V})\log n}{n^{3/2}}\right)}_{\textbf{Approximation error}} + \underbrace{O_{\kappa^2, H_U^2}\left(\frac{1}{n}\right)}_{\textbf{Failure event MSE}}.$$

Proof. Recall that the $\widehat{\mathbf{\Sigma}}_\Gamma$ be the empirical co-variance after Γ timesteps. Then Algorithm 1 pulls each action $a \in \mathcal{A}$ exactly $\left\lceil (n-\Gamma)\widehat{\mathbf{b}}(a)\right\rceil$ times for some $\sqrt{n} > A$ and computes the least squares estimator $\widehat{\boldsymbol{\theta}}_n$. Recall that the estimate $\widehat{\boldsymbol{\theta}}_n$ only uses the $(n-\Gamma)$ data sampled under $\widehat{\mathbf{b}}$. Also recall we actually use $\widehat{\mathbf{\Sigma}}_\Gamma$ as input for optimization problem (3), where $\Gamma = \sqrt{n}$. We first define the good event $\xi_\delta(n-\Gamma)$ as follows:

$$\xi_\delta(n-\Gamma) := \left\{ \left(\sum_{a=1}^A \mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_*)\right)^2 \leq \min\left\{\sqrt{\frac{(8d\lambda_1(\mathbf{V}) + \alpha_0 + \alpha)\log(1/\delta)}{n-\Gamma}}, \frac{(8d\lambda_1(\mathbf{V}) + \alpha_0 + \alpha)\log(1/\delta)}{n-\Gamma}\right\}\right\}$$

where, α_0, and α will be defined later. Also, define the good variance event as follows:

$$\xi^{var}_\delta(\Gamma) := \left\{\forall a, \left|\mathbf{x}(a)^\top\left(\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_*\right)\mathbf{x}(a)\right| < \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\Gamma}\right\}. \tag{20}$$

Then we can bound the loss of the SPEED as follows:

$$\overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma) = \mathbb{E}_{\mathcal{D}}\left[\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top \left(\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_*\right)\right)^2\right]$$
$$= \mathbb{E}_{\mathcal{D}}\left[\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top \left(\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_*\right)\right)^2 \mathbb{I}\{\xi_\delta(n-\Gamma)\}\mathbb{I}\{\xi_\delta^{var}(\Gamma)\}\right] + \mathbb{E}_{\mathcal{D}}\left[\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top \left(\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_*\right)\right)^2 \mathbb{I}\{\xi_\delta^{c}(n-\Gamma)\}\right]$$
$$+ \mathbb{E}_{\mathcal{D}}\left[\left(\sum_{a=1}^{A} \mathbf{w}(a)^\top \left(\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_*\right)\right)^2 \mathbb{I}\{(\xi_\delta^{var}(\Gamma))^c\}\right]. \tag{21}$$

Now we bound the first term of the (21). Note that using weighted least square estimates we have

$$\widehat{\boldsymbol{\theta}}_{n-\Gamma} \overset{(a)}{=} \widehat{\boldsymbol{\theta}}_n := \arg\min_{\boldsymbol{\theta}} \sum_{t=\Gamma+1}^{n} \frac{1}{\sigma^2(a_t)}(r_t - \mathbf{x}(a_t)^\top \boldsymbol{\theta})^2$$

where, in (a) we a_t is the action sampled at timestep t. Recall that the $\mathbf{diag}(\widehat{\mathbf{\Sigma}}_\Gamma) = [\widehat{\sigma}_\Gamma^2(a_1), \widehat{\sigma}_\Gamma^2(a_2), \ldots, \widehat{\sigma}_\Gamma^2(a_n)]$, where $a_1, a_2, \ldots, a_{n-\Gamma}$ are the actions pulled at time $t = \Gamma+1, 2, \ldots, n$. We have that:

$$\widehat{\boldsymbol{\theta}}_{n-\Gamma} = (\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{R}_n = (\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} (\mathbf{X}_{n-\Gamma}\boldsymbol{\theta}_* + \eta)$$
$$\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_* = (\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \eta$$

where the noise vector $\eta \sim \mathcal{SG}(0, \mathbf{\Sigma}_{n-\Gamma})$ where $\mathbf{diag}(\mathbf{\Sigma}_n) = [\sigma^2(a_1), \sigma^2(a_2), \ldots, \sigma^2(a_{n-\Gamma})]$. For any $\mathbf{z} := \sum_a \mathbf{w}(a) \in \mathbb{R}^d$ we have

$$\mathbf{z}^\top(\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_*) = \mathbf{z}^\top(\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \eta. \tag{22}$$

It implies from (22) that

$$\left(\mathbf{z}^\top(\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_*)\right)^2 \sim \mathcal{SE}\left(0, \mathbf{z}^\top(\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbb{E}\left[\eta\eta^\top\right] \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma} (\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{z}\right) \tag{23}$$

where $\mathcal{SE}$ denotes the sub-exponential distribution. Hence to bound the quantity $\left(\mathbf{z}^\top(\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_*)\right)^2$ we need to bound the variance. We first begin by rewriting the loss function for $n \geq \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\sigma_{\min}^2 \Gamma}$ as follows

$$\mathbb{E}\left[\left(\mathbf{z}^\top(\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_*)\right)^2\right] = \mathbf{z}^\top(\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbb{E}\left[\eta\eta^\top\right] \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma} (\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{z}$$
$$\overset{(a)}{=} \mathbf{z}^\top(\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{\Sigma}_n \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma} (\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{z}$$
$$= \mathbf{z}^\top(\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-\frac{1}{2}} \widehat{\mathbf{\Sigma}}_\Gamma^{-\frac{1}{2}} \mathbf{\Sigma}_n \widehat{\mathbf{\Sigma}}_\Gamma^{-\frac{1}{2}} \widehat{\mathbf{\Sigma}}_\Gamma^{-\frac{1}{2}} \mathbf{X}_{n-\Gamma} (\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{z}$$
$$\overset{(b)}{=} \underbrace{\mathbf{z}^\top(\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-\frac{1}{2}}}_{\mathbf{m}^\top \in \mathbb{R}^{n-\Gamma}} \widehat{\mathbf{\Sigma}}_\Gamma^{-\frac{1}{2}} \mathbf{\Sigma}_n \widehat{\mathbf{\Sigma}}_\Gamma^{-\frac{1}{2}} \underbrace{\widehat{\mathbf{\Sigma}}_\Gamma^{-\frac{1}{2}} \mathbf{X}_{n-\Gamma} (\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{z}}_{\mathbf{m} \in \mathbb{R}^{n-\Gamma}}$$
$$\overset{(c)}{\leq} \mathbf{z}^\top(\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1/2} \left(\left(1 + 2C_{\Gamma, \sigma_{\min}^2}(\delta)\right) \mathbf{I}_n\right) \widehat{\mathbf{\Sigma}}_\Gamma^{-1/2} \mathbf{X}_{n-\Gamma} (\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{z}$$
$$\overset{(d)}{=} \left(1 + 2C_{\Gamma, \sigma_{\min}^2}(\delta)\right) \mathbf{z}^\top(\mathbf{X}_{n-\Gamma}^\top \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{X}_{n-\Gamma})^{-1} \mathbf{z} \tag{24}$$

where, (a) follows as $\mathbb{E}\left[\eta\eta^\top\right] = \mathbf{\Sigma}_n$, in (b) $\mathbf{m}$ is a vector in $\mathbb{R}^{n-\Gamma}$. The (c) follows by first observing that

$$\widehat{\mathbf{\Sigma}}_\Gamma^{-\frac{1}{2}} \mathbf{\Sigma}_n \widehat{\mathbf{\Sigma}}_\Gamma^{-\frac{1}{2}} = \widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{\Sigma}_n = \mathbf{diag}(\widehat{\mathbf{\Sigma}}_\Gamma^{-1} \mathbf{\Sigma}_n) = \left[\frac{\sigma^2(I_1)}{\widehat{\sigma}_\Gamma^2(I_1)}, \frac{\sigma^2(I_2)}{\widehat{\sigma}_\Gamma^2(I_2)}, \ldots, \frac{\sigma^2(I_n)}{\widehat{\sigma}_\Gamma^2(I_n)}\right].$$

Then note that using Corollary 2 we have

$$\frac{\sigma^2(I_t)}{\widehat{\sigma}^2_\Gamma(I_t)} \leq 1 + 2 \cdot \underbrace{\frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma}}_{:= C_{\Gamma,\sigma^2_{\min}}(\delta)}$$

for each $t \in [n]$, and (d) follows as $1 + 2C_{\Gamma,\sigma^2_{\min}}(\delta)$ is not a random variable. Let $\widehat{\mathbf{b}}^*$ be the empirical PE-Optimal design returned by the approximator after it is supplied with $\widehat{\mathbf{\Sigma}}_\Gamma$. Now observe that the quantity of the samples collected (following $\widehat{\mathbf{b}}^*$) after exploration is as follows:

$$\left(\widetilde{\mathbf{X}}^\top_{n-\Gamma}\widehat{\mathbf{\Sigma}}^{-1}_\Gamma\widetilde{\mathbf{X}}_{n-\Gamma}\right)^{-1} = \left(\sum_a \left\lceil (n-\Gamma)\widehat{\mathbf{b}}^*(a)\widehat{\sigma}^{-2}_\Gamma(a)\right\rceil \mathbf{w}(a)\mathbf{w}(a)^\top\right)^{-1} = \frac{1}{n-\Gamma}\mathbf{A}^{-1}_{\widehat{\mathbf{b}}^*,\widehat{\mathbf{\Sigma}}_\Gamma}.$$

Hence we use the loss function

$$\mathcal{L}'_{n-\Gamma}(\pi,\widehat{\mathbf{b}},\widehat{\mathbf{\Sigma}}_\Gamma) := \left(1+2C_{\Gamma,\sigma^2_{\min}}(\delta)\right)\mathbf{z}^\top(\widetilde{\mathbf{X}}^\top_{n-\Gamma}\widehat{\mathbf{\Sigma}}^{-1}_\Gamma\widetilde{\mathbf{X}}_{n-\Gamma})^{-1}\mathbf{z} = \frac{\left(1+2C_{\Gamma,\sigma^2_{\min}}(\delta)\right)}{n-\Gamma}\sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\widehat{\mathbf{b}}^*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a'). \tag{25}$$

Also recall that we define

$$\mathcal{L}_n(\pi,\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma) = \frac{1}{n}\sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a').$$

So to minimize the quantity $\mathbb{E}\left[\left(\sum_a \mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_{n-\Gamma}-\boldsymbol{\theta}_*)\right)^2\right]$ is minimizing the quantity $\frac{\left(1+2C_{\Gamma,\sigma^2_{\min}}(\delta)\right)}{n-\Gamma}\sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\widehat{\mathbf{b}}^*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a')$. Further recall that we can show that from Assumption 3 (approximation oracle) and Kiefer-Wolfowitz theorem in Corollary 1 that for the proportion $\mathbf{b}_*$ and any arbitrary positive semi-definite matrix $\widehat{\mathbf{\Sigma}}_\Gamma$ the following holds

$$\begin{aligned}\sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a') &= \mathbf{Tr}\left(\sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a')\right) = \mathbf{Tr}\Big(\mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\underbrace{\sum_{a,a'}\mathbf{w}(a)\mathbf{w}(a')^\top}_{\mathbf{V}}\Big)\\ &= \mathbf{Tr}\left(\mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{V}\right) \leq d\lambda_1(\mathbf{V}).\end{aligned} \tag{26}$$

Then we can decompose the loss as follows:

$$\begin{aligned}\mathcal{L}'_{n-\Gamma}(\pi,\widehat{\mathbf{b}},\widehat{\mathbf{\Sigma}}_\Gamma) &= \mathcal{L}'_{n-\Gamma}(\pi,\widehat{\mathbf{b}},\widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}'_{n-\Gamma}(\pi,\widehat{\mathbf{b}}^*,\widehat{\mathbf{\Sigma}}_\Gamma) + \mathcal{L}'_{n-\Gamma}(\pi,\widehat{\mathbf{b}}^*,\widehat{\mathbf{\Sigma}}_\Gamma)\\ &= \underbrace{\mathcal{L}'_{n-\Gamma}(\pi,\widehat{\mathbf{b}},\widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}'_{n-\Gamma}(\pi,\widehat{\mathbf{b}}^*,\widehat{\mathbf{\Sigma}}_\Gamma)}_{\textbf{Approximation error}} + \underbrace{\mathcal{L}'_{n-\Gamma}(\pi,\widehat{\mathbf{b}}^*,\widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}_n(\pi,\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma)}_{\textbf{Comparing two diff loss}} + \mathcal{L}_n(\pi,\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma).\end{aligned} \tag{27}$$

For the approximation error we need access to an oracle (see Assumption 3) that gives ϵ approximation error. Then setting $\epsilon = \frac{1}{\sqrt{n}}$ we have that

$$\begin{aligned}\mathcal{L}'_{n-\Gamma}(\pi,\widehat{\mathbf{b}},\widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}'_{n-\Gamma}(\pi,\widehat{\mathbf{b}}^*,\widehat{\mathbf{\Sigma}}_\Gamma) &= \frac{\left(1+2C_{\Gamma,\sigma^2_{\min}}(\delta)\right)}{n-\Gamma}\underbrace{\mathbf{Tr}\left(\sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\widehat{\mathbf{b}},\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a') - \sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\widehat{\mathbf{b}}^*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a')\right)}_{\epsilon}\\ &\overset{(a)}{\leq} O_{\kappa^2,H^2_U}\left(\frac{d^2\sigma^2_{\max}\log(A/\delta)}{n^{3/2}}\right)\end{aligned} \tag{28}$$

where, (a) follows by setting $\Gamma = \sqrt{n}$, $\epsilon = 1/\sqrt{n}$ and $C_{\Gamma,\sigma^2_{\min}}(\delta) = \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma} = \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\sqrt{n}}$. Let us define $\mathbf{K}_1 \coloneqq \mathbf{Tr}(\sum_{a,a'}\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\widehat{\mathbf{b}}^*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a'))$, and $\mathbf{K}_2 \coloneqq \mathbf{Tr}(\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a'))$. For the second part of comparing the two losses we can show that

$$
\begin{aligned}
\mathcal{L}'_{n-\Gamma}(\pi,\widehat{\mathbf{b}}^*,\widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}_n(\pi,\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma) &= \frac{1}{(n-\Gamma)}\mathbf{Tr}\left(\left(1+2C_{\Gamma,\sigma^2_{\min}}(\delta)\right)K_1\right) - \frac{1}{n}K_2 \\
&= \frac{(1+2C_{\Gamma,\sigma^2_{\min}}(\delta))\mathbf{K}_1}{n-\Gamma} - \frac{(1+2C_{\Gamma,\sigma^2_{\min}}(\delta))\mathbf{K}_2}{n-\Gamma} + \frac{(1+2C_{\Gamma,\sigma^2_{\min}}(\delta))\mathbf{K}_2}{n-\Gamma} - \frac{1}{n}\mathbf{K}_2 \\
&= \frac{(1+2C_{\Gamma,\sigma^2_{\min}}(\delta))}{n-\Gamma}(\mathbf{K}_1-\mathbf{K}_2) + \frac{2C_{\Gamma,\sigma^2_{\min}}(\delta)\mathbf{K}_2}{n-\Gamma} + \frac{1}{n-\Gamma}\mathbf{K}_2 - \frac{1}{n}\mathbf{K}_2 \\
&\overset{(a)}{=} \frac{\Gamma}{n(n-\Gamma)}\underbrace{\mathbf{Tr}\left(\sum_{a,a'}\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\widehat{\mathbf{b}}^*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a') - \sum_{a,a'}\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a')\right)}_{\leq 0} \\
&\quad + \frac{2C_{\Gamma,\sigma^2_{\min}}(\delta)}{n-\Gamma}\mathbf{Tr}\left(\sum_{a,a'}\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a')\right) + \frac{\Gamma}{n(n-\Gamma)}\mathbf{Tr}\left(\sum_{a,a'}\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a')\right) \\
&\overset{(b)}{\leq} O_{\kappa^2,H_U^2}\left(\frac{d^3\sigma^2_{\max}\lambda_1(\mathbf{V})\log(A/\delta)}{\sigma^2_{\min}n^{3/2}}\right)
\end{aligned} \tag{29}
$$

where, (a) follows by substituting the definition of $\mathbf{K}_1$ and $\mathbf{K}_2$. The (b) follows by setting $\Gamma = \sqrt{n}$, $C_{\Gamma,\sigma^2_{\min}}(\delta) = \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma} = \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\sqrt{n}}$, and $\mathbf{Tr}\left(\sum_{a,a'}\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a')\right) \leq d\lambda_1(\mathbf{V})$.

Now we combine all parts together in (27) using (26), (28) and (29). First we define the quantity

$$
\alpha \coloneqq 2C_{\Gamma,\sigma^2_{\min}}(\delta)\mathbf{Tr}\left(\sum_{a,a'}\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a')\right) + \frac{\Gamma}{n}\mathbf{Tr}\left(\sum_{a,a'}\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a')\right).
$$

It follows then that (27) can be written as

$$
\begin{aligned}
\frac{1+2C_{\Gamma,\sigma^2_{\min}}(\delta)}{n-\Gamma}\sum_{a,a'}\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\widehat{\mathbf{b}},\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a') &\leq \underbrace{\frac{(1+2C_\Gamma(\delta))\epsilon}{(n-\Gamma)}}_{\textbf{Approximation error}} + \frac{\alpha}{n-\Gamma} + \frac{1}{n}\sum_{a,a'}\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a') \\
\implies (1+2C_{\Gamma,\sigma^2_{\min}}(\delta))\sum_{a,a'}\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\widehat{\mathbf{b}},\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a') &\leq \underbrace{(1+2C_\Gamma(\delta))\epsilon}_{\alpha_0} + \alpha + \frac{n-\Gamma}{n}\sum_{a,a'}\mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\mathbf{\Sigma}}_\Gamma}\mathbf{w}(a') \\
&\overset{(a)}{\leq} \alpha_0 + \alpha + d\lambda_1(\mathbf{V})
\end{aligned} \tag{30}
$$

where, (a) follows from Assumption 3, Corollary 1 and (26). Also observe that from (23) we have that $\left(\sum_{a=1}^A \mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*)\right)^2$ is a sub-exponential random variable. Then using the sub-exponential concentration inequality we have with probability at least $1-\delta$

$$\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_{n-\Gamma}-\boldsymbol{\theta}_*)\right)^2 \le \min\Bigg\{\sqrt{(1+2C_{\Gamma,\sigma^2_{\min}}(\delta))\sum_{a,a'}\mathbf{w}(a)\left(\mathbf{X}_{n-\Gamma}^\top\widehat{\boldsymbol{\Sigma}}_\Gamma^{-1}\mathbf{X}_{n-\Gamma}\right)^{-1}\mathbf{w}(a')2\log(1/\delta)},$$
$$(1+2C_\Gamma(\delta))\sum_{a,a'}\mathbf{w}(a)^\top\left(\mathbf{X}_{n-\Gamma}^\top\widehat{\boldsymbol{\Sigma}}_\Gamma^{-1}\mathbf{X}_{n-\Gamma}\right)^{-1}\mathbf{w}(a')2\log(1/\delta)\Bigg\}$$
$$= \min\Bigg\{\frac{1}{\sqrt{n-\Gamma}}\sqrt{(1+2C_{\Gamma,\sigma^2_{\min}}(\delta))\sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\mathbf{b},\widehat{\boldsymbol{\Sigma}}_\Gamma}\mathbf{w}(a')2\log(1/\delta)},$$
$$\frac{(1+2C_{\Gamma,\sigma^2_{\min}}(\delta))}{n-\Gamma}\sum_{a,a'}\mathbf{w}(a)^\top\mathbf{A}^{-1}_{\mathbf{b},\widehat{\boldsymbol{\Sigma}}_\Gamma}\mathbf{w}(a')2\log(1/\delta)\Bigg\}$$
$$\overset{(a)}{\le}\min\left\{\sqrt{\frac{(8d\lambda_1(\mathbf{V})+\alpha_0+\alpha)\log(1/\delta)}{n-\Gamma}},\frac{(8d\lambda_1(\mathbf{V})+\alpha_0+\alpha)\log(1/\delta)}{n-\Gamma}\right\}$$

where, (a) follows from (30), and we have taken at most $n-\Gamma$ pulls to estimate $\widehat{\boldsymbol{\theta}}_n$ after forced exploration and $\sqrt{n}>d$. Thus, for any $\delta\in(0,1)$ we have

$$\mathbb{P}\left(\left\{\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_*)\right)^2>\min\left\{\sqrt{\frac{(8d\lambda_1(\mathbf{V})+\alpha_0+\alpha)\log(1/\delta)}{n-\Gamma}},\frac{(8d\lambda_1(\mathbf{V})+\alpha_0+\alpha)\log(1/\delta)}{n-\Gamma}\right\}\right\}\right)\le\delta. \tag{31}$$

This gives us a bound on the first term of (21). Combining everything in (21) we can bound the loss of the SPEED as follows:

$$\overline{\mathcal{L}}_n(\pi,\widehat{\mathbf{b}},\widehat{\boldsymbol{\Sigma}}_\Gamma)\le\mathbb{E}_{\mathcal{D}}\left[\left(\sum_{a=1}^{A}\mathbf{w}(a)^\top\left(\widehat{\boldsymbol{\theta}}_{n-\Gamma}-\boldsymbol{\theta}_*\right)\right)^2\mathbb{I}\{\xi_\delta(n-\Gamma)\}\mathbb{I}\{\xi^{var}_\delta(\Gamma)\}\right]+\sum_{t=1}^{n}AH_U^2\eta^2\mathbb{P}(\xi^c_\delta(n-\Gamma))$$
$$+\sum_{t=1}^{n}AH_U^2\eta^2\mathbb{P}\left((\xi^{var}_\delta(\Gamma))^c\right)$$
$$\le\min\left\{\frac{2Cd^2\log(A/\delta)}{\Gamma},\sqrt{\frac{(8d\lambda_1(\mathbf{V})+\alpha_0+\alpha)\log(A/\delta)}{n-\Gamma}},\frac{(8d\lambda_1(\mathbf{V})+\alpha_0+\alpha)\log(A/\delta)}{n-\Gamma}\right\}$$
$$+\sum_{t=1}^{n}AH_U^2\eta^2\mathbb{P}(\xi^c_\delta(n-\Gamma))+\sum_{t=1}^{n}AH_U^2\eta^2\mathbb{P}\left((\xi^{var}_\delta(\Gamma))^c\right)$$
$$\overset{(a)}{\le}\min\left\{\frac{8Cd^2\sigma^2_{\max}\log(nA)}{\sqrt{n}},\sqrt{\frac{48(d\lambda_1(\mathbf{V})+\alpha_0+\alpha)\log(nA)}{n}},\frac{48(d\lambda_1(\mathbf{V})+\alpha_0+\alpha)\log(nA)}{n}\right\}+O\left(\frac{1}{n}\right)$$
$$\le\frac{48d^2\sigma^2_{\max}\lambda_1(\mathbf{V})\log(nA)}{n}+\frac{48\alpha\log(nA)}{n}+\frac{48\alpha_0\log(nA)}{n}+O\left(\frac{1}{n}\right)$$
$$\overset{(b)}{\le}\frac{48d^2\sigma^2_{\max}\lambda_1(\mathbf{V})\log(nA)}{n}+\frac{144d\lambda_1(\mathbf{V})C_{\Gamma,\sigma^2_{\min}}(\delta)\log(nA)}{n}+\frac{48d\lambda_1(\mathbf{V})\Gamma\log(nA)}{n^{3/2}}+\frac{48\epsilon\log(nA)}{n}+O\left(\frac{1}{n}\right)$$

where (a) follows as Proposition 6 and setting $\delta=1/n^3$ and noting that $\sqrt{n}>d$. The (b) follows by setting $(1+2C_\Gamma(\delta))\epsilon$ and the definition of α. Recall that for $\Gamma=\sqrt{n}$ we have that $C_{\Gamma,\sigma^2_{\min}}(\delta)=\frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma}=\frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\sqrt{n}}$. Then setting $\epsilon=1/\sqrt{n}$ we can bound the loss of the following PE-Optimal $\widehat{\mathbf{b}}$ as

$$\overline{\mathcal{L}}_n(\pi,\widehat{\mathbf{b}},\widehat{\boldsymbol{\Sigma}}_\Gamma)\le O_{\kappa^2,H_U^2}\left(\frac{d^3\sigma^2_{\max}\lambda_1(\mathbf{V})\log(nA)}{\sigma^2_{\min}n}\right)+O_{\kappa^2,H_U^2}\left(\frac{d^2\sigma^2_{\max}\lambda_1(\mathbf{V})\log(nA)}{n^{3/2}}\right)+O_{\kappa^2,H_U^2}\left(\frac{1}{n}\right).$$

The claim of the proposition follows. □

Remark 3. (Discussion on loss) Observe that from Proposition 6 that the MSE for policy evaluation setting scales as $O(\frac{d^3\log(n)}{n})$. We contrast this result with Chaudhuri et al. [2017] who obtain a bound on the MSE $\mathbb{E}_{\mathcal{D}}[\|\boldsymbol{\theta}_* - \widehat{\boldsymbol{\theta}}_n\|^2] \leq O(\frac{d\log(n)}{n})$ in a related setting. Note that Chaudhuri et al. [2017] only considers the setting when $\boldsymbol{\Sigma}_*$ is rank 1. We make no such assumption and get an additional factor of d in our result due to exploration in d^2 dimension to estimate $\boldsymbol{\Sigma}_*$. Finally we get the scaling as d^3 due to $\sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma} \mathbf{w}(a') \leq d\lambda_1(\mathbf{V})$ from Corollary 1. Also observe that we estimate $\mathbb{E}_{\mathcal{D}}[\sum_a \mathbf{w}(a)^\top(\boldsymbol{\theta}_* - \widehat{\boldsymbol{\theta}}_n)^2]$ as opposed to $\mathbb{E}_{\mathcal{D}}[\|\boldsymbol{\theta}_* - \widehat{\boldsymbol{\theta}}_n\|^2]$ in Chaudhuri et al. [2017].

B.4 Regret of Algorithm 1

Corollary 3. *For, $n \geq 16C^2d^4\sigma^4_{\max}\log^2(A/\delta)/\sigma^4_{\min}$ we have that for all action a, $|\widehat{\sigma}^2_\Gamma(a) - \sigma^2(a)| \leq \sigma^2_{\min}/2$.*

Proof. From the Lemma 1, we know that $\left|\mathbf{x}(a)^\top(\widehat{\boldsymbol{\Sigma}}_\Gamma - \boldsymbol{\Sigma}_*)\mathbf{x}(a)\right| \leq \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\Gamma}$ with probability $1-8\delta$. Hence we can show that

$$|\widehat{\sigma}^2_\Gamma(a) - \sigma^2(a)| \leq \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\Gamma} = \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sqrt{n}} \overset{(a)}{\leq} \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sqrt{16C^2d^4\sigma^4_{\max}\log^2(A/\delta)/\sigma^4_{\min}}} = \frac{\sigma^2_{\min}}{2},$$

where (a) follows for $n \geq 16C^2d^4\sigma^4_{\max}\log^2(A/\delta)/\sigma^4_{\min}$. The claim of the corollary follows. □

Lemma 8. ***(Loss Concentration of design matrix)*** *Let $\widehat{\boldsymbol{\Sigma}}_\Gamma$ be the empirical estimate of $\boldsymbol{\Sigma}_*$. Define $\mathbf{V} = \sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top$. We have that for any arbitrary proportion $\mathbf{b}$ the following*

$$\mathbb{P}\left(\left|\sum_{a,a'} \mathbf{w}(a)^\top(\mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma} - \mathbf{A}^{-1}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*})\mathbf{w}(a')\right| \leq \frac{2CB^*d^3\log(A/\delta)}{\Gamma}\right) \geq 1-\delta$$

where B^* is a problem-dependent quantity such that

$$B^* = \left(\left\|\mathbf{A}^{-1}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*}\mathbf{w}\right\|^2 \left\|\sum_{a=1}^A \mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top H_U^2\right\| \left\|\left(\sum_{a=1}^A \frac{\mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top}{\sigma^2(a) + \frac{2Cd^2\sigma^2_{\max}\log(9H_U^2/\delta)}{\sqrt{n}}}\right)^{-1}\mathbf{w}\right\|\right)$$

and $C > 0$ is a universal constant.

Proof. We have the following

$$\left|\sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma}\mathbf{w}(a') - \sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*}\mathbf{w}(a')\right| = \left|\underbrace{\sum_a \mathbf{w}(a)^\top}_{\mathbf{w}}\left(\mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma} - \mathbf{A}^{-1}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*}\right)\underbrace{\sum_a \mathbf{w}(a)}_{\mathbf{w}}\right|$$

$$= \left|\mathbf{w}^\top\left(\mathbf{A}^{-1}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*}\left(\mathbf{A}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*} - \mathbf{A}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma}\right)\mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma}\right)\mathbf{w}\right| = \left|\underbrace{\mathbf{w}^\top(\,\mathbf{A}^{-1}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*}}_{\mathbf{u}}\left(\mathbf{A}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*} - \mathbf{A}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma}\right)\underbrace{\mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma}\,)\,\mathbf{w}}_{\mathbf{v}}\right|$$

$$= \left|\mathbf{u}\left(\mathbf{A}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*} - \mathbf{A}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma}\right)\mathbf{v}\right| \overset{(a)}{\leq} \|\mathbf{u}\|\underbrace{\left\|\mathbf{A}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*} - \mathbf{A}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma}\right\|}_{\Delta}\|\mathbf{v}\| \tag{32}$$

where, (a) follows by Cauchy-Schwarz inequality. Now observe that the vector $\mathbf{u} \in \mathbb{R}^d$ is a problem dependent

quantity. We now bound the Δ in (32) as follows

$$\begin{aligned}
\Delta &= \left\| \sum_{a=1}^{A} \frac{\mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top}{\mathbf{x}(a)^\top \boldsymbol{\Sigma}_* \mathbf{x}(a)} - \sum_{a=1}^{A} \frac{\mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top}{\mathbf{x}(a)^\top \widehat{\boldsymbol{\Sigma}}_\Gamma \mathbf{x}(a)} \right\| \\
&\overset{(a)}{=} \left\| \sum_{a} \frac{\mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top}{\sigma^2(a)} - \sum_{a} \frac{\mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top}{\widehat{\sigma}^2_\Gamma(a)} \right\| \\
&= \left\| \sum_{a} \mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top \left(\frac{1}{\sigma^2(a)} - \frac{1}{\widehat{\sigma}^2_\Gamma(a)} \right) \right\| \\
&= \left\| \sum_{a} \mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top \left(\frac{\widehat{\sigma}^2_\Gamma(a) - \sigma^2(a)}{\widehat{\sigma}^2_\Gamma(a)\sigma^2(a)} \right) \right\| \\
&\overset{(b)}{\leq} \left\| \sum_{a} \mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top \left(\frac{\widehat{\sigma}^2_\Gamma(a) - \sigma^2(a)}{\sigma^4_{\min}} \right) \right\| \\
&= \left\| \frac{1}{\sigma^4_{\min}} \sum_{a} \mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top \left(\mathbf{x}(a)^\top \widehat{\boldsymbol{\Sigma}}_\Gamma \mathbf{x}(a) - \mathbf{x}(a)^\top \boldsymbol{\Sigma}_* \mathbf{x}(a) \right) \right\| \\
&= \frac{1}{\sigma^4_{\min}} \left\| \sum_{a=1}^{A} \underbrace{\mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top}_{\textbf{Problem dependent quantity}} \underbrace{\left(\mathbf{x}(a)^\top \left(\widehat{\boldsymbol{\Sigma}}_\Gamma - \boldsymbol{\Sigma}_* \right) \mathbf{x}(a) \right)}_{\textbf{Random Quantity}} \right\|
\end{aligned}$$

where, (a) follows $\widehat{\sigma}^2_\Gamma(a) = \mathbf{x}(a)^\top \widehat{\boldsymbol{\Sigma}}_\Gamma \mathbf{x}(a)$ and $\sigma^2(a) = \mathbf{x}(a)^\top \boldsymbol{\Sigma}_* \mathbf{x}(a)$, ad (b) follows from Corollary 3. Now observe from Lemma 7 that we can bound the quantity

$$\|\widehat{\boldsymbol{\Sigma}}_\Gamma - \boldsymbol{\Sigma}_*\| \leq \frac{2Cd^2\sigma^2_{\max}\lambda^{-1}_{\min}(\mathbf{Y})\log(A/\delta)}{\Gamma}.$$

Then we also have that the spread of maximum eigenvalue of $\|\widehat{\boldsymbol{\Sigma}}_\Gamma - \boldsymbol{\Sigma}_*\|_2$ is controlled which implies

$$\begin{aligned}
&\frac{1}{\sigma^4_{\min}} \left\| \sum_{a=1}^{A} \underbrace{\mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top}_{\textbf{Problem dependent quantity}} \underbrace{\left(\mathbf{x}(a)^\top \left(\widehat{\boldsymbol{\Sigma}}_\Gamma - \boldsymbol{\Sigma}_* \right) \mathbf{x}(a) \right)}_{\textbf{Random Quantity}} \right\| \\
&\qquad \overset{(a)}{\leq} \left\| \sum_{a=1}^{A} \mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top (\mathbf{x}(a)^\top \mathbf{x}(a)) \right\| \frac{2Cd^2\sigma^2_{\max}\lambda^{-1}_{\min}(\mathbf{Y})\log(A/\delta)}{\Gamma}
\end{aligned}$$

where, (a) follows by Lemma 7. Next for the third quantity in (32) we can bound as follows

$$\|\mathbf{v}\| = \|\mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma}\mathbf{w}\| = \left\| \left(\sum_{a=1}^{A} \frac{\mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top}{\widehat{\sigma}^2_\Gamma(a)} \right)^{-1} \mathbf{w} \right\| \overset{(a)}{\leq} \left\| \left(\sum_{a=1}^{A} \frac{\mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top}{\sigma^2(a) + \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sqrt{n}}} \right)^{-1} \mathbf{w} \right\|$$

where, (a) follows as

$$\widehat{\sigma}^2(a) \leq \sigma^2(a) + \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\Gamma}$$

from Lemma 1. Finally observe that the first part of (32) we have that $\mathbf{w}^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*}$ is a problem dependent

parameter. Finally, plugging back everything in (32) we get

$$
\begin{aligned}
&\|\mathbf{u}\| \left\| \mathbf{A}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*} - \mathbf{A}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma} \right\| \|\mathbf{v}\| \\
&\quad \le \left\| \mathbf{A}^{-1}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*} \mathbf{w} \right\| \left\| \sum_{a=1}^{A} \mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top (\mathbf{x}(a)^\top \mathbf{x}(a)) \right\| \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^4_{\min}\Gamma} \left\| \left(\sum_{a=1}^{A} \frac{\mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top}{\sigma^2(a) + \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\Gamma}} \right)^{-1} \mathbf{w} \right\| \\
&\quad \le \underbrace{\left(\left\| \mathbf{A}^{-1}_{\mathbf{b}_*,\boldsymbol{\Sigma}_*} \mathbf{w} \right\|^2 \left\| \sum_{a=1}^{A} \mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top H_U^2 \right\| \left\| \left(\sum_{a=1}^{A} \frac{\mathbf{b}_*(a)\mathbf{w}(a)\mathbf{w}(a)^\top}{\sigma^2(a) + \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\Gamma}} \right)^{-1} \mathbf{w} \right\| \right)}_{B^*} \frac{2Cd^3\log(A/\delta)}{\Gamma} \\
&\quad \overset{(a)}{=} \frac{2CB^* d^3 \sigma^2_{\max} \lambda^{-1}_{\min}(\mathbf{Y}) \log(A/\delta)}{\Gamma}
\end{aligned}
$$

where, (a) follows by substituting the value of B^*. □

B.5 Regret Bound of SPEED

Theorem 1. ***(formal)*** *Running Algorithm 1 with budget $n \ge 16C^2d^4\log^2(A/\delta)/\sigma^4_{\min}$ the resulting regret satisfies*

$$
\begin{aligned}
\mathcal{R}_n &\le \frac{1}{n^{3/2}} + O_{\kappa^2,H_U^2}\left(\frac{d^2\sigma^2_{\max}\log(n)}{\sigma^2_{\min} n^{3/2}} \right) + \frac{2B^*Cd^3\sigma^2_{\max}\log(n)}{\sigma^2_{\min} n^{3/2}} + \frac{d^2}{n^2}\mathbf{Tr}\left(\sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top \right) + \frac{2AH_U^2\kappa^2}{n^2} \\
&= O_{\kappa^2,H_U^2}\left(\frac{B^* d^3 \sigma^2_{\max}\log(n)}{\sigma^2_{\min} n^{3/2}} \right).
\end{aligned}
$$

Proof. We follow the same steps as in Proposition 6. Observe that $\frac{16C^2d^4\sigma^4_{\max}\log^2(A/\delta)}{\sigma^4_{\min}} > \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma}$. Hence for $\mathbf{z} = \sum_a \mathbf{w}(a)$ the loss function for $n \ge \frac{2Cd^2\sigma^2_{\max}\log(A/\delta)}{\sigma^2_{\min}\Gamma}$ as follows

$$
\overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\Sigma}}_\Gamma) \coloneqq \mathbb{E}\left[\left(\mathbf{z}^\top(\widehat{\boldsymbol{\theta}}_{n-\Gamma} - \boldsymbol{\theta}_*) \right)^2 \right] \overset{(a)}{\le} \left(1 + 2C_{\Gamma,\sigma^2_{\min}}(\delta) \right) \mathbf{z}^\top (\widetilde{\mathbf{X}}^\top_{n-\Gamma} \widehat{\boldsymbol{\Sigma}}^{-1}_\Gamma \widetilde{\mathbf{X}}_{n-\Gamma})^{-1} \mathbf{z}.
$$

where, (a) follows from (24). Recall that the quantity of the samples collected (following $\widehat{\mathbf{b}}^*$) after exploration is as follows:

$$
\left(\widetilde{\mathbf{X}}^\top_{n-\Gamma} \widehat{\boldsymbol{\Sigma}}^{-1}_\Gamma \widetilde{\mathbf{X}}_{n-\Gamma} \right)^{-1} = \left(\sum_a \left\lceil (n-\Gamma)\widehat{\mathbf{b}}^*(a)\widehat{\sigma}^{-2}_\Gamma(a) \right\rceil \mathbf{w}(a)\mathbf{w}(a)^\top \right)^{-1} = \frac{1}{n-\Gamma} \mathbf{A}^{-1}_{\widehat{\mathbf{b}}^*,\widehat{\boldsymbol{\Sigma}}_\Gamma}.
$$

Hence we use the loss function

$$
\mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\Sigma}}_\Gamma) \coloneqq (1 + 2C_\Gamma(\delta))\, \mathbf{z}^\top (\widetilde{\mathbf{X}}^\top_{n-\Gamma} \widehat{\boldsymbol{\Sigma}}^{-1}_\Gamma \widetilde{\mathbf{X}}_{n-\Gamma})^{-1} \mathbf{z} = \frac{\left(1 + 2C_{\Gamma,\sigma^2_{\min}}(\delta) \right)}{n-\Gamma} \sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}^{-1}_{\widehat{\mathbf{b}}^*,\widehat{\boldsymbol{\Sigma}}_\Gamma} \mathbf{w}(a').
$$

Also, recall that we define

$$
\mathcal{L}_n(\pi, \mathbf{b}_*, \widehat{\boldsymbol{\Sigma}}_\Gamma) = \frac{1}{n} \sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}^{-1}_{\mathbf{b}_*,\widehat{\boldsymbol{\Sigma}}_\Gamma} \mathbf{w}(a').
$$

Then we can decompose the regret as follows:

$$
\begin{aligned}
\mathcal{R}_n &= \overline{\mathcal{L}}_n(\pi, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\Sigma}}_\Gamma) - \mathcal{L}^*_n(\pi, \mathbf{b}_*, \boldsymbol{\Sigma}_*) \\
&\le \mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\Sigma}}_\Gamma) - \mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}^*, \widehat{\boldsymbol{\Sigma}}_\Gamma) + \mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}^*, \widehat{\boldsymbol{\Sigma}}_\Gamma) - \mathcal{L}^*_n(\pi, \mathbf{b}_*, \boldsymbol{\Sigma}_*) \\
&= \underbrace{\mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\Sigma}}_\Gamma) - \mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}^*, \widehat{\boldsymbol{\Sigma}}_\Gamma)}_{\textbf{Approximation error}} + \underbrace{\mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}^*, \widehat{\boldsymbol{\Sigma}}_\Gamma) - \mathcal{L}_n(\pi, \mathbf{b}_*, \widehat{\boldsymbol{\Sigma}}_\Gamma)}_{\textbf{Comparing two diff loss}} + \underbrace{\mathcal{L}_n(\pi, \mathbf{b}_*, \widehat{\boldsymbol{\Sigma}}_\Gamma) - \mathcal{L}^*_n(\pi, \mathbf{b}_*, \boldsymbol{\Sigma}_*)}_{\textbf{Estimation error of } \boldsymbol{\Sigma}_*}
\end{aligned}
$$

First recall that the good variance event as follows:

$$\xi_\delta^{var}(\Gamma) \coloneqq \left\{ \forall a, \left| \mathbf{x}(a)^\top \left(\widehat{\mathbf{\Sigma}}_\Gamma - \mathbf{\Sigma}_* \right) \mathbf{x}(a) \right| < \frac{2Cd^2\sigma_{\max}^2 \log(A/\delta)}{\Gamma} \right\}.$$

Now first observe that $n \geq 16C^2d^4\sigma_{\max}^4 \log^2(A/\delta)/\sigma_{\min}^4$ is a larger regime than $n \geq \frac{2Cd^2\sigma_{\max}^2\log(A/\delta)}{\sigma_{\min}^2\Gamma}$ required for Proposition 6. Then under the good variance event, following the same steps as Proposition 6 we can bound the approximation error setting $\delta = 1/n^3$ as follows

$$\begin{aligned}
\mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}^*, \widehat{\mathbf{\Sigma}}_\Gamma) &\leq O_{\kappa^2, H_U^2}\left(\frac{d^2\sigma_{\max}^2 \log(A/\delta)}{\sigma_{\min}^2 n^{3/2}} \right) \mathbb{I}\{\xi_\delta^{var}(\Gamma)\} + \sum_{t=1}^n AH_U^2\kappa^2 \mathbb{P}\left((\xi_\delta^{var}(\Gamma))^c \right) \\
&\leq O_{\kappa^2, H_U^2}\left(\frac{d^2\sigma_{\max}^2 \log(A/\delta)}{\sigma_{\min}^2 n^{3/2}} \right) + \frac{AH_U^2\kappa^2}{n^2}
\end{aligned}$$

and the second part of comparing the two losses as

$$\begin{aligned}
\mathcal{L}'_{n-\Gamma}(\pi, \widehat{\mathbf{b}}^*, \widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}_n(\pi, \mathbf{b}_*, \widehat{\mathbf{\Sigma}}_\Gamma) &\leq O_{\kappa^2, H_U^2}\left(\frac{d^2\sigma_{\max}^2 \log(A/\delta)}{\sigma_{\min}^2 n^{3/2}} \right) \mathbb{I}\{\xi_\delta^{var}(\Gamma)\} + \sum_{t=1}^n AH_U^2\kappa^2 \mathbb{P}\left((\xi_\delta^{var}(\Gamma))^c \right) \\
&\leq O_{\kappa^2, H_U^2}\left(\frac{d^2\sigma_{\max}^2 \log(A/\delta)}{\sigma_{\min}^2 n^{3/2}} \right) + \frac{AH_U^2\kappa^2}{n^2}
\end{aligned}$$

We define the good estimation event as follows:

$$\xi_\delta^{est}(\Gamma) \coloneqq \left\{ \left| \sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b}_*, \widehat{\mathbf{\Sigma}}_\Gamma}^{-1} \mathbf{w}(a') - \sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b}_*, \mathbf{\Sigma}_*}^{-1} \mathbf{w}(a') \right| \leq \frac{2CB^*d^3\sigma_{\max}^4 \log(9H_U^2/\delta)}{\sigma_{\min}^4 \Gamma} \right\}$$

Under the good estimation event $\xi^{est}(\Gamma)$ and using Lemma 2 we can show that the estimation error is given by

$$\begin{aligned}
\mathcal{L}_n(\pi, \mathbf{b}_*, \widehat{\mathbf{\Sigma}}_\Gamma) - \mathcal{L}_n(\pi, \mathbf{b}_*, \mathbf{\Sigma}_*) &\leq \left(\frac{1}{n} \sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b}_*, \widehat{\mathbf{\Sigma}}_\Gamma}^{-1} \mathbf{w}(a') - \frac{1}{n} \sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b}_*, \mathbf{\Sigma}_*}^{-1} \mathbf{w}(a') \right) \mathbb{I}\{\xi_\delta^{est}(\Gamma)\} \\
&\quad + \left(\frac{1}{n} \sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b}_*, \widehat{\mathbf{\Sigma}}_\Gamma}^{-1} \mathbf{w}(a') - \frac{1}{n} \sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b}_*, \mathbf{\Sigma}_*}^{-1} \mathbf{w}(a') \right) \mathbb{I}\{\xi_\delta^{est}(\Gamma)^C\} \\
&= \left(\frac{1}{n} \sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b}_*, \widehat{\mathbf{\Sigma}}_\Gamma}^{-1} \mathbf{w}(a') - \frac{1}{n} \sum_{a,a'} \mathbf{w}(a)^\top \mathbf{A}_{\mathbf{b}_*, \mathbf{\Sigma}_*}^{-1} \mathbf{w}(a') \right) \mathbb{I}\{\xi_\delta^{est}(\Gamma)\} \\
&\quad + \frac{1}{n} \mathbf{Tr}\left(\left(\mathbf{A}_{\mathbf{b}_*, \widehat{\mathbf{\Sigma}}_\Gamma}^{-1} - \mathbf{A}_{\mathbf{b}_*, \mathbf{\Sigma}_*}^{-1} \right) \left(\sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top \right) \right) \mathbb{I}\{\xi_\delta^{est}(\Gamma)^C\} \\
&\overset{(a)}{\leq} \frac{1}{n} 2B^* \frac{Cd^3\sigma_{\max}^2 \log(1/\delta)}{\sigma_{\min}^2 \Gamma} + \frac{1}{n} \mathbf{Tr}\left(\mathbf{A}_{\mathbf{b}_*, \mathbf{\Sigma}_*}^{-1} \right) \mathbf{Tr}\left(\mathbf{A}_{\mathbf{b}_*, \widehat{\mathbf{\Sigma}}_\Gamma}^{-1} \right) \mathbf{Tr}\left(\sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top \right) \delta \\
&\overset{(b)}{\leq} \frac{1}{n} 2B^* \frac{Cd^3\sigma_{\max}^2 \log(n)}{\sigma_{\min}^2 \sqrt{n}} + \frac{d^2}{n^2} \mathbf{Tr}\left(\sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top \right) = \frac{2B^*Cd^3\sigma_{\max}^2 \log(n)}{\sigma_{\min}^2 n^{3/2}} + \frac{d^2}{n^2} \mathbf{Tr}\left(\sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top \right)
\end{aligned}$$

where, (a) follows from Lemma 2, (b) follows as $\Gamma = \sqrt{n}$ and setting $\delta = \frac{1}{n^3}$. Combining everything we have the following regret as

$$\begin{aligned}
\mathcal{R}_n &\leq \frac{1}{n^{3/2}} + O_{\kappa^2, H_U^2}\left(\frac{d^2\sigma_{\max}^2 \log(n)}{\sigma_{\min}^2 n^{3/2}} \right) + \frac{2B^*Cd^3\sigma_{\max}^2 \log(n)}{\sigma_{\min}^2 n^{3/2}} + \frac{d^2}{n^2} \mathbf{Tr}\left(\sum_{a,a'} \mathbf{w}(a)\mathbf{w}(a')^\top \right) + \frac{2AH_U^2\kappa^2}{n^2} \\
&= O_{\kappa^2, H_U^2}\left(\frac{B^*d^3\sigma_{\max}^2 \log(n)}{\sigma_{\min}^2 n^{3/2}} \right)
\end{aligned}$$

where, $B^* = \left(\left\| \mathbf{A}^{-1}_{\mathbf{b}_*, \mathbf{\Sigma}_*} \mathbf{w} \right\|^2 \left\| \sum_{a=1}^{A} \mathbf{b}_*(a) \mathbf{w}(a) \mathbf{w}(a)^\top H_U^2 \right\| \left\| \left(\sum_{a=1}^{A} \frac{\mathbf{b}_*(a) \mathbf{w}(a) \mathbf{w}(a)^\top}{\sigma^2(a) + \frac{2Cd^3 \log(9H_U^2/\delta)}{\sqrt{n}}} \right)^{-1} \mathbf{w} \right\| \right)$. The claim of the theorem follows. □

Remark 4. (Discussion on Sample regime and B_*): Observe that combining Proposition 5 and Theorem 1 we can have a loss of SPEED that scales as

$$O_{\kappa^2, H_U^2, \sigma^2_{\max}, \sigma^2_{\min}} \left(\frac{d \log(n)}{n} \right) + O_{\kappa^2, H_U^2, \sigma^2_{\max}, \sigma^2_{\min}} \left(\frac{B^\star d^3 \log(n)}{n^{3/2}} \right)$$

which seems to contradict the loss bound in Proposition 6.

However, this is not the case. Observe that the B_* is a problem-dependent quantity that depends on a number of samples n. We define it as

$$B_* := \left\| \mathbf{A}^{-1}_{\mathbf{b}_*, \mathbf{\Sigma}_*} \mathbf{w} \right\|^2 \left\| \sum_{a=1}^{A} \mathbf{b}_*(a) \mathbf{w}(a) \mathbf{w}(a)^\top H_U^2 \right\| \left\| \left(\sum_{a=1}^{A} \frac{\mathbf{b}_*(a) \mathbf{w}(a) \mathbf{w}(a)^\top}{\sigma^2(a) + \frac{2Cd^2 \log(A/\delta)}{\sqrt{n}}} \right)^{-1} \mathbf{w} \right\|.$$

However, there are two regimes when $n \leq \frac{16C^2 d^4 \sigma^4_{\max} \log^2(A/\delta)}{\sigma^4_{\min}}$ then $B_* = \Theta(\sqrt{n})$ and for $n > \frac{16C^2 d^4 \sigma^4_{\max} \log^2(A/\delta)}{\sigma^4_{\min}}$ then $B_* = o(\sqrt{n})$. In the first case when $n \leq \frac{16C^2 d^4 \sigma^4_{\max} \log^2(A/\delta)}{\sigma^4_{\min}}$ with $B_* = \Theta(\sqrt{n})$ we have the loss that scales as

$$O_{\kappa^2, H_U^2, \sigma^2_{\max}, \sigma^2_{\min}} \left(\frac{d \log(n)}{n} \right) + O_{\kappa^2, H_U^2, \sigma^2_{\max}, \sigma^2_{\min}} \left(\frac{B_* d^3 \log(n)}{n^{3/2}} \right) = O_{\kappa^2, H_U^2, \sigma^2_{\max}, \sigma^2_{\min}} \left(\frac{d^3 \log(n)}{n} \right)$$

This is the regime of Proposition 6 as it holds for all $n \geq \frac{2Cd^2 \sigma^2_{\max} \log(A/\delta)}{\sigma^2_{\min} \Gamma}$ for $\Gamma \geq 1$. Note that $\frac{2Cd^2 \sigma^2_{\max} \log(A/\delta)}{\sigma^2_{\min} \Gamma}$ is less than $\frac{16C^2 d^4 \sigma^4_{\max} \log^2(A/\delta)}{\sigma^4_{\min}}$.

In the second case when $n > \frac{16C^2 d^4 \sigma^4_{\max} \log^2(A/\delta)}{\sigma^4_{\min}}$ with $B_* = o(\sqrt{n})$ we have a tighter bound as the first term dominates and we have the loss scaling as

$$O_{\kappa^2, H_U^2, \sigma^2_{\max}, \sigma^2_{\min}} \left(\frac{d \log(n)}{n} \right) + O_{\kappa^2, H_U^2, \sigma^2_{\max}, \sigma^2_{\min}} \left(\frac{B^\star d^3 \log(n)}{n^{3/2}} \right) = O_{\kappa^2, H_U^2, \sigma^2_{\max}, \sigma^2_{\min}} \left(\frac{d \log(n)}{n} \right)$$

Intuitively this is a larger sample regime where the SPEED has a good estimation of $\mathbf{\Sigma}_*$ and the design matrix estimation has also concentrated. Combining both the regimes we can show that for $n \geq \frac{2Cd^2 \log(A/\delta)}{\sigma_\min^2 \Gamma}$ the loss of SPEED scales by

$$\max \left\{ O_{\kappa^2, H_U^2, \sigma^2_{\max}, \sigma^2_{\min}} \left(\frac{d \log(n)}{n} \right), O_{\kappa^2, H_U^2, \sigma^2_{\max}, \sigma^2_{\min}} \left(\frac{d^3 \log(n)}{n} \right) \right\} = O_{\kappa^2, H_U^2, \sigma^2_{\max}, \sigma^2_{\min}} \left(\frac{d^3 \log(n)}{n} \right)$$

which is the bound of Proposition 6. So in summary Proposition 6 is a more general bound for a larger regime size than Theorem 1 and does not contradict the theorem statement.

C Regret Lower Bound

Theorem 2. ***(Lower Bound)*** *Let $|\mathbf{\Theta}| = 2^d$ and $\boldsymbol{\theta}_* \in \mathbf{\Theta}$. Then any δ-PAC policy π following the design $\mathbf{b} \in \triangle(\mathcal{A})$ satisfies $\mathcal{R}'_n = \mathcal{L}_n(\pi, \widehat{\mathbf{b}}, \mathbf{\Sigma}_*) - \mathcal{L}_n(\pi, \mathbf{b}^*, \mathbf{\Sigma}_*) \geq \Omega \left(\frac{d^2 \lambda_d(\mathbf{V}) \log(n)}{n^{3/2}} \right)$ for the environment in (33).*

Proof. **Step 1 (Define Environment):** We define an environment model B_j consisting of A actions and J

hypotheses with true hypothesis $\boldsymbol{\theta}_* = \boldsymbol{\theta}_j$ (j-th column) as follows:

$$\begin{array}{cccccc} \boldsymbol{\theta} & = & \boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 & \boldsymbol{\theta}_3 & \dots & \boldsymbol{\theta}_J \\ \hline \mu_1(\boldsymbol{\theta}) & = & \beta & \beta-\frac{\beta}{J} & \beta-\frac{2\beta}{J} & \dots & \beta-\frac{(J-1)\beta}{J} \\ \mu_2(\boldsymbol{\theta}) & = & \iota_{21} & \iota_{22} & \iota_{23} & \dots & \iota_{2J} \\ & \vdots & & & \vdots & & \\ \mu_A(\boldsymbol{\theta}) & = & \iota_{A1} & \iota_{A2} & \iota_{A3} & \dots & \iota_{AJ} \end{array} \tag{33}$$

where, each ι_{ij} is distinct and satisfies $\iota_{ij} < \beta/4J$. $\boldsymbol{\theta}_1$ is the optimal hypothesis in B_1, $\boldsymbol{\theta}_2$ is the optimal hypothesis in B_2 and so on such that for each B_j and $j \in [J]$ we have column j as the optimal hypothesis.

Finally, assume that $\boldsymbol{\Sigma} = \boldsymbol{\theta}\boldsymbol{\theta}^\top$ is a rank one matrix. To distinguish between the covariance matrix between two distributions we denote $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{\theta}\boldsymbol{\theta}^\top$. Therefore we have that $\sigma_i^2(\boldsymbol{\theta}) = \mathbf{x}_i^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \mathbf{x}_i = (\mathbf{x}_i^\top \boldsymbol{\theta})^2 = \mu_i^2(\boldsymbol{\theta})$. Hence for any algorithm, identifying the co-variance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}$ is the same as identifying the $\boldsymbol{\theta}_*$. Also assume that $\pi(a) = \frac{1}{A}$. Hence each action is equally weighted by the target policy.

This is a general hypothesis testing setting where the functions $\mu_a(\boldsymbol{\theta})$ can be thought of as linear functions of $\boldsymbol{\theta}$ such that $\mu_a(\boldsymbol{\theta}) = \mathbf{x}(a)^\top \boldsymbol{\theta}$. Assume that $0 < \mu_a(\boldsymbol{\theta}) \leq 1$, and $\log(\mu_a(\boldsymbol{\theta})/\mu_a(\boldsymbol{\theta}')) > 1/4$.

Now observe that between any two hypothesis $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ we have the following

$$\begin{aligned} \mathrm{KL}\Big(\mathcal{N}(\mu_i(\boldsymbol{\theta}), \mathbf{x}_i^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \mathbf{x}_i)) \big|\big| \mathcal{N}(\mu_i(\boldsymbol{\theta}'), \mathbf{x}_i^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}'} \mathbf{x}_i))\Big) &= 2\log(\frac{\sigma_i(\boldsymbol{\theta}')}{\sigma_i(\boldsymbol{\theta})}) + \frac{\sigma_i^2(\boldsymbol{\theta}) + (\mu_i(\boldsymbol{\theta}) - \mu_i(\boldsymbol{\theta}'))^2}{2\sigma_i^2(\boldsymbol{\theta}')} - \frac{1}{2} \\ &\overset{(a)}{=} 2\log(\frac{\mu_i(\boldsymbol{\theta}')}{\mu_i(\boldsymbol{\theta})}) + \frac{\mu_i^2(\boldsymbol{\theta}) + (\mu_i(\boldsymbol{\theta}) - \mu_i(\boldsymbol{\theta}'))^2}{2\mu_i^2(\boldsymbol{\theta}')} - \frac{1}{2} \overset{(a)}{\geq} \frac{(\mu_i(\boldsymbol{\theta}) - \mu_i(\boldsymbol{\theta}'))^2}{8} \end{aligned} \tag{34}$$

where, (a) follows from the condition that $0 < \mu_a(\boldsymbol{\theta}) \leq 1$, and $\log(\mu_a(\boldsymbol{\theta})/\mu_a(\boldsymbol{\theta}')) > 1/4$.

Step 2 (Minimum samples to verify $\boldsymbol{\theta}_*$): Let, Λ_1 be the set of alternate models having a different optimal hypothesis than $\boldsymbol{\theta}^* = \boldsymbol{\theta}_1$ such that all models having different optimal hypothesis than $\boldsymbol{\theta}_1$ such as $B_2, B_3, \dots B_J$ are in Λ_1. Let τ_δ be the stopping time for any δ-PAC policy $\mathbf{b}$. That is τ_δ is the time that any algorithm stops and outputs its estimate $\widehat{\boldsymbol{\theta}}_{\tau_\delta}$. Let $T_t(a)$ denote the number of times the action a has been sampled till round t. Let $\widehat{\boldsymbol{\theta}}_{\tau_\delta}$ be the predicted optimal hypothesis at round τ_δ. We first consider the model B_1. Define the event $\xi = \{\widehat{\boldsymbol{\theta}}_{\tau_\delta} \neq \boldsymbol{\theta}_*\}$ as the error event in model B_1. Let the event $\xi' = \{\widehat{\boldsymbol{\theta}}_{\tau_\delta} \neq \boldsymbol{\theta}'^*\}$ be the corresponding error event in model B_2. Note that $\xi^{\complement} \subset \xi'$. Now since $\mathbf{b}$ is δ-PAC policy we have $\mathbb{P}_{B_1,\mathbf{b}}(\xi) \leq \delta$ and $\mathbb{P}_{B_2,\mathbf{b}}(\xi^{\complement}) \leq \delta$. Hence we can show that,

$$\begin{aligned} 2\delta \geq \mathbb{P}_{B_1,\mathbf{b}}(\xi) + \mathbb{P}_{B_2,\mathbf{b}}(\xi^{\complement}) &\overset{(a)}{\geq} \frac{1}{2}\exp\left(-\mathrm{KL}\left(P_{B_1,\mathbf{b}} || P_{B_2,\mathbf{b}}\right)\right) \\ \mathrm{KL}\left(P_{B_1,\mathbf{b}} || P_{B_2,\mathbf{b}}\right) &\geq \log\left(\frac{1}{4\delta}\right) \\ \frac{1}{8}\sum_{i=1}^{A} \mathbb{E}_{B_1,\mathbf{b}}[T_{\tau_\delta}(i)] \cdot \left(\mu_i(\boldsymbol{\theta}_*) - \mu_i(\boldsymbol{\theta}_*')\right)^2 &\overset{(b)}{\geq} \log\left(\frac{1}{4\delta}\right) \\ \frac{1}{8}\left(\beta - \beta + \frac{\beta}{J}\right)^2 \mathbb{E}_{B_1,\mathbf{b}}[T_{\tau_\delta}(1)] + \frac{1}{8}\sum_{i=2}^{A} (\iota_{i1} - \iota_{i2})^2 \mathbb{E}_{B_1,\mathbf{b}}[T_{\tau_\delta}(i)] &\overset{(c)}{\geq} \log\left(\frac{1}{4\delta}\right) \\ \frac{1}{8}\left(\frac{1}{J}\right)^2 \beta^2 \mathbb{E}_{B_1,\mathbf{b}}[T_{\tau_\delta}(1)] + \frac{1}{8}\sum_{i=2}^{A} (\iota_{i1} - \iota_{i2})^2 \mathbb{E}_{B_1,\mathbf{b}}[T_{\tau_\delta}(i)] &\geq \log\left(\frac{1}{4\delta}\right) \\ \frac{1}{8}\left(\frac{1}{J}\right)^2 \beta^2 \mathbb{E}_{B_1,\mathbf{b}}[T_{\tau_\delta}(1)] + \frac{1}{8}\sum_{i=2}^{A} \frac{\beta^2}{4J^2} \mathbb{E}_{B_1,\mathbf{b}}[T_{\tau_\delta}(i)] &\overset{(d)}{\geq} \log\left(\frac{1}{4\delta}\right) \end{aligned} \tag{35}$$

where, (a) follows from Lemma 10, (b) follows from Lemma 9, (c) follows from the construction of the bandit environments and (34), and (d) follows as $(\iota_{ij} - \iota_{ij'})^2 \leq \frac{\beta^2}{4J^2}$ for any i-th action and j-th hypothesis.

Now, we consider the alternate model B_3. Again define the event $\xi = \{\widehat{\boldsymbol{\theta}}_{\tau_\delta} \neq \boldsymbol{\theta}_*\}$ as the error event in model B_1 and the event $\xi' = \{\widehat{\boldsymbol{\theta}}_{\tau_\delta} \neq \boldsymbol{\theta}_*^{''}\}$ be the corresponding error event in model B_3. Note that $\xi^{\complement} \subset \xi'$. Now since $\mathbf{b}$ is δ-PAC policy we have $\mathbb{P}_{B_1,\mathbf{b}}(\xi) \leq \delta$ and $\mathbb{P}_{B_3,\mathbf{b}}(\xi^{\complement}) \leq \delta$. Following the same way as before we can show that,

$$\frac{1}{8}\left(\frac{2}{J}\right)^2 \beta^2 \mathbb{E}_{B_3,\mathbf{b}}[T_{\tau_\delta}(1)] + \frac{1}{8}\sum_{i=2}^{A} \frac{\beta^2}{4J^2}\mathbb{E}_{B_3,\mathbf{b}}[T_{\tau_\delta}(i)] \overset{(d)}{\geq} \log\left(\frac{1}{4\delta}\right). \tag{36}$$

Similarly, we get the equations for all the other $(J-2)$ alternate models in Λ_1. Now consider an optimization problem (ignoring the constant factor of $\frac{1}{8}$ across all the constraints)

$$\begin{aligned}
&\min_{t_i : i \in [A]} \sum t_i \\
s.t. \quad &\left(\frac{1}{J}\right)^2 \beta^2 t_1 + \frac{\beta^2}{4J^2}\sum_{i=2}^{A} t_i \geq \log(1/4\delta) \\
&\left(\frac{2}{J}\right)^2 \beta^2 t_1 + \frac{\beta^2}{4J^2}\sum_{i=2}^{A} t_i \geq \log(1/4\delta) \\
&\vdots \\
&\left(\frac{J-1}{J}\right)^2 \beta^2 t_1 + \frac{\beta^2}{4J^2}\sum_{i=2}^{A} t_i \geq \log(1/4\delta) \\
&t_i \geq 0, \forall i \in [A]
\end{aligned}$$

where the optimization variables are t_i. It can be seen that the optimum objective value is $J^2\beta^{-2}\log(1/4\delta)$. Interpreting $t_i = \mathbb{E}_{B_1,\mathbf{b}}[T_{\tau_\delta}(i)]$ for all i, we get that $\mathbb{E}_{B_1,\mathbf{b}}[\tau_\delta] = \sum_i t_i = t_1 \geq J^2\beta^{-2}\log(1/4\delta)$ which gives us the required lower bound to the number of pulls of action 1. Observe that the optimum objective value is reached by substituting $t_1 = J^2\beta^{-2}\log(1/4\delta)$ and $t_2 = \ldots = t_A = 0$. It follows that for verifying any hypothesis $\boldsymbol{\theta}_j \neq \boldsymbol{\theta}_*$ the verification proportion is given by $\mathbf{b}_{\boldsymbol{\theta}_j} = (1, \underbrace{0, 0, \ldots, 0}_{\text{(A-1) zeros}})$. Observe setting $\beta = J\sqrt{\log(1/4\delta)/n}$ recovers $\tau_\delta = n$ which implies that a budget of n samples is required for verifying hypothesis $\boldsymbol{\theta}_j = \boldsymbol{\theta}_*$. For the remaining steps we take $\beta = J\sqrt{\log(1/4\delta)/n}$.

Step 3 (Lower Bounding Regret): Then we can show that the MSE of any hypothesis $\boldsymbol{\theta}_j = \boldsymbol{\theta}_*$

$$\mathbb{E}_{\mathcal{D}}\left[\left(\sum_a \pi(a)\mathbf{x}(a)^\top(\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_n)\right)^2\right] = \frac{1}{n}\sum_{a,a'}\mathbf{w}(a)\mathbf{A}^{-1}_{\mathbf{b}_{\boldsymbol{\theta}_j},\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}}\mathbf{w}(a') = \frac{1}{n}\mathbf{Tr}\Big(\mathbf{A}^{-1}_{\mathbf{b}_{\boldsymbol{\theta}_j},\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}} \underbrace{\sum_{a,a'}\mathbf{w}(a)\mathbf{w}(a')}_{\mathbf{V}}\Big)$$

where, $\mathbf{b}_{\boldsymbol{\theta}_j}(a)$ is the number of samples allocated to action a. First we will bound the loss of the oracle for this environment given by $\mathcal{L}_n(\pi, \mathbf{b}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}) = \frac{1}{n}\mathbf{Tr}(\mathbf{A}^{-1}_{\mathbf{b}_{\boldsymbol{\theta}_j},\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}}\mathbf{V})$. Note that the oracle has access to the $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}$, so it only need to verify whether $\boldsymbol{\theta}_j = \boldsymbol{\theta}_*$ by following $\mathbf{b}_{\boldsymbol{\theta}_j}$. Then we have that

$$\begin{aligned}
\mathbf{A}_{\mathbf{b}_{\boldsymbol{\theta}_j},\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}} &= \sum_a \mathbf{b}_{\boldsymbol{\theta}_j}(a)\frac{\mathbf{x}(a)\mathbf{x}(a)^\top}{\sigma^2(a)} = \frac{\mathbf{w}(1)\mathbf{w}(1)^\top}{(\mathbf{x}(1)^\top\boldsymbol{\theta}_j)^2} = \frac{\mathbf{w}(1)\mathbf{w}(1)^\top}{(\beta - \frac{j\beta}{J})^2} \\
\implies \mathbf{Tr}(\mathbf{A}^{-1}_{\mathbf{b}_{\boldsymbol{\theta}_j},\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}}) &= \frac{(\beta - \frac{j\beta}{J})^2}{\mathbf{Tr}(\mathbf{w}(1)\mathbf{w}(1)^\top)}
\end{aligned}$$

Now we will bound the loss of the algorithm that uses $\widehat{\boldsymbol{\Sigma}}_\beta$ to estimate $\widehat{\mathbf{b}}$. It then collects the $\mathcal{D}$ and uses it to estimate $\boldsymbol{\theta}_*$ following the WLS estimation using $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}$.

Denote the number of times the algorithm samples each action i be $T'_n(i)$. Let the algorithm allocate $T'_n(1) = J^2\beta^{-2}\log(1/4\delta) - d$ samples to action 1 and to any other action i' it allocates $T'_n(i') = d$ samples such that $d \geq 1$. WLOG let $i' = 2$. Finally let $T'_n(3) = \ldots = T'_n(A) = 0$. Hence the optimal action 1 is under-allocated and the sub-optimal action 2 is over-allocated. The loss of such an algorithm now is given by

$$\mathcal{L}_n(\pi, \widehat{\mathbf{b}}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}) = \frac{1}{n}\mathbf{Tr}(\mathbf{A}^{-1}_{\widehat{\mathbf{b}}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}}\mathbf{V}).$$

Hence it follows by setting $\delta = 1/(nJ)$ that

$$\begin{aligned}
\mathbf{A}_{\widehat{\mathbf{b}}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}} = \frac{1}{n}\sum_a n\widehat{\mathbf{b}}(a)\frac{\mathbf{x}(a)\mathbf{x}(a)^\top}{\sigma^2(a)} &= \frac{1}{n}\sum_a T'_n(a)\frac{\mathbf{x}(a)\mathbf{x}(a)^\top}{\sigma^2(a)} \\
&= \frac{1}{n}T'_n(1)\frac{\mathbf{x}(1)\mathbf{x}(1)^\top}{\sigma^2(1)} + \underbrace{\frac{1}{n}T'_n(2)\frac{\mathbf{x}(2)\mathbf{x}(2)^\top}{\sigma^2(2)}}_{\geq 0} \\
&\geq \frac{1}{n}T'_n(1)\frac{\mathbf{w}(1)\mathbf{w}(1)^\top}{(\mathbf{x}(1)^\top\boldsymbol{\theta}_j)^2} \\
&\overset{(a)}{=} \frac{J^2\beta^{-2}\log(nJ) - d}{n}\frac{\mathbf{w}(1)\mathbf{w}(1)^\top}{(\beta - \frac{j\beta}{J})^2}
\end{aligned}$$

where, (a) follows by substituting the value of T'_n. Then we have that

$$\begin{aligned}
\mathbf{Tr}(\mathbf{A}^{-1}_{\widehat{\mathbf{b}}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}}) \geq \frac{n}{J^2\beta^{-2}\log(nJ) - d}\frac{(\beta - \frac{j\beta}{J})^2}{\mathbf{Tr}(\mathbf{w}(1)\mathbf{w}(1)^\top)} &= \frac{n}{J^2\beta^{-2}(\log(nJ) - \frac{d}{J^2\beta^{-2}})}\frac{(\beta - \frac{j\beta}{J})^2}{\mathbf{Tr}(\mathbf{w}(1)\mathbf{w}(1)^\top)} \\
&\overset{(a)}{\geq} \frac{\beta^2\log(nJ) + \frac{d}{J^2\beta^{-2}}}{J^2}\frac{(\beta - \frac{j\beta}{J})^2}{\mathbf{Tr}(\mathbf{w}(1)\mathbf{w}(1)^\top)} \\
&\geq \frac{\beta^2\log(nJ)}{J^2}\frac{(\beta - \frac{j\beta}{J})^2}{\mathbf{Tr}(\mathbf{w}(1)\mathbf{w}(1)^\top)}
\end{aligned}$$

where, (a) follows as for $d \geq 1$ we have that

$$n - (\log(nJ))^2 \geq -\frac{d^2}{(J^2\beta^{-2})^2} \implies (\log(nJ) - \frac{d}{J^2\beta^{-2}})^{-1} \geq \log(nJ) + \frac{d}{J^2\beta^{-2}}.$$

Step 4 (Lower Bound regret): Hence we have the regret for verifying any hypothesis $\boldsymbol{\theta}_j = \boldsymbol{\theta}_*$ as follows:

$$\begin{aligned}
\mathcal{R}'_n &= \mathcal{L}_n(\pi, \widehat{\mathbf{b}}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}) - \mathcal{L}_n(\pi, \mathbf{b}^*, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}) \\
&\geq \frac{1}{n}\mathbf{Tr}\left(\mathbf{A}^{-1}_{\widehat{\mathbf{b}}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}}\mathbf{V}\right) - \frac{1}{n}\mathbf{Tr}\left(\mathbf{A}^{-1}_{\mathbf{b}_{\boldsymbol{\theta}_j}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}}\mathbf{V}\right) = \frac{1}{n}\mathbf{Tr}\left(\left(\mathbf{A}^{-1}_{\widehat{\mathbf{b}}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}} - \mathbf{A}^{-1}_{\mathbf{b}_{\boldsymbol{\theta}_j}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}}\right)\mathbf{V}\right) \\
&\geq \frac{\lambda_d(\mathbf{V})}{n}\mathbf{Tr}\left(\mathbf{A}^{-1}_{\widehat{\mathbf{b}}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}} - \mathbf{A}^{-1}_{\mathbf{b}_{\boldsymbol{\theta}_j}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}}\right) \\
&= \frac{\lambda_d(\mathbf{V})}{n}\left[\mathbf{Tr}\left(\mathbf{A}^{-1}_{\widehat{\mathbf{b}}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}}\right) - \mathbf{Tr}\left(\mathbf{A}^{-1}_{\mathbf{b}_{\boldsymbol{\theta}_j}, \mathbf{\Sigma}_{\boldsymbol{\theta}_*}}\right)\right] \\
&= \frac{\lambda_d(\mathbf{V})}{n}\left[\frac{\beta^2\log(nJ)}{J^2}\frac{(\beta - \frac{j\beta}{J})^2}{\mathbf{Tr}(\mathbf{w}(1)\mathbf{w}(1)^\top)} - \frac{(\beta - \frac{j\beta}{J})^2}{\mathbf{Tr}(\mathbf{w}(1)\mathbf{w}(1)^\top)}\right] \\
&= \frac{\lambda_d(\mathbf{V})\beta^2(\beta - \frac{j\beta}{J})^2}{n\mathbf{Tr}(\mathbf{w}(1)\mathbf{w}(1)^\top)}\left[\frac{\log(nJ)}{J^2} - 1\right]
\end{aligned}$$

$$\overset{(a)}{\geq} \frac{\lambda_d(\mathbf{V})\beta^2(\beta-\frac{j\beta}{J})^2}{n\mathbf{Tr}(\mathbf{w}(1)\mathbf{w}(1)^\top)}\left[\frac{\log(nJ)}{2J^2}\right]$$
$$\overset{(b)}{\geq} \frac{d\lambda_d(\mathbf{V})\beta^2}{n^{3/2}\mathbf{Tr}(\mathbf{w}(1)\mathbf{w}(1)^\top)}\left[\frac{\log(nJ)}{2J^2}\right]$$
$$\overset{(c)}{\geq} \frac{d^2\lambda_d(\mathbf{V})\beta^2}{n^{3/2}\mathbf{Tr}(\mathbf{w}(1)\mathbf{w}(1)^\top)}\log(2n)$$
$$= \Omega\left(\frac{d^2\lambda_d(\mathbf{V})\log(n)}{n^{3/2}}\right)$$

where, (a) follows as $\frac{\log(nJ)}{J^2} - 1 \geq \frac{\log(nJ)}{2J^2}$, (b) follows as gap $(\beta - \frac{j\beta}{J})^2 \geq \frac{d}{\sqrt{n}}$ for any $\boldsymbol{\theta}_j$, and (c) follows by substituting $|\boldsymbol{\Theta}| = J = 2^d$. □

Lemma 9. ***(Restatement of Lemma 15.1 in Lattimore and Szepesvári [2020a], Divergence Decomposition)*** *Let B and B' be two bandit models having different optimal hypothesis $\boldsymbol{\theta}_*$ and $\boldsymbol{\theta}'^*$ respectively. Fix some policy π and round n. Let $\mathbb{P}_{B,\pi}$ and $\mathbb{P}_{B',\pi}$ be two probability measures induced by some n-round interaction of π with B and π with B' respectively. Then*

$$\mathrm{KL}\left(\mathbb{P}_{B,\pi}||\mathbb{P}_{B',\pi}\right) = \sum_{i=1}^{A}\mathbb{E}_{B,\pi}[T_n(i)]\cdot\mathrm{KL}(\mathcal{N}(\mu_i(\boldsymbol{\theta}),1)||\mathcal{N}(\mu_i(\boldsymbol{\theta}_*),1))$$

where, $\mathrm{KL}\left(.||.\right)$ *denotes the Kullback-Leibler divergence between two probability measures and $T_n(i)$ denotes the number of times action i has been sampled till round n.*

Lemma 10. ***(Restatement of Lemma 2.6 in Tsybakov [2008])*** *Let $\mathbb{P},\mathbb{Q}$ be two probability measures on the same measurable space $(\Omega,\mathcal{F})$ and let $\xi\subset\mathcal{F}$ be any arbitrary event then*

$$\mathbb{P}(\xi)+\mathbb{Q}\left(\xi^{\complement}\right)\geqslant\frac{1}{2}\exp\left(-\mathrm{KL}(\mathbb{P}||\mathbb{Q})\right)$$

where $\xi^{\complement}$ denotes the complement of event ξ and $\mathrm{KL}(\mathbb{P}||\mathbb{Q})$ *denotes the Kullback-Leibler divergence between $\mathbb{P}$ and $\mathbb{Q}$.*

Environment $\mathcal{E}$: Consider the environment $\mathcal{E}$ which consist of 3 actions in $\mathbb{R}^2$ such that $\mathbf{x}(1)=[1,0]$ is along x-axis, $\mathbf{x}(2)=[0,1]$ is along y-axis and $\mathbf{x}(3)=[1/\sqrt{2},1/\sqrt{2}]$. Let $\boldsymbol{\theta}_*=[1,0]$ and so the optimal action is action 1. Let the target policy $\pi=[0.9,0.1,0.0]$. Finally, let the variances be $\sigma^2(1)=5/100$, $\sigma^2(2)=1.0$ and $\sigma^2(3)=5/100$.

Proposition 8. ***(On-policy regret)*** *Let the On-policy algorithm have access to the variance in environment $\mathcal{E}$. Then the regret of On-policy scales as $O\left(\frac{\lambda_1(\mathbf{V})}{n}\right)$.*

Proof. Recall that in $\mathcal{E}$, there are 3 actions in $\mathbb{R}^2$ such that $\mathbf{x}(1)=[1,0]$ is along x-axis, $\mathbf{x}(2)=[0,1]$ is along y-axis and $\mathbf{x}(3)=[1/\sqrt{2},1/\sqrt{2}]$. The $\boldsymbol{\theta}_*=[1,0]$ and so the optimal action is action 1. The target policy $\pi=[0.9,0.1,0.0]$. Finally, let the variances be $\sigma^2(1)=1.0$, $\sigma^2(2)=1.0$ and $\sigma^2(3)=5/100$. Hence, PE-Optimal design results in $\mathbf{b}^*=[0.5,0.5,0.0]$.

$$\mathbf{A}_{\pi,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}} = \sum_a \pi(a)\frac{\mathbf{x}(a)\mathbf{x}(a)^\top}{\sigma^2(a)} = \frac{9}{10}\cdot\mathbf{x}(1)\mathbf{x}(1)^\top + \frac{1}{10}\mathbf{x}(2)\mathbf{x}(2)^\top$$
$$\mathbf{A}_{\mathbf{b}_*,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}} = \sum_a \mathbf{b}_*(a)\frac{\mathbf{x}(a)\mathbf{x}(a)^\top}{\sigma^2(a)} = \frac{1}{2}\cdot\mathbf{x}(1)\mathbf{x}(1)^\top + \frac{1}{2}\mathbf{x}(2)\mathbf{x}(2)^\top$$

Recall that $\mathbf{V}=\sum_a\mathbf{w}(a)\mathbf{w}(a)^\top$. Hence, the regret scales as

$$\mathcal{R}_n = \mathcal{L}_n(\pi,\pi,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}) - \mathcal{L}_n(\pi,\mathbf{b}^*,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}) \leq \frac{1}{n}\mathbf{Tr}\left(\mathbf{A}^{-1}_{\pi,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}}\mathbf{V}\right) - \frac{1}{n}\mathbf{Tr}\left(\mathbf{A}^{-1}_{\mathbf{b}_*,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}}\mathbf{V}\right) = \frac{1}{n}\mathbf{Tr}\left(\left(\mathbf{A}^{-1}_{\pi,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}} - \mathbf{A}^{-1}_{\mathbf{b}_*,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}}\right)\mathbf{V}\right)$$
$$\overset{(a)}{\leq} O\left(\frac{\lambda_1(\mathbf{V})}{n}\right)$$

where, (a) follows by substituting the value of $\mathbf{A}_{\pi,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}}$ and $\mathbf{A}_{\mathbf{b}_*,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}}$. □

D Additional Experiments

In this section, we state additional experimental details.

Unit Ball: This experiment consists of a set of 4 actions that are arranged in a unit ball in $\mathbb{R}^2$, and $\|\mathbf{x}(a)\| = 1$ for all $a \in \mathcal{A}$. We consider three groups of actions: **a)** the reward-maximizing action in the direction of $\boldsymbol{\theta}^*$, **b)** the informative action (orthogonal to optimal action) that maximally reduces the uncertainty of $\widehat{\boldsymbol{\theta}}_t$ and **c)** the less-informative actions as shown in Figure 1 (Top-Left). The variance of the most informative action is chosen to be high (0.35), but the target probability is set as low 0.1, which forces the on-policy algorithm to sample the high variance action less. Figure 1 (Top-Right) shows that SPEED outperforms On-policy, G-Optimal , and A-Optimal . Note that we experiment with A-Optimal design [Fontaine et al., 2021] because this criterion results in minimizing the average variance of the estimates of the regression coefficients and is most closely aligned with our goal than G-, or, D-optimal designs [Jamieson and Jain, 2022].

Air Quality: We perform this experiment on real-world dataset Air Quality from UCI datasets. The Air quality dataset consists of 1500 samples each of which consists of 6 features. We first select 400 samples which are the actions in our setting. We then fit a weighted least square estimate to the original dataset and get an estimate of $\boldsymbol{\theta}_*$ and $\boldsymbol{\Sigma}_*$. The reward model is linear and given by $\mathbf{x}_{I_t}^\top \boldsymbol{\theta}_* +$ noise where $\mathbf{x}_{I_t}$ is the observed action at round t, and the noise is a zero-mean additive noise with variance scaling as $\mathbf{x}_{I_t}^\top \boldsymbol{\Sigma}_* \mathbf{x}_{I_t}$. Hence the variance of each action depends on their feature vectors and $\boldsymbol{\Sigma}_*$. Finally, we set a level τ, such that 30 actions having variance crossing τ are set with low target probability, and the remaining probability mass is uniformly distributed among the rest 370 action. Hence, again high variance actions are set with a low target probability, which forces the on-policy algorithm to sample the high-variance action less number of times. We apply SPEED to this problem and compare it to baselines A-Optimal , G-Optimal , and the On-policy algorithm.

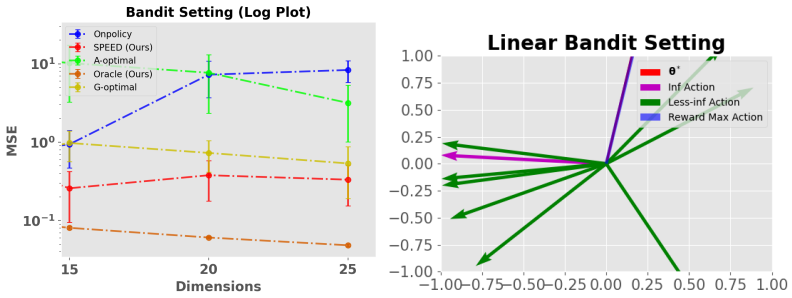

Figure 2: 10 action unit ball environment

Red Wine Quality: The UCI Red Wine Quality dataset consist of 1600 samples of red wine with each sample i having feature $\mathbf{x}_i \in \mathbb{R}^{11}$. We first fit a weighted least square estimate to the original dataset and get an estimate of $\boldsymbol{\theta}^*$ and $\boldsymbol{\Sigma}_*$. The reward model is linear and given by $\mathbf{x}_{I_t}^T \boldsymbol{\theta}^* +$ noise where $\mathbf{x}_{I_t}$ is the observed action at round t, and the noise is a zero-mean additive noise with variance scaling as $\mathbf{x}_{I_t}^\top \boldsymbol{\Sigma}_* \mathbf{x}_{I_t}$. Note that we consider the 1600 samples as actions. Then we run each of our benchmark algorithms on this dataset and reward model. Finally, we set a level τ, such that 40 actions having variance crossing τ are set with low target probability, and the remaining probability mass is uniformly distributed among the rest 1560 action. Hence, again high variance actions are set with a low target probability, which forces the on-policy algorithm to sample the high-variance action less number of times. We apply SPEED to this problem and compare it to baselines A-Optimal , G-Optimal , and the On-policy algorithm.

Movielens: We experiment with a movie recommendation problem on the MovieLens 1M dataset [Lam and Herlocker, 2016]. This dataset contains one million ratings given by 6 040 users to 3 952 movies. We first apply a low-rank factorization to the rating matrix to obtain 5-dimensional representations: $\boldsymbol{\theta}_j \in \mathbb{R}^5$ for user $j \in [6\,040]$ and $\mathbf{x}(a) \in \mathbb{R}^5$ for movie $a \in [3\,952]$. In each run, we choose one user $\boldsymbol{\theta}_j$ and 100 movies $\mathbf{x}(a)$ randomly, and they represent the unknown model parameter and known feature vectors of actions, respectively.

Increasing Dimension: We perform this experiment to show how the MSE of SPEED scales with increasing dimensions and number of actions. We choose dimension $d \in \{15, 20, 25\}$. For each dimension $d \in \{15, 20, 25\}$ we choose the number of actions $|\mathcal{A}| = d^2 + 20$. Hence we ensure that the number of actions are greater than d^2 dimensions. We also choose the horizon as $T \in \{13000, 18000, 25000\}$ for each $d \in \{15, 20, 25\}$. We choose the

same environment as the unit ball experiment. So the actions arranged in a unit ball in $\mathbb{R}^2$ and $\|\mathbf{x}(a)\| = 1$ for all $a \in \mathcal{A}$. Again we consider three groups of actions: **a)** the reward-maximizing action in the direction of $\boldsymbol{\theta}^*$, **b)** the informative action (orthogonal to optimal action) that maximally reduces the uncertainty of $\widehat{\boldsymbol{\theta}}_t$ and **c)** the less-informative actions as shown in Figure 2 but scaled to a larger set of actions. For each case of dimension $d \in \{15, 20, 25\}$, the variance of the most informative actions along the directions orthogonal to the reward maximizing action are chosen to be high, but the target probability is set as low, which forces the on-policy algorithm to sample the high variance action less. We again show the performance in Figure 1 (Bottom-left). We observe that with increasing dimensions d the SPEED outperforms on-policy. Also, observe that the oracle with knowledge of $\boldsymbol{\Sigma}_*$ performs the best.

E Table of Notations

Notations	**Definition**
$\pi(a)$	Target policy probability for action a
$\mathbf{b}(a)$	Behavior policy probability for action a
$\mathbf{x}(a)$	Feature of action a
$\boldsymbol{\theta}_*$	Optimal mean parameter
$\widehat{\boldsymbol{\theta}}_n$	Estimate of $\boldsymbol{\theta}_*$
$\mu(a) = \mathbf{x}^\top \boldsymbol{\theta}_*$	Mean of action a
$\widehat{\mu}_t(a) = \mathbf{x}^\top \widehat{\boldsymbol{\theta}}_t$	Empirical mean of action a at time t
$R_t(a)$	Reward for action a at time t
$\boldsymbol{\Sigma}_*$	Optimal co-variance matrix
$\widehat{\boldsymbol{\Sigma}}_t$	Empirical co-variance matrix at time t
$\sigma^2(a) = \mathbf{x}(a)^\top \boldsymbol{\Sigma}_* \mathbf{x}(a)$	Variance of action a
$\widehat{\sigma}_t^2(a) = \mathbf{x}(a)^\top \widehat{\boldsymbol{\Sigma}}_t \mathbf{x}(a)$	Empirical variance of action a at time t
n	Total budget
$T_n(a)$	Total Samples of action a after n timesteps

Table 1: Table of Notations