Multi-task Representation Learning for Pure Exploration in Linear Bandits

Yihan Du 1 Longbo Huang 1 Wen Sun 2

Abstract
Despite the recent success of representation learning in sequential decision making, the study of the pure exploration scenario (i.e., identify the best option and minimize the sample complexity) is still limited. In this paper, we study multi-task representation learning for best arm identification in linear bandits (RepBAI-LB) and best policy identification in contextual linear bandits (RepBPI-CLB), two popular pure exploration settings with wide applications, e.g., clinical trials and web content optimization. In these two problems, all tasks share a common low-dimensional linear representation, and our goal is to leverage this feature to accelerate the best arm (policy) identification process for all tasks. For these problems, we design computationally and sample efficient algorithms DouExpDes and C-DouExpDes, which perform double experimental designs to plan optimal sample allocations for learning the global representation. We show that by learning the common representation among tasks, our sample complexity is significantly better than that of the native approach which solves tasks independently. To the best of our knowledge, this is the first work to demonstrate the benefits of representation learning for multi-task pure exploration.

1. Introduction
Multi-task representation learning (Caruana, 1997) is an important problem which aims to learn a common low-dimensional representation from multiple related tasks. Representation learning has received extensive attention in both empirical applications (Ando et al., 2005; Bengio et al., 2013; Li et al., 2014) and theoretical study (Maurer et al., 2016; Du et al., 2021a; Tripuraneni et al., 2021).

Recently, an emerging number of works (Yang et al., 2021; 2022; Hu et al., 2021; Cella et al., 2022b) investigate representation learning for sequential decision making, and show that if all tasks share a joint low-rank representation, then by leveraging such a joint representation, it is possible to learn faster than treating each task independently. Despite the accomplishments of these works, they mainly focus on the regret minimization setting, where the performance is measured by the cumulative reward gap between the optimal option and the actually chosen options.

However, in real-world applications where obtaining a sample is expensive and time-consuming, e.g., clinical trials (Zhang et al., 2012), it is often desirable to identify the optimal option using as few samples as possible, i.e., we face the pure exploration scenario rather than regret minimization. Moreover, in many decision-making applications, we often need to tackle multiple related tasks, e.g., treatment planning for different diseases (Bragman et al., 2018) and content optimization for multiple websites (Agarwal et al., 2009), and there usually exists a common representation among these tasks, e.g., the features of drugs and the representations of website items. Thus, we desire to exploit the shared representation among tasks, e.g., the features of drugs and the representations of website items. Thus, we desire to make use of the shared representation and reduce the number of samples required.

Motivated by the above fact, in this paper, we study representation learning for multi-task pure exploration in sequential decision making. Following prior works (Yang et al., 2021; 2022; Hu et al., 2021), we consider the linear bandit setting, which is one of the most popular settings in sequential decision making and has various applications such as clinical trials and recommendation systems. Specifically, we investigate two pure exploration problems, i.e., representation learning for best arm identification in linear bandits (RepBAI-LB) and best policy identification in contextual linear bandits (RepBPI-CLB).

In RepBAI-LB, an agent is given a confidence parameter $\delta$, an arm set $\mathcal{X} := \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ and $M$ tasks. For each task $m \in [M]$, the expected reward of each arm $x \in \mathcal{X}$
Multi-task Representation Learning for Pure Exploration in Linear Bandits

is generated by \( x^\top \theta_m \), where \( \theta_m \in \mathbb{R}^d \) is an underlying reward parameter. There exists an unknown global feature extractor \( B \in \mathbb{R}^{d \times k} \) and an underlying prediction parameter \( w_m \) such that \( \theta_m = Bw_m \) for any \( m \in [M] \), where \( M \gg d \gg k \). We can understand the problem as that all tasks share a joint representation \( f(x) := B^\top x \) for arms, where the dimension of \( f(x) \) is much smaller than that of \( x \). The agent sequentially selects arms and tasks to sample, and observes noisy rewards. The goal of the agent is to identify the best arm with the maximum expected reward for each task with confidence \( 1 - \delta \), using as few samples as possible.

The RepBPI-CLB problem is an extension of RepBAI-LB to environments with random and varying contexts. In RepBPI-CLB, there are a context space \( S \), an action space \( A \), a known feature mapping \( \phi : S \times A \rightarrow \mathbb{R}^d \) and an unknown context distribution \( D \). For each task \( m \in [M] \), the expected reward of each context-action pair \( (s, a) \in S \times A \) is generated by \( \phi(s, a)^\top \theta_m \), where \( \theta_m = Bw_m \). We can similarly interpret the problem as that all tasks share a low-dimensional context-action representation \( B^\top \phi(s, a) \in \mathbb{R}^k \). At each timestep, the agent first observes a context drawn from \( D \), and chooses an action and a task to sample, and then observes a random reward. Given a confidence parameter \( \delta \) and an accuracy parameter \( \epsilon \), the agent aims to identify an \( \epsilon \)-optimal policy (i.e., a mapping \( S \rightarrow A \) that gives suboptimality within \( \epsilon \)) for each task with confidence \( 1 - \delta \), while minimizing the number of samples used.

In contrast to existing representation learning works (Yang et al., 2021; 2022; Hu et al., 2021; Cella et al., 2022b), we focus on the pure exploration scenario and face several unique challenges: (i) The sample complexity minimization objective requires us to plan an optimal sample allocation for recovering the low-rank representation, in order to save samples to the highest degree. (ii) Unlike prior works which either assume that the arm set is an ellipsoid/sphere (Yang et al., 2021; 2022) or are computationally inefficient (Hu et al., 2021), we allow an arbitrary arm set that spans \( \mathbb{R}^d \), which poses challenges on how to efficiently schedule samples according to the shapes of arms. (iii) Different from prior works (Huang et al., 2015; Li et al., 2022), we do not assume prior knowledge of the context distribution. This imposes additional difficulties in sample allocation planning and estimator construction. To handle these challenges, we design computationally and sample efficient algorithms, which effectively estimate the context distribution and employ the experimental design approaches to plan samples.

We summarize our contributions in this paper as follows.

- We formulate the problems of multi-task representation learning for best arm identification in linear bandits (RepBAI-LB) and best policy identification in contextual linear bandits (RepBPI-CLB). To the best of our knowledge, this is the first work to study representation learning in the multi-task pure exploration scenario.
- For RepBAI-LB, we propose an efficient algorithm DouExpDes equipped with double experimental designs. The first design optimally schedules samples to learn the joint representation according to arm shapes, and the second design minimizes the estimation error for rewards using low-dimensional representations. Furthermore, we establish a sample complexity guarantee \( O(\frac{Mk}{\Delta_{\min}^2}) \), which shows superiority over the baseline result \( O(\frac{Md}{\Delta_{\min}^2}) \) (i.e., solving each task independently). Here \( \Delta_{\min} \) denotes the minimum reward gap.
- For RepBPI-CLB, we develop C-DouExpDes, an algorithm which efficiently estimates the context distribution and conducts double experimental designs under the estimated context distribution to learn the global representation. A sample complexity result \( O(\frac{Md^2}{\Delta_{\min}^2}) \) is also provided for C-DouExpDes, which significantly outperforms the baseline result \( O(\frac{Md^2}{\Delta_{\min}^2}) \), and demonstrates the power of representation learning.

2. Related Work

In this section, we introduce two lines of related works, and defer a more complete literature review to Appendix A.

**Representation Learning.** The study of representation learning has been initiated and developed in the supervised learning setting, e.g., (Baxter, 2000; Ando et al., 2005; Maurer et al., 2016; Du et al., 2021a; Tripuraneni et al., 2021).

Recently, representation learning for sequential decision making has attracted extensive attention. Lale et al. (2019); Jun et al. (2019); Lu et al. (2021b); Huang et al. (2021) study linear bandits with a hidden low-rank structure (e.g., bilinear bandits), which is very related to the problem of representation learning. Yang et al. (2021; 2022); Hu et al. (2021); Cella et al. (2022b) consider multi-task representation learning for linear bandits with the regret minimization objective. Yang et al. (2021; 2022) assume that the arm set is an ellipsoid or sphere. Hu et al. (2021) relax this assumption and allow arbitrary arm sets, but their algorithms that build upon a multi-task joint least-square estimator are computationally inefficient. Cella et al. (2022b) design algorithms that do not need to know the dimension of the underlying representation. There are also other works (Lu et al., 2021a; 2022; Pacchiano et al., 2022; Zhang & Wang, 2021; Cheng et al., 2022; Agarwal et al., 2022) which investigate representation learning for reinforcement learning.

Different from the above works which consider regret minimization, we study representation learning for (contextual) linear bandits with the pure exploration objective, which brings unique challenges on how to optimally allocate samples to learn the feature extractor, and motivates us to design
where $\theta \in \mathbb{M}_{x \times 2019}$; Katz-Samuels et al., 2020; Degenne et al., 2020). For identification tasks. Without loss of generality, we assume $X$ set of arms in Linear Bandits (RepBAI-LB).

In this section, we present the formal problem formulations of RepBAI-LB and RepBPI-CLB. Before describing the formulations, we first introduce some useful notations.

**Notations.** We use bold lower-case letters to denote vectors and bold upper-case letters to denote matrices. For any matrix $A$, $\|A\|$ denotes the spectral norm of $A$, and $\sigma_{\min}(A)$ denotes the minimum singular value of $A$. For any positive semi-definite matrix $A \in \mathbb{R}^{d \times d}$ and vector $x \in \mathbb{R}^d$, $\|x\|_A := \sqrt{x^\top A x}$. We use polylog$(\cdot)$ to denote a polylogarithmic function in given parameters, and $O(\cdot)$ to denote an expression that hides polylogarithmic factors in all problem parameters except $\delta$ and $\varepsilon$.

**Representation Learning for Best Arm Identification in Linear Bandits (RepBAI-LB).** An agent is given a set of arms $\mathcal{X} := \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ and $M$ best arm identification tasks. Without loss of generality, we assume that $\mathcal{X}$ spans $\mathbb{R}^d$, as done in many prior works (Fiez et al., 2019; Katz-Samuels et al., 2020; Degenne et al., 2020). For any $x \in \mathcal{X}$, $\|x\| \leq L_x$ for some constant $L_x$. For each task $m \in [M]$, the expected reward of each arm $x \in \mathcal{X}$ is $x^\top \theta_m$, where $\theta_m \in \mathbb{R}^d$ is an unknown reward parameter. Among all tasks, there exists a common underlying feature extractor $B \in \mathbb{R}^{d \times k}$, which satisfies that for each task $m \in [M], \theta_m = B w_m$. Here $B$ has orthonormal columns, $w_m \in \mathbb{R}^k$ is an unknown prediction parameter, and $M \gg d \gg k$. For any $m \in [M], \|w_m\| \leq L_w$ for some constant $L_w$.

At each timestep $t$, the agent chooses an arm $x \in \mathcal{X}$ and a task $m \in [M]$, to sample arm $x$ in task $m$. Then, she observes a random reward $r_{tm} = x^\top \theta_m + \eta_t = x^\top B w_m + \eta_t$, where $\eta_t$ is an independent, zero-mean and sub-Gaussian noise. For simplicity of analysis, we assume that $\mathbb{E}[\eta_t^2] = 1$, which can be easily relaxed by using a more carefully-designed estimator in our algorithm. Given a confidence parameter $\delta \in (0, 1)$, the agent aims to identify the best arms $x^*_m := \arg\max_{x \in \mathcal{X}} x^\top \theta_m$ for all tasks $m \in [M]$ with probability at least $1 - \delta$, using as few samples as possible. We define sample complexity as the total number of samples used over all tasks, which is the performance metric considered in our paper.

To efficiently learn the underlying low-dimensional representation, we make the following standard assumptions.

**Assumption 3.1 (Diverse Tasks).** We assume that $\sigma_{\min}(\frac{1}{M} \sum_{m=1}^M w_m w_m^\top) = \Omega(\frac{1}{d})$.

This assumption indicates that the prediction parameters $w_1, \ldots, w_M$ are uniformly spread out in all directions of $\mathbb{R}^k$, which was also assumed in (Du et al., 2021a; Tripuraneni et al., 2021; Yang et al., 2021), and is necessary for recovering the feature extractor $B$.

For any distribution $\lambda \in \triangle_{\mathcal{X}}$ and $B \in \mathbb{R}^{d \times k}$, let $A(\lambda, B) := \sum_{i=1}^n \lambda(x_i) B^\top x_i x_i^\top B$. For any task $m \in [M]$, let

$$\lambda_m^* := \arg\min_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \in \mathcal{X} \setminus \{x_m^*\}} \frac{\|B^\top(x_m^* - x)^\top\|^2}{((x_m^* - x)^\top \theta_m)^2}.$$

Here $\lambda_m^*$ denotes the optimal sample allocation that minimizes prediction error of arms (i.e., the solution of G-optimal design (Pukelsheim, 2006)) under the underlying low-dimensional representation.

**Assumption 3.2 (Eigenvalue of G-optimal Design Matrix).** For any task $m \in [M]$, $\sigma_{\min}(A(\lambda_m^*, B)) \geq \omega$ for some constant $\omega > 0$.

This assumption implies that the covariance matrix $A(\lambda_m^*, B)$ under the optimal sample allocation is invertible, which is necessary for estimating $w_m$. Note that the quantities introduced in Assumptions 3.1 and 3.2, i.e., $\sigma_{\min}(\frac{1}{M} \sum_{m=1}^M w_m w_m^\top)$ and $\sigma_{\min}(A(\lambda_m^*, B))$, are both defined on the low-dimensional subspace, which scale as $k$ instead of $d$.

**Representation Learning for Best Policy Identification in Contextual Linear Bandits (RepBPI-CLB).** In this problem, there are a context space $\mathcal{S}$, an action space $\mathcal{A}$, a feature mapping $\phi(\cdot, \cdot) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ and an unknown context distribution $D \in \triangle S$. For any $(s, a) \in \mathcal{S} \times \mathcal{A}, \|\phi(s, a)\| \leq L_\phi$ for some constant $L_\phi$. An agent needs to solve $M$ best
policy identification tasks. For each task \( m \in [M] \), the expected reward of each context-action pair \((s, a) \in S \times A\) is \( \phi(s, a) \theta_m \), where \( \theta_m \in \mathbb{R}^d \) is an unknown reward parameter. Similar to RepBAI-LB, there exists a global feature extractor \( B \in \mathbb{R}^{d \times k} \) with orthonormal columns, such that for each task \( m \in [M] \), \( \theta_m = B w_m \). Here \( w_m \in \mathbb{R}^k \) is an unknown prediction parameter, \( |w_m| \leq L_w \) for any \( m \in [M] \), and \( M \gg d \gg k \).

At each timestep \( t \), the agent first observes a random context \( s_t \), which is i.i.d. drawn from \( \mathcal{D} \). Then, she selects an action \( a_t \in A \) and a task \( m \in [M] \), to sample action \( a_t \) in context \( s_t \) under task \( m \). After sampling, she observes a random reward \( r_t = \phi(s_t, a_t) \theta_m + \eta_t = \phi(s_t, a_t) B w_m + \eta_t \), where \( \eta_t \) is an independent, zero-mean and 1-sub-Gaussian noise.

We define a policy \( \pi \) as a mapping from \( S \) to \( A \). For each task \( m \in [M] \), we say a policy \( \hat{\pi}_m \) is \( \varepsilon \)-optimal if

\[
\mathbb{E}_{s \sim \mathcal{D}} \left[ \max_{a \in A} (\phi(s, a) - \phi(s, \hat{\pi}_m(s))^\top \theta_m) \right] \leq \varepsilon.
\]

Given a confidence parameter \( \delta \in (0, 1) \) and an accuracy parameter \( \varepsilon > 0 \), the goal of the agent is to identify an \( \varepsilon \)-optimal policy \( \hat{\pi}_m \) for each task \( m \in [M] \) with probability at least \( 1 - \delta \), and minimize the number of samples used, i.e., sample complexity.

We also make two standard assumptions for RepBPI-CLB: Assumption 3.1 and the following assumption on the context distribution and context-action features.

**Assumption 3.3.** There exists some \( \lambda \in \Delta_A \) such that

\[
\sigma_{\min} \left( \sum_{a \in A} \lambda(a) \mathbb{E}_{s \sim \mathcal{D}} [\phi(s, a) \phi(s, a)^\top] \right) \geq \nu
\]

for some constant \( \nu > 0 \).

Assumption 3.3 manifests that there exists at least one sample allocation, under which the expected covariance matrix with respect to random contexts is invertible. This assumption enables one to reveal the feature extractor \( B \), despite stochastic and varying contexts. Note that Assumption 3.3 only assumes the existence of a feasible sample allocation, rather than the knowledge of this sample allocation.

It is worth mentioning that in this work, we do not assume that we can sample arbitrary vectors in an ellipsoid/sphere as in (Yang et al., 2021; 2022), or assume that each arm (action) has zero mean and identity covariance as in (Tripathani et al., 2021). In contrast, we allow arbitrary shapes of arms (actions), and efficiently allocate samples according to their different shapes. Moreover, we do not assume prior knowledge of the context distribution as in (Huang et al., 2015; Li et al., 2022). Instead, we design an effective scheme to estimate the context distribution, and carefully bound the estimation error in our analysis.

Below we will introduce our algorithms and results. We defer all our proofs to Appendix due to space limit.

### 4. Representation Learning for Best Arm Identification in Linear Bandits

In this section, we design a computationally efficient algorithm DouExpDes for RepBPI-LB, which performs double delicate experimental designs to recover the feature extractor and distinguish the best arms using low-rank representations. Furthermore, we provide sample complexity guarantees that mainly depend on the underlying low dimension.

To better describe our algorithm, we first introduce the notion of *experimental design*. Experimental design is an important problem in statistics (Pukelsheim, 2006). Consider a set of feature vectors and an unknown linear regression parameter. Sampling each feature vector will produce a noisy feedback of the inner-product of this feature vector and the unknown parameter. Experimental design investigates how to schedule samples to maximize the statistical power of estimating the unknown parameter. In our algorithm, we mainly use two popular types of experimental design, i.e., E-optimal design, which minimizes the spectral norm of the inverse of sample covariance matrix, and G-optimal design, which minimizes the maximum prediction error for feature vectors.

#### 4.1. Algorithm DouExpDes

Now we present our algorithm DouExpDes, whose pseudocode is provided in Algorithm 1. DouExpDes is a phased elimination algorithm, which first conducts the E-optimal design to optimally schedule samples for learning the feature extractor \( B \), and then performs the G-optimal design with low-dimensional representations to eliminate suboptimal arms.

DouExpDes uses a *rounding procedure* \( \text{ROUND} \) (Allen-Zhu et al., 2017; Fiez et al., 2019), which transforms a given continuous sample allocation (design) into a discrete sample sequence and maintains important properties (e.g., E-optimality and G-optimality) of the design. \( \text{ROUND}([\{(q_i, Q_i)\}_{i=1}^{n'}, \lambda, \zeta, N]) \) takes \( n' \) arm-matrix pairs \( (q_1, Q_1), \ldots, (q_{n'}, Q_{n'}) \in X \times \mathbb{R}^{d \times d} \), a distribution \( \lambda \in \Delta_{\{q_1, \ldots, q_{n'}\}} \), a rounding approximation parameter \( \zeta > 0 \), and the number of samples \( N \) such that \( N \geq \frac{180d^2}{\varepsilon} \) as inputs. It will return a sample sequence \( s_1, \ldots, s_N \in \mathcal{X} \), which correspond to feature matrices \( S_1, \ldots, S_N \in \{Q_1, \ldots, Q_{n'}\} \), and \( \sum_{j=1}^{N} S_j \) has similar properties as the covariance matrix of the inputted design \( N \sum_{i=1}^{n'} \lambda(q_i)Q_i \) (see Appendix B for more details).
Multi-task Representation Learning for Pure Exploration in Linear Bandits

Algorithm 1 DouExpDes (Double Experimental Design)

1: **Input:** $\mathcal{X}$, $\delta$, rounding procedure ROUND, rounding approximation parameter $\zeta := \frac{1}{H}$, and the size of sample batch $p := \frac{18d}{\delta^2}$

2: Let $\lambda^E$ and $p^E$ be the optimal solution and the optimal value of the E-optimal design optimization:

$$\min_{\lambda \in \Delta_\mathcal{X}} \left\| \sum_{i=1}^{n} \lambda(x_i)x_i^T \right\|_1$$

3: $\vec{x}_1, \ldots, \vec{x}_p \leftarrow$ ROUND($\{(x_i, x_i^T)\}_{i=1}^n$, $\lambda^E$, $p^E$)

4: $\vec{X}_{1,m} \leftarrow \mathcal{X}$ for any $m \in [M]$, $\delta_t := \frac{\delta}{\sqrt{t}}$ for any $t \geq 1$

5: for phase $t = 1, 2, \ldots$

6: $T_t \leftarrow \lceil c_1(1+\zeta)^2p^E\lambda^E \rceil \log\left(\frac{M}{\delta_t^2}\right)$, polylog($\zeta, p^E, \delta_t, L_x, L_w, \frac{1}{\delta_t^2}$), where $c_1$ is an absolute constant

7: $B_t \leftarrow$ FeatRecover($T_t, \{\vec{x}_i\}_{i \in [p]}$)

8: $\{\vec{X}_{t+1,m}\}_{m \in [M]} \leftarrow$ EliLowRep($T, \{\vec{X}_{t,m}\}_{m \in [M]}$, $\delta_t$, ROUND, $\zeta$, $B_t$)

9: if $|\vec{X}_{t+1,m}| = 1, \forall m \in [M]$ then

10: return $\vec{X}_{t+1,m}$ for all tasks $m \in [M]$

12: end for

The procedure of DouExpDes is as follows. At the beginning, DouExpDes performs the E-optimal design with raw representations, to plan an optimal sample allocation $\lambda^E$ for the purpose of recovering the feature extractor $B$ (Line 2). Then, DouExpDes calls ROUND to convert the E-optimal sample allocation $\lambda^E$ into a discrete sample batch $\vec{x}_1, \ldots, \vec{x}_p$, which satisfies that

$$\left\| \sum_{j=1}^{p} \vec{x}_j\vec{x}_j^T \right\| \leq (1+\zeta)\left\| \sum_{i=1}^{n} \lambda^E(x_i)x_i^T \right\|.$$\]

Next, DouExpDes enters multiple phases, and maintains a candidate arm set $\vec{X}_{t,m}$ for each task. The specific value of $T_t$ in Line 6 is presented in Eq. (8) of Appendix C.2.

In each phase $t$, DouExpDes first calls subroutine FeatRecover to recover the feature extractor $B$. In FeatRecover (Algorithm 2), we repeatedly sample $\vec{x}_1, \ldots, \vec{x}_p$ in all tasks, and construct an estimator $Z$ for $\frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \bar{x}^T$, which contains the information of underlying reward parameters (Line 9). Then, we perform SVD on $Z$ and obtain the estimated feature extractor $\hat{B}$ (Line 10).

Then, DouExpDes calls subroutine EliLowRep to eliminate suboptimal arms using low-dimensional representations. In EliLowRep (Algorithm 3), we conduct the G-optimal design with the reduced-dimensional representations $\hat{B}^T \vec{x}$, and obtain sample allocation $\lambda^G_m$ for each task (Line 2). We further use ROUND to transform $\lambda^G_m$ into a sample sequence $z_m, 1, \ldots, z_m, N_m$, which satisfies that

$$\max_{\vec{x}, \vec{x}' \in \bar{X}_m} \left\| \vec{x} - \vec{x}' \right\|^2 \left(\sum_{j=1}^{N_m} \hat{B}^T z_{m,j} \bar{x}^T \bar{x}^T \hat{B} \right) \leq (1+\zeta) \max_{\vec{x}, \vec{x}' \in \bar{X}_m} \left\| \vec{x} - \vec{x}' \right\|^2 \left(\sum_{j=1}^{N_m} \lambda^G_m(x_i) \hat{B}^T z_{m,j} \bar{x}^T \bar{x}^T \hat{B} \right) \leq \frac{2^t}{2^t}.$$\]

After sampling this sequence, we build estimators $\hat{w}_{1,m}$ and $\hat{\theta}_{1,m}$ for the underlying prediction parameter $w_m$ and reward parameter $\theta_m$, respectively (Lines 7-8). Then, we discard the arms that show large gaps to the estimated optimal arm for each task (Line 9).

4.2. Theoretical Performance of DouExpDes

In this subsection, we provide sample complexity guarantees for DouExpDes. To formally present our sample complexity, complexity,
we first revisit existing results for conventional single-task best arm identification in linear bandits (BAI-LB).

For a single-task BAI-LB instance with arm set \( \mathcal{X} \in \mathbb{R}^d \) and underlying reward parameter \( \theta \in \mathbb{R}^d \), the instance-dependent hardness is defined as (Fiez et al., 2019)

\[
\rho^S(\mathcal{X}, \theta) := \min_{x \in \Delta \mathcal{X}, x \in \mathcal{X} \setminus \{x^*\}} \frac{\|x^* - x\|^2}{\left(\sum_{i=1}^{m} \lambda(x_i) x_i^2\right)^{-1}},
\]

and the best known sample complexity result is \( \tilde{O}(\rho^S(\mathcal{X}, \theta) \log \frac{1}{\delta}) = \tilde{O}\left(\frac{d}{\Delta_{\text{min}}} \log (\frac{1}{\delta})\right) \) (Fiez et al., 2019). Here \( x^* := \max_{x \in \mathcal{X}} x^\top \theta \) denotes the best arm, and \( \Delta_{\text{min}} := \min_{x \in \mathcal{X} \setminus \{x^*\}} (x^* - x)^\top \theta \) refers to the minimum reward gap.

It can be seen that a naive algorithm for RepBAI-LB is to run an existing single-task BAI-LB algorithm (Fiez et al., 2019; Katz-Samuels et al., 2020) to solve \( M \) tasks independently. Then, the sample complexity of such naive algorithm is

\[
\tilde{O}\left( \sum_{m=1}^{M} \rho^S(\mathcal{X}, \theta_m) \log \left( \frac{1}{\delta} \right) \right) = \tilde{O}\left( \frac{M d}{\Delta_{\text{min}}} \log \left( \frac{1}{\delta} \right) \right),
\]

where \( \Delta_{\text{min}} := \min_{m \in [M], x \in \mathcal{X} \setminus \{x_m^*\}} (x_m^* - x)^\top \theta_m \) denotes the minimum reward gap among all tasks. In the following, we take Eq. (1) as the baseline to demonstrate the power of representation learning.

Now we state the sample complexity for DouExpDes.

**Theorem 4.1.** With probability at least \( 1 - \delta \), algorithm DouExpDes returns the best arms \( x_m^* \) for all tasks \( m \in [M] \), and the number of samples used is bounded by

\[
\tilde{O}\left( \sum_{m=1}^{M} \min_{\lambda \in \Delta \mathcal{X}} \max_{x \in \mathcal{X} \setminus \{x_m^*\}} \frac{\|B^\top (x_m^* - x)\|^2}{\|x_m^*-x\|^2} \log \left( \frac{1}{\delta} \right) \right. \\
+ (\rho^E)^2 \bar{d}^4 \ell^2 L_x^2 L_u D \log \left( \frac{1}{\delta} \right) \\
= \tilde{O}\left( \frac{M k}{\Delta_{\text{min}}} \log \left( \frac{1}{\delta} \right) \right)
\]

where \( D := \max\left(\frac{1}{\Delta_{\text{min}}}, \frac{L_x^2}{\omega} \right) \).

**Remark 1.** In Theorem 4.1, the factors that have implicit dimensional dependency include

\[
\min_{\lambda \in \Delta \mathcal{X}} \max_{x \in \mathcal{X} \setminus \{x_m^*\}} \frac{\|B^\top (x_m^* - x)\|^2}{\|x_m^*-x\|^2} \quad \omega \text{ and } \rho^E,
\]

which scale as \( k^{1/2} \) and \( d \), respectively.

In our sample complexity bound (Eq. (2)), the first term,

\[
\sum_{m=1}^{M} \min_{\lambda \in \Delta \mathcal{X}} \max_{x \in \mathcal{X} \setminus \{x_m^*\}} \frac{\|B^\top (x_m^* - x)\|^2}{\|x_m^*-x\|^2} = \tilde{O}\left( \frac{M k}{\Delta_{\text{min}}} \right),
\]

represents the hardness of \( M \) k-dimensional linear bandit instances with arm set \( \mathcal{B}^\top x : x \in \mathcal{X} \) and underlying reward parameters \( w_1, \ldots, w_M \). This term only depends on the reduced dimension \( k \), instead of \( d \). In other words, it is an essential price that is needed for solving \( M \) low-dimensional tasks, even if one knows the feature extractor \( B \).

The second term \( (\rho^E)^2 \bar{d}^4 \ell^2 L_x^2 L_u D \), which depends on the raw dimension \( d \), is a cost paid for learning the feature extractor. Note that since this term does not contain \( M \), the cost for learning the underlying features is paid only once, rather than for all tasks.

When \( M \gg d \gg k \), the first term dominates the bound, which only depends on the low dimension \( k \). This indicates that algorithm DouExpDes effectively learns the low-dimensional representation, and exploits the intrinsic problem structure to reduce the sample complexity from \( \tilde{O}\left( \frac{M d}{\Delta_{\text{min}}} \log \left( \frac{1}{\delta} \right) \right) \) (i.e., learning each task independently) to only \( \tilde{O}\left( \frac{M k}{\Delta_{\text{min}}} \log \left( \frac{1}{\delta} \right) \right) \). Our result corroborates the benefits of representation learning for multi-task pure exploration.

**Technical Novelty.** We highlight the novelty in the analysis of Theorem 4.1 as follows. (i) Prior low-rank bandit works (Jun et al., 2019; Lu et al., 2021b) use arbitrary sample distributions to recover the low-dimensional subspace, and their results depend on the eigenvalue of an arbitrary sample distribution \( \|X^{-1}\| \), where \( X = [x^{(1)}, \ldots, x^{(d_k)}] \) is a collection of arbitrary \( d_k \) arms from the arm set. By contrast, we utilize the E-optimality of the sample batch \( \bar{x}_1, \ldots, \bar{x}_p \) to obtain an optimized dependency \( \rho^E \approx \min_{x^{(1)}, \ldots, x^{(d_k)}} \|X^{-1}\| \), which is the best one can achieve at the subspace recovery stage. (ii) If one naively applies existing single-task BAI-LB analysis (Fiez et al., 2019; Katz-Samuels et al., 2020) in the estimated subspace \( \hat{B}_t \), one can only obtain a sample complexity \( \|B_t^\top (x - x^\top)\|^2 \left(\sum_{i=1}^{d_k} \lambda_i^*(x_i) B_t^\top x_i x_i^\top B_t \right)^{-1} \) dependent on \( \hat{B}_t \), but this is not a valid upper bound. To tackle this challenge, we connect the low-dimensional sample complexity under the estimated subspace \( \|B_t^\top (x - x^\top)\|^2 \left(\sum_{i=1}^{d_k} \lambda_i^*(x_i) B_t^\top x_i x_i^\top B_t \right)^{-1} \) with that under the true subspace \( \|B^\top (x - x^\top)\|^2 \left(\sum_{i=1}^{d_k} \lambda_i^*(x_i) B^\top x_i x_i^\top B \right)^{-1} \), and drive a tight sample complexity.

**Lower Bound Conjecture.** We conjecture that the lower bound for RepBAI-LB is \( \Omega\left( \sum_{m=1}^{M} \rho^S(\mathcal{X}, \theta_m) \log \left( \frac{1}{\delta} \right) \right) \). We describe the preliminary idea below.

First, the lower bound for single-task BAI-LB with arm set \( \mathcal{X} \) and underlying reward parameter \( \theta_m \) is \( \Omega\left( \rho^S(\mathcal{X}, \theta_m) \log \left( \frac{1}{\delta} \right) \right) \) (Fiez et al., 2019). If the global feature extractor \( B \) is known, then the RepBAI-LB problem will reduce to \( M \) \( k \)-dimensional BAI-LB instances with arm set \( \{B^\top x : x \in \mathcal{X}\} \) and underlying reward parameters \( w_1, \ldots, w_M \). Therefore, we conjecture that the lower bound for RepBAI-LB is \( \Omega\left( \sum_{m=1}^{M} \rho^S(\mathcal{X}, \theta_m) \log \left( \frac{1}{\delta} \right) \right) \), which is the cost of solving \( M \) \( k \)-dimensional BAI-LB instances. However, it is challenging to rigorously analyze
5. Representation Learning for Best Policy Identification in Contextual Linear Bandits

In this section, we turn to contextual linear bandits. Different from prior contextual linear bandit works, e.g., (Huang et al., 2015; Li et al., 2022), here we do not assume any knowledge of context distribution. As a result, our RepBPI-CLB problem faces several unique challenges: (i) how to plan an efficient sample allocation for recovering the feature extractor in advance under an unknown context distribution, and (ii) how to construct an estimator for the feature extractor with a partially observed context space.

We propose algorithm C-DouExpDes, which first (i) efficiently estimates the context distribution and conducts experimental designs under the estimated context distribution, and then (ii) builds a delicate estimator for the feature extractor using instantaneous contexts. Moreover, we also establish a sample complexity guarantee for C-DouExpDes, which mainly depends on the low dimension of the common representation among tasks.

5.1. Algorithm C-DouExpDes

Algorithm 4 presents the pseudo-code of C-DouExpDes. At the beginning, C-DouExpDes uses $T_0$ samples to estimate the context distribution $D$ (Lines 3-6). Then, it performs the E-optimal design under the estimated context distribution $D$, and obtains an efficient sample allocation $\lambda_E^p$ for the purpose of recovering the feature extractor $B$ (Line 7).

Further, C-DouExpDes calls the rounding procedure ROUND.
to transform $\lambda^D_{\xi}$ into a sample batch $\bar{a}_1, \ldots, \bar{a}_p$, such that
$$
\left\| \left( \sum_{j=1}^{p} \mathbb{E}_{s \sim \mathcal{D}} \left[ \phi(s, \bar{a}_j)\phi(s, \bar{a}_j)^\top \right] \right)^{-1} \right\| 
\leq (1 + \zeta) \left\| \left( \sum_{a \in \mathcal{A}} \lambda^D_{\xi}(a)\mathbb{E}_{s \sim \mathcal{D}} \left[ \phi(s, a)\phi(s, a)^\top \right] \right)^{-1} \right\|.
$$

The specific values of $p$ and $T$ in Lines 1, 9 are provided in Eq. (19) of Appendix D.1 and Eq. (29) of Appendix D.2, respectively.

Next, C-DouExpDes runs subroutine C-FeatRecover to estimate the feature extractor $B$ using the sample batch $\bar{a}_1, \ldots, \bar{a}_p$. In C-FeatRecover (Algorithm 5), we repeatedly sample $\bar{a}_1, \ldots, \bar{a}_p$ in all tasks with random contexts. In Lines 4-5, we sample this batch twice, and the superscripts $1, \ldots, n_{\text{tr}}$ denote the first and second samples, respectively. After sampling, we carefully establish an estimator $Z$ for the reward parameter related matrix $\frac{1}{M} \sum_{m=1}^{M} \theta_m \theta_m^\top$, using instantaneous context-action features $\bar{\phi}(s^{(t)}_{m,j,i}, \bar{a}_i)^\top$. Then, we perform SVD decomposition on $Z$ to obtain the estimated feature extractor $B$ (Lines 11-12).

Then, C-DouExpDes calls subroutine EstLowRep, which adapts existing reward-free-exploration algorithm in (Zanette et al., 2021) with low-rank representations to estimate $\theta_m$. In EstLowRep (Algorithm 6), we employ the estimated representation $B^\top \phi(s, a)$ to sample the actions with the maximum uncertainty under the observed contexts. After that, we construct estimators $\bar{w}_{m,t}$ and $\theta_{m,t}$ for the prediction parameter $\bar{w}_m$ and reward parameter $\theta_m$ (Lines 8-9). At last, C-DouExpDes returns the greedy policy with respect to the estimated reward parameter $\theta_{m,N}$ for each task.

### 5.2. Theoretical Performance of C-DouExpDes

Next, we establish sample complexity guarantees for algorithm C-DouExpDes. In order to illustrate the advantages of representation learning, we first review existing results for traditional single-task best policy identification in contextual linear bandits (BPI-CLB). For a single BPI-CLB instance with context-action features $\phi(s, a) \in \mathbb{R}^d$ and reward parameter $\theta \in \mathbb{R}^d$, the best known sample complexity is $O\left(\frac{d^2 \log \left(\frac{2}{\delta}\right)}{\varepsilon^2}\right)$ (Zanette et al., 2021; Li et al., 2022).

Apparently, if one naively solves the RepBPI-CLB problem by running single-task BPI-CLB algorithms to tackle $M$ tasks independently, one will have a sample complexity $O\left(\frac{M d^2}{\varepsilon^2} \log \left(\frac{1}{\delta}\right)\right)$, which heavily depends on the raw dimension $d$ of context-action features. The goal of representation learning is to leverage the common representation among tasks to alleviate the dependency of dimension and save samples.

Now we present the sample complexity for C-DouExpDes.

**Theorem 5.1.** With probability at least $1 - \delta$, C-DouExpDes returns an $\varepsilon$-optimal policy $\pi_m$ such that $\mathbb{E}_{s \sim \mathcal{D}}[\max_{a \in \mathcal{A}} \phi(s, a) - \phi(s, \pi_m(s))^\top \theta_m] \leq \varepsilon$ for each task $m \in [M]$, and the number of samples used is

$$
\tilde{O}\left(\frac{M(k^2 + \gamma k L_w^2)}{\varepsilon^2} + \frac{k^4 L_w^4 L_D^4}{\nu^4 \varepsilon^2}\right).
$$

**Remark 2.** In this result, only factor $\nu$ has implicit dimensional dependency, which scales as $\frac{1}{\varepsilon^2}$. The first term $\frac{M(k^2 + \gamma k L_w^2)}{\varepsilon^2}$ is a cost of identifying optimal policies for $M$ tasks with $k$-dimensional features $B^\top \phi(s, a)$. The second term $\frac{k^4 L_w^4 L_D^4}{\nu^4 \varepsilon^2}$ is a price paid for learning global feature extractor $B$ and does not depend on $M$. This indicates that we only need to pay this price once, and then enjoy the benefits of dimension reduction for all $M$ tasks.

When $M \gg \frac{1}{\nu} \gg k$, this result becomes $\tilde{O}(\frac{M k^2}{\varepsilon^2})$ and only depends on the low dimension $k$, which implies that C-DouExpDes performs as well as an oracle that knows the underlying low-rank subspace $B$. This sample complexity significantly outperforms the baseline result $\tilde{O}(\frac{M d^2}{\varepsilon^2})$ (i.e., solving $M$ tasks independently), and demonstrates the power of representation learning.

**Analytical Novelty.** Below we elaborate the novelty in the proof of Theorem 5.1. (i) We carefully bound the deviation between the context-action features under the estimated context distribution $\mathbb{E}_{s \sim \mathcal{D}}[\phi(s, \bar{a}_i)\phi(s, \bar{a}_i)^\top]$ and those under the true context distribution $\mathbb{E}_{s \sim \mathcal{D}}[\phi(s, a_1)\phi(s, a_1)^\top]$. We further bound the distance between $\mathbb{E}_{s \sim \mathcal{D}}[\phi(s, \bar{a}_i)\phi(s, \bar{a}_i)^\top]$ and the context-action features under actual instantaneous contexts $\phi(s^{(t)}_{m,j,i}, \bar{a}_i)\phi(s^{(t)}_{m,j,i}, \bar{a}_i)^\top$. (ii) We leverage the E-optimality of the sample batch $\bar{a}_1, \ldots, \bar{a}_p$ to bound $\left\| (\sum_{i=1}^{p} \phi^{(t)}_{m,j,i,\bar{a}_i})^{-1} \right\|$. Then, we establish a concentration inequality for $\|Z - \frac{1}{M} \sum_{m=1}^{M} \theta_m \theta_m^\top\|$ using the bounded $\left\| (\sum_{i=1}^{p} \phi^{(t)}_{m,j,i,\bar{a}_i})^{-1}\right\|$ and matrix Bernstein inequality with truncated noises. (iii) Furthermore, we decompose the prediction error $\phi(s, a)^\top (\theta_{m,t} - \theta_m)$ into three components, including the sample variance and bias of $\bar{w}_{m,t}$, and the estimation error of $B$. This prediction error is bounded via self-normalized concentration inequalities with the reduced dimension $k$.

### 6. Experiments

In this section, we present experiments to evaluate the empirical performance of our algorithms.

In our experiments, we set $\delta = 0.005$, $d = 5$, $k = 2$ and $M \in [50, 230]$, where $k$ divides $M$. In RepBAI-LB, $\lambda$ is
the sample complexity of our algorithms mainly depend on the low dimension of the underlying joint representation among tasks, instead of the raw high dimension. Our theoretical and experimental results demonstrate the benefit of representation learning for pure exploration in multi-task bandits. There are many interesting directions for further exploration. One direction is to establish lower bounds to validate the optimality of our algorithms. Another direction is to extend this work to more complex (nonlinear) representation settings.

Acknowledgements

The work of Yihan Du and Longbo Huang is supported by the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grant 2020AAA0108400 and 2020AAA0108403 and the Tsinghua Precision Medicine Foundation 10001020109. Wen Sun acknowledges funding support from NSF IIS-2154711.

References


Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. IEEE Transac-
Multi-task Representation Learning for Pure Exploration in Linear Bandits


Appendix

A. Related Work

In this section, we present a full literature review for two lines of related works, i.e., representation learning and pure exploration in (contextual) linear bandits.

**Representation Learning.** The study of representation learning has been initiated and developed in the supervised learning setting, e.g., (Baxter, 2000; Ben-David & Schuller, 2003; Ando et al., 2005; Maurer, 2006; Cavallanti et al., 2010; Maurer et al., 2016; Du et al., 2021a; Tripuraneni et al., 2021). A most related work is (Tripuraneni et al., 2021), which proposes a method-of-moments estimator for recovering the feature extractor, and establishes error guarantees for transferring the learned representation from past tasks to a new task.

Recently, representation learning for sequential decision making (bandits and reinforcement learning) has attracted extensive attention. We first introduce several works on low-rank bandits, which is a very similar topic to representation learning for bandits. Lale et al. (2019) study linear bandits with a hidden low-rank structure, and provide a regret bound dependent on the eigenvalue of the action distribution covariance. Jun et al. (2019); Lu et al. (2021b) also investigate low-rank linear bandits (bilinear bandits), and design algorithms which run traditional linear bandit algorithm LinUCB (Abbasi-Yadkori et al., 2011) in the estimated low-dimensional subspace. Lattimore & Hao (2021) consider an instantiation of low-rank bandits, called bandit phase retrieval. Huang et al. (2021) study a large family of bandit problems with non-concave reward functions, including low-rank linear bandits. They design a stochastic gradient-based algorithm that achieves an improved regret bound over those in (Jun et al., 2019; Lu et al., 2021b).

Now we introduce related works on representation learning for bandits. Yang et al. (2021; 2022) study multi-task representation learning for linear bandits with the regret minimization objective, and assume that the action set at each timestep is an ellipsoid or sphere. Hu et al. (2021) further relax this assumption and allow arbitrary action sets, but their algorithms equipped with a multi-task joint least-square estimator are computationally inefficient. Cella et al. (2022a;b) also investigate the problem in (Yang et al., 2021) and propose algorithms which do not need to know the dimension of the underlying representation. Qin et al. (2022) study multi-task representation learning for linear bandits in a non-stationary environment, and develop algorithms that learn and transfer non-stationary representations adaptively.

There are also other works studying multi-task representation learning for reinforcement learning (RL). Lu et al. (2021a; 2022) consider multi-task representation learning for linear MDPs, where the agent learns a shared representation function from a given function class. Pacchiano et al. (2022) investigate multi-task RL with a joint low-dimensional linear representation, and design a computationally efficient algorithm using a bilinear optimization oracle. Zhang & Wang (2021) consider multi-task (multi-player) RL in tabular MDPs, where the relatedness of MDPs are measured by the similarity of reward functions and transition distributions. Cheng et al. (2022); Agarwal et al. (2022) study multi-task representation learning and representational transfer for low-rank MDPs, where multiple low-rank MDPs share a common state-action feature mapping.

Different from the above works which consider regret minimization, we study representation learning for (contextual) linear bandits with the pure exploration objective, which imposes unique challenges on how to optimally allocate samples to learn the feature extractor, and motivates us to design algorithms based on double experimental designs.

**Pure Exploration in (Contextual) Linear Bandits.** Most linear bandit studies consider regret minimization, e.g., (Dani et al., 2008; Rusmevichientong & Tsitsiklis, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011). Recently, there is a surge of interests in pure exploration for (contextual) linear bandits, e.g., (Soare et al., 2014; Tao et al., 2018; Xu et al., 2018; Fiez et al., 2019; Katz-Samuels et al., 2020; Degenne et al., 2020; Jedra & Proutiere, 2020; Du et al., 2021b; Zanette et al., 2021; Li et al., 2022). For linear bandits, Soare et al. (2014) firstly apply the G-optimal design to identify the best arm, and provide a sample complexity result that heavily depends on the minimum reward gap. Tao et al. (2018) design a novel randomized estimator for the underlying reward parameter, and achieve tighter sample complexity which depends on the reward gaps of the best $d$ arms. Du et al. (2021b) further extend the algorithm in (Tao et al., 2018) to develop a polynomial-time algorithm for combinatorially large arm sets. Xu et al. (2018) propose a fully-adaptive algorithm which changes the arm selection strategy at each timestep. Fiez et al. (2019) establish the first near-optimal sample complexity upper and lower bounds for best arm identification in linear bandits. Katz-Samuels et al. (2020) further extend the algorithm in (Fiez et al., 2019) and use empirical processes to avoid an explicit union bound over the number of arms. Degenne et al. (2020); Jedra & Proutiere (2020) develop asymptotically optimal algorithms using the track-and-stop approaches. For contextual linear
Multi-task Representation Learning for Pure Exploration in Linear Bandits

bandits, Zanette et al. (2021) design a single non-adaptive policy to collect a dataset, from which a near-optimal policy can be computed. Li et al. (2022) build the first instance-dependent upper and lower bounds for best policy identification in contextual linear bandits, with the prior knowledge of the context distribution. By contrast, our work studies multi-task best arm/policy identification in (contextual) linear bandits with a shared representation among tasks, and does not assume any prior knowledge of the context distribution.

B. Rounding Procedure

In this section, we introduce the rounding procedure ROUND in detail.

Let $\mathcal{X}^+ := \mathcal{X} \cup \mathcal{A}$ denote the union space of arm set $\mathcal{X}$ and action space $\mathcal{A}$. There are $n$ arms or actions $p_1, \ldots, p_n \in \mathcal{X}^+$ and $n$ positive semi-definite matrices $Q_1, \ldots, Q_n \in \mathbb{S}_+^d$, where $Q_i$ represents the feature of arm or action $p_i$ for any $i \in [n]$. Denote $\mathcal{P} := \{p_1, \ldots, p_n\}$ and $\mathcal{Q} := \{Q_1, \ldots, Q_n\}$.

The rounding procedure $\text{ROUND}((\{p_i, Q_i\})_{i=1}^n, \lambda, \zeta, N)$ (Allen-Zhu et al., 2017; Fiez et al., 2019) takes $n$ arm-matrix or action-matrix pairs $(p_1, Q_1), \ldots, (p_n, Q_n) \in \mathcal{X}^+ \times \mathbb{S}_+^d$, a distribution $\lambda \in \Delta_\mathcal{P}$ (or equivalently, $\lambda \in \Delta_\mathcal{Q}$), an approximation parameter $\zeta > 0$, and the number of samples $N$ which satisfies that $N \geq \frac{180d}{\zeta^2}$ as inputs. Roughly speaking, it will find a $N$-length discrete arm or action sequence whose associated feature matrices maintain the similar property (e.g., G-optimality and E-optimality) as the continuous sample allocation $\lambda$.

Formally, $\text{ROUND}((\{p_i, Q_i\})_{i=1}^n, \lambda, \zeta, N)$ returns a discrete sample sequence $s_1, \ldots, s_N \in \mathcal{P}^N$ associated with feature matrices $S_1, \ldots, S_N \in \mathbb{Q}^N$, which satisfy the following properties:

(i) If $\lambda$ is an E-optimal design, i.e., $\lambda$ is the optimal solution of the optimization

$$
\min_{\lambda \in \Delta_\mathcal{Q}} \left\| \sum_{i=1}^n \lambda(Q_i)Q_i^{-1} \right\|,
$$

then $S_1, \ldots, S_N$ satisfy that

$$
\left\| \left( \sum_{j=1}^N S_j \right)^{-1} \right\| \leq (1 + \zeta) \left\| \left( N \sum_{i=1}^n \lambda(Q_i)Q_i^{-1} \right)^{-1} \right\|.
$$

(ii) If $\lambda$ is a G-optimal design, i.e., for a given prediction set $\mathcal{Y} \subseteq \mathbb{R}^d$, $\lambda$ is the optimal solution of the optimization

$$
\min_{\lambda \in \Delta_\mathcal{Q}} \max_{y \in \mathcal{Y}} \frac{\|y\|^2}{\left( \sum_{i=1}^n \lambda(Q_i)Q_i^{-1} \right)^{-1}},
$$

then $S_1, \ldots, S_N$ satisfy that

$$
\max_{y \in \mathcal{Y}} \frac{\|y\|^2}{\left( \sum_{j=1}^N S_j \right)^{-1}} \leq (1 + \zeta) \max_{y \in \mathcal{Y}} \frac{\|y\|^2}{\left( N \sum_{i=1}^n \lambda(Q_i)Q_i^{-1} \right)^{-1}}.
$$

We implement $\text{ROUND}$ by setting $\pi^* = N\lambda$, $k = r = N$ and $x; x_i^T = \left( \sum_{i=1}^n \pi^*(Q_i)Q_i^{-1} \right)^{-\frac{1}{2}} Q_i \left( \sum_{i=1}^n \pi^*(Q_i)Q_i^{-1} \right)^{-\frac{1}{2}}$ for any $i \in [n]$ in Algorithm 1 of (Allen-Zhu et al., 2017). Note that Algorithm 1 in (Allen-Zhu et al., 2017) only needs to access the feature matrix $x, x_i^T$ rather than the separate feature vector $x_i$, which allows us to apply it to our problem. We refer interested readers to (Allen-Zhu et al., 2017) and Appendix B in (Fiez et al., 2019) for more implementation details of this rounding procedure.

C. Proofs for Algorithm DouExpDes

In this section, we provide the proofs for Algorithm DouExpDes.

Throughout our proofs, we use $L_\theta$ to denote the upper bound of $\|\theta_m\|$ for any $m \in [M]$. Since $\theta_m = Bw_m$ for any $m \in [M]$, we have that $\|\theta_m\| \leq \|B\|\|w_m\| \leq \|w_m\| \leq L_w$, and thus $L_\theta \leq L_w$. 

13
C.1. Sample Batch Planning

Recall that

$\lambda^E := \arg\min_{\lambda \in \Delta} \left\| \left( \sum_{i=1}^{n} \lambda(x_i) x_i x_i^\top \right)^{-1} \right\|$

and

$\rho^E := \min_{\lambda \in \Delta} \left\| \left( \sum_{i=1}^{n} \lambda(x_i) x_i x_i^\top \right)^{-1} \right\|$

are the optimal solution and the optimal value of the E-optimal design optimization, respectively (Line 2 in Algorithm 1). $\bar{x}_1, \ldots, \bar{x}_p$ is an arm sequence generated according to sample allocation $\lambda^E$ via rounding procedure ROUND (Line 3 in Algorithm 1).

Let

$X_{\text{batch}} := \begin{bmatrix} \bar{x}_1 \top \\ \vdots \\ \bar{x}_p \top \end{bmatrix}$

and

$X_{\text{batch}}^+ := (X_{\text{batch}} \top X_{\text{batch}})^{-1} X_{\text{batch}}^\top$. According to the fact that $X$ spans $\mathbb{R}^d$, the definition of E-optimal design and the guarantee of ROUND, we have that $X_{\text{batch}} X_{\text{batch}}$ is invertible.

Now, we first give an upper bound of $\|X_{\text{batch}}^+\|$.

**Lemma C.1.** It holds that

$$\|X_{\text{batch}}^+\| \leq \sqrt{\frac{(1 + \zeta) \rho^E}{p}}.$$

**Proof of Lemma C.1.** We have

$$\|X_{\text{batch}}^+\| = \left\| (X_{\text{batch}} \top X_{\text{batch}})^{-1} X_{\text{batch}}^\top \right\|$$

$$= \sqrt{\left\| (X_{\text{batch}} \top X_{\text{batch}})^{-1} X_{\text{batch}} X_{\text{batch}} (X_{\text{batch}} \top X_{\text{batch}})^{-1} \right\|}$$

$$= \sqrt{\left\| (X_{\text{batch}} \top X_{\text{batch}})^{-1} \right\|}$$

$$= \sqrt{\left\| \left( \sum_{i=1}^{p} \bar{x}_i \bar{x}_i^\top \right)^{-1} \right\|}$$

$$\leq (1 + \zeta) \left\| \left( \sum_{i=1}^{n} \lambda^E(x_i) x_i x_i^\top \right)^{-1} \right\|$$

$$= \sqrt{\frac{(1 + \zeta) \rho^E}{p}}.$$

$\square$
C.2. Global Feature Extractor Recovery

For clarity of notation, we add subscript $t$ to the notations in subroutine $\text{FeatRecover}$ to denote the quantities generated in phase $t$. Specifically, we use $\alpha_{t,m,j,i}$, $\theta_{t,m,j}$, $Z_t$ and $\hat B_t$ to denote the random reward, estimator of reward parameter, estimator of $\frac{1}{MT} \sum_{t=1}^{MT} \theta_m \theta_m^\top$ and estimator of feature extractor in phase $t$, respectively.

For any phase $t > 0$, task $m \in [M]$, round $j \in [T_t]$ and arm $i \in [p]$, let $\eta_{t,m,j,i}$ denote the noise of the sample on arm $\hat x_i$ in the $j$-th round for task $m$, during the execution of $\text{FeatRecover}$ in phase $t$ (Line 4 in Algorithm 2). The noise $\eta_{t,m,j,i}$ is zero-mean and sub-Gaussian, and has variance $\frac{1}{T_t}$. Specifically, we use $\alpha_t := \alpha_{t,m,j,i}$. During the execution of $\text{FeatRecover}$, let $\alpha_t$ in phase $t$, for task $m$, be the $j$-th round for task $m$.

For any phase $t > 0$, task $m \in [M]$, round $j \in [T_t]$, let $\alpha_{t,m,j} := \begin{bmatrix} \alpha_{t,m,j,1}, \ldots, \alpha_{t,m,j,p} \end{bmatrix}^\top$. Then, we have that

$$\theta_{t,m,j} = X_{\text{batch}} \alpha_{t,m,j},$$

and

$$Z_t = \frac{1}{MT_t} \sum_{m=1}^{M} \sum_{j=1}^{T_t} \theta_{t,m,j}(\theta_{t,m,j})^\top - X_{\text{batch}} + (X_{\text{batch}})^\top.$$

**Lemma C.2** (Expectation of $Z_t$). It holds that

$$E[Z_t] = \frac{1}{M} \sum_{m=1}^{M} \theta_m \theta_m^\top.$$

**Proof of Lemma C.2.** $Z_t$ can be written as

$$Z_t = \frac{1}{MT_t} \sum_{m=1}^{M} \sum_{j=1}^{T_t} \theta_{t,m,j}(\theta_{t,m,j})^\top - X_{\text{batch}} + (X_{\text{batch}})^\top$$

$$= \frac{1}{MT_t} \sum_{m=1}^{M} \sum_{j=1}^{T_t} X_{\text{batch}}^\top \begin{bmatrix} \alpha_{t,m,j,1} & \vdots & \alpha_{t,m,j,p} \end{bmatrix} \begin{bmatrix} \alpha_{t,m,j,1} & \vdots & \alpha_{t,m,j,p} \end{bmatrix}^\top (X_{\text{batch}})^\top - X_{\text{batch}} + (X_{\text{batch}})^\top$$

$$= \frac{1}{MT_t} \sum_{m=1}^{M} \sum_{j=1}^{T_t} X_{\text{batch}}^\top \begin{bmatrix} x_1^\top \theta_m + \eta_{t,m,j,1} & \ldots & x_p^\top \theta_m + \eta_{t,m,j,p} \end{bmatrix} \begin{bmatrix} x_1^\top \theta_m + \eta_{t,m,j,1} & \ldots & x_p^\top \theta_m + \eta_{t,m,j,p} \end{bmatrix}^\top (X_{\text{batch}})^\top - X_{\text{batch}} + (X_{\text{batch}})^\top$$

$$= \frac{1}{MT_t} \sum_{m=1}^{M} \sum_{j=1}^{T_t} X_{\text{batch}}^\top \left[ \begin{bmatrix} (x_1^\top \theta_m)^2 & \ldots & x_1^\top \theta_m x_p^\top \theta_m \\ x_2^\top \theta_m x_m^\top \theta_m & \ldots & (x_p^\top \theta_m)^2 \end{bmatrix} + \begin{bmatrix} 2x_1^\top \theta_m \eta_{t,m,j,1} & \ldots & x_1^\top \theta_m \eta_{t,m,j,p} + x_p^\top \theta_m \eta_{t,m,j,1} \\ \ldots & \ldots & \ldots \\ x_1^\top \theta_m \eta_{t,m,j,p} + x_p^\top \theta_m \eta_{t,m,j,1} & \ldots & 2x_p^\top \theta_m \eta_{t,m,j,p} \end{bmatrix} + \begin{bmatrix} (\eta_{t,m,j,1})^2 & \ldots & \eta_{t,m,j,1} \eta_{t,m,j,p} \\ \ldots & \ldots & \ldots \\ \eta_{t,m,j,1} \eta_{t,m,j,p} & \ldots & (\eta_{t,m,j,p})^2 \end{bmatrix} \right] (X_{\text{batch}})^\top - X_{\text{batch}} + (X_{\text{batch}})^\top.$$

Then, taking the expectation on $Z_t$, we have

$$E[Z_t] = \frac{1}{MT_t} \sum_{m=1}^{M} \sum_{j=1}^{T_t} X_{\text{batch}}^\top \left[ \begin{bmatrix} (x_1^\top \theta_m)^2 & \ldots & x_1^\top \theta_m x_p^\top \theta_m \\ x_2^\top \theta_m x_m^\top \theta_m & \ldots & (x_p^\top \theta_m)^2 \end{bmatrix} + I_d \right] (X_{\text{batch}})^\top - X_{\text{batch}} + (X_{\text{batch}})^\top.$$
Lemma C.3

Define the following matrices:

$$
\begin{align*}
A_{t,m,j} & := \frac{1}{MT_t} \sum_{m=1}^M \sum_{j=1}^{T_t} X_{\text{batch}}^+ \begin{bmatrix}
2\theta_1^\top \eta_{t,m,j,1} & \cdots & 2\theta_1^\top \eta_{t,m,j,p} + \theta_1^\top \eta_{t,m,j,1} \\
\vdots & \ddots & \vdots \\
2\theta_1^\top \eta_{t,m,j,p} + \theta_1^\top \eta_{t,m,j,1} & \cdots & 2\theta_1^\top \eta_{t,m,j,p}
\end{bmatrix}
\end{align*}
$$

$$
A_t := \sum_{m=1}^M \sum_{j=1}^{T_t} A_{t,m,j},
$$

Recall that for any $t > 0$, $\delta_t := \frac{\delta}{2^{t^2}}$.

For any phase $t > 0$, define events

$$
E_t := \left\{ \|Z_t - \mathbb{E}[Z_t]\| \leq \frac{96 \|X_{\text{batch}}^+\|^2 pL_2 L_2 \log \left( \frac{16p}{\delta_t} \frac{1}{16pMT_t} \right)}{\sqrt{MT_t}} \log \left( \frac{16pMT_t}{\delta_t} \right) \right\},
$$

and

$$
E := \cap_{t=1}^\infty E_t.
$$

Lemma C.3 (Concentration of $Z_t$). It holds that

$$
\Pr[E] \geq \frac{\delta}{2}.
$$

Proof of Lemma C.3. According to Eq. (3), we have

$$
Z_t - \mathbb{E}[Z_t] = \frac{1}{MT_t} \sum_{m=1}^M \sum_{j=1}^{T_t} X_{\text{batch}}^+ \left( \begin{bmatrix}
2\theta_1^\top \eta_{t,m,j,1} & \cdots & 2\theta_1^\top \eta_{t,m,j,p} + \theta_1^\top \eta_{t,m,j,1} \\
\vdots & \ddots & \vdots \\
2\theta_1^\top \eta_{t,m,j,p} + \theta_1^\top \eta_{t,m,j,1} & \cdots & 2\theta_1^\top \eta_{t,m,j,p}
\end{bmatrix} - \mathbb{E} \begin{bmatrix}
2\theta_1^\top \eta_{t,m,j,1} & \cdots & 2\theta_1^\top \eta_{t,m,j,p} + \theta_1^\top \eta_{t,m,j,1} \\
\vdots & \ddots & \vdots \\
2\theta_1^\top \eta_{t,m,j,p} + \theta_1^\top \eta_{t,m,j,1} & \cdots & 2\theta_1^\top \eta_{t,m,j,p}
\end{bmatrix} \right) (X_{\text{batch}}^+)^\top.
$$

Define the following matrices:

$$
A_{t,m,j} := \frac{1}{MT_t} \sum_{m=1}^M \sum_{j=1}^{T_t} X_{\text{batch}}^+ \begin{bmatrix}
2\theta_1^\top \eta_{t,m,j,1} & \cdots & 2\theta_1^\top \eta_{t,m,j,p} + \theta_1^\top \eta_{t,m,j,1} \\
\vdots & \ddots & \vdots \\
2\theta_1^\top \eta_{t,m,j,p} + \theta_1^\top \eta_{t,m,j,1} & \cdots & 2\theta_1^\top \eta_{t,m,j,p}
\end{bmatrix},
$$

$$
A_t := \sum_{m=1}^M \sum_{j=1}^{T_t} A_{t,m,j}.
$$
Then, we have:

\[ C_{t,m,j} := \frac{1}{MT_t} \begin{bmatrix} \eta_{t,m,j,1} \cdots \eta_{t,m,j,p} \\ \eta_{t,m,j,1} \cdots \eta_{t,m,j,p} \cdots \\ \eta_{t,m,j,1} \cdots (\eta_{t,m,j,p})^2 \end{bmatrix}, \]

\[ C_t := \sum_{m=1}^{M} \sum_{j=1}^{T_t} C_{t,m,j}. \]

Then, we can write \( Z_t - \mathbb{E}[Z_t] \) as:

\[ Z_t - \mathbb{E}[Z_t] = X_{\text{batch}}^+ (A_t - \mathbb{E}[A_t] + C_t - \mathbb{E}[C_t]) (X_{\text{batch}}^+)\top, \]

and thus,

\[ \| Z_t - \mathbb{E}[Z_t] \| \leq \| X_{\text{batch}}^+ \| \| (A_t - \mathbb{E}[A_t]) + (C_t - \mathbb{E}[C_t]) \| \].

Next, we analyze \( \| A_t - \mathbb{E}[A_t] \| \) and \( \| C_t - \mathbb{E}[C_t] \| \). In order to use the truncated matrix Bernstein inequality (Lemma E.2), we define the truncated noise and truncated matrices as follows.

Let \( R > 0 \) be a truncation level of noises, which will be chosen later. For any \( t > 0, m \in [M], j \in [T_t] \) and \( i \in [p] \), let \( \bar{\eta}_{t,m,j,i} = \eta_{t,m,j,i} \mathbb{I}\{ \| \eta_{t,m,j,i} \| \leq R \} \) denote the truncated noise. Then, we define the following truncated matrices:

\[ \tilde{A}_{t,m,j} := \frac{1}{MT_t} \begin{bmatrix} 2\bar{x}_{1}\theta_m \bar{\eta}_{t,m,j,1} \cdots & \cdots & \bar{x}_1 \theta_m \bar{\eta}_{t,m,j,p} + \bar{x}_p \theta_m \bar{\eta}_{t,m,j,1} \\ \cdots & \cdots & \cdots \\ \bar{x}_1 \theta_m \bar{\eta}_{t,m,j,p} + \bar{x}_p \theta_m \bar{\eta}_{t,m,j,1} \cdots & \cdots & 2\bar{x}_p \theta_m \bar{\eta}_{t,m,j,p} \end{bmatrix} \]

\[ \tilde{A}_t := \sum_{m=1}^{M} \sum_{j=1}^{T_t} \tilde{A}_{t,m,j}, \]

\[ \tilde{C}_{t,m,j} := \frac{1}{MT_t} \begin{bmatrix} \bar{\eta}_{t,m,j,1} \cdots \bar{\eta}_{t,m,j,1} \bar{\eta}_{t,m,j,p} \\ \cdots \cdots \cdots \\ \bar{\eta}_{t,m,j,1} \cdots (\bar{\eta}_{t,m,j,p})^2 \end{bmatrix} \]

\[ \tilde{C}_t := \sum_{m=1}^{M} \sum_{j=1}^{T_t} \tilde{C}_{t,m,j}. \]

First, we bound \( \| A_t - \mathbb{E}[A_t] \| \). Since for any \( t > 0, m \in [M], j \in [T_t] \) and \( i \in [p], |\bar{\eta}_{t,m,j,i}| \leq R \) and \( |\bar{x}_i \theta_m| \leq L_x L_\theta \), we have \( \| \tilde{A}_{t,m,j} \| \leq \frac{1}{MT_t} \cdot 2pL_x L_\theta R \).

Recall that for any \( t > 0, m \in [M], j \in [T_t] \) and \( i \in [p], \eta_{t,m,j,i} \) is 1-sub-Gaussian. Using a union bound over \( i \in [p] \), we have that for any \( t > 0, m \in [M], j \in [T_t] \), with probability at least \( 1 - 2p \exp(-\frac{R^2}{2}) \), \( |\eta_{t,m,j,i}| \leq R \) for all \( i \in [p] \). Thus, with probability at least \( 1 - 2p \exp(-\frac{R^2}{2}) \), \( \| A_{t,m,j} \| \leq \frac{1}{MT_t} \cdot 2pL_x L_\theta R \).

Then, we have

\[ \left\| \mathbb{E}[A_{t,m,j}] - \mathbb{E}[\tilde{A}_{t,m,j}] \right\| \leq \mathbb{E} \left[ \left\| A_{t,m,j} \cdot 1 \left\{ \| A_{t,m,j} \| \geq \frac{2pL_x L_\theta R}{MT_t} \right\} \right\| \right] \]

\[ = \mathbb{E} \left[ \left\| A_{t,m,j} \right\| \cdot 1 \left\{ \| A_{t,m,j} \| \geq \frac{2pL_x L_\theta R}{MT_t} \right\} \right] \]

\[ = \mathbb{E} \left[ \left\{ \left( \frac{2pL_x L_\theta R}{MT_t} \right) \cdot 1 \left\{ \| A_{t,m,j} \| \geq \frac{2pL_x L_\theta R}{MT_t} \right\} \right\} \right] \]

\[ = \frac{2pL_x L_\theta R}{MT_t} \cdot \mathbb{P} \left[ \| A_{t,m,j} \| \geq \frac{2pL_x L_\theta R}{MT_t} \right] + \int_0^{\infty} \mathbb{P} \left[ \| A_{t,m,j} \| - \frac{2pL_x L_\theta R}{MT_t} > x \right] dx \]
we have that with probability at least 
\[ \frac{2pL_xL_0R}{MT_t} \cdot 2p \cdot \exp \left( -\frac{R^2}{2} \right) + \frac{2pL_xL_0}{MT_t} \int_{\mathbb{R}} \Pr \left( \| A_{t,m,j} \| > \frac{2pL_xL_0y}{MT_t} \right) \, dy \]
\[ \leq \frac{2pL_xL_0R}{MT_t} \cdot 2p \cdot \exp \left( -\frac{R^2}{2} \right) + \frac{2pL_xL_0}{MT_t} \int_{\mathbb{R}} 2p \exp \left( -\frac{y^2}{2} \right) \, dy \]
\[ \leq \frac{2pL_xL_0R}{MT_t} \cdot 2p \cdot \exp \left( -\frac{R^2}{2} \right) + \frac{2pL_xL_0}{MT_t} \cdot 2p \cdot \exp \left( -\frac{R^2}{2} \right) \]
\[ = \frac{2pL_xL_0}{MT_t} \cdot 2p \left( R + \frac{1}{R} \right) \exp \left( -\frac{R^2}{2} \right). \]

Let \( \delta' \in (0, 1) \) be a confidence parameter which will be chosen later. Using the truncated matrix Bernstein inequality (Lemma E.2) with \( n = MT_t, R = \sqrt{2 \log \left( \frac{2pMT_t}{\delta'} \right)} \), \( n \Pr[\| A_{t,m,j} \| \geq \frac{4p}{MT_t} \cdot 2pL_xL_0R] \leq \delta' \), \( U = \frac{2pL_xL_0\sqrt{2 \log \left( \frac{2pMT_t}{\delta'} \right)}}{MT_t} \), \( \sigma^2 = \frac{MT_tU^2}{\sqrt{MT_t}} \), we have that with probability at least \( 1 - 2\delta' \),

\[ \| A_t - E[A_t] \| \leq \frac{4 \cdot 2pL_xL_0\sqrt{2 \log \left( \frac{2pMT_t}{\delta'} \right)} \log \left( \frac{2p}{\delta'} \right)}{\sqrt{MT_t}} + \frac{4 \cdot 2pL_xL_0\sqrt{2 \log \left( \frac{2pMT_t}{\delta'} \right)} \log \left( \frac{2p}{\delta'} \right)}{MT_t} \]
\[ \leq \frac{8 \cdot 2pL_xL_0\sqrt{2 \log \left( \frac{2pMT_t}{\delta'} \right)} \log \left( \frac{2p}{\delta'} \right)}{\sqrt{MT_t}} \]  
(6)

Now we investigate \( \| C_t - E[C_t] \| \). Recall that in Eq. (5), for any \( t > 0, m \in [M], j \in [T_t] \) and \( i \in [p], |\eta_{t,m,j,i}| \leq R \). Then, we have \( \| \hat{C}_{t,m,j} \| \leq \frac{1}{MT_t} \cdot pR^2 \).

Recall that for any \( t > 0, m \in [M] \) and \( j \in [T_t] \), with probability at least \( 1 - 2p \exp(-\frac{R^2}{2}) \), \( |\eta_{t,m,j,i}| \leq R \) for all \( i \in [p] \). Thus, with probability at least \( 1 - 2p \exp(-\frac{R^2}{2}) \), \( \| C_{t,m,j} \| \leq \frac{1}{MT_t} \cdot pR^2 \). Then, we have

\[ \| E[C_{t,m,j}] - E[\hat{C}_{t,m,j}] \| \leq \left\| E \left[ C_{t,m,j} \cdot 1 \left\{ \| C_{t,m,j} \| \geq \frac{pR^2}{MT_t} \right\} \right] \right\| \]
\[ \leq E \left[ \| C_{t,m,j} \| \cdot 1 \left\{ \| C_{t,m,j} \| \geq \frac{pR^2}{MT_t} \right\} \right] \]
\[ = E \left[ \frac{pR^2}{MT_t} \cdot 1 \left\{ \| C_{t,m,j} \| \geq \frac{pR^2}{MT_t} \right\} \right] + \int_0^\infty \Pr \left[ \| C_{t,m,j} \| > \frac{pR^2}{MT_t} \right] \, dx \]
\[ \leq \frac{pR^2}{MT_t} \cdot 2p \cdot \exp \left( -\frac{R^2}{2} \right) + \frac{2p}{MT_t} \int_{\mathbb{R}} y \cdot \Pr \left[ \| C_{t,m,j} \| > \frac{dy}{MT_t} \right] \, dy \]
\[ \leq \frac{pR^2}{MT_t} \cdot 2p \cdot \exp \left( -\frac{R^2}{2} \right) + \frac{2p}{MT_t} \int_{\mathbb{R}} y \cdot 2p \exp \left( -\frac{y^2}{2} \right) \, dy \]
\[ \leq \frac{pR^2}{MT_t} \cdot 2p \cdot \exp \left( -\frac{R^2}{2} \right) + \frac{2p}{MT_t} \cdot 2p \cdot \exp \left( -\frac{R^2}{2} \right) \]
\[ = \frac{p}{MT_t} \cdot 2p \cdot (R^2 + 2) \exp \left( -\frac{R^2}{2} \right). \]

Using the truncated matrix Bernstein inequality (Lemma E.2) with \( n = MT_t, R = \sqrt{2 \log \left( \frac{2pMT_t}{\delta'} \right)} \), \( n \Pr[\| C_{t,m,j} \| \geq \frac{4p}{MT_t} \cdot 2pL_xL_0] \leq \delta' \), \( \sigma^2 = \frac{32p}{MT_t} \cdot \tau = \frac{4p}{MT_t} \cdot 2pL_xL_0/\sqrt{MT_t} \) and \( \Delta = \frac{1}{MT_t} \cdot pR^2 \) ≤ \( \delta' \), \( U = \frac{p^2 \log \left( \frac{2pMT_t}{\delta'} \right)}{MT_t} \), \( \sigma^2 = \frac{32p}{MT_t} \).
According to Assumption 3.1, there exists an absolute constant \( \Pr \), which implies that

\[
\| C_t - E[C_t] \| \leq \frac{4 \cdot 2p \log \left( \frac{2pMT}{\delta} \right) \log \left( \frac{2p}{\delta} \right)}{\sqrt{MT_t}} + \frac{8 \cdot 2p \log \left( \frac{2pMT}{\delta} \right) \log \left( \frac{2p}{\delta} \right)}{\sqrt{MT_t}}
\]

Plugging Eqs. (6) and (7) into Eq. (4), we have that with probability at least \( 1 - 2\delta' \),

\[
\| Z_t - E[Z_t] \| \leq \| X_{batch}^+ \|^2 \left( \| A_t - E[A_t] \| + \| C_t - E[C_t] \| \right)
\]

\[
\leq \| X_{batch}^+ \|^2 \left( \frac{8 \cdot 2pL_x L_\theta \sqrt{2 \log \left( \frac{2pMT}{\delta} \right) \log \left( \frac{2p}{\delta} \right)}}{\sqrt{MT_t}} + \frac{8 \cdot 2p \log \left( \frac{2pMT}{\delta} \right) \log \left( \frac{2p}{\delta} \right)}{\sqrt{MT_t}} \right)
\]

\[
\leq \frac{96 \| X_{batch}^+ \|^2 pL_x L_\theta \log \left( \frac{2p}{\delta} \right) \log \left( \frac{16pMT}{\delta_t} \right)}{\sqrt{MT_t}}
\]

Let \( \delta' = \frac{\delta}{8} \). Then, we obtain that with probability at least \( 1 - \frac{\delta}{2} \),

\[
\| Z_t - E[Z_t] \| \leq \frac{96 \| X_{batch}^+ \|^2 pL_x L_\theta \log \left( \frac{16p}{\delta_t} \right)}{\sqrt{MT_t}} \log \left( \frac{16pMT}{\delta_t} \right),
\]

which implies that \( \Pr[\mathcal{E}_t] \geq 1 - \frac{\delta}{2} \).

Taking a union bound over all phases \( t \geq 1 \) and recalling \( \delta_t := \frac{\delta}{2^{t+1}} \), we obtain

\[
\Pr[\mathcal{E}] \geq 1 - \sum_{t=1}^{\infty} \Pr[\mathcal{E}_t]
\]

\[
\geq 1 - \sum_{t=1}^{\infty} \frac{\delta_t}{2}
\]

\[
= 1 - \sum_{t=1}^{\infty} \frac{\delta}{4t^2}
\]

\[
\geq 1 - \frac{\delta}{2}
\]

For any matrix \( A \in \mathbb{R}^{m \times n} \) with \( m \geq n \), let \( \sigma_{\max}(A) \) and \( \sigma_{\min}(A) \) denote the maximum and minimum singular values of \( A \), respectively. For any \( i \in [m] \), let \( \sigma_i(A) \) denote the \( i \)-th singular value of \( A \).

For any matrix \( A \in \mathbb{R}^{m \times n} \) with \( m \geq n \), let \( A_\perp \) denote the orthogonal complement matrix of \( A \), where the columns of \( A_\perp \) are the orthogonal complement of those of \( A \). Then, it holds that \( AA^\top + A_\perp A_\perp^\top = I_m \), where \( I_m \) is the \( m \times m \) identity matrix.

According to Assumption 3.1, there exists an absolute constant \( c_0 \) which satisfies that \( \sigma_{\min}\left( \frac{1}{M} \sum_{m=1}^{M} \theta_m \theta_m^\top \right) = \sigma_{\min}\left( \frac{1}{M} \sum_{m=1}^{M} \theta_m \theta_m^\top \right) \geq \frac{c_0}{\delta^2} \).

**Lemma C.4** (Concentration of \( \overline{B}_t \)). Suppose that event \( \mathcal{E} \) holds. Then, for any phase \( t > 0 \),

\[
\| \overline{B}_{t,\perp} \| \leq \frac{192 \| X_{batch}^+ \|^2 kpL_x L_\theta \log \left( \frac{16p}{\sigma_t} \right)}{\sqrt{MT_t}} \log \left( \frac{16pMT}{\delta_t} \right).
\]
Furthermore, for any phase $t > 0$, if

$$
T_t = \left[ \frac{68 \cdot 192^2 \cdot 8^2 (1 + \zeta)^3 (\rho^E)^2 k^4 \ell_\ell^4 k^2 \ell_\ell^2 L_w}{c_0^2 M} \cdot \max \left\{ 2^{2t}, \frac{L_\ell^4}{\omega^2} \right\} \cdot \log^2 \left( \frac{16p}{\delta_t} \right) \right] \\
\log^2 \left( \frac{192 \cdot 16 \cdot 8 (1 + \zeta)^3 \rho^E k^2 p^2 \ell_\ell^2 L_w L_w}{c_0} \cdot \max \left\{ 2^t, \frac{L_\ell^2}{\omega} \right\} \cdot \frac{1}{\delta_t} \cdot \log \left( \frac{16p}{\delta_t} \right) \right],
$$

then

$$
\left\| B_{t,\perp}^T B \right\| \leq \min \left\{ \frac{1}{8k L_w L_w \cdot 2^t \sqrt{1 + \zeta}}, \frac{\omega}{6 L_\ell^2} \right\}.
$$

**Proof of Lemma C.4.** From Assumption 3.1, $\sigma_k(\mathbb{E}[Z_t]) - \sigma_k(1(\mathbb{E}[Z_t]) = \sigma_{\min}(\frac{1}{\pi} \sum_{m=1}^M \theta_m \theta_m^T) \geq \frac{c_0}{\kappa}$. Using the Davis-Kahan sin $\theta$ Theorem (Bhatia, 2013) and letting $T_t$ be large enough to satisfy $\|Z_t - \mathbb{E}[Z_t]\| \leq \frac{c_0}{2k}$, we have

$$
\left\| B_{t,\perp}^T B \right\| \leq \frac{\|Z_t - \mathbb{E}[Z_t]\|}{\sigma_k(\mathbb{E}[Z_t]) - \sigma_k(1(\mathbb{E}[Z_t]) - \|Z_t - \mathbb{E}[Z_t]\|} \leq \frac{2k}{c_0} \|Z_t - \mathbb{E}[Z_t]\|
$$

$$
\leq \frac{192 \left\| X_{\text{batch}}^+ \right\|^2 k L_x L_\theta \log \left( \frac{16p}{\pi} \right)}{c_0 \sqrt{MT_t}} \log \left( \frac{16pMT_t}{\delta_t} \right).
$$

where inequality (a) uses the definition of event $\mathcal{E}$.

Using Lemma E.3 with $A = \frac{192 \left\| X_{\text{batch}}^+ \right\|^2 k L_x L_\theta}{c_0} \log \left( \frac{16p}{\pi} \right)$, $B = \frac{16p}{\pi}$ and $\kappa = \min \left\{ \frac{1}{8k L_w L_w \cdot 2^t \sqrt{1 + \zeta}} \cdot \frac{\omega}{6 L_\ell^2} \right\}$, we have that if

$$
MT_t \geq 68 \left( \frac{192 \left\| X_{\text{batch}}^+ \right\|^2 k L_x L_\theta}{c_0} \log \left( \frac{16p}{\pi} \right) \right)^2 \cdot \max \left\{ \left( 8k L_w L_w \cdot 2^t \sqrt{1 + \zeta} \right)^2, \frac{6^2 L_\ell^4}{\omega^2} \right\}.
$$

$$
\log^2 \left( \frac{192 \cdot 16 \cdot 8 (1 + \zeta)^3 \rho^E k^2 p^2 \ell_\ell^2 L_w L_w}{c_0} \cdot \max \left\{ 2^t, \frac{L_\ell^2}{\omega} \right\} \cdot \frac{1}{\delta_t} \cdot \log \left( \frac{16p}{\delta_t} \right) \right),
$$

then

$$
\left\| B_{t,\perp}^T B \right\| \leq \min \left\{ \frac{1}{8k L_w L_w \cdot 2^t \sqrt{1 + \zeta}}, \frac{\omega}{6 L_\ell^2} \right\}.
$$

According to Lemma C.1, we have $\left\| X_{\text{batch}}^+ \right\| \leq \sqrt{\frac{(1 + \zeta)^3}{p}}.

Then, further enlarging $MT_t$, we have that if

$$
MT_t \geq 68 \cdot 192^2 \cdot 8^2 (1 + \zeta)^3 (\rho^E)^2 k^4 \ell_\ell^4 k^2 \ell_\ell^2 L_w^2 \cdot \max \left\{ 2^{2t}, \frac{L_\ell^4}{\omega^2} \right\} \cdot \log^2 \left( \frac{16p}{\delta_t} \right) \\
\log^2 \left( \frac{192 \cdot 16 \cdot 8 (1 + \zeta)^3 \rho^E k^2 p^2 \ell_\ell^2 L_w L_w}{c_0} \cdot \max \left\{ 2^t, \frac{L_\ell^2}{\omega} \right\} \cdot \frac{1}{\delta_t} \cdot \log \left( \frac{16p}{\delta_t} \right) \right),
$$

then

$$
\left\| B_{t,\perp}^T B \right\| \leq \min \left\{ \frac{1}{8k L_w L_w \cdot 2^t \sqrt{1 + \zeta}}, \frac{\omega}{6 L_\ell^2} \right\}.
$$

\[\square\]
C.3. Elimination with Low-dimensional Representations

For clarity of notation, we also add subscript $t$ to the notations in subroutine $\text{EliLowRep}$ to denote the quantities generated in phase $t$. Specifically, we use the notations $\hat{B}_t, \hat{\lambda}_{t,m}, \lambda_t^G, \rho_{t,m}, \{z_{t,m,i}\}_{i \in [N_{t,m}]}, \{r_{t,m,i}\}_{i \in [N_{t,m}]}$, $\tilde{w}_{t,m}$ and $\hat{\theta}_{t,m}$ to denote the corresponding quantities used in $\text{EliLowRep}$ in phase $t$.

Before analyzing the sample complexity of $\text{EliLowRep}$, we first prove that there exists a sample allocation $\lambda \in \Delta_X$ such that $\sum_{i=1}^n \lambda(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t$ is invertible, i.e., the G-optimal design optimization with $\hat{B}_t$ is non-vacuous (Line 2 in Algorithm 3).

For any task $m \in [M]$, let

$$\lambda^*_m := \arg\min_{\lambda \in \Delta_X} \max_{x \in X \setminus \{x_m^*\}} \frac{\|B^\top x_m^* - B^\top x\|^2}{\left(\sum_{i=1}^n \lambda(x_i) B^\top x_i x_i^\top B\right)^{-1} \left(\|x_m^* - x\|^2\right)}.$$ 

$\lambda^*_m$ is the optimal solution of the G-optimal design optimization with true feature extractor $B$.

**Lemma C.5.** For any phase $t > 0$ and task $m \in [M]$, if $\|\hat{B}_t^\top B\| \leq \frac{\omega}{6L_x^2}$, we have

$$\sigma_{\min} \left( \sum_{i=1}^n \lambda^*_m(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t \right) > 0.$$ 

**Proof of Lemma C.5.** For any task $m \in [M]$, let $A_m := \sum_{i=1}^n \lambda^*_m(x_i) x_i x_i^\top$. Then, for any phase $t > 0$ and task $m \in [M]$, we have

$$\sum_{i=1}^n \lambda^*_m(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t = \hat{B}_t^\top A_m \hat{B}_t$$

$$= \hat{B}_t^\top (BB^\top + B \perp B^\top \perp) A_m (BB^\top + B \perp B^\top \perp) \hat{B}_t$$

$$= \hat{B}_t^\top BB^\top A_m BB^\top \hat{B}_t + \hat{B}_t^\top BB^\top A_m B \perp B^\top \perp \hat{B}_t$$

$$+ \hat{B}_t^\top B \perp B^\top \perp A_m BB^\top \hat{B}_t + \hat{B}_t^\top B \perp B^\top \perp A_m B \perp B^\top \perp \hat{B}_t.$$

Hence, we have

$$\sigma_{\min} \left( \sum_{i=1}^n \lambda^*_m(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t \right) \geq \sigma_{\min} \left( \hat{B}_t^\top BB^\top A_m BB^\top \hat{B}_t \right) - \sigma_{\max} \left( \hat{B}_t^\top BB^\top A_m B \perp B^\top \perp \hat{B}_t \right)$$

$$- \sigma_{\max} \left( \hat{B}_t^\top B \perp B^\top \perp A_m BB^\top \hat{B}_t \right) - \sigma_{\max} \left( \hat{B}_t^\top B \perp B^\top \perp A_m B \perp B^\top \perp \hat{B}_t \right)$$

$$\geq \sigma_{\min} \left( \hat{B}_t^\top B \right) \sigma_{\min} \left( B^\top A_m B \right) \sigma_{\min} \left( B^\top B_t \right) - \left\| B_t^\top B \right\| \| A_m \|$$

$$- \left\| B_t^\top B \right\| \| A_m \| - \left\| B_t^\top B \right\| \| A_m \|$$

$$\geq \sigma_{\min}^2 \left( \hat{B}_t^\top B \right) \sigma_{\min} \left( B^\top A_m B \right) - 3 \left\| B_t^\top B \right\| L_x^2$$

$$\geq \left( 1 - \left\| B_t^\top B \right\| L_x^2 \right) \sigma_{\min} \left( B^\top B \right) B \left( \sum_{i=1}^n \lambda^*_m(x_i) \right) \hat{B}_t^\top x_i x_i^\top \hat{B}_t \hat{B}_t$$

$$= \left( 1 - \left\| B_t^\top B \right\| L_x^2 \right) \omega - \frac{\omega^3}{36L_x^4}.$$ 

where inequality (a) uses the fact that $\hat{B}_t^\top BB^\top \hat{B}_t + \hat{B}_t^\top B \perp B^\top \perp \hat{B}_t = \hat{B}_t^\top (BB^\top + B \perp B^\top \perp) \hat{B}_t = \hat{B}_t^\top \hat{B}_t = I_k$, and thus, $\sigma_{\min}^2 \left( \hat{B}_t^\top B \right) = 1 - \left\| B_t^\top B \right\| L_x^2$.

Let $\| \hat{B}_t^\top B \| \leq \frac{\omega}{6L_x^2}$. Then, we have

$$\sigma_{\min} \left( \sum_{i=1}^n \lambda^*_m(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t \right) \geq \left( 1 - \frac{\omega^2}{36L_x^4} \right) \omega - \frac{\omega^3}{2}$$

$$= \frac{\omega^2}{2} - \frac{\omega^3}{36L_x^4}.$$
where the last inequality is due to $\omega \leq L^2 x < \sqrt{18} L^2 x$. \hfill \square

Next, we bound the optimal value $\rho_{t,m}^G$ of the G-optimal design optimization with the estimated feature extractor $\Tilde{B}_t$.

For any $Z \subseteq \mathcal{X}$, let $\mathcal{Y}(Z) := \{x - x' : \forall x, x' \in Z, x \neq x'\}$. Recall that in Line 2 of Algorithm 3, for any phase $t > 0$ and task $m \in [M]$,

$$
\rho_{t,m}^G := \min_{\lambda \in \Delta \mathcal{X}} \max_{y \in \mathcal{Y}(\hat{\mathcal{X}}_{t,m})} \left\| \Tilde{B}_t^\top y \right\|^2 \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1}.
$$

**Lemma C.6.** For any phase $t > 0$ and task $m \in [M]$,

$$
\rho_{t,m}^G \leq 4k.
$$

**Proof of Lemma C.6.** For any phase $t > 0$ and task $m \in [M]$, we have that $\hat{\mathcal{X}}_{t,m} \subseteq \mathcal{X}$ and $\mathcal{Y}(\hat{\mathcal{X}}_{t,m}) \subseteq \mathcal{Y}(\mathcal{X})$.

For any fixed $\lambda \in \Delta \mathcal{X}$,

$$
\max_{y \in \mathcal{Y}(\hat{\mathcal{X}}_{t,m})} \left\| \Tilde{B}_t^\top y \right\|^2 \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1} \leq \max_{y \in \mathcal{Y}(\mathcal{X})} \left\| \Tilde{B}_t^\top y \right\|^2 \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1}
$$

$$
= \left\| \Tilde{B}_t^\top \left( x'_1 - x'_2 \right) \right\|^2 \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1}
$$

$$
\leq \left( \left\| \Tilde{B}_t^\top x'_1 \right\| \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1} + \left\| \Tilde{B}_t^\top x'_2 \right\| \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1} \right)^2
$$

$$
\leq 2 \left\| \Tilde{B}_t^\top x'_1 \right\|^2 \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1} + 2 \left\| \Tilde{B}_t^\top x'_2 \right\|^2 \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1}
$$

$$
\leq 4 \max_{x \in \mathcal{X}} \left\| \Tilde{B}_t^\top x \right\|^2 \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1},
$$

where $x'_1$ and $x'_2$ are the arms which satisfy $y = x'_1 - x'_2$ achieves the maximum value $\max_{y \in \mathcal{Y}(\mathcal{X})} \left\| \Tilde{B}_t^\top y \right\|^2 \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1}$.

Since $\Tilde{B}_t^\top x \in \mathbb{R}^k$, according to the Equivalence Theorem in (Kiefer & Wolfowitz, 1960), we have

$$
\min_{\lambda \in \Delta \mathcal{X}} \max_{x \in \mathcal{X}} \left\| \Tilde{B}_t^\top x \right\|^2 \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1} = k.
$$

Therefore, we have

$$
4k = 4 \min_{\lambda \in \Delta \mathcal{X}} \max_{x \in \mathcal{X}} \left\| \Tilde{B}_t^\top x \right\|^2 \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1}
$$

$$
= 4 \max_{x \in \mathcal{X}} \left\| \Tilde{B}_t^\top x \right\|^2 \left( \sum_{i=1}^n \lambda'(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1}
$$

$$
\geq \max_{y \in \mathcal{Y}(\hat{\mathcal{X}}_{t,m})} \left\| \Tilde{B}_t^\top y \right\|^2 \left( \sum_{i=1}^n \lambda'(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1}
$$

$$
\geq \min_{\lambda \in \Delta \mathcal{X}} \max_{y \in \mathcal{Y}(\hat{\mathcal{X}}_{t,m})} \left\| \Tilde{B}_t^\top y \right\|^2 \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1}
$$

$$
= \rho_{t,m}^G,
$$

where $\lambda' := \arg\min_{\lambda \in \Delta \mathcal{X}} \max_{x \in \mathcal{X}} \left\| \Tilde{B}_t^\top x \right\|^2 \left( \sum_{i=1}^n \lambda(x_i) \Tilde{B}_t^\top x_i x_i^\top B_t \right)^{-1}$. \hfill \square

Now we analyze the estimation error of the estimated reward parameter $\hat{\theta}_{t,m} = \Tilde{B}_t \hat{w}_{t,m}$ in $\text{ElilLowRep}$.

For any phase $t > 0$, task $m \in [M]$ and arm $j \in [N_{t,m}]$, let $\xi_{t,m,j}$ denote the noise of the sample on arm $z_{t,m,j}$ for task $m$, during the execution of $\text{ElilLowRep}$ in phase $t$ (Line 5 in Algorithm 3).
For any phase $t > 0$, define events
\[
F_t := \left\{ y^\top \tilde{B}_t \left( \sum_{j=1}^{N_{t,m}} \tilde{B}_t^\top z_{t,m,j} z_{t,m,j}^\top \tilde{B}_t \right)^{-1} \sum_{j=1}^{N_{t,m}} \tilde{B}_t^\top z_{t,m,j} \cdot \xi_{t,m,j} \right\}
\]
\leq \left\| \tilde{B}_t^\top y \right\| \left( \sum_{j=1}^{N_{t,m}} \tilde{B}_t^\top z_{t,m,j} z_{t,m,j}^\top \tilde{B}_t \right)^{-1} \sqrt{2 \log \left( \frac{4n^2 M}{\delta_t} \right)}, \forall m \in [M], \forall y \in \mathcal{Y}(\hat{X}_{t,m}) \right\}
\]
and
\[
F := \cap_{t=1}^\infty F_t.
\]

**Lemma C.7 (Concentration of the Variance Term).** It holds that
\[
\Pr[F] \geq 1 - \frac{\delta}{2}.
\]

**Proof of Lemma C.7.** Let $\Sigma_{t,m} := \sum_{j=1}^{N_{t,m}} \tilde{B}_t^\top z_{t,m,j} z_{t,m,j}^\top \tilde{B}_t$. Then, we can write
\[
y^\top \tilde{B}_t \left( \sum_{j=1}^{N_{t,m}} \tilde{B}_t^\top z_{t,m,j} z_{t,m,j}^\top \tilde{B}_t \right)^{-1} \sum_{j=1}^{N_{t,m}} \tilde{B}_t^\top z_{t,m,j} \cdot \xi_{t,m,j} = \sum_{j=1}^{N_{t,m}} y^\top \tilde{B}_t \Sigma_{t,m}^{-1} \tilde{B}_t^\top z_{t,m,j} \cdot \xi_{t,m,j}.
\]

For any phase $t > 0$, task $m \in [M]$ and arm $j \in [N_{t,m}]$, $\tilde{B}_t, \Sigma_{t,m}$ and $\{z_{t,m,j}\}_{j=1}^{N_{t,m}}$ are fixed before the sampling in EliLowRep, and the noise $\xi_{t,m,j}$ is 1-sub-Gaussian (Line 5 in Algorithm 3). Thus, we have that for any $t > 0$, $m \in [M]$ and $j \in [N_{t,m}]$, $y^\top \tilde{B}_t \Sigma_{t,m}^{-1} \tilde{B}_t^\top z_{t,m,j} \cdot \xi_{t,m,j}$ is $(y^\top \tilde{B}_t \Sigma_{t,m}^{-1} \tilde{B}_t^\top z_{t,m,j})$-sub-Gaussian.

Using Hoeffding’s inequality and taking a union bound over all $m \in [M]$ and $y \in \mathcal{Y}(\hat{X}_{t,m})$, we have that with probability at least $1 - \frac{\delta}{2}$,
\[
\sum_{j=1}^{N_{t,m}} y^\top \tilde{B}_t \Sigma_{t,m}^{-1} \tilde{B}_t^\top z_{t,m,j} \cdot \xi_{t,m,j}
\]
\leq \frac{2 \sum_{j=1}^{N_{t,m}} \left( y^\top \tilde{B}_t \Sigma_{t,m}^{-1} \tilde{B}_t^\top z_{t,m,j} \right)^2 \cdot \log \left( \frac{4n^2 M}{\delta_t} \right)}{2 \sum_{j=1}^{N_{t,m}} y^\top \tilde{B}_t \Sigma_{t,m}^{-1} \tilde{B}_t^\top z_{t,m,j} \cdot z_{t,m,j}^\top \tilde{B}_t \Sigma_{t,m}^{-1} \tilde{B}_t^\top y \cdot \log \left( \frac{4n^2 M}{\delta_t} \right)}
\]
\leq 2y^\top \tilde{B}_t \Sigma_{t,m}^{-1} \left( \tilde{B}_t^\top \sum_{j=1}^{N_{t,m}} z_{t,m,j} \cdot z_{t,m,j}^\top \tilde{B}_t \right) \Sigma_{t,m}^{-1} \tilde{B}_t^\top y \cdot \log \left( \frac{4n^2 M}{\delta_t} \right)
\]
\leq 2y^\top \tilde{B}_t \Sigma_{t,m}^{-1} \tilde{B}_t^\top y \cdot \log \left( \frac{4n^2 M}{\delta_t} \right)
\]
which implies that
\[
\Pr[F] \geq 1 - \frac{\delta_t}{2}.
\]
Taking a union bound over all phases \( t \geq 1 \) and recalling \( \delta_t := \frac{\delta}{2t^2} \), we obtain

\[
\Pr [F] \geq 1 - \sum_{t=1}^{\infty} \Pr [\tilde{F}_t] \\
\geq 1 - \sum_{t=1}^{\infty} \frac{\delta_t}{2} \\
= 1 - \sum_{t=1}^{\infty} \frac{\delta}{4t^2} \\
\geq 1 - \frac{\delta}{2}.
\]

**Lemma C.8 (Concentration of \( \hat{\theta}_{t,m} \)).** Suppose that event \( E \cap F \) holds. Then, for any phase \( t > 0 \), task \( m \in [M] \) and \( y \in \mathcal{Y}(\hat{\chi}_{t,m}) \),

\[
\left| y^\top (\hat{\theta}_{t,m} - \theta_m) \right| \leq \frac{1}{2t}.
\]

**Proof of Lemma C.8.** For any phase \( t > 0 \), task \( m \in [M] \) and \( y \in \mathcal{Y}(\hat{\chi}_{t,m}) \),

\[
y^\top (\hat{\theta}_{t,m} - \theta_m) = y^\top B_t \hat{w}_{t,m} - y^\top (B_t \hat{B}_t^\top + \hat{B}_{t,\perp} \hat{B}_{t,\perp}^\top) \theta_m \\
= y^\top B_t \left( \hat{w}_{t,m} - \hat{B}_t \theta_m \right) - y^\top B_{t,\perp} \hat{B}_{t,\perp} \theta_m. \tag{10}
\]

Here, \( \hat{w}_{t,m} \) can be written as

\[
\hat{w}_{t,m} = \left( \sum_{j=1}^{N_{t,m}} B_{t,m,j} z_{t,m,j} \right) \left( \sum_{j=1}^{N_{t,m}} \hat{B}_{t,m,j} \right)^{-1} \left( \sum_{j=1}^{N_{t,m}} \hat{B}_{t,m,j} z_{t,m,j} \right)^\top \theta_m + \xi_{t,m,j}
\]

\[
= \frac{\hat{B}_t^\top \theta_m + \left( \sum_{j=1}^{N_{t,m}} B_{t,m,j} z_{t,m,j} \right) \left( \sum_{j=1}^{N_{t,m}} \hat{B}_{t,m,j} \right)^{-1} \left( \sum_{j=1}^{N_{t,m}} \hat{B}_{t,m,j} z_{t,m,j} \right)^\top \theta_m + \xi_{t,m,j}}{\sum_{j=1}^{N_{t,m}} B_{t,m,j} z_{t,m,j} \cdot \hat{B}_{t,m,j} \hat{B}_{t,\perp} \hat{B}_{t,\perp} \theta_m}
\]

\[
+ \left( \sum_{j=1}^{N_{t,m}} B_{t,m,j} z_{t,m,j} \right) \left( \sum_{j=1}^{N_{t,m}} \hat{B}_{t,m,j} \right)^{-1} \left( \sum_{j=1}^{N_{t,m}} \hat{B}_{t,m,j} \right)^\top \theta_m. \tag{11}
\]

Plugging Eq. (11) into Eq. (10), we can decompose the estimation error of \( \hat{\theta}_{t,m} \) in EliLowRep into three parts as

\[
y^\top (\hat{\theta}_{t,m} - \theta_m) = y^\top \hat{B}_t \left( \sum_{j=1}^{N_{t,m}} \hat{B}_{t,m,j} z_{t,m,j} \right) \left( \sum_{j=1}^{N_{t,m}} \hat{B}_{t,m,j} \right)^{-1} \left( \sum_{j=1}^{N_{t,m}} \hat{B}_{t,m,j} z_{t,m,j} \right)^\top \theta_m + \xi_{t,m,j}
\]

\[
\text{Bias}
\]

\[
+ \left( \sum_{j=1}^{N_{t,m}} B_{t,m,j} z_{t,m,j} \right) \left( \sum_{j=1}^{N_{t,m}} \hat{B}_{t,m,j} \right)^{-1} \left( \sum_{j=1}^{N_{t,m}} \hat{B}_{t,m,j} \right)^\top \theta_m. \tag{11}
\]
Taking the absolute value on both sides, and using the Cauchy–Schwarz inequality and definition of event $\mathcal{F}$ (Eq. (9)), we have

$$
\left| y^\top \theta_{t,m} - y^\top \theta_m \right|
\leq \left\| \mathbf{B}_t^\top y \right\| \left( \sum_{j=1}^{N_{t,m}} \mathbf{B}_t^\top z_{t,m,j} z_{t,m,j}^\top \mathbf{B}_t \right)^{-1} \cdot \left\| \sum_{j=1}^{N_{t,m}} \mathbf{B}_t^\top z_{t,m,j} \mathbf{B}_t^\top \mathbf{B}_t^\top Bw_m \right\| 
+ \left\| \mathbf{B}_t^\top y \right\| \left( \sum_{j=1}^{N_{t,m}} \mathbf{B}_t^\top z_{t,m,j} z_{t,m,j}^\top \mathbf{B}_t \right)^{-1} \cdot \left\| \sum_{j=1}^{N_{t,m}} \mathbf{B}_t^\top z_{t,m,j} \right\| 
\leq \frac{\sqrt{1 + \zeta}}{\sqrt{N_{t,m}}} \left\| \mathbf{B}_t^\top y \right\| \left( \sum_{i=1}^{n} \lambda_{t,m}^G(x_i) \mathbf{B}_t^\top x_i x_i^\top \mathbf{B}_t \right)^{-1} \cdot L_x L_w \left\| \mathbf{B}_t^\top_{t,m,j} \mathbf{B}_t^\top \mathbf{B}_t^\top Bw_m \right\| 
+ \frac{\sqrt{1 + \zeta}}{\sqrt{N_{t,m}}} \left\| \mathbf{B}_t^\top y \right\| \left( \sum_{i=1}^{n} \lambda_{t,m}^G(x_i) \mathbf{B}_t^\top x_i x_i^\top \mathbf{B}_t \right)^{-1} \cdot \sqrt{k} \cdot \left\| \mathbf{B}_t^\top_{t,m,j} \mathbf{B}_t^\top \mathbf{B}_t^\top Bw_m \right\|
$$

Here inequality (a) is due to the guarantee of rounding procedure \textsc{Round} and the triangle inequality. Inequality (b) uses Lemma E.5, and inequality (c) follows from Lemma C.6. Inequality (d) comes from Lemma C.4 and $N_{t,m} := \max \left\{ [32 \cdot 2^{2t}(1 + \zeta) \rho_{t,m}^G \log(\frac{4n^2M}{\delta t})], \frac{180k}{\zeta^2} \right\}$. $\square$

For any task $m \in [M]$ and arm $x \in \mathcal{X}$, let $\Delta_m(x) := (x_m^* - x)^\top \theta_m$ denote the reward gap between the optimal arm $x_m^*$ and arm $x$. 

\[ 
\begin{align*}
\text{Multi-task Representation Learning for Pure Exploration in Linear Bandits} \quad \\
\frac{1}{4} & \cdot \frac{1}{4} + \frac{1}{2} + \frac{1}{4} \\
\leq & \frac{1}{2} \\
\end{align*} \]
and arm $x$ in task $m$. For any phase $t > 0$ and task $m \in [M]$, let $Z_{t,m} := \{x \in \mathcal{X} : \Delta_m(x) \leq 4 \cdot 2^{-t}\}$.

**Lemma C.9.** Suppose that event $E \cap F$ holds. For any phase $t > 0$ and task $m \in [M]$, 
\[ x^*_m \in \hat{X}_{t,m}, \]
and for any phase $t \geq 2$ and task $m \in [M]$, 
\[ \hat{X}_{t,m} \subseteq Z_{t,m}. \]

**Proof of Lemma C.9.** This proof follows a similar analytical procedure as that of Lemma 2 in (Fiez et al., 2019).

First, we prove $x^*_m \in \hat{X}_{t,m}$ for any phase $t > 0$ and task $m \in [M]$ by contradiction.

Suppose that for some $t > 0$ and some $m \in [M]$, $x^*_m$ is eliminated from $\hat{X}_{t,m}$ in phase $t$. Then, we have that there exists some $x' \in \hat{X}_{t,m}$ such that 
\[ (x' - x^*_m)\top \theta_{t,m} > 2^{-t}. \]

Then, we have 
\[ (x' - x^*_m)\top \theta_{t,m} = (x' - x^*_m)\top \hat{\theta}_{t,m} - (x' - x^*_m)\top (\hat{\theta}_{t,m} - \theta_{t,m}) \]
\[ \geq (x' - x^*_m)\top \hat{\theta}_{t,m} - 2^{-t} \]
\[ > 2^{-t} - 2^{-t} \]
\[ = 0, \]
which contradicts the definition of $x^*_m$. Thus, we obtain that $x^*_m \in \hat{X}_{t,m}$ for any phase $t > 0$ and task $m \in [M]$.

Next, we prove $\hat{X}_{t,m} \subseteq Z_{t,m}$ for any phase $t \geq 2$ and task $m \in [M]$, i.e., each $x \in \hat{X}_{t,m}$ satisfies that $\Delta_m(x) \leq 4 \cdot 2^{-t}$.

Suppose that there exists some phase $t$, some task $m$, and some $x \in \hat{X}_{t,m}$ such that $\Delta_m(x) > 4 \cdot 2^{-t}$. Then, in phase $t - 1 \geq 1$, we have 
\[ (x_m^* - x)\top \hat{\theta}_{t-1,m} = (x_m^* - x)\top \theta_m - (x_m^* - x)\top (\theta_m - \hat{\theta}_{t-1,m}) \]
\[ \geq (x_m^* - x)\top \theta_m - 2^{-t} \]
\[ > 4 \cdot 2^{-t} - 2^{-t} \]
\[ = 2^{-t}, \]
which implies that $x$ should have been eliminated from $\hat{X}_{t,m}$ in phase $t - 1$, and contradicts our supposition. Thus, we complete the proof.

**C.4. Proof of Theorem 4.1**

Before proving Theorem 4.1, we first introduce a useful lemma.

For any task $m \in [M]$, let 
\[ \lambda_m^* := \arg\min_{\lambda \in \Delta_x} \max_{x \in \mathcal{X} \setminus \{x_m^*\}} \frac{\|B\top x_m^* - B\top x\|^2}{\sum_{i=1}^m \lambda(x_i) B\top B^{-1}} \]
\[ \times \frac{((x_m^* - x)\top \theta_m)^2}{\sum_{i=1}^m \lambda(x_i) B\top x_i x_i\top B^{-1}}. \]

and 
\[ \rho_m^* := \min_{\lambda \in \Delta_x} \max_{x \in \mathcal{X} \setminus \{x_m^*\}} \frac{\|B\top x_m^* - B\top x\|^2}{\sum_{i=1}^m \lambda(x_i) B\top B^{-1}} \]
\[ \times \frac{((x_m^* - x)\top \theta_m)^2}{\sum_{i=1}^m \lambda(x_i) B\top x_i x_i\top B^{-1}}. \]

$\lambda_m^*$ and $\rho_m^*$ are the optimal solution and the optimal value of the G-optimal design optimization with true feature extractor $B$, respectively.
Lemma C.10. Suppose that event $\mathcal{E} \cap \mathcal{F}$ holds. For any task $m \in [M]$ and $y \in \mathbb{R}^d$, 

$$\|\hat{B}_t^\top y\|^2\left(\sum_{i=1}^n \lambda_m^*(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t\right)^{-1} \leq \|B^\top y\|^2\left(\sum_{i=1}^n \lambda_m^*(x_i) B^\top x_i x_i^\top B\right)^{-1} + \frac{11L_s^2}{k \omega^2 \cdot 2^t}.$$ 

Proof of Lemma C.10. We first handle the term $\left(\sum_{i=1}^n \lambda_m^*(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t\right)^{-1}$. 

For any task $m \in [M]$, we have 

$$\sum_{i=1}^n \lambda_m^*(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t = \sum_{i=1}^n \lambda_m^*(x_i) \left(\hat{B}_t^\top B B^\top x_i + \hat{B}_t^\top B \perp B^\top x_i\right) \cdot \left(\hat{B}_t^\top B B^\top x_i + \hat{B}_t^\top B \perp B^\top x_i\right)^\top$$ 

$$= \sum_{i=1}^n \lambda_m^*(x_i) \left(\hat{B}_t^\top B B^\top x_i\right) \cdot \left(\hat{B}_t^\top B B^\top x_i\right)^\top + \left(\hat{B}_t^\top B \perp B^\top x_i\right) \cdot \left(\hat{B}_t^\top B \perp B^\top x_i\right)^\top + \left(\hat{B}_t^\top B \perp B^\top x_i\right) \cdot \left(\hat{B}_t^\top B \perp B^\top x_i\right)^\top$$ 

$$= \sum_{i=1}^n \lambda_m^*(x_i) \left(\hat{B}_t^\top B B^\top x_i\right) \cdot \left(\hat{B}_t^\top B B^\top x_i\right)^\top + \sum_{i=1}^n \lambda_m^*(x_i) \left(\hat{B}_t^\top B \perp B^\top x_i\right) \cdot \left(\hat{B}_t^\top B \perp B^\top x_i\right)^\top$$ 

$$+ \left(\hat{B}_t^\top B \perp B^\top x_i\right) \cdot \left(\hat{B}_t^\top B \perp B^\top x_i\right)^\top.$$ 

Let $P_t := \sum_{i=1}^n \lambda_m^*(x_i) (\hat{B}_t^\top B B^\top x_i) \cdot (\hat{B}_t^\top B B^\top x_i)^\top$. Let $Q_t := \sum_{i=1}^n \lambda_m^*(x_i) (\hat{B}_t^\top B B^\top x_i) \cdot (\hat{B}_t^\top B \perp B^\top x_i) + (\hat{B}_t^\top B \perp B^\top x_i) \cdot (\hat{B}_t^\top B \perp B^\top x_i)^\top$. Then, we have $\sum_{i=1}^n \lambda_m^*(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t = P_t + Q_t$. 

From Assumption 3.2, we have that for any task $m \in [M]$, $\sum_{i=1}^n \lambda_m^*(x_i) B^\top x_i x_i^\top B$ is invertible. Since $\hat{B}_t^\top B$ is also invertible, we have that $P_t$ is invertible. According to Lemmas C.4 and C.5, we have that $\sum_{i=1}^n \lambda_m^*(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t$ is also invertible. Thus, we can write $\left(\sum_{i=1}^n \lambda_m^*(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t\right)^{-1}$ as follows. 

$$\left(\sum_{i=1}^n \lambda_m^*(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t\right)^{-1} = (P_t - (P_t + Q_t))^{-1} Q_t (P_t)^{-1}$$ 

Hence, for any task $m \in [M]$ and $y \in \mathbb{R}^d$, we have 

$$\|\hat{B}_t^\top y\|^2\left(\sum_{i=1}^n \lambda_m^*(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t\right)^{-1} = \left(\hat{B}_t^\top y\right)^\top \left(\sum_{i=1}^n \lambda_m^*(x_i) \hat{B}_t^\top x_i x_i^\top \hat{B}_t\right)^{-1} \hat{B}_t^\top y$$ 

$$= \left(\hat{B}_t^\top y\right)^\top (P_t)^{-1} \hat{B}_t^\top y - \left(\hat{B}_t^\top y\right)^\top (P_t + Q_t)^{-1} Q_t (P_t)^{-1} \hat{B}_t^\top y.$$

From Lemma C.4, we have 

$$\|\hat{B}_t^\top B\| \leq \min \left\{\frac{1}{8k \cdot 2^t \cdot \sqrt{1 + \zeta}}, \frac{\omega}{6L_2^2}\right\} \leq \min \left\{\frac{1}{8k \cdot 2^t}, \frac{\omega}{6L_2^2}\right\}.$$ 

Since $B^\top \hat{B}_t \hat{B}_t^\top B + B^\top \hat{B}_t, \hat{B}_t^\top B = B^\top (\hat{B}_t \hat{B}_t^\top + \hat{B}_t, \hat{B}_t^\top) B = B^\top B = I_k$, we have $\sigma_{\min}^2(\hat{B}_t^\top B) = 1 - \|\hat{B}_t^\top B\|^2$. 

27
Thus, we have

$$\sigma_{\min}(B^T_i B) = \sqrt{1 - \|B^T_i B\|^2} \geq \sqrt{1 - \min \left\{ \frac{1}{64k^2 \cdot 2 \omega t}, \frac{\omega^2}{36L^2} \right\}} > 0,$$

which implies that $\hat{B}^T_i B$ is invertible.

Now, we first analyze Term 1 in Eq. (12).

$$\text{Term 1} = \left( \hat{B}^T_i y \right)^T P_t^{-1} \hat{B}^T_i y$$

$$= \left( \hat{B}^T_i BB^T y + \hat{B}^T_i B \ll B^T_{\perp} y \right)^T P_t^{-1} \left( \hat{B}^T_i BB^T y + \hat{B}^T_i B \ll B^T_{\perp} y \right)$$

$$= \left( \hat{B}^T_i BB^T y \right)^T P_t^{-1} \left( \hat{B}^T_i BB^T y \right) + \left( \hat{B}^T_i BB^T y \right)^T P_t^{-1} \left( \hat{B}^T_i B \ll B^T_{\perp} y \right)$$

$$+ \left( \hat{B}^T_i B \ll B^T_{\perp} y \right)^T P_t^{-1} \left( \hat{B}^T_i BB^T y \right) + \left( \hat{B}^T_i B \ll B^T_{\perp} y \right)^T P_t^{-1} \left( \hat{B}^T_i B \ll B^T_{\perp} y \right).$$

In the following, we bound Terms 1-1, 1-2, 1-3 and 1-4, respectively.

First, we have

$$\text{Term 1-1} = \left( \hat{B}^T_i BB^T y \right)^T \left( \sum_{i=1}^n \lambda^*_m(x_i) \left( \hat{B}^T_i BB^T x_i \right) \cdot \left( \hat{B}^T_i BB^T x_i \right)^T \right)^{-1} \hat{B}^T_i BB^T y$$

$$= \left( \hat{B}^T_i BB^T y \right)^T \hat{B}^T_i B \left( \sum_{i=1}^n \lambda^*_m(x_i) B^T x_i x_i^T B \right) \left( \hat{B}^T_i B \right)^T \hat{B}^T_i BB^T y$$

$$= \left( \hat{B}^T_i BB^T y \right)^T \left( \hat{B}^T_i B \right)^{-1} \left( \sum_{i=1}^n \lambda^*_m(x_i) B^T x_i x_i^T B \right)^{-1} \left( \hat{B}^T_i B \right)^{-1} \hat{B}^T_i BB^T y$$

$$= \left( B^T y \right)^T \left( \sum_{i=1}^n \lambda^*_m(x_i) B^T x_i x_i^T B \right)^{-1} B^T y$$

$$= \|B^T y\|^2 \left( \sum_{i=1}^n \lambda^*_m(x_i) B^T x_i x_i^T B \right)^{-1}. $$

We note that since $\hat{B}^T_i BB^T \hat{B}_i + \hat{B}^T_i B \ll B^T_{\perp} \hat{B}_i = \hat{B}^T_i (BB^T + B \ll B^T_{\perp}) \hat{B}_i = \hat{B}^T_i \hat{B}_i = I_k$, $\sigma^2_{\min}(\hat{B}^T_i B) = 1 - \|\hat{B}^T_i B_{\perp}\|^2$. In addition, $\left\| (\hat{B}^T_i B)^{-1} \right\| = \frac{1}{\sigma_{\min}(\hat{B}^T_i B)} = \frac{1}{\sqrt{1 - \|\hat{B}^T_i B_{\perp}\|^2}}.$

Then, second, we have

$$\text{Term 1-2} = \left( \hat{B}^T_i BB^T y \right)^T \left( \sum_{i=1}^n \lambda^*_m(x_i) \left( \hat{B}^T_i BB^T x_i \right) \cdot \left( \hat{B}^T_i BB^T x_i \right)^T \right)^{-1} \hat{B}^T_i B \ll B^T_{\perp} y$$

$$= \left( \hat{B}^T_i BB^T y \right)^T \left( \hat{B}^T_i B \left( \sum_{i=1}^n \lambda^*_m(x_i) B^T x_i x_i^T B \right) \left( \hat{B}^T_i B \right)^T \right)^{-1} \hat{B}^T_i B \ll B^T_{\perp} y$$

$$= \left( \hat{B}^T_i BB^T y \right)^T \left( \hat{B}^T_i B \right)^{-1} \left( \sum_{i=1}^n \lambda^*_m(x_i) B^T x_i x_i^T B \right)^{-1} \left( \hat{B}^T_i B \right)^{-1} \hat{B}^T_i B \ll B^T_{\perp} y$$
Third, we have

$$\begin{align*}
\text{Term 1-3} & = (B_i^\top y)^\top \left( \sum_{i=1}^n \lambda_m^*(x_i, B_i^\top x_i) \right)^{-1} \left( \hat{B}_i^\top B_i \right)^{-1} \hat{B}_i^\top B_i \leq \lambda_{2i}^2 \cdot \frac{1}{\omega} \cdot \frac{1}{\sqrt{1 - \| \hat{B}_i^\top B_i \|^2}} \cdot 2L_x \cdot \frac{1}{\omega} \cdot 2L_x \\
& \leq \lambda_{2i}^2 \cdot \frac{1}{\omega} \cdot \frac{1}{\sqrt{1 - \left( \frac{1}{\omega^2} \right)^2}} \cdot \frac{1}{8k \cdot 2^t} \\
& \leq \lambda_{2i}^2 \cdot \frac{1}{\omega} \cdot \frac{1}{\sqrt{1 - \frac{3}{4}}} \cdot \frac{1}{8k \cdot 2^t} \\
& = \frac{L_x^2}{k \omega \cdot 2^t}.
\end{align*}$$

Finally, we have

$$\begin{align*}
\text{Term 1-4} & = (B_i^\top y)^\top \left( \sum_{i=1}^n \lambda_m^*(x_i, B_i^\top x_i) \right)^{-1} \left( \hat{B}_i^\top B_i \right)^{-1} \hat{B}_i^\top B_i \leq \lambda_{2i}^2 \cdot \frac{1}{\omega} \cdot \frac{1}{\sqrt{1 - \| \hat{B}_i^\top B_i \|^2}} \cdot 2L_x \cdot \frac{1}{\omega} \cdot 2L_x \\
& \leq \lambda_{2i}^2 \cdot \frac{1}{\omega} \cdot \frac{1}{\sqrt{1 - \left( \frac{1}{\omega^2} \right)^2}} \cdot \frac{1}{8k \cdot 2^t} \\
& \leq \lambda_{2i}^2 \cdot \frac{1}{\omega} \cdot \frac{1}{\sqrt{1 - \frac{3}{4}}} \cdot \frac{1}{8k \cdot 2^t} \\
& = \frac{L_x^2}{k \omega \cdot 2^t}.
\end{align*}$$
Thus, we have

\[
\begin{align*}
\leq & \left( \frac{1}{\sqrt{1 - \|B_i B_\perp\|}} \cdot \left\| B_i B_\perp \right\| : 2L_x \right)^2 \cdot \frac{1}{\omega} \\
\leq & \left( 2L_x \cdot \frac{1}{\sqrt{1 - \left(1 - \frac{1}{8k \cdot 2^t}\right)^2}} \right)^2 \cdot \frac{1}{\omega} \\
\leq & \left( 2L_x \cdot \frac{1}{\sqrt{1 - \frac{3}{4} \cdot \frac{1}{8k \cdot 2^t}}} \right)^2 \cdot \frac{1}{\omega} \\
= & \frac{L_x^2}{4k^2 \omega \cdot 2^{2t}}
\end{align*}
\]

Thus, we have

\[
\text{Term 1} = \left( \hat{B}_i^\top y \right)^\top P_t^{-1} \hat{B}_i^\top y \leq \left\| B^\top y \right\|^2 \left( \sum_{i=1}^n \lambda_i(x_i) B^\top x_i x_i^\top B \right)^{-1} + \frac{2L_x^2}{k \omega \cdot 2^t} + \frac{L_x^2}{4k^2 \omega \cdot 2^{2t}} \leq \left\| B^\top y \right\|^2 \left( \sum_{i=1}^n \lambda_i(x_i) B^\top x_i x_i^\top B \right)^{-1} + \frac{3L_x^2}{2k \omega \cdot 2^t},
\]

(13)

Next, we investigate Term 2. In order to bound Term 2, we first bound the minimum singular value of \( P_t \) and the maximum singular value of \( Q_t \).

Since \( P_t = \hat{B}_i^\top B (\sum_{i=1}^n \lambda_i(x_i) B^\top x_i x_i^\top B) (\hat{B}_i^\top B)^\top \), we have

\[
\sigma_{\min} (P_t) \geq \sigma_{\min} (\hat{B}_i^\top B) \cdot \omega = \left( 1 - \| \hat{B}_i^\top B_\perp \|^2 \right) \omega \geq \left( 1 - \frac{1}{8^2 k^2 \cdot 2^t} \right) \omega \geq \frac{3}{4} \omega.
\]

Since \( Q_t = \hat{B}_i^\top B (\sum_{i=1}^n \lambda_i(x_i) B^\top x_i x_i^\top B) (\hat{B}_i^\top B_\perp)^\top + \hat{B}_i^\top B_\perp (\sum_{i=1}^n \lambda_i(x_i) B^\top x_i x_i^\top B) (\hat{B}_i^\top B)^\top + \hat{B}_i^\top B_\perp (\sum_{i=1}^n \lambda_i(x_i) B^\top x_i x_i^\top B) (\hat{B}_i^\top B_\perp)^\top \), we have

\[
\sigma_{\max} (Q_t) \leq 3L_x^2 \left\| \hat{B}_i^\top B_\perp \right\| \leq \min \left\{ \frac{3L_x^2}{8k \cdot 2^t}, \frac{\omega}{2} \right\}.
\]

Then, we can bound Term 2 as

\[
\text{Term 2} = \left( \hat{B}_i^\top y \right)^\top (P_t + Q_t)^{-1} Q_t P_t^{-1} \hat{B}_i^\top y \leq \left\| \hat{B}_i^\top y \right\|^2 \cdot \left\| (P_t + Q_t)^{-1} \right\| \cdot \left\| Q_t \right\| \cdot \left\| P_t^{-1} \right\| \leq \frac{4L_x^2 \cdot \sigma_{\max} (Q_t)}{\sigma_{\min} (P_t + Q_t) \cdot \sigma_{\min} (P_t)} \leq \frac{4L_x^2 \cdot \sigma_{\max} (Q_t)}{(\sigma_{\min} (P_t) - \sigma_{\max} (Q_t)) \cdot \sigma_{\min} (P_t)}
\]

30
We first prove the correctness.

(ii) In all phases (Line 4 in Algorithm 2, Line 5 in Algorithm 3), we have that the total number of samples is bounded by

\[ \frac{3L_x^2}{8KL^2} \left( \frac{3}{4} \omega - \frac{3}{4} \omega \right) \cdot \frac{1}{2} \omega \]

\[ = \frac{8L_x^4}{k\omega^2 \cdot 2^t}. \]

Plugging Eqs. (13) and (14) into Eq. (12), we have

\[ \|B_t\|^2 \leq \frac{3L_x^2}{k\omega^2 \cdot 2^t} \]

Below we prove the sample complexity for algorithm DouExpDes (Theorem 4.1).

**Proof of Theorem 4.1.** According to Lemmas C.3 and C.7, we have \( \Pr[\mathcal{E} \cap \mathcal{F}] \geq 1 - \delta \). Below, supposing that event \( \mathcal{E} \cap \mathcal{F} \) holds, we prove the correctness and sample complexity.

We first prove the correctness.

For any task \( m \in [M] \), let \( t^*_m \) denote the first phase which satisfies \( |\tilde{X}_{t,m}| = 1 \). Let \( t_* = \max_{m \in [M]} t^*_m \) denote the total number of phases used. For any task \( m \in [M] \), let \( \Delta_{m, \min} := \min_{x \in \mathcal{X}} (x^* - x) \) denote the minimum reward gap for task \( m \). We have that \( \Delta_{m, \min} \) denote the minimum reward gap among all tasks.

From Lemma C.9, we can obtain the following facts: (i) For any task \( m \in [M] \), the optimal arm \( x^*_m \) will never be eliminated. (ii) \( t^*_m \leq \lceil \log \left( \frac{1}{\Delta_{m, \min}} \right) \rceil + 1 \), and thus, \( t_* \leq \lceil \log \left( \frac{1}{\Delta_{m, \min}} \right) \rceil + 1 \). Therefore, after at most \( \lceil \log \left( \frac{1}{\Delta_{m, \min}} \right) \rceil + 1 \) phases, algorithm DouExpDes will return the optimal arms \( x^*_m \) for all tasks \( m \in [M] \).

Now we prove the sample complexity. In the following, we first prove that the sample complexity of algorithm DouExpDes is bounded by

\[ O \left( \frac{M}{\Delta_{m, \min}} \log (\delta^{-1}) + (\rho^E)^2 dk^4 L_x^2 L_\theta^2 D \log^4 (\delta^{-1}) \right). \]

Recall that \( p = \frac{180kd}{\epsilon^2} \) and \( \zeta = \frac{1}{10} \). Then, summing the number of samples used in subroutines FeatureRec and EliLowRep in all phases (Line 4 in Algorithm 2, Line 5 in Algorithm 3), we have that the total number of samples is

\[
\sum_{t=1}^{t_*} pMT_t + \sum_{m=1}^{M} \sum_{t=1}^{t^*_m} N_{t,m}
\]

\[
= \sum_{t=1}^{t_*} p \cdot O \left( (1 + \zeta)^3 (\rho^E)^2 k^4 L_x^2 L_\theta^2 \max \left\{ 2^{2t}, \frac{L_x^4}{\omega^2} \right\} \log \left( \frac{p}{\delta_t} \right) \right)
\]

\[
+ \sum_{m=1}^{M} \sum_{t=1}^{t^*_m} O \left( 2^{2t} (1 + \zeta) \rho^C L_x \max \left\{ 2^{2t}, \frac{L_x^4}{\omega^2} \right\} \frac{1}{\delta_t} \log \left( \frac{p}{\delta_t} \right) \right)
\]

\[
= O \left( \rho^E k^4 dL_x^2 L_\theta \max \left\{ 2^{2t}, \frac{L_x^4}{\omega^2} \right\} \log \left( \frac{d \log (\Delta_{m, \min})}{\delta} \right) \right).
\]

\[
= \sum_{t=1}^{t_*} O \left( (\rho^E)^2 k^4 dL_x^2 L_\theta \max \left\{ 2^{2t}, \frac{L_x^4}{\omega^2} \right\} \log \left( \frac{d \log (\Delta_{m, \min})}{\delta} \right) \right).
\]

\[
= \sum_{m=1}^{M} \sum_{t=1}^{t^*_m} O \left( 2^{2t} \rho^C \max \left\{ \frac{L_x^4}{\omega^2} \right\} \frac{1}{\delta_t} \log \left( \frac{d \log (\Delta_{m, \min})}{\delta} \right) \right)
\]

\[
+ \sum_{m=1}^{M} \sum_{t=1}^{t^*_m} O \left( 2^{2t} \rho^C \max \left\{ \frac{L_x^4}{\omega^2} \right\} \frac{1}{\delta_t} \log \left( \frac{d \log (\Delta_{m, \min})}{\delta} \right) + k \right)
\]

\[
\leq 4L_x^2 \frac{3L_x^2}{8KL^2} \left( \frac{3}{4} \omega - \frac{3}{4} \omega \right) \cdot \frac{1}{2} \omega
\]

\[ = \frac{8L_x^4}{k\omega^2 \cdot 2^t}. \]
Next, we prove that the sample complexity of algorithm DouExpDes is bounded by

\[ O\left( \sum_{m=1}^{M} \min_{\lambda \in \Delta^k} \max_{y \in \mathcal{Y}(x^*_m, \theta^*)} \| B^\top \mathbf{y} \|_2^2 \left( \sum_{i=1}^{m} \lambda(x_i) B^\top_i x \lambda(x_i) B_i \right)^{-1} \right) \leq \rho \cdot \log \left( \frac{d}{\delta} \right), \]

\[ + O\left( M k \Delta^{-2} \log \left( \frac{n^2 M \log(\Delta^{-1})}{\delta} \right) \right), \]

for any $Z \subseteq X$, $\mathcal{Y}(Z) := \{ x - x' : \forall x, x' \in Z, x \neq x' \}$ and $\mathcal{Y}_m(Z) := \{ x^*_m - x : \forall x \in Z, x \neq x^*_m \}$. Then, we have that for any task $m \in [M]$ and phase $t \geq 2$,

\[
(2t)^2 \rho_{t,m}^G = (2t)^2 \min_{\lambda \in \Delta^k} \max_{y \in \mathcal{Y}(x^*_m, \theta^*)} \| \hat{B}_t^\top \mathbf{y} \|_2^2 \left( \sum_{i=1}^{t} \lambda(x_i) \hat{B}_i^\top x \lambda(x_i) \hat{B}_i \right)^{-1} \\
\leq (2t)^2 \max_{y \in \mathcal{Y}(x^*_m, \theta^*)} \| \hat{B}_t^\top \mathbf{y} \|_2^2 \left( \sum_{i=1}^{t} \lambda(x_i) \hat{B}_i^\top x \lambda(x_i) \hat{B}_i \right)^{-1} \\
\leq (2t)^2 \max_{y \in \mathcal{Y}(z^*_m, \theta^*)} \| \hat{B}_t^\top \mathbf{y} \|_2^2 \left( \sum_{i=1}^{t} \lambda(x_i) \hat{B}_i^\top x \lambda(x_i) \hat{B}_i \right)^{-1} \\
\leq 4 (2t)^2 \max_{y \in \mathcal{Y}(z^*_m, \theta^*)} \| \hat{B}_t^\top \mathbf{y} \|_2^2 \left( \sum_{i=1}^{t} \lambda(x_i) \hat{B}_i^\top x \lambda(x_i) \hat{B}_i \right)^{-1} \\
\leq 4 (2t)^2 \left( \max_{y \in \mathcal{Y}(z^*_m, \theta^*)} \| B^\top \mathbf{y} \|_2^2 \left( \sum_{i=1}^{t} \lambda(x_i) B^\top x \lambda(x_i) B_i \right)^{-1} + \frac{11 L^4}{\omega^2 k} \right) \\
\leq 4 \left( \frac{16 \max_{y \in \mathcal{Y}(z^*_m, \theta^*)} \| B^\top \mathbf{y} \|_2^2 \left( \sum_{i=1}^{t} \lambda(x_i) B^\top x \lambda(x_i) B_i \right)^{-1}}{(4 \cdot 2^{-t})^2} + \frac{11 L^4}{\omega^2 k} \right) \\
\leq 4 \left( \frac{16 \max_{y \in \mathcal{Y}(x^*_m, \theta^*)} \| B^\top \mathbf{y} \|_2^2 \left( \sum_{i=1}^{t} \lambda(x_i) B^\top x \lambda(x_i) B_i \right)^{-1}}{(y^\top \theta)^2} + \frac{11 L^4}{\omega^2 k} \right) \\
\leq 4 \left( \frac{16 \max_{y \in \mathcal{Y}(x^*_m, \theta^*)} \| B^\top \mathbf{y} \|_2^2 \left( \sum_{i=1}^{t} \lambda(x_i) B^\top x \lambda(x_i) B_i \right)^{-1}}{(y^\top \theta)^2} + \frac{11 L^4}{\omega^2 k} \right) \\
\leq 4 \left( \frac{16 \max_{\lambda \in \Delta^k} \max_{y \in \mathcal{Y}_m(x)} \| B^\top \mathbf{y} \|_2^2 \left( \sum_{i=1}^{t} \lambda(x_i) B^\top x \lambda(x_i) B_i \right)^{-1}}{(y^\top \theta)^2} + \frac{11 L^4}{\omega^2 k} \right). \tag{17}
\]

Here inequality (a) is due to $\hat{X}_{t,m} \subseteq Z_{t,m}$ (from Lemma C.9). Inequality (b) uses the fact that for any $y = x_i - x_j \in \mathcal{Y}(Z_{t,m})$, we can write $y = (x^*_m - x_j) - (x^*_m - x_i)$, and the triangle inequality. Inequality (c) follows from Lemma C.10, and inequality (d) is due to that for any $y \in \mathcal{Y}_m(Z_{t,m})$, $y^\top \theta_m \leq 4 \cdot 2^{-t}$ (from the definition of $Z_{t,m}$). Equality (e) comes from the definition of $\lambda^*_m$.\]
Multi-task Representation Learning for Pure Exploration in Linear Bandits

Let $L := \log^2 \left( \frac{d \log(\Delta^{-1}_m)}{\delta} \right) \cdot \log^2 \left( \rho^E kdL_x L_\theta \max \left\{ \Delta^{-1}_m, \frac{L^x}{\omega} \right\} \log \left( \frac{d \log(\Delta^{-1}_m)}{\delta} \right) \log(\frac{d \log(\Delta^{-1}_m)}{\delta}) \right)$. Plugging Eq. (17) into Eq. (16), we have that with probability $1 - \delta$, the number of samples used by algorithm DouExpDes is bounded by

$$O \left( \sum_{m=1}^M \sum_{t=1}^{2^t \rho_{t,m}^E} \log \left( \frac{n^2 M \log(\Delta^{-1}_m)}{\delta} \right) + Mk \log(\Delta^{-1}_m) + \sum_{t=1}^{\max(\Delta^{-1}_m)} (\rho^E)^2 k^4 dL_x^2 L_\theta^2 \max \left\{ 2^{2t}, \frac{L^x}{\omega^2} \right\} L \right)$$

Equality (a) uses Lemma C.6.

When $L_x = \omega = \Theta(1)$, we have that with probability $1 - \delta$, the sample complexity of algorithm DouExpDes is bounded by

$$\tilde{O} \left( \sum_{m=1}^M \min_{\lambda \in \Delta, y \in \mathcal{Y}_n^y(\lambda)} \frac{||B^T y||^2}{(\sum_{i=1}^n \lambda_i)^{\frac{1}{2}}} \cdot \log \left( \frac{n^2 M \log(\Delta^{-1}_m)}{\delta} \right) \cdot \log(\Delta^{-1}_m) \right) + \frac{ML_x^4}{\omega^2 k} \cdot \log \left( \frac{n^2 M \log(\Delta^{-1}_m)}{\delta} \right) + Mk \cdot \log \left( \frac{n^2 M \log(\Delta^{-1}_m)}{\delta} \right) \cdot \log(\Delta^{-1}_m) + (\rho^E)^2 k^4 dL_x^2 L_\theta^2 \max \left\{ \Delta^{-1}_m, \frac{L^x}{\omega^2} \right\} \cdot L, \right)$$

where equality (a) uses Lemma C.6.

D. Proofs for Algorithm C-DouExpDes

In this section, we present the proofs for Algorithm C-DouExpDes.

D.1. Context Distribution Estimation and Sample Batch Planning

Define $\lambda^E_D$ and $\rho^E_D$ as the optimal solution and the optimal value of the following E-optimal design optimization:

$$\min_{\lambda \in \Delta, A} \left\| \sum_{a \in A} \lambda(a) \mathbb{E}_{s \sim D} \left[ \phi(s, a) \phi(s, a)^T \right] \right\|^{-1}.$$  \hspace{1cm} (18)

**Lemma D.1.** It holds that

$$\rho^E_D \leq \frac{1}{\nu}.$$  

**Proof of Lemma D.1.** The optimization in Eq. (18) is equivalent to maximize the minimum singular value of the matrix

$$\sum_{a \in A} \lambda(a) \mathbb{E}_{s \sim D} \left[ \phi(s, a) \phi(s, a)^T \right].$$

Thus, $\lambda^E_D$ is the optimal solution of the following optimization:

$$\max_{\lambda \in \Delta, A} \sigma_{\min} \left( \sum_{a \in A} \lambda(a) \mathbb{E}_{s \sim D} \left[ \phi(s, a) \phi(s, a)^T \right] \right).$$

Using Assumption 3.3, we have

$$\sigma_{\min} \left( \sum_{a \in A} \lambda^E_D \mathbb{E}_{s \sim D} \left[ \phi(s, a) \phi(s, a)^T \right] \right) \geq \nu.$$  

33
Then, we have
\[
\rho_D^E = \left\| \sum_{a \in A} \lambda_D^E \mathbb{E}_{s \sim D} [\phi(s, a)\phi(s, a)^\top] \right\|^{-1} = \frac{1}{\sigma_{\min} \left( \sum_{a \in A} \lambda_D^E \mathbb{E}_{s \sim D} [\phi(s, a)\phi(s, a)^\top] \right)} \leq \frac{1}{\nu}.
\]

Define event
\[
K := \left\{ \left\| \mathbb{E}_{s \sim \hat{D}} [\phi(s, a)\phi(s, a)^\top] - \mathbb{E}_{s \sim D} [\phi(s, a)\phi(s, a)^\top] \right\| \leq \frac{8L^2_\phi \log \left( \frac{20d|A|}{\delta} \right)}{\sqrt{T_0}}, \forall a \in A \right\}.
\]

**Lemma D.2.** It holds that
\[
\Pr[K] \geq 1 - \frac{\delta}{5}.
\]
Furthermore, if event $K$ holds and
\[
T_0 = \left\lceil \frac{32^2 (1 + \zeta)^2 L^4_\phi}{\nu^2 \log^2 \left( \frac{20d|A|}{\delta} \right)} \right\rceil,
\]
we have that for any $a \in A$,
\[
\left\| \mathbb{E}_{s \sim \hat{D}} [\phi(s, a)\phi(s, a)^\top] - \mathbb{E}_{s \sim D} [\phi(s, a)\phi(s, a)^\top] \right\| \leq \frac{\nu}{4(1 + \zeta)}.
\]

**Proof of Lemma D.2.** For any $(s, a) \in S \times A$, $\|\phi(s, a)\phi(s, a)^\top\| \leq L^2_\phi$. Then, using the matrix Bernshtet inequality (Lemma E.2) and a union bound over $a \in A$, we have that with probability $1 - \frac{\delta}{5}$, for any $a \in A$,
\[
\left\| \mathbb{E}_{s \sim \hat{D}} [\phi(s, a)\phi(s, a)^\top] - \mathbb{E}_{s \sim D} [\phi(s, a)\phi(s, a)^\top] \right\| \leq 4L^2_\phi \log \left( \frac{10 - 2d|A|}{\delta} \right) T_0 + 4L^2_\phi \log \left( \frac{10 - 2d|A|}{\delta} \right) T_0 \leq \frac{8L^2_\phi \log \left( \frac{20d|A|}{\delta} \right)}{\sqrt{T_0}}.
\]

If $T_0 \geq 32^2 (1 + \zeta)^2 \nu^{-2} L^4_\phi \log^2 \left( \frac{20d|A|}{\delta} \right)$, we have
\[
\left\| \mathbb{E}_{s \sim \hat{D}} [\phi(s, a)\phi(s, a)^\top] - \mathbb{E}_{s \sim D} [\phi(s, a)\phi(s, a)^\top] \right\| \leq \frac{\nu}{4(1 + \zeta)},
\]
which completes the proof.

Define event
\[
L := \left\{ \left\| \sum_{i=1}^p \phi(s^{(f)}_{m,j,i}, a_i)\phi(s^{(f)}_{m,j,i}, a_i)^\top - \sum_{i=1}^p \mathbb{E}_{s \sim \hat{D}} [\phi(s, a_i)\phi(s, a_i)^\top] \right\| \leq 8L^2_\phi \sqrt{p} \log \left( \frac{40dMT}{\delta} \right), \forall m \in [M], \forall j \in [T], \forall f \in \{1, 2\} \right\}.
\]
**Lemma D.3.** It holds that

\[ \Pr[\mathcal{L}] \geq 1 - \frac{\delta}{5}. \]

Furthermore, if event $\mathcal{L}$ holds and

\[ p = \left[ \frac{32^2(1 + \zeta)^2 L_\phi^4 \log^2 \left( \frac{40dMT}{\delta} \right)}{\nu^2} \right], \tag{19} \]

we have that for any $m \in [M], j \in [T]$ and $\ell \in \{1, 2\}$,

\[
\left\| \sum_{i=1}^{p} \phi(s_{m,j,i}^{(\ell)}, \bar{a}_i)\phi(s_{m,j,i}^{(\ell)}, \bar{a}_i)^\top - \sum_{i=1}^{p} \mathbb{E}_{\bar{s} \sim \mathcal{D}} [\phi(s, \bar{a}_i)\phi(s, \bar{a}_i)^\top] \right\| \leq \frac{p\nu}{4(1 + \zeta)}.
\]

Here, the value of $T$ is specified in Eq. (29).

**Proof of Lemma D.3.** For any $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\|\phi(s, a)\phi(s, a)^\top\| \leq L_\phi^4$. Then, using the matrix Bernstein inequality (Lemma E.2) and a union bound over $m \in [M], j \in [T]$ and $\ell \in \{1, 2\}$, we have that with probability $1 - \frac{\delta}{5}$, for any $m \in [M], j \in [T]$ and $\ell \in \{1, 2\}$,

\[
\left\| \sum_{i=1}^{p} \phi(s_{m,j,i}^{(\ell)}, \bar{a}_i)\phi(s_{m,j,i}^{(\ell)}, \bar{a}_i)^\top - \sum_{i=1}^{p} \mathbb{E}_{\bar{s} \sim \mathcal{D}} [\phi(s, \bar{a}_i)\phi(s, \bar{a}_i)^\top] \right\| \leq 4L_\phi^2 \sqrt{p \log \left( \frac{10 \cdot 4dMT}{\delta} \right)} + 4L_\phi^2 \log \left( \frac{10 \cdot 4dMT}{\delta} \right)
\]

\[
\leq 8L_\phi^2 \sqrt{p \log \left( \frac{40dMT}{\delta} \right)}.
\]

In addition, if $p \geq 32^2(1 + \zeta)^2 \nu^{-2} L_\phi^4 \log^2 \left( \frac{40dMT}{\delta} \right)$, we have that

\[
8L_\phi^2 \sqrt{p \log \left( \frac{40dMT}{\delta} \right)} \leq \frac{p\nu}{4(1 + \zeta)}
\]

and thus,

\[
\left\| \sum_{i=1}^{p} \phi(s_{m,j,i}^{(\ell)}, \bar{a}_i)\phi(s_{m,j,i}^{(\ell)}, \bar{a}_i)^\top - \sum_{i=1}^{p} \mathbb{E}_{\bar{s} \sim \mathcal{D}} [\phi(s, \bar{a}_i)\phi(s, \bar{a}_i)^\top] \right\| \leq \frac{p\nu}{4(1 + \zeta)},
\]

which completes the proof.

For any task $m \in [M], j \in [T]$ and $\ell \in \{1, 2\}$, let

\[
\Phi_{m,j}^{(\ell)} = \begin{bmatrix} \phi(s_{m,j,1}^{(\ell)}, \bar{a}_1)^\top \\ \cdots \\ \phi(s_{m,j,p}^{(\ell)}, \bar{a}_p)^\top \end{bmatrix},
\]

and

\[
(\Phi_{m,j}^{(\ell)})^+ = ((\Phi_{m,j}^{(\ell)})^\top \Phi_{m,j}^{(\ell)})^{-1}(\Phi_{m,j}^{(\ell)})^\top.
\]

**Lemma D.4.** Suppose that event $\mathcal{K} \cap \mathcal{L}$ holds. Then, for any $m \in [M], j \in [T]$ and $\ell \in \{1, 2\}$,

\[
\left\| (\Phi_{m,j}^{(\ell)})^+ \right\| \leq 2 \sqrt{\frac{(1 + \zeta)}{p\nu}}.
\]
\textbf{Proof of Lemma D.4}. We first assume that \((\Phi_{m,j}^{(\ell)})^T \Phi_{m,j}^{(\ell)}\) is invertible. In our later analysis, we will prove that as long as \(T_0\) and \(p\) are large enough, \((\Phi_{m,j}^{(\ell)})^T \Phi_{m,j}^{(\ell)}\) is invertible.

For any \(m \in [M], j \in [T]\) and \(\ell \in \{1, 2\}\), we have

\[
\|\left(\Phi_{m,j}^{(\ell)}\right)^+\| = \left\|\left((\Phi_{m,j}^{(\ell)})^T \Phi_{m,j}^{(\ell)}\right)^{-1}(\Phi_{m,j}^{(\ell)})^T\right\|
\]

\[
= \sqrt{\left\|\left((\Phi_{m,j}^{(\ell)})^T \Phi_{m,j}^{(\ell)}\right)^{-1}\right\|}
\]

\[
= \frac{1}{\sqrt{\sigma_{\min} \left((\Phi_{m,j}^{(\ell)})^T \Phi_{m,j}^{(\ell)}\right)}}.
\]

(20)

In addition, we have

\[
\sigma_{\min} \left((\Phi_{m,j}^{(\ell)})^T \Phi_{m,j}^{(\ell)}\right)
\]

\[
= \sigma_{\min} \left(\sum_{i=1}^{p} \phi(s_{m,j,i}^{(\ell)}, \bar{a}_i) \phi(s_{m,j,i}^{(\ell)}, \bar{a}_i)^T\right)
\]

\[
= \sigma_{\min} \left(\sum_{i=1}^{p} \mathbb{E}_{s \sim \mathcal{D}} \left[\phi(s, \bar{a}_i) \phi(s, \bar{a}_i)^T\right] + \sum_{i=1}^{p} \phi(s_{m,j,i}^{(\ell)}, \bar{a}_i) \phi(s_{m,j,i}^{(\ell)}, \bar{a}_i)^T - \sum_{i=1}^{p} \mathbb{E}_{s \sim \mathcal{D}} \left[\phi(s, \bar{a}_i) \phi(s, \bar{a}_i)^T\right]\right)
\]

\[
\geq \sigma_{\min} \left(\sum_{i=1}^{p} \mathbb{E}_{s \sim \mathcal{D}} \left[\phi(s, \bar{a}_i) \phi(s, \bar{a}_i)^T\right] - \sum_{i=1}^{p} \mathbb{E}_{s \sim \mathcal{D}} \left[\phi(s, \bar{a}_i) \phi(s, \bar{a}_i)^T\right]\right)
\]

\[
\geq \sigma_{\min} \left(\sum_{i=1}^{p} \mathbb{E}_{s \sim \mathcal{D}} \left[\phi(s, \bar{a}_i) \phi(s, \bar{a}_i)^T\right] - \sum_{i=1}^{p} \mathbb{E}_{s \sim \mathcal{D}} \left[\phi(s, \bar{a}_i) \phi(s, \bar{a}_i)^T\right]\right)
\]

\[
\geq \frac{\sigma_{\min} \left(\sum_{i=1}^{p} \mathbb{E}_{s \sim \mathcal{D}} \left[\phi(s, \bar{a}_i) \phi(s, \bar{a}_i)^T\right]\right)}{4(1 + \zeta)} - \frac{p \nu}{4(1 + \zeta)},
\]

(21)

where the last inequality uses Lemmas D.2 and D.3.

In the following, we analyze \(\sigma_{\min} \left(\sum_{i=1}^{p} \mathbb{E}_{s \sim \mathcal{D}} \left[\phi(s, \bar{a}_i) \phi(s, \bar{a}_i)^T\right]\right)\). According to the guarantee of the rounding procedure \textsc{round}, we have

\[
\left\|\left(\sum_{i=1}^{p} \mathbb{E}_{s \sim \mathcal{D}} \left[\phi(s, \bar{a}_i) \phi(s, \bar{a}_i)^T\right]\right)^{-1}\right\| \leq (1 + \zeta) \left\|\left(p \sum_{a \in \mathcal{A}} \lambda_a \mathbb{E}_{s \sim \mathcal{D}} \left[\phi(s, a) \phi(s, a)^T\right]\right)^{-1}\right\|
\]

36
\[ \leq (1 + \zeta) \left\| \left( p \sum_{a \in \mathcal{A}} \lambda^{E}(a) E_{s \sim \mathcal{D}} \left[ \phi(s, a) \phi(s, a)^{\top} \right] \right)^{-1} \right\| , \]

which implies that

\[
\begin{align*}
\sigma_{\min} \left( \sum_{i=1}^{p} E_{s \sim \mathcal{D}} \left[ \phi(s, \bar{a}_{i}) \phi(s, \bar{a}_{i})^{\top} \right] \right) \\
\geq \frac{p}{1 + \zeta} \sigma_{\min} \left( \sum_{a \in \mathcal{A}} \lambda^{E}(a) E_{s \sim \mathcal{D}} \left[ \phi(s, a) \phi(s, a)^{\top} \right] \right) \\
\geq \frac{p}{1 + \zeta} \sigma_{\min} \left( \sum_{a \in \mathcal{A}} \lambda^{E}(a) E_{s \sim \mathcal{D}} \left[ \phi(s, a) \phi(s, a)^{\top} \right] \right) \\
+ \sum_{a \in \mathcal{A}} \lambda^{E}(a) E_{s \sim \mathcal{D}} \left[ \phi(s, a) \phi(s, a)^{\top} \right] - \sum_{a \in \mathcal{A}} \lambda^{E}(a) E_{s \sim \mathcal{D}} \left[ \phi(s, a) \phi(s, a)^{\top} \right] \\
\geq \frac{p}{1 + \zeta} \left( \sigma_{\min} \left( \sum_{a \in \mathcal{A}} \lambda^{E}(a) E_{s \sim \mathcal{D}} \left[ \phi(s, a) \phi(s, a)^{\top} \right] \right) \right) \\
- \left\| \sum_{a \in \mathcal{A}} \lambda^{E}(a) E_{s \sim \mathcal{D}} \left[ \phi(s, a) \phi(s, a)^{\top} \right] - \sum_{a \in \mathcal{A}} \lambda^{E}(a) E_{s \sim \mathcal{D}} \left[ \phi(s, a) \phi(s, a)^{\top} \right] \right\| \right) \right) \\
\geq \frac{p}{1 + \zeta} \left( \nu - \frac{\nu}{4(1 + \zeta)} \right) \\
= \frac{3p\nu}{4(1 + \zeta)},
\end{align*}
\]

where inequality (a) uses Lemmas D.1 and D.2.

Plugging Eq. (22) into Eq. (21), we have

\[
\begin{align*}
\sigma_{\min} \left( \left( \Phi_{m,j}^{(T)} \right)^{\top} \Phi_{m,j}^{(T)} \right) \\
\geq \frac{3p\nu}{4(1 + \zeta)} - \frac{p\nu}{4(1 + \zeta)} - \frac{p\nu}{4(1 + \zeta)} \\
= \frac{p\nu}{4(1 + \zeta)}.
\end{align*}
\]

Equations (21) and (23) show that if \( T_{0} \) and \( p \) are large enough to satisfy that \( \left\| E_{s \sim \mathcal{D}} \left[ \phi(s, a) \phi(s, a)^{\top} \right] - E_{s \sim \mathcal{D}} \left[ \phi(s, a) \phi(s, a)^{\top} \right] \right\| \leq \frac{\nu}{4(1 + \zeta)} \) for any \( a \in \mathcal{A} \) and \( \sum_{i=1}^{p} \phi(s_{m,j,i}, \bar{a}_{i}) \phi(s_{m,j,i}, \bar{a}_{i})^{\top} - \sum_{i=1}^{p} E_{s \sim \mathcal{D}} \left[ \phi(s, \bar{a}_{i}) \phi(s, \bar{a}_{i})^{\top} \right] \right\| \leq \frac{\nu}{4(1 + \zeta)} \) for any \( m \in [M], j \in [T] \) and \( \ell \in \{1, 2\} \), respectively, then we have that \( \left( \Phi_{m,j}^{(T)} \right)^{\top} \Phi_{m,j}^{(T)} \) is invertible.

Continuing with Eq. (20), we have

\[
\left\| \left( \Phi_{m,j}^{(T)} \right)^{\top} \right\| \leq 2 \frac{(1 + \zeta)}{p\nu}.
\]

\[\square\]

D.2. Global Feature Extractor Recovery with Stochastic Contexts

In subroutine C-FeatRecover, for any \( m \in [M], j \in [T], i \in [p] \) and \( \ell \in \{1, 2\} \), let \( s_{m,j,i}^{(\ell)} \) and \( \eta_{m,j,i}^{(\ell)} \) denote the random context and noise of the \( \ell \)-th sample on action \( \bar{a}_{i} \) in the \( j \)-th round for task \( m \), respectively. Here, the superscript \( \ell \in \{1, 2\} \) refers to the first sample (Line 4 in Algorithm 5) or the second sample (Line 5 in Algorithm 5) on an action \( \bar{a}_{i} \).
In C-FeatRecover, for any $m \in [M]$, $j \in [T]$, $i \in [p]$ and $\ell \in \{1, 2\}$, let $\alpha^{(\ell)}_{m,j} \leftarrow [\alpha^{(\ell)}_{m,j,1}, \ldots, \alpha^{(\ell)}_{m,j,p}]^\top$, and then, $\bar{\theta}^{(\ell)}_{m,j} = (\Phi^{(\ell)}_{m,j})^\top \alpha^{(\ell)}_{m,j}$. Recall that $Z = \frac{1}{MT} \sum_{m=1}^{M} \sum_{j=1}^{T} \bar{\theta}^{(1)}_{m,j} (\bar{\theta}^{(2)}_{m,j})^\top$.

**Lemma D.5 (Expectation of $Z$).** It holds that

$$
\mathbb{E}[Z] = \frac{1}{M} \sum_{m=1}^{M} \theta_m \theta_m^\top.
$$

**Proof of Lemma D.5.** $Z$ can be written as

$$
Z = \frac{1}{MT} \sum_{m=1}^{M} \sum_{j=1}^{T} \bar{\theta}^{(1)}_{m,j} (\bar{\theta}^{(2)}_{m,j})^\top
\quad = \frac{1}{MT} \sum_{m=1}^{M} \sum_{j=1}^{T} (\Phi^{(1)}_{m,j})^+ \left[\begin{array}{c}
\alpha^{(1)}_{m,j,1} \\
\vdots \\
\alpha^{(1)}_{m,j,p}
\end{array}\right] \left[\begin{array}{c}
\alpha^{(2)}_{m,j,1}, \ldots, \\
\alpha^{(2)}_{m,j,p}
\end{array}\right] (\Phi^{(2)}_{m,j})^+)^\top
\quad = \frac{1}{MT} \sum_{m=1}^{M} \sum_{j=1}^{T} \left[\begin{array}{c}
(\phi(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \theta_m) \\
(\phi(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \theta_m) \\
\vdots \\
(\phi(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \theta_m)
\end{array}\right] \left[\begin{array}{c}
\phi(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \theta_m \\
\phi(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \theta_m \\
\vdots \\
\phi(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \theta_m)
\end{array}\right] (\Phi^{(2)}_{m,j})^+)^\top.
$$

For any task $m \in [M]$, $j \in [T]$, $i \in [p]$, the sample on action $a_i$ in the first round (i.e., $s^{(1)}_{m,j,i}$ and $\eta^{(1)}_{m,j,i}$) is independent of that in the second round (i.e., $s^{(2)}_{m,j,i}$ and $\eta^{(2)}_{m,j,i}$). Hence, taking the expectation on $Z$, we obtain

$$
\mathbb{E}[Z] = \frac{1}{MT} \sum_{m=1}^{M} \sum_{j=1}^{T} \mathbb{E}\left[ (\Phi^{(1)}_{m,j})^+ \right]
\quad = \frac{1}{MT} \sum_{m=1}^{M} \sum_{j=1}^{T} \mathbb{E}\left[ (\Phi^{(1)}_{m,j})^\top \Phi^{(1)}_{m,j} \right]^{-1} (\Phi^{(1)}_{m,j})^\top
\quad = \frac{1}{MT} \sum_{m=1}^{M} \sum_{j=1}^{T} \mathbb{E}\left[ (\Phi^{(1)}_{m,j})^\top \Phi^{(1)}_{m,j} \right]^{-1} \cdot \Phi^{(1)}_{m,j} \theta_m \theta_m^\top \Phi^{(2)}_{m,j} (\Phi^{(2)}_{m,j})^\top
\quad = \frac{1}{MT} \sum_{m=1}^{M} \sum_{j=1}^{T} \theta_m \theta_m^\top.
$$

38
From Eq. (24), we can bound $\|Z - \mathbb{E}[Z]\|$ as

$$
= \frac{1}{M} \sum_{m=1}^{M} \theta_m \theta_m^T.
$$

Define event

$$
\mathcal{G} := \left\{ \|Z - \mathbb{E}[Z]\| \leq \frac{256(1 + \zeta)L_0L_0 \log \left( \frac{50d}{\delta} \right)}{\nu \sqrt{MT}} \log \left( \frac{100pMT}{\delta} \right) \right\}.
$$

Lemma D.6 (Concentration of $Z$). Suppose that $\mathcal{K} \cap \mathcal{L}$ holds. Then, it holds that

$$
\Pr[\mathcal{G}] \geq 1 - \frac{\delta}{5}.
$$

Proof of Lemma D.6. Define the following matrices:

$$
D_{m,j} := \frac{1}{MT}(\Phi_{m,j}^{(1)})^+.
$$

$$
E_{m,j} := \frac{1}{MT}(\Phi_{m,j}^{(1)})^+.
$$

$$
F_{m,j} := \frac{1}{MT}(\Phi_{m,j}^{(1)})^+.
$$

$$
\begin{bmatrix}
\phi(s_{m,j,1}^{(1)}, \bar{a}_1)^{\top} \theta_m & \phi(s_{m,j,1}^{(2)}, \bar{a}_1)^{\top} \theta_m & \cdots & \phi(s_{m,j,1}^{(2)}, \bar{a}_1)^{\top} \theta_m \\
\phi(s_{m,j,1}^{(1)}, \bar{a}_1)^{\top} \theta_m & \phi(s_{m,j,1}^{(1)}, \bar{a}_1)^{\top} \theta_m & \cdots & \phi(s_{m,j,1}^{(2)}, \bar{a}_1)^{\top} \theta_m \\
\phi(s_{m,j,p}^{(1)}, \bar{a}_p)^{\top} \theta_m & \phi(s_{m,j,p}^{(2)}, \bar{a}_p)^{\top} \theta_m & \cdots & \phi(s_{m,j,p}^{(2)}, \bar{a}_p)^{\top} \theta_m
\end{bmatrix}
$$

$$
\begin{bmatrix}
\eta_{m,j,1}^{(1)}, \eta_{m,j,1}^{(1)}, \cdots, \eta_{m,j,1}^{(2)} \eta_{m,j,1}^{(1)} & \cdots & \eta_{m,j,p}^{(1)}, \eta_{m,j,p}^{(1)}, \cdots, \eta_{m,j,p}^{(2)} \eta_{m,j,p}^{(2)}
\end{bmatrix}
$$

$$
\begin{bmatrix}
\phi(s_{m,j,1}^{(1)}, \bar{a}_1)^{\top} \theta_m & \phi(s_{m,j,1}^{(2)}, \bar{a}_1)^{\top} \theta_m & \cdots & \phi(s_{m,j,1}^{(2)}, \bar{a}_1)^{\top} \theta_m \\
\phi(s_{m,j,1}^{(1)}, \bar{a}_1)^{\top} \theta_m & \phi(s_{m,j,1}^{(1)}, \bar{a}_1)^{\top} \theta_m & \cdots & \phi(s_{m,j,1}^{(2)}, \bar{a}_1)^{\top} \theta_m \\
\phi(s_{m,j,p}^{(1)}, \bar{a}_p)^{\top} \theta_m & \phi(s_{m,j,p}^{(2)}, \bar{a}_p)^{\top} \theta_m & \cdots & \phi(s_{m,j,p}^{(2)}, \bar{a}_p)^{\top} \theta_m
\end{bmatrix}
$$

$$
\begin{bmatrix}
\phi(s_{m,j,1}^{(1)}, \bar{a}_1)^{\top} \theta_m & \phi(s_{m,j,1}^{(2)}, \bar{a}_1)^{\top} \theta_m & \cdots & \phi(s_{m,j,1}^{(2)}, \bar{a}_1)^{\top} \theta_m \\
\phi(s_{m,j,1}^{(1)}, \bar{a}_1)^{\top} \theta_m & \phi(s_{m,j,1}^{(1)}, \bar{a}_1)^{\top} \theta_m & \cdots & \phi(s_{m,j,1}^{(2)}, \bar{a}_1)^{\top} \theta_m \\
\phi(s_{m,j,p}^{(1)}, \bar{a}_p)^{\top} \theta_m & \phi(s_{m,j,p}^{(2)}, \bar{a}_p)^{\top} \theta_m & \cdots & \phi(s_{m,j,p}^{(2)}, \bar{a}_p)^{\top} \theta_m
\end{bmatrix}
$$

From Eq. (24), we can bound $\|Z - \mathbb{E}[Z]\|$ as

$$
\|Z - \mathbb{E}[Z]\| \leq \|D - \mathbb{E}[D]\| + \|E - \mathbb{E}[E]\| + \|F - \mathbb{E}[F]\|. \quad (25)
$$

Similar to the proof of Lemma C.3, in order to use the truncated matrix Bernstein inequality (Lemma E.2), we define the truncated noise and some truncated matrices as follows.

Let $R > 0$ be a truncation parameter of noises which will be chosen later. For any $m \in [M]$, $j \in [T]$, $i \in [p]$ and $\ell \in \{1, 2\}$, let $\eta_{m,j,i}^{(\ell)} = \eta_{m,j,i}^k \mathbb{1}\{|\eta_{m,j,i}| \leq R\}$ denote the truncated noise. Furthermore, we define the following matrices with truncated noises:

$$
\tilde{E}_{m,j} := \frac{1}{MT}(\Phi_{m,j}^{(1)})^+.
$$
We first analyze \( \| \mathbf{D} - \mathbf{E}[\mathbf{D}] \| \). Since \( |\phi(s_{m,j}^{(1)}, \bar{a}_i)^+ \Theta_m| \leq L_\phi L_\theta \) for any \( m \in [M], j \in [T], i \in [p] \) and \( \ell \in \{1, 2\} \), we have that \( \| \mathbf{D}_{m,j} \| \leq \frac{1}{\sqrt{T}} \cdot pL_\phi L_\theta B_\phi^2 \) and \( \sum_{m=1}^M \sum_{j=1}^T \| \mathbf{D}_{m,j} \|^2 \leq MT \cdot \frac{1}{MT^2} \cdot p^2 L_\phi^2 L_\theta^2 B_\phi^4 = \frac{1}{MT} \cdot p^2 L_\phi^2 L_\theta^2 B_\phi^4 \) for any \( m \in [M] \) and \( j \in [T] \).

Let \( \delta' \in (0, 1) \) be a confidence parameter which will be chosen later. Using the matrix Bernstein inequality (Lemma E.2), we have that with probability at least \( 1 - \delta' \),

\[
\| \mathbf{D} - \mathbf{E}[\mathbf{D}] \| \leq 4\sqrt{\frac{p^2 L_\phi^2 L_\theta^2 B_\phi^4 \log \left( \frac{2d}{\delta} \right)}{MT}} + 4pL_\phi L_\theta B_\phi^2 \log \left( \frac{2d}{\delta} \right) \leq 8 \cdot 4pL_\phi L_\theta B_\phi^2 \log \left( \frac{2d}{\delta} \right) \sqrt{MT} \tag{26}
\]

Next, we bound \( \| \mathbf{E} - \mathbf{E}[\mathbf{E}] \| \). Since \( |\phi(s_{m,j,i}^{(1)}, \bar{a}_i)^+ \Theta_m| \leq L_\phi L_\theta \) and \( |\tilde{\eta}_{m,j,i}^{(1)}| \leq R \) for any \( m \in [M], j \in [T], i \in [p] \) and \( \ell \in \{1, 2\} \), we have that \( \| \mathbf{E}_{m,j} \| \leq \frac{1}{\sqrt{T}} \cdot 2pR L_\phi L_\theta B_\phi^2 \) and \( \| \sum_{m=1}^M \sum_{j=1}^T \mathbf{E}_{m,j} \|^2 \leq \frac{1}{MT} \cdot 4p^2 R^2 L_\phi^2 L_\theta^2 B_\phi^4 \) for any \( m \in [M] \) and \( j \in [T] \).

Since \( \tilde{\eta}_{m,j,i}^{(1)} \) is 1-sub-Gaussian for any \( m \in [M], j \in [T], i \in [p] \) and \( \ell \in \{1, 2\} \), using a union bound over \( i \in [p] \) and \( \ell \in \{1, 2\} \), we have that for any \( m \in [M] \) and \( j \in [T] \), with probability at least \( 1 - 4p \exp(-\frac{R^2}{2}) \), \( \| \mathbf{E}_{m,j} \| \leq \frac{1}{\sqrt{T}} \cdot 2pR L_\phi L_\theta B_\phi^2 \). Then, we have

\[
\| \mathbf{E}_{m,j} - \mathbf{E}[\mathbf{E}_{m,j}] \| \\
\leq \mathbb{E} \left[ \left| \mathbb{E}_{m,j} \cdot 1 \left\{ \| \mathbf{E}_{m,j} \| \geq \frac{2pR L_\phi L_\theta B_\phi^2 \sqrt{MT}}{\sqrt{T}} \right\} \right\} \right] \\
\leq \mathbb{E} \left[ \left| \mathbb{E}_{m,j} \cdot 1 \left\{ \| \mathbf{E}_{m,j} \| \geq \frac{2pR L_\phi L_\theta B_\phi^2 \sqrt{MT}}{\sqrt{T}} \right\} \right\} \right] \\
= \frac{2pR L_\phi L_\theta B_\phi^2 \sqrt{MT}}{\sqrt{T}} \cdot \Pr \left[ \| \mathbf{E}_{m,j} \| \geq \frac{2pR L_\phi L_\theta B_\phi^2 \sqrt{MT}}{\sqrt{T}} \right] \\
+ \int_0^\infty \Pr \left[ \| \mathbf{E}_{m,j} \| - \frac{2pR L_\phi L_\theta B_\phi^2 \sqrt{MT}}{\sqrt{T}} > x \right] \, dx
\]
\[
\begin{align*}
\frac{2pRL_0B_{\Phi}^2}{MT} \cdot 4p \cdot \exp \left( -\frac{R^2}{2} \right) &+ \frac{2pL_0L_0B_{\Phi}^2}{MT} \int_R^{\infty} \Pr \left[ \|E_{m,j}\| > \frac{2pL_0L_0B_{\Phi}^2y}{MT} \right] dy \\
\frac{2pRL_0B_{\Phi}^2}{MT} \cdot 4p \cdot \exp \left( -\frac{R^2}{2} \right) &+ \frac{2pL_0L_0B_{\Phi}^2}{MT} \int_R^{\infty} 4p \exp \left( -\frac{y^2}{2} \right) dy \\
&\leq \frac{2pL_0L_0B_{\Phi}^2}{MT} \cdot 4p \cdot \exp \left( -\frac{R^2}{2} \right) \\
\frac{2pL_0L_0B_{\Phi}^2}{MT} \cdot 4p \cdot \exp \left( -\frac{R^2}{2} \right) &+ \frac{2pL_0L_0B_{\Phi}^2}{MT} \cdot 4p \cdot \frac{1}{R} \cdot \exp \left( -\frac{R^2}{2} \right) \\
&\leq \frac{2pL_0L_0B_{\Phi}^2}{MT} \cdot 4p \cdot \left( R + \frac{1}{R} \right) \exp \left( -\frac{R^2}{2} \right).
\end{align*}
\]

Using the truncated matrix Bernstein inequality (Lemma E.2) with \( n = MT, R = \sqrt{2 \log \left( \frac{4pMT}{\delta'} \right)} \),

\[
U = \frac{2pL_0L_0B_{\Phi}^2}{MT} \sqrt{2 \log \left( \frac{4pMT}{\delta'} \right)}, \quad \sigma^2 = \frac{2pL_0L_0B_{\Phi}^2}{MT} \sqrt{2 \log \left( \frac{4pMT}{\delta'} \right)}, \quad \tau = \frac{4 \sqrt{2pL_0L_0B_{\Phi}^2 \left( 2 \log \left( \frac{4pMT}{\delta'} \right) \right)}}{MT}, \quad \Delta = \frac{2pL_0L_0B_{\Phi}^2}{MT} \cdot \frac{2 \left( 2 \log \left( \frac{4pMT}{\delta'} \right) \right)}{\delta' MT}, \quad \text{we have that with probability at least } 1 - 2\delta',
\]

\[
\left\| E - E[E] \right\| \leq \frac{8 \cdot 2pL_0L_0B_{\Phi}^2}{\sqrt{MT}} \left( \frac{2 \log \left( \frac{4pMT}{\delta'} \right)}{\delta'} \right) \cdot \log \left( \frac{2\delta'}{\delta'} \right).
\] (27)

Now we investigate \( \| F - E[F] \| \). Since \( |z_{m,j}^{(t)}| \leq R \) for any \( m \in [M], j \in [T], i \in [p] \) and \( \ell \in \{1, 2\} \), we have that

\[
\| F_{m,j} \| \leq \frac{1}{M} \cdot pR^2 B_{\Phi}^2 \text{ and } \| \sum_{m=1}^{M} \sum_{j=1}^{T} E[F_{m,j}] \| \leq \frac{1}{MT} \cdot p^2 R^4 B_{\Phi}^2.
\]

Recall that for any \( m \in [M] \) and \( j \in [T] \), with probability at least \( 1 - 4p \exp \left( -\frac{R^2}{2} \right) \), \( |z_{m,j}^{(t)}| \leq R \) for all \( i \in [p] \) and \( \ell \in \{1, 2\} \), and thus, \( \| F_{m,j} \| \leq \frac{1}{MT} \cdot p^2 R^4 \). Then, we have

\[
\left\| E[F_{m,j}] - E[F_{m,j}] \right\| \leq \left\| E \left[ F_{m,j}, 1 \left\{ \| F_{m,j} \| \geq \frac{pB_{\Phi}^2 R^2}{MT} \right\} \right] \right\|
\leq E \left[ \left\| F_{m,j} \right\| \cdot 1 \left\{ \| F_{m,j} \| \geq \frac{pB_{\Phi}^2 R^2}{MT} \right\} \right]
\leq E \left[ \frac{pB_{\Phi}^2 R^2}{MT} \cdot 1 \left\{ \| F_{m,j} \| \geq \frac{pB_{\Phi}^2 R^2}{MT} \right\} \right] + \left[ \left\| F_{m,j} \right\| - \frac{pB_{\Phi}^2 R^2}{MT} \right] \cdot 1 \left\{ \| F_{m,j} \| \geq \frac{pB_{\Phi}^2 R^2}{MT} \right\}
\leq \frac{pB_{\Phi}^2 R^2}{MT} \cdot 4p \cdot \exp \left( -\frac{R^2}{2} \right) + \frac{2pB_{\Phi}^2}{MT} \int_R^{\infty} \Pr \left[ \| F_{m,j} \| > \frac{pB_{\Phi}^2 y^2}{MT} \right] dy
\leq \frac{pB_{\Phi}^2 R^2}{MT} \cdot 4p \cdot \exp \left( -\frac{R^2}{2} \right) + \frac{2pB_{\Phi}^2}{MT} \int_R^{\infty} y \cdot 4p \exp \left( -\frac{y^2}{2} \right) dy
\leq \frac{pB_{\Phi}^2 R^2}{MT} \cdot 4p \cdot \exp \left( -\frac{R^2}{2} \right) + \frac{2pB_{\Phi}^2}{MT} \cdot 4p \cdot \exp \left( -\frac{R^2}{2} \right)
\leq \frac{pB_{\Phi}^2}{MT} \cdot 4p \cdot \left( R^2 + 2 \right) \exp \left( -\frac{R^2}{2} \right).
\]

Using the truncated matrix Bernstein inequality (Lemma E.2) with \( n = MT, R = \sqrt{2 \log \left( \frac{4pMT}{\delta'} \right)} \),

\[
\sigma^2 = \left( \frac{pB_{\Phi}^2 \cdot 2 \log \left( \frac{4pMT}{\delta'} \right)}{MT} \right)^2, \quad \tau = \frac{4 \sqrt{pB_{\Phi}^2 \cdot 2 \log \left( \frac{4pMT}{\delta'} \right) \cdot \log \left( \frac{2\delta'}{\delta'} \right)}}{MT}, \quad \Delta = \frac{pB_{\Phi}^2 \cdot 2 \log \left( \frac{4pMT}{\delta'} \right)}{MT}, \quad \delta' MT,
\]

we have that with probability at least \( 1 - 2\delta' \),

\[
\left\| F - E[F] \right\| \leq \frac{8 \cdot pB_{\Phi}^2 \cdot 2 \log \left( \frac{4pMT}{\delta'} \right) \cdot \log \left( \frac{2\delta'}{\delta'} \right)}{\sqrt{MT}}.
\] (28)
Plugging Eqs. (26)-(28) into Eq. (25), we have that with probability at least $1 - 5\delta'$,
\[
\|Z - \mathbb{E}[Z]\| \leq \|D - \mathbb{E}[D]\| + \|E - \mathbb{E}[E]\| + \|F - \mathbb{E}[F]\| \\
\leq \frac{64pL\sigma L_B^2 \log \left( \frac{4pMT}{\delta'} \right) \log \left( \frac{2d}{\delta} \right)}{\sqrt{MT}}.
\]

Let $\delta' = \frac{\delta}{25}$. Recall that $B_\psi := 2\sqrt{(1 + \zeta)\nu\sigma^2}$. Then, we obtain that with probability at least $1 - \frac{\delta}{5}$,
\[
\|Z - \mathbb{E}[Z]\| \leq \frac{256(1 + \zeta)L\sigma L_\theta \log \left( \frac{50d}{\delta} \right)}{\nu \sqrt{MT}} \log \left( \frac{100pMT}{\delta} \right),
\]
which implies that $\Pr[\mathcal{G}] \geq 1 - \frac{\delta}{5}$.

According to Assumption 3.1, there exists an absolute constant $c_0$ which satisfies that $\sigma_{\min}(\frac{1}{M} \sum_{m=1}^{M} \theta_m \theta_m^\top) = \sigma_{\min}(\frac{1}{M} \sum_{m=1}^{M} \theta_m \theta_m^\top) \geq \frac{c_0}{\kappa}$. 

**Lemma D.7** (Concentration of $\hat{B}$). Suppose that event $\mathcal{G}$ holds. Then,
\[
\left\| \hat{B}_{t+1} \right\| \leq \frac{2048(1 + \zeta)kL\sigma L_\theta \log \left( \frac{50d}{\delta} \right)}{c_0 \nu \sqrt{MT}} \log \left( \frac{135(1 + \zeta)dL\sigma MT}{\nu \delta} \right).
\]

Furthermore, if
\[
T = \left[ \frac{68 \cdot 2048^2 \cdot 96^2(1 + \zeta)^2k^4L^4\sigma^2 L_\sigma^2 L_\theta^2}{c_0^2 \nu^2 \varepsilon^2 M^4} \right] \log \delta \left( \frac{2048 \cdot 135 \cdot 96 \cdot 50 \cdot 5(1 + \zeta)^2k^2d^2L^4\sigma L_\sigma L_\theta N}{c_0 \nu^2 \delta^3 \varepsilon} \right),
\]
we have
\[
\left\| \hat{B}_{t+1} \right\| \leq \frac{\varepsilon}{96k \log \left( \frac{52}{\delta} \right) L_\sigma L_\theta}.
\]

**Proof of Lemma D.7.** First, we have that $\sigma_k(\mathbb{E}[Z]) - \sigma_{k+1}(\mathbb{E}[Z]) = \sigma_{\min}(\frac{1}{M} \sum_{m=1}^{M} \theta_m \theta_m^\top) \geq \frac{c_0}{\kappa}$. Let $p := \lceil 32^2(1 + \zeta)^2\nu^{-2}L^4\sigma^2 \log^2 \left( \frac{40dMT}{\delta} \right) \rceil$. Then, using the Davis-Kahan sin $\theta$ Theorem (Bhatia, 2013) and letting $T$ be large enough to satisfy that $\|Z - \mathbb{E}[Z]\| \leq \frac{\varepsilon}{2k}$, we have
\[
\left\| \hat{B}_{t+1} \right\| \leq \left\| \sigma_k(\mathbb{E}[Z]) - \sigma_{k+1}(\mathbb{E}[Z]) - \|Z - \mathbb{E}[Z]\| \right\|
\leq \frac{2k}{c_0} \|Z - \mathbb{E}[Z]\| \\
\leq \frac{512(1 + \zeta)kL\sigma L_\theta \log \left( \frac{50d}{\delta} \right)}{c_0 \nu \sqrt{MT}} \log \left( \frac{100pMT}{\delta} \right) \\
\leq \frac{512(1 + \zeta)kL\sigma L_\theta \log \left( \frac{50d}{\delta} \right)}{c_0 \nu \sqrt{MT}} \log \left( \frac{100MT \cdot 2 \cdot 32^2(1 + \zeta)^2L^4\sigma \delta^3 \log^2 (40dMT \delta) \right) \\
\leq \frac{512(1 + \zeta)kL\sigma L_\theta \log \left( \frac{50d}{\delta} \right)}{c_0 \nu \sqrt{MT}} \log \left( \frac{2 \cdot 100 \cdot 32^2 \cdot 40^2(1 + \zeta)^2d^2L^4\sigma M^3T^3}{\nu^2 \delta^3} \right) \\
\leq \frac{2048(1 + \zeta)kL\sigma L_\theta \log \left( \frac{50d}{\delta} \right)}{c_0 \nu \sqrt{MT}} \log \left( \frac{135(1 + \zeta)dL\sigma MT}{\nu \delta} \right).
\]
Using Lemma E.3 with $A = 2048(1 + \zeta)k\varepsilon_{\min}^{-1}L_\psi^{-1}L_\sigma L_\theta \log \left( \frac{50d}{\delta} \right)$, $B = \frac{135(1 + \zeta)dL\sigma}{\nu \delta}$ and $\kappa = \frac{2048 \log \left( \frac{50}{\delta} \right)L_\sigma L_\theta}{96k}$, we have that if
\[
MT \geq \frac{68 \cdot 2048^2 \cdot 96^2(1 + \zeta)^2k^4L^4\sigma^2 L_\sigma^2 L_\theta^2 L_\psi^2}{c_0^2 \nu^2 \varepsilon^2},
\]

then \[ \| \hat{B}_{t, \perp} \| \leq \frac{\varepsilon}{96k \log \left( \frac{5N}{\delta} \right) L_\phi L_w}. \]

Further enlarging \( MT \), if
\[
MT \geq 68 \cdot 2048^2 \cdot 96^2 (1 + \zeta)^4 L_\phi^4 L_w^2 \log^6 \left( \frac{2048 \cdot 135 \cdot 96 \cdot 50 \cdot 5 (1 + \zeta)^2 d^2 L_\phi^3 L_w N}{c_0 \nu^2 \delta^3 \varepsilon} \right),
\]
then
\[
\| \hat{B}_{t, \perp} \| \leq \frac{\varepsilon}{96k \log \left( \frac{5N}{\delta} \right) L_\phi L_w}.
\]

D.3. Estimation with Low-dimensional Representations

**Lemma D.8.** In subroutine \( \text{EstLowRep} \) (Algorithm 6), for any \( m \in [M] \) and \( t > 0 \), we have
\[
\log \left( \frac{\det \left( \gamma I + \sum_{\tau=1}^t \hat{B}^\top \phi(s_{m, \tau}, a_{m, \tau}) \phi(s_{m, \tau}, a_{m, \tau})^\top \hat{B} \right)}{\det (\gamma I)} \right) \leq k \log \left( 1 + \frac{t}{\gamma k} \right).
\]

**Proof of Lemma D.8.** This proof uses a similar idea as Lemma 11 in (Abbasi-Yadkori et al., 2011). It holds that
\[
\log \left( \frac{\det \left( \gamma I + \sum_{\tau=1}^t \hat{B}^\top \phi(s_{m, \tau}, a_{m, \tau}) \phi(s_{m, \tau}, a_{m, \tau})^\top \hat{B} \right)}{\det (\gamma I)} \right) \leq \log \left( \frac{\text{Trace} \left( \gamma I + \sum_{\tau=1}^t \hat{B}^\top \phi(s_{m, \tau}, a_{m, \tau}) \phi(s_{m, \tau}, a_{m, \tau})^\top \hat{B} \right)^k}{\gamma^k} \right)
\]
\[
= k \log \left( \frac{\text{Trace} \left( \gamma I + \sum_{\tau=1}^t \hat{B}^\top \phi(s_{m, \tau}, a_{m, \tau}) \phi(s_{m, \tau}, a_{m, \tau})^\top \hat{B} \right)}{\gamma^k} \right) \]
\[
= k \log \left( \frac{\gamma^k + \sum_{\tau=1}^t \| \hat{B}^\top \phi(s_{m, \tau}, a_{m, \tau}) \|^2}{\gamma^k} \right) \]
\[
\leq k \log \left( 1 + \frac{t}{\gamma k} \right).
\]

**Lemma D.9.** In subroutine \( \text{EstLowRep} \) (Algorithm 6), for any \( m \in [M] \) and \( t \geq 0 \), we have
\[
\mathbb{E}_{s \sim D} \left[ \max_{a \in A} \| \hat{B}^\top \phi(s, a) \|_{\Sigma_{m,t}^{-1}} \right] \geq \mathbb{E}_{s \sim D} \left[ \max_{a \in A} \| \hat{B}^\top \phi(s, a) \|_{\Sigma_{m,t+1}^{-1}} \right].
\]

**Proof of Lemma D.9.** This proof is similar to that of Lemma 6 in (Zanette et al., 2021).
For any $m \in [M]$ and $t \geq 0$, since $\Sigma_{m,t+1} \succeq \Sigma_{m,t}$, we have $\Sigma_{m,t+1}^{-1} \succeq \Sigma_{m,t}^{-1}$. Hence, for any $m \in [M]$, $t \geq 0$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have

$$\phi(s,a)\top \hat{B} \Sigma_{m,t}^{-1} \hat{B} \phi(s,a) \geq \phi(s,a)\top B \Sigma_{m,t+1}^{-1} B \phi(s,a),$$

which implies that

$$\left\| \hat{B} \phi(s,a) \right\|_{\Sigma_{m,t}^{-1}} \geq \left\| B \phi(s,a) \right\|_{\Sigma_{m,t+1}^{-1}} .$$

Therefore, for any $m \in [M]$ and $t \geq 0$, we have

$$\mathbb{E}_{s \sim D} \left[ \max_{a \in \mathcal{A}} \left\| \hat{B} \phi(s,a) \right\|_{\Sigma_{m,t}^{-1}} \right] \geq \mathbb{E}_{s \sim D} \left[ \max_{a \in \mathcal{A}} \left\| B \phi(s,a) \right\|_{\Sigma_{m,t+1}^{-1}} \right] .$$

In subroutine $\text{EstLowRep}$, for any $m \in [M]$ and $t > 0$, let $\xi_{m,t}$ denote the noise of the sample at timestep $t$ for task $m$ (Line 6 in Algorithm 6).

Define event

$$\mathcal{H} := \left\{ \left\| \sum_{\tau=1}^{t} \hat{B} \phi(s_{m,\tau}, a_{m,\tau}) \xi_{m,\tau} \right\| \leq k \log \left( 1 + \frac{t}{\gamma k} \right) + 2 \log \left( \frac{5}{\delta} \right) , \forall m \in [M], \forall t > 0 \right\} .$$

**Lemma D.10** (Martingale Concentration of the Variance Term). It holds that

$$\Pr \left[ \mathcal{H} \right] \geq 1 - \frac{\delta}{5} .$$

**Proof of Lemma D.10.** Let $\delta'$ be a confidence parameter which will be chosen later. Since $\hat{B}$ is fixed before sampling $(s_{m,\tau}, a_{m,\tau})$ for all $m \in [M]$ and $\tau > 0$, using Lemma E.7, we have that with probability at least $1 - \delta'$, for any task $m \in [M]$ and $t > 0$,

$$\left\| \sum_{\tau=1}^{t} \hat{B} \phi(s_{m,\tau}, a_{m,\tau}) \xi_{m,j} \right\| \leq 2 \log \left( \frac{\det \left( \gamma I + \sum_{\tau=1}^{t} \hat{B} \phi(s_{m,\tau}, a_{m,\tau}) \phi(s_{m,\tau}, a_{m,\tau})\top \hat{B} \right)^{-1}}{\det (\gamma I)^{\frac{1}{2}} \cdot \delta'} \right) \leq \log \left( \frac{\det \left( \gamma I + \sum_{\tau=1}^{t} \hat{B} \phi(s_{m,\tau}, a_{m,\tau}) \phi(s_{m,\tau}, a_{m,\tau})\top \hat{B} \right)}{\det (\gamma I)^{\frac{1}{2}} \cdot \delta'} \right) \leq k \log \left( 1 + \frac{t}{\gamma k} \right) + 2 \log \left( \frac{1}{\delta'} \right) ,$$

where inequality (a) uses Lemma D.8.

Letting $\delta' = \frac{\delta}{5}$, we obtain this lemma.

Define event

$$\mathcal{J} := \left\{ \sum_{t=1}^{N} \mathbb{E}_{s \sim D} \left[ \max_{a \in \mathcal{A}} \left\| \hat{B} \phi(s,a) \right\|_{\Sigma_{m,t}^{-1}} \right] \leq$$
Lemma D.11. It holds that
\[ \Pr [J] \geq 1 - \frac{\delta}{5}. \]

Proof of Lemma D.11. Using Lemma E.8, we can obtain this lemma.

Lemma D.12. Suppose that event \( K \cap L \cap G \cap H \cap J \) holds. For any task \( m \in [M] \), we have
\[
\mathbb{E}_{s \sim D} \left[ \max_{a \in A} \left| \phi(s, a)^\top \left( \hat{\theta}_{m,N} - \theta_m \right) \right| \right] \leq \left( 2 \sqrt{\frac{2k \log \left( 1 + \frac{N}{\gamma k} \right)}{N}} + 8 \log \left( \frac{\gamma}{N} \right) \right) \cdot \left( \| \hat{B}_\perp B \| \sqrt{Nk} + k \log \left( 1 + \frac{N}{\gamma k} \right) + 2 \log \left( \frac{\gamma}{N} \right) + \sqrt{\gamma} \right) + \| \hat{B}_\perp B \|.
\]
Furthermore, if
\[
N = \left\lceil \frac{4^2 \cdot 26^4 \cdot 24^2 \cdot 2 (k^2 + k\gamma L_2^2) \log^4 \left( \frac{240(k + \sqrt{k\gamma L_2})}{\epsilon^2} \right)}{\delta^2} \right\rceil,
\]
then
\[ \mathbb{E}_{s \sim D} \left[ \max_{a \in A} \left| \phi(s, a)^\top \left( \hat{\theta}_{m,N} - \theta_m \right) \right| \right] \leq \frac{\epsilon}{2}. \]

Proof of Lemma D.12. For any task \( m \in [M] \) and \( t \in [N] \),
\[
\hat{w}_{m,t} = \sum_{m,j=1}^{t-1} \hat{B}_\perp^\top \phi(s_{m,j}, a_{m,j}) \tau_{m,j}
\]
\[
= \left( \gamma I + \sum_{\tau=1}^{t} \hat{B}_\perp^\top \phi(s_{m,\tau}, a_{m,\tau}) \phi(s_{m,\tau}, a_{m,\tau})^\top \hat{B} \right)^{-1} \sum_{\tau=1}^{t} \hat{B}_\perp^\top \phi(s_{m,\tau}, a_{m,\tau}) \phi(s_{m,\tau}, a_{m,\tau})^\top \theta_m + \xi_{m,j}
\]
\[
= \left( \gamma I + \sum_{\tau=1}^{t} \hat{B}_\perp^\top \phi(s_{m,\tau}, a_{m,\tau}) \phi(s_{m,\tau}, a_{m,\tau})^\top \hat{B} \right)^{-1} \sum_{\tau=1}^{t} \hat{B}_\perp^\top \phi(s_{m,\tau}, a_{m,\tau}) \phi(s_{m,\tau}, a_{m,\tau})^\top \theta_m + \xi_{m,j}
\]
\[
+ \gamma \left( \gamma I + \sum_{\tau=1}^{t} \hat{B}_\perp^\top \phi(s_{m,\tau}, a_{m,\tau}) \phi(s_{m,\tau}, a_{m,\tau})^\top \hat{B} \right)^{-1} \hat{B}_\perp^\top \theta_m
\]
\[
- \gamma \left( \gamma I + \sum_{\tau=1}^{t} \hat{B}_\perp^\top \phi(s_{m,\tau}, a_{m,\tau}) \phi(s_{m,\tau}, a_{m,\tau})^\top \hat{B} \right)^{-1} \hat{B}_\perp^\top \theta_m
\]
\[
= \hat{B}_\perp^\top \theta_m + \left( \gamma I + \sum_{\tau=1}^{t} \hat{B}_\perp^\top \phi(s_{m,\tau}, a_{m,\tau}) \phi(s_{m,\tau}, a_{m,\tau})^\top \hat{B} \right)^{-1} \hat{B}_\perp^\top \theta_m
\]
\[
= \sum_{\tau=1}^{t} \hat{B}_\perp^\top \phi(s_{m,\tau}, a_{m,\tau}) \phi(s_{m,\tau}, a_{m,\tau})^\top \hat{B}_\perp^\top \hat{B}_\perp \hat{B} \hat{w}_m
\]

45
Taking the absolute value on both sides and using the Cauchy–Schwarz inequality, we obtain that for any \( \phi \in \mathcal{H} \),
\[
\xi_{m,j} = \sum_{\tau=1}^{t} \phi(s_m, a_m, \tau) \phi(s_m, a_m, \tau)^\top B - \sum_{\tau=1}^{t} \phi(s_m, a_m, \tau) \xi_{m,j}.
\]
Hence, for any task \( m \in [M] \) and \( s, a \) \( \in \mathcal{S} \times \mathcal{A} \),
\[
\phi(s, a)^\top \left( \hat{\theta}_{m,t} - \theta_m \right) = \phi(s, a)^\top \hat{B}w_{m,t} - \phi(s, a)^\top \left( \hat{B}B^\top + \hat{B} \hat{B}^\top \right) \theta_m
\]
\[
= \phi(s, a)^\top \hat{B} \left( \hat{w}_{m,t} - \hat{B}^\top \theta_m \right) - \phi(s, a)^\top \hat{B} \hat{B}^\top \theta_m
\]
\[
= \phi(s, a)^\top \hat{B} \left( \gamma I + \sum_{\tau=1}^{t} \hat{B}^\top \phi(s_m, a_m, \tau) \phi(s_m, a_m, \tau)^\top \hat{B} \right)^{-1} \sum_{\tau=1}^{t} \hat{B}^\top \phi(s_m, a_m, \tau) \phi(s_m, a_m, \tau)^\top \hat{B} \xi_{m,j}
\]
\[
+ \gamma \phi(s, a)^\top \hat{B} \left( \gamma I + \sum_{\tau=1}^{t} \hat{B}^\top \phi(s_m, a_m, \tau) \phi(s_m, a_m, \tau)^\top \hat{B} \right)^{-1} \sum_{\tau=1}^{t} \hat{B}^\top \phi(s_m, a_m, \tau) \xi_{m,j}
\]
\[
- \gamma \phi(s, a)^\top \hat{B} \left( \gamma I + \sum_{\tau=1}^{t} \hat{B}^\top \phi(s_m, a_m, \tau) \phi(s_m, a_m, \tau)^\top \hat{B} \right)^{-1} \hat{B}^\top \theta_m - \phi(s, a)^\top \hat{B} \hat{B}^\top Bw_m.
\]
For any \( m \in [M] \), let \( \Sigma_{m,0} := \gamma I \). For any \( m \in [M] \) and \( t \geq 1 \), let \( \Sigma_{m,t} := \gamma I + \sum_{\tau=1}^{t} \hat{B}^\top \phi(s_m, a_m, \tau) \phi(s_m, a_m, \tau)^\top \hat{B} \).

Taking the absolute value on both sides and using the Cauchy–Schwarz inequality, we obtain that for any \( m \in [M] \) and \( t \in [N] \),
\[
\left| \phi(s, a)^\top \left( \hat{\theta}_{m,t} - \theta_m \right) \right| \leq \left\| \hat{B}^\top \phi(s, a) \right\|_{\Sigma_{m,t}^{-1}} \left\| \sum_{\tau=1}^{t} \hat{B}^\top \phi(s_m, a_m, \tau) \phi(s_m, a_m, \tau)^\top \hat{B} \hat{B}^\top Bw_m \right\|_{\Sigma_{m,t}^{-1}}
\]
\[
+ \left\| \hat{B}^\top \phi(s, a) \right\|_{\Sigma_{m,t}^{-1}} \left\| \sum_{\tau=1}^{t} \hat{B}^\top \phi(s_m, a_m, \tau) \xi_{m,j} \right\|_{\Sigma_{m,t}^{-1}}
\]
\[
+ \gamma \left\| \hat{B}^\top \phi(s, a) \right\|_{\Sigma_{m,t}^{-1}} \left\| \hat{B}^\top \theta_m \right\|_{\Sigma_{m,t}}
\]
\[
+ \phi(s, a)^\top \hat{B} \hat{B}^\top Bw_m \right\|_{\Sigma_{m,t}} \leq \left\| \hat{B}^\top \phi(s, a) \right\|_{\Sigma_{m,t}^{-1}} \sum_{\tau=1}^{t} \left| \phi(s_m, a_m, \tau)^\top \hat{B} \hat{B}^\top Bw_m \right| \cdot \left\| \hat{B}^\top \phi(s_m, a_m, \tau) \right\|_{\Sigma_{m,t}^{-1}}
\]
\[
+ \left\| \hat{B}^\top \phi(s, a) \right\|_{\Sigma_{m,t}^{-1}} \sqrt{k \log \left( 1 + \frac{t}{\gamma k} \right) + 2 \log \left( \frac{5}{\delta} \right)}
\]
\[
+ \gamma \left\| \hat{B}^\top \phi(s, a) \right\|_{\Sigma_{m,t}^{-1}} \cdot \frac{1}{\sqrt{\gamma}} \cdot \left\| \hat{B}^\top \theta_m \right\| + \left\| \hat{B}^\top \hat{B} \right\| L_\phi L_w
\]
\[
\leq \left\| \hat{B}^\top \phi(s, a) \right\|_{\Sigma_{m,t}^{-1}} \cdot \left\| \hat{B} \right\| L_\phi L_w \cdot \sum_{\tau=1}^{t} \left\| \hat{B}^\top \phi(s_m, a_m, \tau) \right\|_{\Sigma_{m,t}^{-1}}.
\]
where inequality (a) is due to the definition of event $\mathcal{J}$, and inequality (b) is due to Lemma E.6.

Taking the maximum over $a \in A$ and taking the expectation on $s \sim D$, we have that for any task $m \in [M],$

$$
\mathbb{E}_{s \sim D} \left[ \max_{a \in A} \left( \phi(s, a) \right) \right] \leq \mathbb{E}_{s \sim D} \left[ \max_{a \in A} \left( \hat{B}^\top \phi(s, a) \right) \right].
$$

According to Lemma D.9, $\mathbb{E}_{s \sim D} \left[ \max_{a \in A} \left( \hat{B}^\top \phi(s, a) \right) \right]$ is non-increasing with respect to $t$. Hence, we have

$$
\mathbb{E}_{s \sim D} \left[ \max_{a \in A} \left( \hat{B}^\top \phi(s, a) \right) \right] \leq \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{s \sim D} \left[ \max_{a \in A} \left( \hat{B}^\top \phi(s, a) \right) \right].
$$

$$
\leq \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{s \sim D} \left[ \max_{a \in A} \left( \hat{B}^\top \phi(s, a) \right) \right].
$$

$$
\leq \frac{1}{4N} \left( 2 \sqrt{\log \frac{5}{\delta}} \right)^2 + \sqrt{4 \log \frac{5}{\delta} + 4 \left( \sum_{i=1}^{N} \max_{a \in A} \left( \hat{B}^\top \phi(s_t, a) \right) \right)^2}
$$

$$
= \frac{1}{4N} \left( 2 \sqrt{\log \frac{5}{\delta}} \right)^2 + \sqrt{4 \log \frac{5}{\delta} + 4 \left( \sum_{i=1}^{N} \max_{a \in A} \left( \hat{B}^\top \phi(s_t, a) \right) \right)^2},
$$

where inequality (a) is due to the definition of event $\mathcal{J}$.

In addition, we have

$$
\sum_{i=1}^{N} \left| \hat{B}^\top \phi(s_t, a_t) \right| \leq \sqrt{N} \sum_{i=1}^{N} \left( \hat{B}^\top \phi(s_t, a_t) \right)^2
$$
Combining Eqs. (31) and (32), we have

\[
E_{s \sim D} \left[ \max_{a \in A} \left\| \hat{B}^T \phi(s, a) \right\|_{\Sigma^{-1}} \right] \leq \frac{1}{4N} \left( 2 \sqrt{2 \log \left( \frac{5}{\delta} \right)} + 4 \log \left( \frac{5}{\delta} \right) + 4 \left( \sqrt{2Nk \log \left( 1 + \frac{N}{\gamma k} \right)} + 2 \log \left( \frac{5}{\delta} \right) \right) \right)^2
\]

\[
\leq \frac{1}{2N} \left( 4 \log \left( \frac{5}{\delta} \right) + 4 \log \left( \frac{5}{\delta} \right) + 4 \left( \sqrt{2Nk \log \left( 1 + \frac{N}{\gamma k} \right)} + 2 \log \left( \frac{5}{\delta} \right) \right) \right)
\]

\[
= \frac{1}{N} \left( 2 \sqrt{2Nk \log \left( 1 + \frac{N}{\gamma k} \right)} + 8 \log \left( \frac{5}{\delta} \right) \right)
\]

\[
= 2 \sqrt{\frac{2k \log \left( 1 + \frac{N}{\gamma k} \right)}{N}} + 8 \log \left( \frac{5}{\delta} \right),
\]

where inequality (a) uses Lemma E.9, and inequality (b) is due to Lemma D.8.

Furthermore, plugging Eq. (33) into Eq. (30) and using \( \gamma \geq 1 \), we have that for \( N \geq 1 \) and \( \sqrt{k} \log(2N) \geq 1 \),

\[
E_{s \sim D} \left[ \max_{a \in A} \phi(s, a)^T \left( \hat{\theta}_{m,N} - \theta_m \right) \right] \leq \left( 2 \sqrt{\frac{2k \log \left( 1 + \frac{N}{\gamma k} \right)}{N}} + 8 \log \left( \frac{5}{\delta} \right) \right).
\]

\[
\leq 12 \sqrt{k} \log \left( \frac{5N}{\delta} \right) \left( \left\| \hat{B}^T \right\|_{L_\phi L_w \sqrt{N}} + k \log \left( 1 + \frac{N}{\gamma k} \right) + 2 \log \left( \frac{5}{\delta} \right) + \sqrt{k} L_{\theta} \right) + \left\| \hat{B}^T \right\|_{L_\phi L_w}
\]

\[
\leq \frac{(24k + 12 \sqrt{k} L_{\theta}) \log^2 \left( \frac{5N}{\delta} \right)}{\sqrt{N}} + 24k \log \left( \frac{5N}{\delta} \right) \left\| \hat{B}^T \right\|_{L_\phi L_w}.
\]

Using Lemma E.4 with \( A = 24k + 12 \sqrt{k} L_{\theta} \), \( B = \frac{5}{\delta} \) and \( \kappa = \frac{\xi}{4} \), we have that if

\[
N \geq \frac{26^4 \left( 24k + 12 \sqrt{k} L_{\theta} \right)^2 \log^4 \left( \frac{2.5 \left( 24k + 12 \sqrt{k} L_{\theta} \right)}{\varepsilon \delta} \right)}{\left( \frac{\xi}{4} \right)^2},
\]

then \( \frac{(24k + 12 \sqrt{k} L_{\theta}) \log^2 \left( \frac{5N}{\delta} \right)}{\sqrt{N}} \leq \frac{\xi}{4} \).

Further enlarging \( N \), if

\[
N \geq \frac{4^2 \cdot 26^4 \cdot 24^2 \cdot 2 \left( k^2 + k \sqrt{\gamma L_{\theta}} \right) \log^4 \left( \frac{240 \left( k + \sqrt{\gamma L_{\theta}} \right)}{\varepsilon \delta} \right)}{\varepsilon^2},
\]

\[
(36)
\]
then
\[
\frac{(24k + 12\sqrt{k}L\theta) \log^2 \left( \frac{5N}{\delta} \right)}{\sqrt{N}} \leq \frac{\varepsilon}{4}.
\]

According to Lemma D.7, we have
\[
\left\| \hat{B}_t^T B \right\| \leq \frac{96k \log \left( \frac{5N}{\delta} \right)L\theta Lw}{\varepsilon}.
\]

Thus, setting \( N \) as the value in Eq. (36), and continuing with Eq. (35), we have
\[
\mathbb{E}_{s \sim D} \left[ \max_{a \in A} \phi(s, a)^\top \left( \hat{\theta}_{m,N} - \theta_m \right) \right] \leq \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}.
\]

D.4. Proof of Theorem 5.1

Proof of Theorem 5.1. Combining Lemmas D.2, D.3, D.6, D.10 and D.11, we have that
\[
\Pr \left[ K \cap L \cap G \cap H \cap J \right] \geq 1 - \delta.
\]

Suppose that event \( K \cap L \cap G \cap H \cap J \) holds.

First, we use a similar analytical procedure as that in (Zanette et al., 2021) to prove the correctness.

Using Lemma D.12, we have that for any task \( m \in [M] \),
\[
\mathbb{E}_{s \sim D} \left[ \max_{a \in A} \phi(s, a)^\top \left( \hat{\theta}_{m,N} - \theta_m \right) \right] \leq \frac{\varepsilon}{2}.
\]

For any \( m \in [M] \) and \( s \in \mathcal{S} \), let \( \beta_m(s) := \max_{a \in A} \phi(s, a)^\top (\hat{\theta}_{m,N} - \theta_m) \) and \( \pi_m^*(s) := \arg\max_{a \in A} \phi(s, a)^\top \theta_m \).

For any \( m \in [M] \) and \( s \in \mathcal{S} \), we have
\[
\phi(s, \pi_m(s))^\top \theta_m \geq \phi(s, \pi_m(s))^\top (\hat{\theta}_{m,N} - \beta_m(s)) \geq (a) \frac{\phi(s, \pi_m^*(s))^\top \beta_m(s)}{\varepsilon} \geq \phi(s, \pi_m^*(s))^\top (\theta_m - 2\beta_m(s)),
\]

where inequality (a) is due to that \( \pi_m(s) \) is greedy with respect to \( \hat{\theta}_{m,N} \).

Rearranging the above equation and taking the expectation of \( s \) on both sides, we have
\[
\mathbb{E}_{s \sim D} \left[ \max_{a \in A} (\phi(s, a) - \phi(s, \pi_m(s)))^\top \theta_m \right] \leq 2\mathbb{E}_{s \sim D}[\beta_m(s)] \leq \varepsilon.
\]

Now we prove the sample complexity. Summing the number of samples used in the main algorithm of C-DouExpDes and subroutines C-FeatRecover and EstLowRep (Line ?? in Algorithm 4, Lines 4-5 in Algorithm 5 and Line 6 in Algorithm 6), we have that the total number of samples is bounded by
\[
T_0 + 2MTp + MN = O \left( \frac{L_\phi^4}{\nu^2} \log^2 \left( \frac{|A|}{\delta} \right) + \frac{k^4L_\phi^4L_\theta^2L_w^2}{\nu^2\epsilon^2} \log^6 \left( \frac{kdl_\phi L_\theta L_w N}{\nu \delta \epsilon} \right) \cdot \frac{L_\phi^4}{\nu^2} \log^2 \left( \frac{dMT}{\delta} \right) \right).
\]
Furthermore, we have
\[ + M \cdot \left( \frac{k^2 + k^2 L_0^2}{\varepsilon^2} \log^4 \left( \frac{k + \sqrt{\text{Var}X}}{\varepsilon^2} \right) \right) \]

\[ = \tilde{O} \left( \frac{k^4 L_0^2 L_w^2}{\nu^2 \varepsilon^2} + M \left( \frac{k^2 + k^2 L_0^2}{\varepsilon^2} \right) \right). \]

\[ \square \]

### E. Technical Tools

In this section, we provide some useful technical tools.

**Lemma E.1** (Matrix Bernstein Inequality - Average, Lemma 31 in (Tripuraneni et al., 2021)). Consider a truncation level \( U > 0 \). If \( \{Z_1, \ldots, Z_n\} \) is a sequence of \( d_1 \times d_2 \) independent random matrices and \( Z'_i = Z_i \cdot \mathbf{1} \{\|Z_i\| \leq U\} \) for any \( i \in [n] \), then

\[
\Pr \left[ \left\| \frac{1}{n} \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \right\| \geq t \right] \leq \Pr \left[ \left\| \frac{1}{n} \sum_{i=1}^{n} (Z'_i - \mathbb{E}[Z'_i]) \right\| \geq t - \Delta \right] + n \Pr \|Z_i\| \geq U, \]

where \( \Delta \geq \|\mathbb{E}[Z_i] - \mathbb{E}[Z'_i]\| \) for any \( i \in [n] \).

In addition, for \( t \geq \Delta \), we have

\[
\Pr \left[ \left\| \frac{1}{n} \sum_{i=1}^{n} (Z'_i - \mathbb{E}[Z'_i]) \right\| \geq t - \Delta \right] \leq (d_1 + d_2) \exp \left( -\frac{n^2(t - \Delta)^2}{2\sigma^2 + 2U(n - \Delta)} \right),
\]

where

\[
\sigma^2 = \max \left\{ \left\| \mathbb{E}[(Z'_i - \mathbb{E}[Z'_i])^\top (Z'_i - \mathbb{E}[Z'_i])] \right\|, \left\| \sum_{i=1}^{n} \mathbb{E}[(Z'_i - \mathbb{E}[Z'_i])(Z'_i - \mathbb{E}[Z'_i])^\top] \right\| \right\} \]

\[
\leq \max \left\{ \left\| \sum_{i=1}^{n} \mathbb{E}[Z'_i^\top Z'_i] \right\|, \left\| \sum_{i=1}^{n} \mathbb{E}[Z'_i Z'_i^\top] \right\| \right\}.
\]

Lemma 31 in (Tripuraneni et al., 2021) gives a truncated matrix Bernstein inequality for symmetric random matrices. Here we extend it to general random matrices.

Lemma E.1 can be obtained by combining the truncation argument in the proof of Lemma 31 in (Tripuraneni et al., 2021) and Theorem 6.1.1 in (Tropp et al., 2015) (classic matrix Bernstein inequality for general random matrices).

**Lemma E.2** (Matrix Bernstein Inequality - Summation). Consider a truncation level \( U > 0 \). If \( \{Z_1, \ldots, Z_n\} \) is a sequence of \( d_1 \times d_2 \) independent random matrices, and \( Z'_i = Z_i \cdot \mathbf{1} \{\|Z_i\| \leq U\} \) and \( \Delta \geq \|\mathbb{E}[Z_i] - \mathbb{E}[Z'_i]\| \) for any \( i \in [n] \), then for \( \tau \geq 2n\Delta \),

\[
\Pr \left[ \left\| \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \right\| \geq \tau \right] \leq (d_1 + d_2) \exp \left( -\frac{1}{4} \cdot \frac{\tau^2}{2\sigma^2 + \frac{U\tau}{3}} \right) + n \Pr \|Z_i\| \geq U, \]

where

\[
\sigma^2 = \max \left\{ \left\| \sum_{i=1}^{n} \mathbb{E}[(Z'_i - \mathbb{E}[Z'_i])^\top (Z'_i - \mathbb{E}[Z'_i])] \right\|, \left\| \sum_{i=1}^{n} \mathbb{E}[(Z'_i - \mathbb{E}[Z'_i])(Z'_i - \mathbb{E}[Z'_i])^\top] \right\| \right\} \]

\[
\leq \max \left\{ \left\| \sum_{i=1}^{n} \mathbb{E}[Z'_i^\top Z'_i] \right\|, \left\| \sum_{i=1}^{n} \mathbb{E}[Z'_i Z'_i^\top] \right\| \right\}.
\]

Furthermore, we have

\[
\Pr \left[ \left\| \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \right\| \geq 4 \sqrt{\sigma^2 \log \left( \frac{d_1 + d_2}{\delta} \right)} + 4U \log \left( \frac{d_1 + d_2}{\delta} \right) \right] \leq \delta + n \Pr \|Z_i\| \geq U.
\]
Proof of Lemma E.2. Using Lemma E.1 and defining \( \tau := nt \), we have that for \( \tau > n\Delta \),

\[
\Pr \left[ \left\| \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \right\| \geq \tau \right] \leq (d_1 + d_2) \exp \left( -\frac{(\tau - n\Delta)^2}{2\sigma^2 + 2U(t-n\Delta)} \right) + n \Pr \left[ \|Z_i\| \geq U \right].
\]

If \( \tau > 2n\Delta \), then \( \tau - n\Delta > \frac{1}{2} \tau \) and we have

\[
\Pr \left[ \left\| \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \right\| \geq \tau \right] \leq (d_1 + d_2) \exp \left( -\frac{(\frac{1}{2}\tau)^2}{2\sigma^2 + 2U(\frac{1}{2}\tau)} \right) + n \Pr \left[ \|Z_i\| \geq U \right]
\leq (d_1 + d_2) \exp \left( -\frac{1}{4} \frac{\tau^2}{2\sigma^2 + \tau U} \right) + n \Pr \left[ \|Z_i\| \geq U \right].
\]

Plugging \( \tau = 4\sqrt{\sigma^2 \log \left( \frac{d_1 + d_2}{\delta} \right) + 4U \log \left( \frac{d_1 + d_2}{\delta} \right)} \) into Eq. (37), we have

\[
\Pr \left[ \left\| \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \right\| \geq 4\sqrt{\sigma^2 \log \left( \frac{d_1 + d_2}{\delta} \right) + 4U \log \left( \frac{d_1 + d_2}{\delta} \right)} \right] 
\leq (d_1 + d_2) \exp \left( -\frac{1}{4} \cdot 16\sigma^2 \log \left( \frac{d_1 + d_2}{\delta} \right) + 16U^2 \log \left( \frac{d_1 + d_2}{\delta} \right) + 32U \log \left( \frac{d_1 + d_2}{\delta} \right) \sqrt{\sigma^2 \log \left( \frac{d_1 + d_2}{\delta} \right)} \right)
\leq (d_1 + d_2) \exp \left( -\frac{1}{4} \cdot 4 \log \left( \frac{d_1 + d_2}{\delta} \right) \right) + n \Pr \left[ \|Z_i\| \geq U \right]
\leq (d_1 + d_2) \exp \left( -\frac{1}{4} \cdot 4 \log \left( \frac{d_1 + d_2}{\delta} \right) \right) + n \Pr \left[ \|Z_i\| \geq U \right]
= \delta + n \Pr \left[ \|Z_i\| \geq U \right].
\]

\[\Box\]

Lemma E.3. For any \( A, B > 1, \kappa \in (0, 1) \) and \( T > 0 \) such that \( \log \left( \frac{AB}{\kappa} \right) > 1 \) and \( \log(BT) > 2 \), if

\[ T \geq \frac{68A^2 \log^2 \left( \frac{AB}{\kappa} \right)}{\kappa^2}, \]

then

\[ \frac{A}{\sqrt{T}} \log(BT) \leq \kappa. \]

Proof of Lemma E.3. If \( T = \frac{68A^2 \log^2 \left( \frac{AB}{\kappa} \right)}{\kappa^2} \), we have

\[
\frac{A}{\sqrt{T}} \log(BT) = \frac{A\kappa}{\sqrt{68A \log \left( \frac{AB}{\kappa} \right)}} \log \left( \frac{68A^2 B \log^2 \left( \frac{AB}{\kappa} \right)}{\kappa^2} \right)
\leq \frac{\kappa}{\sqrt{68 \log \left( \frac{AB}{\kappa} \right)}} \left( \log(68) + \log \left( \frac{A^2 B}{\kappa^2} \right) + \log \left( \log^2 \left( \frac{AB}{\kappa} \right) \right) \right)
\leq \frac{\kappa}{\sqrt{68 \log \left( \frac{AB}{\kappa} \right)}} \left( \log(68) + 2 \log \left( \frac{AB}{\kappa} \right) + 2 \log \left( \frac{AB}{\kappa} \right) \right)
\leq \frac{\kappa}{\sqrt{68 \log \left( \frac{AB}{\kappa} \right)}} \left( \log(68) + 4 \log \left( \frac{AB}{\kappa} \right) \right)
\leq \kappa.
\]
Let $f(T) = \frac{A}{\sqrt{T}} \log(BT)$. Then, the derivative of $f(T)$ is

$$f'(T) = \frac{2A - A \log(BT)}{2T \sqrt{T}}.$$ 

If $\log(BT) > 2$, then $f'(T) < 0$, and thus $f(T)$ is decreasing with respect to $T$.

Therefore, if $T \geq \frac{68A^2 \log^2(\frac{AB}{\kappa})}{\kappa^2}$, we have

$$\frac{A}{\sqrt{T}} \log(BT) \leq \kappa.$$

\[\square\]

**Lemma E.4.** For any $A, B > 1$ and $\kappa \in (0, 1)$ such that $\log(\frac{AB}{\kappa}) > 1$ and $\log(BN) > 4$, if

$$N \geq \frac{26A^2 \log^4(\frac{AB}{\kappa})}{\kappa^2},$$

then

$$\frac{A \log^2(BN)}{\sqrt{N}} \leq \kappa.$$

**Proof of Lemma E.4.** If $N = \frac{26A^2 \log^4(\frac{AB}{\kappa})}{\kappa^2}$, we have $\kappa \sqrt{N} = 26A \log^2(\frac{AB}{\kappa})$, and

$$A \log^2(BN) = A \log \left( \frac{26A^2 B \log^4(\frac{AB}{\kappa})}{\kappa^2} \right) \leq A \log \left( \frac{26A^2 B}{\kappa^2} \cdot \frac{A^4 B^4}{\kappa^4} \right) \leq 36A \log^2 \left( \frac{26AB}{\kappa} \right) = 36A \left( \log(26) + \log \left( \frac{AB}{\kappa} \right) \right)^2 \leq 36A \left( \log(26) \log \left( \frac{AB}{\kappa} \right) + \log \left( \frac{AB}{\kappa} \right) \right)^2 = 36(\log(26) + 1)^2 A \log^2 \left( \frac{AB}{\kappa} \right) \leq 26^2 A \log^2 \left( \frac{AB}{\kappa} \right) = \kappa \sqrt{N},$$

and thus $\frac{A \log^2(BN)}{\sqrt{N}} \leq \kappa$.

Let $f(N) = \frac{A \log^2(BN)}{\sqrt{N}}$. Then, the derivative function of $f(N)$ is

$$f'(N) = \frac{4A \log(BN) - A \log^2(BN)}{2N \sqrt{N}} = \frac{A \log(BN) \cdot (4 - \log(BN))}{2N \sqrt{N}}.$$ 

If $\log(BN) > 4$, then $f'(N) < 0$, and thus $f(N)$ is decreasing with respect to $N$.

Therefore, if $N \geq \frac{26A^2 \log^4(\frac{AB}{\kappa})}{\kappa^2}$, we have $\frac{A \log^2(BN)}{\sqrt{N}} \leq \kappa.$

\[\square\]
Lemma E.5. For any $x_1, \ldots, x_n \in \mathbb{R}^k$, we have

$$\sum_{j=1}^{n} \|x_j\| (\sum_{i=1}^{n} x_i x_i^\top)^{-1} \leq \sqrt{n}k.$$ 

Proof of Lemma E.5. It holds that

$$\sum_{j=1}^{n} \|x_j\| (\sum_{i=1}^{n} x_i x_i^\top)^{-1} = \sum_{j=1}^{n} \left[ x_j^\top \left( \sum_{i=1}^{n} x_i x_i^\top \right)^{-1} x_j \right]$$

$$\leq \sqrt{n} \sum_{j=1}^{n} \left[ x_j^\top \left( \sum_{i=1}^{n} x_i x_i^\top \right)^{-1} x_j \right]$$

$$= \sqrt{n} \sum_{j=1}^{n} \text{Trace} \left( x_j^\top \left( \sum_{i=1}^{n} x_i x_i^\top \right)^{-1} x_j \right)$$

$$= \sqrt{n} \cdot \text{Trace} (I_k)$$

$$= \sqrt{n}k$$ 

\[\square\]

Lemma E.6. For any $x_1, \ldots, x_n \in \mathbb{R}^k$ and $\gamma > 0$, we have

$$\sum_{j=1}^{n} \|x_j\| (\gamma I + \sum_{i=1}^{n} x_i x_i^\top)^{-1} \leq \sqrt{n}k.$$ 

Proof of Lemma E.6. It holds that

$$\sum_{j=1}^{n} \|x_j\| (\gamma I + \sum_{i=1}^{n} x_i x_i^\top)^{-1} = \sum_{j=1}^{n} \left[ x_j^\top \left( \gamma I + \sum_{i=1}^{n} x_i x_i^\top \right)^{-1} x_j \right]$$

$$\leq \sqrt{n} \sum_{j=1}^{n} \left[ x_j^\top \left( \gamma I + \sum_{i=1}^{n} x_i x_i^\top \right)^{-1} x_j \right]$$

$$= \sqrt{n} \sum_{j=1}^{n} \text{Trace} \left( x_j^\top \left( \gamma I + \sum_{i=1}^{n} x_i x_i^\top \right)^{-1} x_j \right)$$

$$= \sqrt{n} \cdot \text{Trace} (\gamma I + \sum_{i=1}^{n} x_i x_i^\top)^{-1}$$

$$= \sqrt{n} \cdot \text{Trace} (I_k)$$

$$= \sqrt{n}k$$ 

\[\square\]
Lemma E.7 (Self-normalized Concentration for Martingales, Theorem 1 in (Abbasi-Yadkori et al., 2011)). Let \( \{F_t\}_{t=0}^{\infty} \) be a filtration such that for any \( t \geq 1 \), the selected action \( X_t \in \mathbb{R}^k \) is \( F_{t-1} \)-measurable, the noise \( \eta_t \in \mathbb{R}^k \) is \( F_t \)-measurable, and conditioning on \( F_{t-1} \), \( \eta_t \) is zero-mean and \( R \)-sub-Gaussian. Let \( V_0 \in \mathbb{R}^{k \times k} \) be a positive definite matrix and let \( V_t = \sum_{i=1}^t X_i X_i^\top \) for any \( t \geq 1 \). Then, for any \( \delta > 0 \), with probability at least \( 1 - \delta \), for all \( t \geq 1 \),

\[
\left\| \sum_{i=1}^t X_i \cdot \eta_i \right\|_{(V_0 + V_t)^{-1}}^2 \leq 2R^2 \log \left( \frac{\det(V_t)^{\frac{1}{2}}}{\det(V_0)^{\frac{1}{2}} \cdot \delta} \right).
\]

Lemma E.8 (Reverse Bernstein Inequality for Martingales, Theorem 3 in (Zanette et al., 2021)). Let \( (\Sigma, \mathcal{F}, \mathbb{P}[-]) \) be a probability space and consider the stochastic process \( \{X_t\} \) adapted to the filtration \( \{F_t\} \). Let \( \mathbb{E}_t[X_t] := \mathbb{E}[X_t | F_{t-1}] \) be the conditional expectation of \( X_t \) given \( F_{t-1} \). If \( 0 \leq X_t \leq 1 \) then it holds that

\[
\Pr \left[ \sum_{i=1}^T \mathbb{E}_t[X_t] \geq \frac{1}{4} \left( 2 \sqrt{\log \left( \frac{1}{\delta} \right)} + \sqrt{4 \log \left( \frac{1}{\delta} \right)} + 4 \left( \sum_{i=1}^T X_i + 2 \log \left( \frac{1}{\delta} \right) \right) \right)^2 \right] \leq \delta.
\]

Lemma E.9 (Elliptical Potential Lemma, Lemma 11 in (Abbasi-Yadkori et al., 2011)). Let \( \{X_t\}_{t=1}^n \) be a sequence in \( \mathbb{R}^k \). Let \( V_0 \) be a \( k \times k \) positive definite matrix and let \( V_t = V_0 + \sum_{i=1}^t X_i X_i^\top \) such that for any \( t \geq 1 \), \( \|X_t\|_{V_t^{-1}}^2 \leq 1 \). Then, we have that

\[
\sum_{t=1}^n \|X_t\|_{V_t^{-1}}^2 \leq 2 \log \frac{\det(V_n)}{\det(V_0)}.
\]

Lemma E.10 (Moments of Sub-Gaussian Random Variables, Proposition 3.2 in (Rivasplata, 2012)). For a \( \sigma^2 \)-sub-Gaussian random variable \( X \) which satisfies

\[
\mathbb{E} [\exp(\mu X)] \leq \exp \left( \frac{\sigma^2 \mu^2}{2} \right), \ \forall \mu \in \mathbb{R},
\]

we have that for any integer \( n \geq 1 \),

\[
\mathbb{E}[|X|^n] \leq (2\sigma^2)^{\frac{n}{2}} n \cdot \Gamma \left( \frac{n}{2} \right),
\]

where \( \Gamma(n) := (n-1)! \) for any integer \( n \geq 1 \).