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).

[1]IIIS, Tsinghua University [2]Cornell University. Correspondence to: Yihan Du <duyh18@mails.tsinghua.edu.cn>, Longbo Huang <longbohuang@tsinghua.edu.cn>, Wen Sun <ws455@cornell.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

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 trails (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 to expedite learning. For example, in clinical treatment planning, we want to identify the optimal treatment for multiple diseases, and there exists a joint representation of treatments. In this case, since conducting a clinical trial and collecting a sample is time-consuming, 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 δ, an arm set $\mathcal{X} := \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subseteq \mathbb{R}^d$ and M tasks. For each task $m \in [M]$, the expected reward of each arm $\boldsymbol{x} \in \mathcal{X}$

is generated by $\boldsymbol{x}^\top \boldsymbol{\theta}_m$, where $\boldsymbol{\theta}_m \in \mathbb{R}^d$ is an underlying reward parameter. There exists an unknown global feature extractor $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ and an underlying prediction parameter $\boldsymbol{w}_m$ such that $\boldsymbol{\theta}_m = \boldsymbol{B}\boldsymbol{w}_m$ for any $m \in [M]$, where $M \gg d \gg k$. We can understand the problem as that all tasks share a joint representation $\boldsymbol{f}(\boldsymbol{x}) := \boldsymbol{B}^\top \boldsymbol{x}$ for arms, where the dimension of $\boldsymbol{f}(\boldsymbol{x})$ is much smaller than that of $\boldsymbol{x}$. 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 $\mathcal{S}$, an action space $\mathcal{A}$, a known feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ and an *unknown* context distribution $\mathcal{D}$. For each task $m \in [M]$, the expected reward of each context-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ is generated by $\boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_m$, where $\boldsymbol{\theta}_m = \boldsymbol{B}\boldsymbol{w}_m$. We can similarly interpret the problem as that all tasks share a low-dimensional context-action representation $\boldsymbol{B}^\top \boldsymbol{\phi}(s,a) \in \mathbb{R}^k$. At each timestep, the agent first observes a context drawn from $\mathcal{D}$, and chooses an action and a task to sample, and then observes a random reward. Given a confidence parameter δ and an accuracy parameter ε, the agent aims to identify an ε-optimal policy (i.e., a mapping $\mathcal{S} \mapsto \mathcal{A}$ that gives suboptimality within ε) 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 $\tilde{O}(\frac{Mk}{\Delta^2_{\min}})$, which shows superiority over the baseline result $\tilde{O}(\frac{Md}{\Delta^2_{\min}})$ (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 $\tilde{O}(\frac{Mk^2}{\varepsilon^2})$ is also provided for `C-DouExpDes`, which significantly outperforms the baseline result $\tilde{O}(\frac{Md^2}{\varepsilon^2})$, 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 algorithms based on double experimental designs.

Pure Exploration in (Contextual) Linear Bandits. Most existing linear bandit works focus on regret minimization, e.g., (Dani et al., 2008; Chu et al., 2011; Abbasi-Yadkori et al., 2011). Recently, there has been a surge of interests in the pure exploration objective for (contextual) linear bandits. For linear bandits, Soare et al. (2014) firstly apply the experimental design approach to distinguish the optimal arm, and establish sample complexity 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. Fiez et al. (2019) provide the first near-optimal sample complexity upper and lower bounds for best arm identification in linear bandits. For contextual linear bandits, Zanette et al. (2021) develop a non-adaptive policy to collect data, from which a near-optimal policy can be computed. Li et al. (2022) build instance-optimal sample complexity for best policy identification in contextual linear bandits, with prior knowledge of the context distribution. By contrast, our work studies a multi-task setting where tasks share a common representation, and does not assume any prior knowledge of the context distribution.

3. Problem Formulation

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 $\boldsymbol{A}$, $\|\boldsymbol{A}\|$ denotes the spectral norm of $\boldsymbol{A}$, and $\sigma_{\min}(\boldsymbol{A})$ denotes the minimum singular value of $\boldsymbol{A}$. For any positive semi-definite matrix $\boldsymbol{A} \in \mathbb{R}^{d' \times d'}$ and vector $\boldsymbol{x} \in \mathbb{R}^{d'}$, $\|\boldsymbol{x}\|_{\boldsymbol{A}} := \sqrt{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}$. We use polylog$(\cdot)$ to denote a polylogarithmic factor in given parameters, and $\tilde{O}(\cdot)$ to denote an expression that hides polylogarithmic factors in all problem parameters except δ and ε.

Representation Learning for Best Arm Identification in Linear Bandits (RepBAI-LB). An agent is given a set of arms $\mathcal{X} := \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subseteq \mathbb{R}^d$ 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 $\boldsymbol{x} \in \mathcal{X}$, $\|\boldsymbol{x}\| \leq L_x$ for some constant L_x. For each task $m \in [M]$, the expected reward of each arm $\boldsymbol{x} \in \mathcal{X}$ is $\boldsymbol{x}^\top \boldsymbol{\theta}_m$, where $\boldsymbol{\theta}_m \in \mathbb{R}^d$ is an unknown reward parameter. Among all tasks, there exists a common underlying feature extractor $\boldsymbol{B} \in \mathbb{R}^{d \times k}$, which satisfies that for each task $m \in [M]$, $\boldsymbol{\theta}_m = \boldsymbol{B}\boldsymbol{w}_m$. Here $\boldsymbol{B}$ has orthonormal columns, $\boldsymbol{w}_m \in \mathbb{R}^k$ is an unknown prediction parameter, and $M \gg d \gg k$. For any $m \in [M]$, $\|\boldsymbol{w}_m\| \leq L_w$ for some constant L_w.

At each timestep t, the agent chooses an arm $\boldsymbol{x} \in \mathcal{X}$ and a task $m \in [M]$, to sample arm $\boldsymbol{x}$ in task m. Then, she observes a random reward $r_t = \boldsymbol{x}^\top \boldsymbol{\theta}_m + \eta_t = \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{w}_m + \eta_t$, where η_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 $\boldsymbol{x}^*_m := \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{x}^\top \boldsymbol{\theta}_m$ 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 $\sigma_{\min}(\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{w}_m \boldsymbol{w}_m^\top) = \Omega(\frac{1}{k})$.

This assumption indicates that the prediction parameters $\boldsymbol{w}_1, \dots, \boldsymbol{w}_M$ 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 $\boldsymbol{B}$.

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

$$\boldsymbol{\lambda}^*_m := \operatorname*{argmin}_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{x} \in \mathcal{X} \setminus \{\boldsymbol{x}^*_m\}} \frac{\|\boldsymbol{B}^\top (\boldsymbol{x}^*_m - \boldsymbol{x})\|^2_{\boldsymbol{A}(\boldsymbol{\lambda}, \boldsymbol{B})^{-1}}}{((\boldsymbol{x}^*_m - \boldsymbol{x})^\top \boldsymbol{\theta}_m)^2}.$$

Here $\boldsymbol{\lambda}^*_m$ 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 $m \in [M]$, $\sigma_{\min}(\boldsymbol{A}(\boldsymbol{\lambda}^*_m, \boldsymbol{B})) \geq \omega$ for some constant $\omega > 0$.

This assumption implies that the covariance matrix $\boldsymbol{A}(\boldsymbol{\lambda}^*_m, \boldsymbol{B})$ under the optimal sample allocation is invertible, which is necessary for estimating $\boldsymbol{w}_m$. Note that the quantities introduced in Assumptions 3.1 and 3.2, i.e., $\sigma_{\min}(\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{w}_m \boldsymbol{w}_m^\top)$ and $\sigma_{\min}(\boldsymbol{A}(\boldsymbol{\lambda}^*_m, \boldsymbol{B}))$, 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 $\boldsymbol{\phi}(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ and an *unknown* context distribution $\mathcal{D} \in \triangle_{\mathcal{S}}$. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\|\boldsymbol{\phi}(s,a)\| \leq L_\phi$ for some constant L_ϕ. An agent needs to solve M best policy identification tasks. For each task $m \in [M]$, the expected reward of each context-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ is $\boldsymbol{\phi}(s, a)^\top \boldsymbol{\theta}_m$, where $\boldsymbol{\theta}_m \in \mathbb{R}^d$ is an unknown reward parameter. Similar to RepBAI-LB, there exists a global feature extractor $\boldsymbol{B} \in \mathbb{R}^{d\times k}$ with orthonormal columns, such that for each task $m \in [M]$, $\boldsymbol{\theta}_m = \boldsymbol{B}\boldsymbol{w}_m$. Here $\boldsymbol{w}_m \in \mathbb{R}^k$ is an unknown prediction parameter, $\|\boldsymbol{w}_m\| \le L_w$ 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 $a_t \in \mathcal{A}$ 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 $r_t = \boldsymbol{\phi}(s_t, a_t)^\top \boldsymbol{\theta}_m + \eta_t = \boldsymbol{\phi}(s_t, a_t)^\top \boldsymbol{B}\boldsymbol{w}_m + \eta_t$, where η_t is an independent, zero-mean and 1-sub-Gaussian noise.

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

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

Given a confidence parameter $\delta \in (0, 1)$ and an accuracy parameter $\varepsilon > 0$, the goal of the agent is to identify an ε-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 $\boldsymbol{\lambda} \in \triangle_{\mathcal{A}}$ such that

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

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 $\boldsymbol{B}$, 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 (Tripuraneni 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 RepBAI-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 pseudo-code 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 $\boldsymbol{B}$, and then performs the G-optimal design with low-dimensional representations to eliminate suboptimal arms.

`DouExpDes` uses a *rounding procedure* `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. $\texttt{ROUND}(\{(\boldsymbol{q}_i, \boldsymbol{Q}_i)\}_{i=1}^{n'}, \boldsymbol{\lambda}, \zeta, N)$ takes n' arm-matrix pairs $(\boldsymbol{q}_1, \boldsymbol{Q}_1), \dots, (\boldsymbol{q}_{n'}, \boldsymbol{Q}_{n'}) \in \mathcal{X} \times \mathbb{R}^{d'\times d'}$, a distribution $\boldsymbol{\lambda} \in \triangle_{\{\boldsymbol{q}_1,\dots,\boldsymbol{q}_{n'}\}}$, a rounding approximation parameter $\zeta > 0$, and the number of samples N such that $N \ge \frac{180d'}{\zeta^2}$ as inputs. It will return a sample sequence $\boldsymbol{s}_1, \dots, \boldsymbol{s}_N \in \mathcal{X}$, which correspond to feature matrices $\boldsymbol{S}_1, \dots, \boldsymbol{S}_N \in \{\mathcal{Q}_1, \dots, \mathcal{Q}_{n'}\}$, and $\sum_{j=1}^{N} \boldsymbol{S}_j$ has similar properties as the covariance matrix of the inputted design $N\sum_{i=1}^{n'} \lambda(\boldsymbol{q}_i)\boldsymbol{Q}_i$ (see Appendix B for more details).

Algorithm 1 `DouExpDes` (Double Experimental Design)

1: **Input:** $\mathcal{X}$, δ, rounding procedure `ROUND`, rounding approximation parameter $\zeta := \frac{1}{10}$, and the size of sample batch $p := \frac{180d}{\zeta^2}$.
2: Let $\boldsymbol{\lambda}^E$ and ρ^E be the optimal solution and the optimal value of the E-optimal design optimization:

$$\min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \left\| \Big(\sum_{i=1}^{n} \lambda(\boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^\top \Big)^{-1} \right\|$$

3: $\bar{\boldsymbol{x}}_1, \dots, \bar{\boldsymbol{x}}_p \leftarrow \texttt{ROUND}(\{(\boldsymbol{x}_i, \boldsymbol{x}_i \boldsymbol{x}_i^\top)\}_{i=1}^n, \boldsymbol{\lambda}^E, \zeta, p)$
4: $\hat{\mathcal{X}}_{1,m} \leftarrow \mathcal{X}$ for any $m \in [M]$. $\delta_t \leftarrow \frac{\delta}{2t^2}$ for any $t \geq 1$
5: **for** phase $t = 1, 2, \dots$ **do**
6: $T_t \leftarrow \lceil \frac{c_1(1+\zeta)^3(\rho^E)^2 k^4 L_x^4 L_w^4}{M} \max\{2^{2t}, \frac{L_x^4}{\omega^2}\} \cdot \mathrm{polylog}(\zeta, \rho^E, p, k, L_x, L_w, \frac{1}{\delta_t}, \frac{1}{\omega}) \rceil$, where c_1 is an absolute constant
7: $\hat{\boldsymbol{B}}_t \leftarrow \texttt{FeatRecover}(T_t, \{\bar{\boldsymbol{x}}_i\}_{i \in [p]})$
8: $\{\hat{\mathcal{X}}_{t+1,m}\}_{m \in [M]} \leftarrow$ $\texttt{EliLowRep}(t, \mathcal{X}, \{\hat{\mathcal{X}}_{t,m}\}_{m \in [M]}, \delta_t, \texttt{ROUND}, \zeta, \hat{\boldsymbol{B}}_t)$
9: **if** $|\hat{\mathcal{X}}_{t+1,m}| = 1, \forall m \in [M]$ **then**
10: **return** $\hat{\mathcal{X}}_{t+1,m}$ for all tasks $m \in [M]$
11: **end if**
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 $\boldsymbol{\lambda}^E$ for the purpose of recovering the feature extractor $\boldsymbol{B}$ (Line 2). Then, `DouExpDes` calls `ROUND` to convert the E-optimal sample allocation $\boldsymbol{\lambda}^E$ into a discrete sample batch $\bar{\boldsymbol{x}}_1, \dots, \bar{\boldsymbol{x}}_p$, which satisfies that

$$\left\| \Big(\sum_{j=1}^{p} \bar{\boldsymbol{x}}_j \bar{\boldsymbol{x}}_j^\top \Big)^{-1} \right\| \leq (1+\zeta) \left\| \Big(p \sum_{i=1}^{n} \lambda^E(\boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^\top \Big)^{-1} \right\|.$$

Next, `DouExpDes` enters multiple phases, and maintains a candidate arm set $\hat{\mathcal{X}}_{t,m}$ 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 $\boldsymbol{B}$. In `FeatRecover` (Algorithm 2), we repeatedly sample $\bar{\boldsymbol{x}}_1, \dots, \bar{\boldsymbol{x}}_p$ in all tasks, and construct an estimator $\boldsymbol{Z}$ for $\frac{1}{M}\sum_{i=1}^{M} \boldsymbol{\theta}_m \boldsymbol{\theta}_m^\top$, which contains the information of underlying reward parameters (Line 9). Then, we perform SVD on $\boldsymbol{Z}$ and obtain the estimated feature extractor $\hat{\boldsymbol{B}}$ (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 $\hat{\boldsymbol{B}}^\top \boldsymbol{x}$, and obtain sample allocation $\boldsymbol{\lambda}_m^G$ for each task (Line 2). We further use `ROUND` to transform $\boldsymbol{\lambda}_m^G$ into a sample sequence

Algorithm 2 `FeatRecover`$(T, \{\bar{\boldsymbol{x}}_i\}_{i \in [p]})$

1: **for** task $m \in [M]$ **do**
2: **for** round $j \in [T]$ **do**
3: **for** arm $i \in [p]$ **do**
4: Sample $\bar{\boldsymbol{x}}_i$, and observe random reward $\alpha_{m,j,i}$
5: **end for**
6: $\tilde{\boldsymbol{\theta}}_{m,j} \leftarrow (\sum_{i=1}^{p} \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^\top)^{-1} \sum_{i=1}^{p} \bar{\boldsymbol{x}}_i \alpha_{m,j,i}$
7: **end for**
8: **end for**
9: $\boldsymbol{Z} \leftarrow \frac{1}{MT} \sum_{m=1}^{M} \sum_{j=1}^{T} \tilde{\boldsymbol{\theta}}_{m,j} (\tilde{\boldsymbol{\theta}}_{m,j})^\top - (\sum_{i=1}^{p} \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^\top)^{-1}$
10: Perform SVD decomposition on $\boldsymbol{Z}$, and let $\hat{\boldsymbol{B}}$ be the top-k left singular vectors of $\boldsymbol{Z}$
11: **return** $\hat{\boldsymbol{B}}$

Algorithm 3 `EliLowRep`$(t, \mathcal{X}, \{\hat{\mathcal{X}}_m\}_{m \in [M]}, \delta', \texttt{ROUND}, \zeta, \hat{\boldsymbol{B}})$

1: **for** task $m \in [M]$ **do**
2: Let $\boldsymbol{\lambda}_m^G$ and ρ_m^G be the optimal solution and the optimal value of the G-optimal design optimization:

$$\operatorname*{argmin}_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{x}, \boldsymbol{x}' \in \hat{\mathcal{X}}_m} \left\| \hat{\boldsymbol{B}}^\top (\boldsymbol{x} - \boldsymbol{x}') \right\|^2_{\boldsymbol{A}(\boldsymbol{\lambda}, \hat{\boldsymbol{B}})^{-1}}$$

3: $N_m \leftarrow \lceil \max\{32(1+\zeta) 2^{2t} \rho_m^G \log(\frac{4n^2M}{\delta'}), \frac{180k}{\zeta^2}\} \rceil$
4: $\boldsymbol{z}_{m,1}, \dots, \boldsymbol{z}_{m,N_m} \leftarrow$ $\texttt{ROUND}(\{(\boldsymbol{x}_i, \hat{\boldsymbol{B}}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}})\}_{i=1}^n, \boldsymbol{\lambda}_m^G, \zeta, N_m)$
5: Sample the arms $\boldsymbol{z}_{m,1}, \dots, \boldsymbol{z}_{m,N_m} \in \mathcal{X}$, and observe random rewards $r_{m,1}, \dots, r_{m,N_m}$
6: Let $\tilde{\boldsymbol{z}}_{m,j} := \hat{\boldsymbol{B}}^\top \boldsymbol{z}_{m,j}$ for any $j \in [N_m]$
7: $\hat{\boldsymbol{w}}_m \leftarrow (\sum_{j=1}^{N_m} \tilde{\boldsymbol{z}}_{m,j} \tilde{\boldsymbol{z}}_{m,j}^\top)^{-1} \sum_{j=1}^{N_m} \tilde{\boldsymbol{z}}_{m,j} r_{m,j}$
8: $\hat{\boldsymbol{\theta}}_m \leftarrow \hat{\boldsymbol{B}} \hat{\boldsymbol{w}}_m$
9: $\hat{\mathcal{X}}'_m \leftarrow \hat{\mathcal{X}}_m \backslash \{\boldsymbol{x} \in \hat{\mathcal{X}}_m \mid \exists \boldsymbol{x}' \in \hat{\mathcal{X}}_m : (\boldsymbol{x}' - \boldsymbol{x})^\top \hat{\boldsymbol{\theta}}_m > 2^{-t}\}$
10: **end for**
11: **return** $\{\hat{\mathcal{X}}'_m\}_{m \in [M]}$

$\boldsymbol{z}_{m,1}, \dots, \boldsymbol{z}_{m,N_m}$, which satisfies that

$$\begin{aligned} &\max_{\boldsymbol{x}, \boldsymbol{x}' \in \hat{\mathcal{X}}_m} \|\boldsymbol{x} - \boldsymbol{x}'\|^2_{(\sum_{j=1}^{N_m} \hat{\boldsymbol{B}}^\top \boldsymbol{z}_{m,j} \boldsymbol{z}_{m,j}^\top \hat{\boldsymbol{B}})^{-1}} \\ \leq &(1+\zeta) \max_{\boldsymbol{x}, \boldsymbol{x}' \in \hat{\mathcal{X}}_m} \|\boldsymbol{x} - \boldsymbol{x}'\|^2_{(N_m \sum_{i=1}^{n} \lambda_m^G(\boldsymbol{x}_i) \hat{\boldsymbol{B}}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}})^{-1}}. \end{aligned}$$

After sampling this sequence, we build estimators $\hat{\boldsymbol{w}}_{t,m}$ and $\hat{\boldsymbol{\theta}}_{t,m}$ for the underlying prediction parameter $\boldsymbol{w}_m$ and reward parameter $\boldsymbol{\theta}_m$, 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, 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 $\boldsymbol{\theta} \in \mathbb{R}^d$, the instance-dependent hardness is defined as (Fiez et al., 2019)

$$\rho^S(\mathcal{X}, \boldsymbol{\theta}) := \min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{x} \in \mathcal{X} \setminus \{\boldsymbol{x}^*\}} \frac{\|\boldsymbol{x}^* - \boldsymbol{x}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}}}{((\boldsymbol{x}^* - \boldsymbol{x})^\top \boldsymbol{\theta})^2},$$

and the best known sample complexity result is $\tilde{O}(\rho^S(\mathcal{X}, \boldsymbol{\theta}) \log(\frac{1}{\delta})) = \tilde{O}(\frac{d}{(\Delta^S_{\min})^2} \log(\frac{1}{\delta}))$ (Fiez et al., 2019). Here $\boldsymbol{x}^* := \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{x}^\top \boldsymbol{\theta}$ denotes the best arm, and $\Delta^S_{\min} := \min_{\boldsymbol{x} \in \mathcal{X} \setminus \{\boldsymbol{x}^*\}} (\boldsymbol{x}^* - \boldsymbol{x})^\top \boldsymbol{\theta}$ 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

$$\tilde{O}\left(\sum_{m=1}^{M} \rho^S(\mathcal{X}, \boldsymbol{\theta}_m) \log\left(\frac{1}{\delta}\right)\right) = \tilde{O}\left(\frac{Md}{\Delta^2_{\min}} \log\left(\frac{1}{\delta}\right)\right), \quad (1)$$

where $\Delta_{\min} := \min_{m \in [M], \boldsymbol{x} \in \mathcal{X} \setminus \{\boldsymbol{x}^*_m\}} (\boldsymbol{x}^*_m - \boldsymbol{x})^\top \boldsymbol{\theta}_m$ 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* $\boldsymbol{x}^*_m$ *for all tasks* $m \in [M]$, *and the number of samples used is bounded by*

$$\begin{aligned} &\tilde{O}\Bigg(\sum_{m=1}^{M} \min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{x} \in \mathcal{X} \setminus \{\boldsymbol{x}^*_m\}} \frac{\|\boldsymbol{B}^\top (\boldsymbol{x}^*_m - \boldsymbol{x})\|^2_{\boldsymbol{A}(\boldsymbol{\lambda}, \boldsymbol{B})^{-1}}}{((\boldsymbol{x}^*_m - \boldsymbol{x})^\top \boldsymbol{\theta}_m)^2} \log\left(\frac{1}{\delta}\right) \\ &\quad + (\rho^E)^2 d k^4 L_x^2 L_w^2 D \log^4\left(\frac{1}{\delta}\right) \Bigg) \quad (2) \\ &= \tilde{O}\Bigg(\frac{Mk}{\Delta^2_{\min}} \log\left(\frac{1}{\delta}\right) + (\rho^E)^2 d k^4 L_x^2 L_w^2 D \log^4\left(\frac{1}{\delta}\right) \Bigg), \end{aligned}$$

where $D := \max\{\frac{1}{\Delta^2_{\min}}, \frac{L_x^4}{\omega^2}\}$.

Remark 1. In Theorem 4.1, the factors that have implicit dimensional dependency include $\min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{x} \in \mathcal{X} \setminus \{\boldsymbol{x}^*_m\}} \frac{\|\boldsymbol{B}^\top (\boldsymbol{x}^*_m - \boldsymbol{x})\|^2_{\boldsymbol{A}(\boldsymbol{\lambda}, \boldsymbol{B})^{-1}}}{((\boldsymbol{x}^*_m - \boldsymbol{x})^\top \boldsymbol{\theta}_m)^2}$, ω and ρ^E, which scale as k, $\frac{1}{k}$ and d, respectively.

In our sample complexity bound (Eq. (2)), the first term, $\sum_{m=1}^{M} \min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{x} \in \mathcal{X} \setminus \{\boldsymbol{x}^*_m\}} \frac{\|\boldsymbol{B}^\top (\boldsymbol{x}^*_m - \boldsymbol{x})\|^2_{\boldsymbol{A}(\boldsymbol{\lambda}, \boldsymbol{B})^{-1}}}{((\boldsymbol{x}^*_m - \boldsymbol{x})^\top \boldsymbol{\theta}_m)^2} = O(\frac{Mk}{\Delta^2_{\min}})$, represents the hardness of M k-dimensional linear bandit instances with arm set $\{\boldsymbol{B}^\top \boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\}$ and underlying reward parameters $\boldsymbol{w}_1, \dots, \boldsymbol{w}_M$. 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 $\boldsymbol{B}$. The second term $(\rho^E)^2 d k^4 L_x^2 L_w^2 D$, 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 $\tilde{O}(\frac{Md}{\Delta^2_{\min}} \log(\frac{1}{\delta}))$ (i.e., learning each task independently) to only $\tilde{O}(\frac{Mk}{\Delta^2_{\min}} \log(\frac{1}{\delta}))$. 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 $\|\boldsymbol{X}^{-1}\|$, where $\boldsymbol{X} = [\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d_1)}]$ is a collection of arbitrary d_1 arms from the arm set. By contrast, we utilize the *E-optimality* of the sample batch $\bar{\boldsymbol{x}}_1, \dots, \bar{\boldsymbol{x}}_p$ to obtain an optimized dependency $\rho^E \approx \min_{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d_1)} \in \mathcal{X}} \|\boldsymbol{X}^{-1}\|$, 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 $\hat{\boldsymbol{B}}_t$, one can only obtain a sample complexity $\|\hat{\boldsymbol{B}}_t^\top (\boldsymbol{x} - \boldsymbol{x}')\|^2_{(\sum_{i=1}^n \lambda^*_m(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t)^{-1}}$ dependent on $\hat{\boldsymbol{B}}_t$, but this is not a valid upper bound. To tackle this challenge, we connect the low-dimensional sample complexity under the estimated subspace $\|\hat{\boldsymbol{B}}_t^\top (\boldsymbol{x} - \boldsymbol{x}')\|^2_{(\sum_{i=1}^n \lambda^*_m(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t)^{-1}}$ with that under the true subspace $\|\boldsymbol{B}^\top (\boldsymbol{x} - \boldsymbol{x}')\|^2_{(\sum_{i=1}^n \lambda^*_m(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B})^{-1}}$, and drive a tight sample complexity.

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

First, the lower bound for single-task BAI-LB with arm set $\mathcal{X}$ and underlying reward parameter $\boldsymbol{\theta}_m$ is $\Omega(\rho^S(\mathcal{X}, \boldsymbol{\theta}_m) \log(\frac{1}{\delta}))$ (Fiez et al., 2019). If the global feature extractor $\boldsymbol{B}$ is known, then the RepBAI-LB problem will reduce to M k-dimensional BAI-LB instances with arm set $\{\boldsymbol{B}^\top \boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\}$ and underlying reward parameters $\boldsymbol{w}_1, \dots, \boldsymbol{w}_M$. Therefore, we conjecture that the lower bound for RepBAI-LB is $\Omega(\sum_{m=1}^{M} \rho^S(\mathcal{X}, \boldsymbol{\theta}_m) \log(\frac{1}{\delta}))$, which is the cost of solving M k-dimensional BAI-LB instances. However, it is challenging to rigorously analyze

Algorithm 4 C-DouExpDes (Contextual Double Experimental Design)

1: **Input:** δ, ε, $\boldsymbol{\phi}(\cdot,\cdot)$, regularization parameter $\gamma \geq 1$, rounding procedure ROUND, rounding approximation parameter $\zeta := \frac{1}{10}$, and the size of sample batch $p := \lceil \frac{c_2(1+\zeta)^2 L_\phi^4}{\nu^2} \text{polylog}(\zeta, M, d, k, L_\phi, L_w, \gamma, \frac{1}{\nu}, \frac{1}{\delta}, \frac{1}{\varepsilon}) \rceil$, where c_2 is an absolute constant.
2: $T_0 \leftarrow \lceil \frac{32^2(1+\zeta)^2 L_\phi^4}{\nu^2} \log^2(\frac{20d|\mathcal{A}|}{\delta}) \rceil$. $\hat{\mathcal{D}} \leftarrow \emptyset$
3: **for** $\tau \in [T_0]$ **do**
4: Observe context s_τ, and randomly sample an action
5: $\hat{\mathcal{D}} \leftarrow \hat{\mathcal{D}} \cup \{s_\tau\}$
6: **end for**
7: Let $\boldsymbol{\lambda}^E_{\hat{\mathcal{D}}}$ and $\rho^E_{\hat{\mathcal{D}}}$ be the optimal solution and the optimal value of the E-optimal design optimization:
$$\min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{A}}} \left\| \Big(\sum_{a \in \mathcal{A}} \lambda(a) \mathbb{E}_{s \sim \hat{\mathcal{D}}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \Big)^{-1} \right\|$$
8: $\{\bar{a}_i\}_{i \in [p]} \leftarrow \texttt{ROUND}(\{(a, \mathbb{E}_{s\sim\hat{\mathcal{D}}}[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\top])\}_{a \in \mathcal{A}}, \boldsymbol{\lambda}^E_{\hat{\mathcal{D}}}, \zeta, p)$
9: $T \leftarrow \lceil \frac{c_3(1+\zeta)^2 k^4 L_\phi^4 L_w^4}{M \nu^2 \varepsilon^2} \text{polylog}(\zeta, d, k, L_\phi, L_w, \gamma, \frac{1}{\nu}, \frac{1}{\delta}, \frac{1}{\varepsilon}) \rceil$, where c_3 is an absolute constant
10: $\hat{\boldsymbol{B}} \leftarrow \texttt{C-FeatRecover}(T, \{\bar{a}_i\}_{i\in[p]})$
11: $N \leftarrow \lceil \frac{(k^2 + \gamma k L_w^2)}{\varepsilon^2} \log^4(\frac{\gamma k L_w}{\varepsilon \delta}) \rceil$
12: $\{\hat{\boldsymbol{\theta}}_{m,N}\}_{m \in [M]} \leftarrow \texttt{EstLowRep}(N, \gamma, \hat{\boldsymbol{B}})$
13: **return** $\hat{\pi}_m(\cdot) := \text{argmax}_{a \in \mathcal{A}} \boldsymbol{\phi}(\cdot, a)^\top \hat{\boldsymbol{\theta}}_{m,N}$ for all tasks $m \in [M]$

the independence of these M k-dimensional instances and drive the summation in our conjectured lower bound. We leave the formal lower bound proof for future work.

When $M \gg d \gg k$, Theorem 4.1 matches our conjectured lower bound, which implies that algorithm DouExpDes performs as well as an oracle that knows the low-rank representation $\boldsymbol{B}$ in advance.

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 ex-

Algorithm 5 C-FeatRecover$(T, \{\bar{a}_i\}_{i\in[p]})$

for task $m \in [M]$ **do**
  for round $j \in [T]$ **do**
    for arm $i \in [p]$ **do**
      Observe context $s^{(1)}_{m,j,i}$, sample action $\bar{a}_i$ in task m, and observe reward $\alpha^{(1)}_{m,j,i}$
      Observe context $s^{(2)}_{m,j,i}$, sample action $\bar{a}_i$ in task m, and observe reward $\alpha^{(2)}_{m,j,i}$
    end for
    Let $\boldsymbol{\phi}^{(\ell)}_{m,j,i} := \boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)$, $\forall i \in [p]$, $\forall \ell \in \{1,2\}$
    $\tilde{\boldsymbol{\theta}}^{(\ell)}_{m,j} \leftarrow (\sum_{i=1}^p \boldsymbol{\phi}^{(\ell)}_{m,j,i} \boldsymbol{\phi}^{(\ell)}_{m,j,i}{}^\top)^{-1} \sum_{i=1}^p \boldsymbol{\phi}^{(\ell)}_{m,j,i} \alpha^{(\ell)}_{m,j,i}$, $\forall \ell \in \{1,2\}$
  end for
end for
$\boldsymbol{Z} \leftarrow \frac{1}{MT} \sum_{m=1}^M \sum_{j=1}^T \tilde{\boldsymbol{\theta}}^{(1)}_{m,j} (\tilde{\boldsymbol{\theta}}^{(2)}_{m,j})^\top$
Perform SVD decomposition on $\boldsymbol{Z}$, and let $\hat{\boldsymbol{B}}$ be the top-k left singular vectors
return $\hat{\boldsymbol{B}}$

Algorithm 6 EstLowRep$(N, \gamma, \hat{\boldsymbol{B}})$

$\boldsymbol{\Sigma}_{m,0} \leftarrow \gamma I$ for any $m \in [M]$
for task $m \in [M]$ **do**
  for timestep $t \in [N]$ **do**
    Observe context $s_{m,t}$
    $a_{m,t} \leftarrow \text{argmax}_{a \in \mathcal{A}} \|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,t}, a)\|_{\boldsymbol{\Sigma}^{-1}_{m,t-1}}$
    Sample action $a_{m,t}$, and observe reward $r_{m,t}$
    $\boldsymbol{\Sigma}_{m,t} \leftarrow \boldsymbol{\Sigma}_{m,t-1} + \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,t}, a_{m,t}) \boldsymbol{\phi}(s_{m,t}, a_{m,t})^\top \hat{\boldsymbol{B}}$
    $\hat{\boldsymbol{w}}_{m,t} \leftarrow \boldsymbol{\Sigma}^{-1}_{m,t} \sum_{\tau=1}^t \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) r_{m,\tau}$
    $\hat{\boldsymbol{\theta}}_{m,t} \leftarrow \hat{\boldsymbol{B}} \hat{\boldsymbol{w}}_{m,t}$
  end for
end for
return $\{\hat{\boldsymbol{\theta}}_{m,N}\}_{m \in [M]}$

perimental 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 $\mathcal{D}$ (Lines 3-6). Then, it performs the E-optimal design under the estimated context distribution $\hat{\mathcal{D}}$, and obtains an efficient sample allocation $\boldsymbol{\lambda}^E_{\hat{\mathcal{D}}}$ for the purpose of recovering the feature extractor $\boldsymbol{B}$ (Line 7). Further, C-DouExpDes calls the rounding procedure ROUND to transform $\boldsymbol{\lambda}^E_{\hat{\mathcal{D}}}$ into a sample batch $\bar{a}_1, \dots, \bar{a}_p$, such that

$$\Big\|\Big(\sum_{j=1}^{p} \mathbb{E}_{s\sim\hat{\mathcal{D}}}\left[\boldsymbol{\phi}(s,\bar{a}_j)\boldsymbol{\phi}(s,\bar{a}_j)^\top\right]\Big)^{-1}\Big\| \\ \leq (1+\zeta)\Big\|\Big(p\sum_{a\in\mathcal{A}} \lambda^E_{\hat{\mathcal{D}}}(a)\mathbb{E}_{s\sim\hat{\mathcal{D}}}\left[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\top\right]\Big)^{-1}\Big\|.$$

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 $\boldsymbol{B}$ using the sample batch $\bar{a}_1, \dots, \bar{a}_p$. In C-FeatRecover (Algorithm 5), we repeatedly sample $\bar{a}_1, \dots, \bar{a}_p$ in all tasks with random contexts. In Lines 4-5, we sample this batch twice, and the superscripts (1) and (2) denotes the first and second samples, respectively. After sampling, we carefully establish an estimator $\boldsymbol{Z}$ for the reward parameter related matrix $\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\theta}_m\boldsymbol{\theta}_m^\top$, using instantaneous context-action features $\boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)^\top$. We then perform SVD decomposition on $\boldsymbol{Z}$ to obtain the estimated feature extractor $\hat{\boldsymbol{B}}$ (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 $\boldsymbol{\theta}_m$. In EstLowRep (Algorithm 6), we employ the estimated representation $\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s,a)$ to sample the actions with the maximum uncertainty under the observed contexts. After that, we construct estimators $\hat{\boldsymbol{w}}_{m,t}$ and $\hat{\boldsymbol{\theta}}_{m,t}$ for the prediction parameter $\hat{\boldsymbol{w}}_m$ and reward parameter $\hat{\boldsymbol{\theta}}_m$ (Lines 8-9). At last, C-DouExpDes returns the greedy policy with respect to the estimated reward parameter $\hat{\boldsymbol{\theta}}_{m,N}$ 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 $\boldsymbol{\phi}(s,a) \in \mathbb{R}^d$ and reward parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, the best known sample complexity is $\tilde{O}(\frac{d^2}{\varepsilon^2}\log(\frac{1}{\delta}))$ (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

$$\tilde{O}\left(\frac{Md^2}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right),$$

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 ε-optimal policy $\hat{\pi}_m$ such that $\mathbb{E}_{s\sim\mathcal{D}}[\max_{a\in\mathcal{A}}(\boldsymbol{\phi}(s,a)-\boldsymbol{\phi}(s,\hat{\pi}_m(s)))^\top\boldsymbol{\theta}_m] \leq \varepsilon$ for each task $m\in[M]$, and the number of samples used is*

$$\tilde{O}\left(\frac{M\left(k^2+\gamma k L_w^2\right)}{\varepsilon^2}+\frac{k^4L_\phi^8L_w^4}{\nu^4\varepsilon^2}\right).$$

Remark 2. In this result, only factor ν has implicit dimensional dependency, which scales as $\frac{1}{d}$. 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 $\boldsymbol{B}^\top\boldsymbol{\phi}(s,a)$. The second term $\frac{k^4L_\phi^8L_w^4}{\nu^4\varepsilon^2}$ is a price paid for learning global feature extractor $\boldsymbol{B}$ 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.

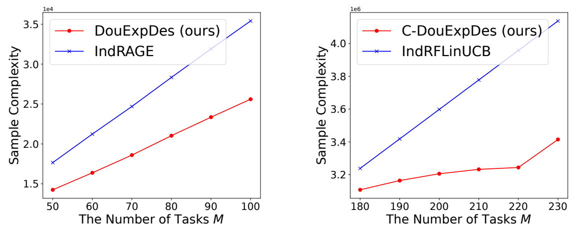
When $M \gg \frac{1}{\nu} \gg k$, this result becomes $\tilde{O}(\frac{Mk^2}{\varepsilon^2})$ and only depends on the low dimension k, which implies that C-DouExpDes performs as well as an oracle that knows the underlying low-rank subspace $\boldsymbol{B}$. This sample complexity significantly outperforms the baseline result $\tilde{O}(\frac{Md^2}{\varepsilon^2})$ (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 $\mathbb{E}_{s\sim\hat{\mathcal{D}}}[\boldsymbol{\phi}(s,\bar{a}_i)\boldsymbol{\phi}(s,\bar{a}_i)^\top]$ and those under the true context distribution $\mathbb{E}_{s\sim\mathcal{D}}[\boldsymbol{\phi}(s,\bar{a}_i)\boldsymbol{\phi}(s,\bar{a}_i)^\top]$. We further bound the distance between $\mathbb{E}_{s\sim\mathcal{D}}[\boldsymbol{\phi}(s,\bar{a}_i)\boldsymbol{\phi}(s,\bar{a}_i)^\top]$ and the context-action features under actual instantaneous contexts $\boldsymbol{\phi}(s^{(\ell)}_{m,j,i},\bar{a}_i)\boldsymbol{\phi}(s^{(\ell)}_{m,j,i},\bar{a}_i)^\top$. (ii) We leverage the E-optimality of the sample batch $\bar{a}_1,\dots,\bar{a}_p$ to bound $\|(\sum_{i=1}^{p}\boldsymbol{\phi}^{(\ell)}_{m,j,i}\boldsymbol{\phi}^{(\ell)}_{m,j,i}{}^\top)^{-1}\|$. Then, we establish a concentration inequality for $\|\boldsymbol{Z}-\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\theta}_m\boldsymbol{\theta}_m^\top\|$ using the bounded $\|(\sum_{i=1}^{p}\boldsymbol{\phi}^{(\ell)}_{m,j,i}\boldsymbol{\phi}^{(\ell)}_{m,j,i}{}^\top)^{-1}\|$ and matrix Bernstern inequality with truncated noises. (iii) Furthermore, we decompose the prediction error $\boldsymbol{\phi}(s,a)^\top(\hat{\boldsymbol{\theta}}_{m,t}-\boldsymbol{\theta}_m)$ into three components, including the sample variance and bias of $\hat{\boldsymbol{w}}_{m,t}$, and the estimation error of $\hat{\boldsymbol{B}}$. 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, $\mathcal{X}$ is



(a) RepBAI-LB (b) RepBPI-CLB

Figure 1. Experimental results for RepBAI-LB and RepBPI-CLB. The two figures compare the sample complexities of our algorithms with the naive algorithms which treat each task independently.

the canonical basis of $\mathbb{R}^d$. In RepBPI-CLB, we set $\varepsilon = 0.1$, $|\mathcal{S}| = 5$ and $|\mathcal{A}| = 5$. $\mathcal{D}$ is the uniform distribution on $\mathcal{S}$. For any $s \in \mathcal{S}$, $\{\boldsymbol{\phi}(s,a)\}_{a\in\mathcal{A}}$ is the canonical basis of $\mathbb{R}^d$. In both problems, $\boldsymbol{B} = [I_k; \boldsymbol{0}]$, where I_k denotes the $k \times k$ identity matrix. $\boldsymbol{w}_1, \dots, \boldsymbol{w}_M$ are divided into k groups, with $\frac{M}{k}$ same members in each group. The members in the i-th group ($i \in [k]$), i.e., $\boldsymbol{w}_{(M/k)\times(i-1)+1}, \dots, \boldsymbol{w}_{(M/k)\times i}$, have 1 in the i-th coordinate and 0 in all other coordinates. For any $m \in [M]$, $\boldsymbol{\theta}_m = \boldsymbol{B}\boldsymbol{w}_m$. We vary M and perform 50 independent runs to report the average sample complexity across runs.

For RepBAI-LB, we compare algorithm `DouExpDes` with the baseline `IndRAGE` which runs the state-of-the-art single-task BAI-LB algorithm `RAGE` (Fiez et al., 2019) to solve M tasks independently. Figure 1(a) shows the empirical results for RepBAI-LB. From Figure 1(a), we can see that `DouExpDes` has a better sample complexity than `IndRAGE`, and as the number of tasks M increases, the sample complexity of `DouExpDes` increases at a lower rate than that of `IndRAGE`. This demonstrates that `DouExpDes` effectively utilize the shared representation among tasks to reduce the number of samples needed for multi-task learning.

For RepBPI-CLB, our algorithm `C-DouExpDes` is compared with the baseline `IndRFLinUCB`, which tackles M tasks independently by calling the state-of-the-art single-task BPI-CLB algorithm `Reward-free LinUCB` (Zanette et al., 2021). As presented in Figure 1(b), `C-DouExpDes` achieves a significantly lower sample complexity than `IndRFLinUCB`. In addition, the slope of the sample complexity curve of `C-DouExpDes` with respect to M is much smaller than that of `IndRFLinUCB`, which validates that `C-DouExpDes` enjoys a lighter dependency on dimension in multi-task learning. These empirical results match our theoretical bounds, and corroborate the power of representation learning.

7. Conclusion and Future Work

In this paper, we investigate representation learning for pure exploration in multi-task (contextual) linear bandits. We propose two efficient algorithms which conduct double experimental designs to optimally allocate samples for learning the low-rank representation. The sample complexities 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

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Agarwal, A., Song, Y., Sun, W., Wang, K., Wang, M., and Zhang, X. Provable benefits of representational transfer in reinforcement learning. *arXiv preprint arXiv:2205.14571*, 2022.

Agarwal, D., Chen, B.-C., and Elango, P. Explore/exploit schemes for web content optimization. In *International Conference on Data Mining*, pp. 1–10. IEEE, 2009.

Allen-Zhu, Z., Li, Y., Singh, A., and Wang, Y. Near-optimal design of experiments via regret minimization. In *International Conference on Machine Learning*, pp. 126–135. PMLR, 2017.

Ando, R. K., Zhang, T., and Bartlett, P. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(11), 2005.

Baxter, J. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

Ben-David, S. and Schuller, R. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pp. 567–580. Springer, 2003.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transac-*

tions on Pattern Analysis and Machine Intelligence, 35 (8):1798–1828, 2013.

Bhatia, R. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Bragman, F. J., Tanno, R., Eaton-Rosen, Z., Li, W., Hawkes, D. J., Ourselin, S., Alexander, D. C., McClelland, J. R., and Cardoso, M. J. Uncertainty in multitask learning: joint representations for probabilistic MR-only radiotherapy planning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 3–11. Springer, 2018.

Caruana, R. Multitask learning. *Machine Learning*, 28(1): 41–75, 1997.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11:2901–2934, 2010.

Cella, L., Lounici, K., and Pontil, M. Meta representation learning with contextual linear bandits. *arXiv preprint arXiv:2205.15100*, 2022a.

Cella, L., Lounici, K., and Pontil, M. Multi-task representation learning with stochastic linear bandits. *arXiv preprint arXiv:2202.10066*, 2022b.

Cheng, Y., Feng, S., Yang, J., Zhang, H., and Liang, Y. Provable benefit of multitask representation learning in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008.

Degenne, R., Ménard, P., Shang, X., and Valko, M. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pp. 2432–2442. PMLR, 2020.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021a.

Du, Y., Kuroki, Y., and Chen, W. Combinatorial pure exploration with full-bandit or partial linear feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7262–7270, 2021b.

Fiez, T., Jain, L., Jamieson, K. G., and Ratliff, L. Sequential experimental design for transductive linear bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Hu, J., Chen, X., Jin, C., Li, L., and Wang, L. Near-optimal representation learning for linear bandits and linear rl. In *International Conference on Machine Learning*, pp. 4349–4358. PMLR, 2021.

Huang, B., Huang, K., Kakade, S., Lee, J. D., Lei, Q., Wang, R., and Yang, J. Optimal gradient-based algorithms for non-concave bandit optimization. *Advances in Neural Information Processing Systems*, 34:29101–29115, 2021.

Huang, T.-K., Agarwal, A., Hsu, D. J., Langford, J., and Schapire, R. E. Efficient and parsimonious agnostic active learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Jedra, Y. and Proutiere, A. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

Jun, K.-S., Willett, R., Wright, S., and Nowak, R. Bilinear bandits with low-rank structure. In *International Conference on Machine Learning*, pp. 3163–3172. PMLR, 2019.

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

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

Lale, S., Azizzadenesheli, K., Anandkumar, A., and Hassibi, B. Stochastic linear bandits with hidden low rank structure. *arXiv preprint arXiv:1901.09490*, 2019.

Lattimore, T. and Hao, B. Bandit phase retrieval. *Advances in Neural Information Processing Systems*, 34:18801–18811, 2021.

Li, J., Zhang, H., Zhang, L., Huang, X., and Zhang, L. Joint collaborative representation with multitask learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(9):5923–5936, 2014.

Li, Z., Ratliff, L., Nassif, H., Jamieson, K., and Jain, L. Instance-optimal PAC algorithms for contextual bandits. *Advances in Neural Information Processing Systems*, 2022.

Lu, R., Huang, G., and Du, S. S. On the power of multitask representation learning in linear MDP. *arXiv preprint arXiv:2106.08053*, 2021a.

Lu, R., Zhao, A., Du, S. S., and Huang, G. Provable general function class representation learning in multitask bandits and MDPs. *Advances in Neural Information Processing Systems*, 2022.

Lu, Y., Meisami, A., and Tewari, A. Low-rank generalized linear bandit problems. In *International Conference on Artificial Intelligence and Statistics*, pp. 460–468. PMLR, 2021b.

Maurer, A. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7:117–139, 2006.

Maurer, A., Pontil, M., and Romera-Paredes, B. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Pacchiano, A., Nachum, O., Tripuraneni, N., and Bartlett, P. Joint representation training in sequential tasks with shared structure. *arXiv preprint arXiv:2206.12441*, 2022.

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

Qin, Y., Menara, T., Oymak, S., Ching, S., and Pasqualetti, F. Non-stationary representation learning in sequential linear bandits. *IEEE Open Journal of Control Systems*, 2022.

Rivasplata, O. Subgaussian random variables: An expository note. *Internet Publication, PDF*, 5, 2012.

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

Soare, M., Lazaric, A., and Munos, R. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

Tao, C., Blanco, S., and Zhou, Y. Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, pp. 4877–4886. PMLR, 2018.

Tripuraneni, N., Jin, C., and Jordan, M. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pp. 10434–10443. PMLR, 2021.

Tropp, J. A. et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Xu, L., Honda, J., and Sugiyama, M. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 843–851. PMLR, 2018.

Yang, J., Hu, W., Lee, J. D., and Du, S. S. Impact of representation learning in linear bandits. In *ICLR*, 2021.

Yang, J., Lei, Q., Lee, J. D., and Du, S. S. Nearly minimax algorithms for linear bandits with shared representation. *arXiv preprint arXiv:2203.15664*, 2022.

Zanette, A., Dong, K., Lee, J. N., and Brunskill, E. Design of experiments for stochastic contextual linear bandits. *Advances in Neural Information Processing Systems*, 34: 22720–22731, 2021.

Zhang, C. and Wang, Z. Provably efficient multi-task reinforcement learning with model transfer. *Advances in Neural Information Processing Systems*, 34:19771–19783, 2021.

Zhang, D., Shen, D., Initiative, A. D. N., et al. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, 59(2):895–907, 2012.

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 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 $p_1, \dots, p_n \in \mathcal{X}^+$ and n positive semi-definite matrices $\boldsymbol{Q}_1, \dots, \boldsymbol{Q}_n \in \mathbb{S}^d_+$, where $\boldsymbol{Q}_i$ represents the feature of arm or action p_i for any $i \in [n]$. Denote $\mathcal{P} := \{p_1, \dots, p_n\}$ and $\mathcal{Q} := \{\boldsymbol{Q}_1, \dots, \boldsymbol{Q}_n\}$.

The rounding procedure $\texttt{ROUND}(\{(p_i, \boldsymbol{Q}_i)\}_{i=1}^n, \boldsymbol{\lambda}, \zeta, N)$ (Allen-Zhu et al., 2017; Fiez et al., 2019) takes n arm-matrix or action-matrix pairs $(p_1, \boldsymbol{Q}_1), \dots, (p_n, \boldsymbol{Q}_n) \in \mathcal{X}^+ \times \mathbb{S}^d_+$, a distribution $\boldsymbol{\lambda} \in \triangle_{\mathcal{P}}$ (or equivalently, $\boldsymbol{\lambda} \in \triangle_{\mathcal{Q}}$), 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 $\boldsymbol{\lambda}$.

Formally, $\texttt{ROUND}(\{(p_i, \boldsymbol{Q}_i)\}_{i=1}^n, \lambda, \zeta, N)$ returns a discrete sample sequence $s_1, \dots, s_N \in \mathcal{P}^N$ associated with feature matrices $\boldsymbol{S}_1, \dots, \boldsymbol{S}_N \in \mathcal{Q}^N$, which satisfy the following properties:

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

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

then $\boldsymbol{S}_1, \dots, \boldsymbol{S}_N$ satisfy that

$$\left\| \left(\sum_{j=1}^N \boldsymbol{S}_j \right)^{-1} \right\| \leq (1+\zeta) \left\| \left(N \sum_{i=1}^n \lambda(\boldsymbol{Q}_i) \boldsymbol{Q}_i \right)^{-1} \right\|.$$

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

$$\min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{Q}}} \max_{\boldsymbol{y} \in \mathcal{Y}} \|\boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{Q}_i) \boldsymbol{Q}_i\right)^{-1}},$$

then $\boldsymbol{S}_1, \dots, \boldsymbol{S}_N$ satisfy that

$$\max_{\boldsymbol{y} \in \mathcal{Y}} \|\boldsymbol{y}\|^2_{\left(\sum_{j=1}^N \boldsymbol{S}_j\right)^{-1}} \leq (1+\zeta) \max_{\boldsymbol{y} \in \mathcal{Y}} \|\boldsymbol{y}\|^2_{\left(N \sum_{i=1}^n \lambda(\boldsymbol{Q}_i) \boldsymbol{Q}_i\right)^{-1}}.$$

We implement `ROUND` by setting $\boldsymbol{\pi}^* = N\boldsymbol{\lambda}$, $k = r = N$ and $\boldsymbol{x}_i \boldsymbol{x}_i^\top = (\sum_{i=1}^n \pi^*(\boldsymbol{Q}_i)\boldsymbol{Q}_i)^{-\frac{1}{2}} \boldsymbol{Q}_i (\sum_{i=1}^n \pi^*(\boldsymbol{Q}_i)\boldsymbol{Q}_i)^{-\frac{1}{2}}$ 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 $\boldsymbol{x}_i \boldsymbol{x}_i^\top$ rather than the separate feature vector $\boldsymbol{x}_i$, 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_θ to denote the upper bound of $\|\boldsymbol{\theta}_m\|$ for any $m \in [M]$. Since $\boldsymbol{\theta}_m = \boldsymbol{B}\boldsymbol{w}_m$ for any $m \in [M]$, we have that $\|\boldsymbol{\theta}_m\| \leq \|\boldsymbol{B}\| \|\boldsymbol{w}_m\| \leq \|\boldsymbol{w}_m\| \leq L_w$, and thus $L_\theta \leq L_w$.

C.1. Sample Batch Planning

Recall that

$$\boldsymbol{\lambda}^E := \operatorname*{argmin}_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \left\| \left(\sum_{i=1}^{n} \lambda(\boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \right\|$$

and

$$\rho^E := \min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \left\| \left(\sum_{i=1}^{n} \lambda(\boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \right\|$$

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

Let

$$\boldsymbol{X}_{\text{batch}} := \begin{bmatrix} \bar{\boldsymbol{x}}_1^\top \\ \dots \\ \bar{\boldsymbol{x}}_p^\top \end{bmatrix},$$

and

$$\boldsymbol{X}_{\text{batch}}^{+} := (\boldsymbol{X}_{\text{batch}}^\top \boldsymbol{X}_{\text{batch}})^{-1} \boldsymbol{X}_{\text{batch}}^\top.$$

According to the fact that $\mathcal{X}$ spans $\mathbb{R}^d$, the definition of E-optimal design and the guarantee of `ROUND`, we have that $\boldsymbol{X}_{\text{batch}}^\top \boldsymbol{X}_{\text{batch}}$ is invertible.

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

Lemma C.1. *It holds that*

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

Proof of Lemma C.1. We have

$$\begin{aligned}
\|\boldsymbol{X}_{\text{batch}}^{+}\| &= \left\| \left(\boldsymbol{X}_{\text{batch}}^\top \boldsymbol{X}_{\text{batch}} \right)^{-1} \boldsymbol{X}_{\text{batch}}^\top \right\| \\
&= \sqrt{\left\| \left(\boldsymbol{X}_{\text{batch}}^\top \boldsymbol{X}_{\text{batch}} \right)^{-1} \boldsymbol{X}_{\text{batch}}^\top \boldsymbol{X}_{\text{batch}} \left(\boldsymbol{X}_{\text{batch}}^\top \boldsymbol{X}_{\text{batch}} \right)^{-1} \right\|} \\
&= \sqrt{\left\| \left(\boldsymbol{X}_{\text{batch}}^\top \boldsymbol{X}_{\text{batch}} \right)^{-1} \right\|} \\
&= \sqrt{\left\| \left(\sum_{i=1}^{p} \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^\top \right)^{-1} \right\|} \\
&\leq \sqrt{(1+\zeta) \left\| \left(p \sum_{i=1}^{n} \lambda^E(\boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \right\|} \\
&= \sqrt{\frac{(1+\zeta)\rho^E}{p}}.
\end{aligned}$$

□

C.2. Global Feature Extractor Recovery

For clarity of notation, we add subscript t to the notations in subroutine `FeatRecover` to denote the quantities generated in phase t. Specifically, we use $\alpha_{t,m,j,i}$, $\tilde{\boldsymbol{\theta}}_{t,m,j}$, $\boldsymbol{Z}_t$ and $\hat{\boldsymbol{B}}_t$ to denote the random reward, estimator of reward parameter, estimator of $\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\theta}_m\boldsymbol{\theta}_m^\top$ 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 $\bar{\boldsymbol{x}}_i$ in the j-th round for task m, during the execution of `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 1. $\eta_{t,m,j,i}$ is independent for different t,m,j,i.

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

$$\tilde{\boldsymbol{\theta}}_{t,m,j}=\boldsymbol{X}_{\text{batch}}^{+}\boldsymbol{\alpha}_{t,m,j},$$

and

$$\boldsymbol{Z}_t=\frac{1}{MT_t}\sum_{m=1}^{M}\sum_{j=1}^{T_t}\tilde{\boldsymbol{\theta}}_{t,m,j}(\tilde{\boldsymbol{\theta}}_{t,m,j})^\top-\boldsymbol{X}_{\text{batch}}^{+}(\boldsymbol{X}_{\text{batch}}^{+})^\top.$$

Lemma C.2 (Expectation of $\boldsymbol{Z}_t$). *It holds that*

$$\mathbb{E}\left[\boldsymbol{Z}_t\right]=\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\theta}_m\boldsymbol{\theta}_m^\top.$$

Proof of Lemma C.2. $\boldsymbol{Z}_t$ can be written as

$$\begin{aligned}
\boldsymbol{Z}_t=&\frac{1}{MT_t}\sum_{m=1}^{M}\sum_{j=1}^{T_t}\tilde{\boldsymbol{\theta}}_{t,m,j}(\tilde{\boldsymbol{\theta}}_{t,m,j})^\top-\boldsymbol{X}_{\text{batch}}^{+}(\boldsymbol{X}_{\text{batch}}^{+})^\top\\
=&\frac{1}{MT_t}\sum_{m=1}^{M}\sum_{j=1}^{T_t}\boldsymbol{X}_{\text{batch}}^{+}\begin{bmatrix}\alpha_{t,m,j,1}\\ \vdots\\ \alpha_{t,m,j,p}\end{bmatrix}[\alpha_{t,m,j,1},\dots,\alpha_{t,m,j,p}]^\top(\boldsymbol{X}_{\text{batch}}^{+})^\top-\boldsymbol{X}_{\text{batch}}^{+}(\boldsymbol{X}_{\text{batch}}^{+})^\top\\
=&\frac{1}{MT_t}\sum_{m=1}^{M}\sum_{j=1}^{T_t}\boldsymbol{X}_{\text{batch}}^{+}\begin{bmatrix}\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m+\eta_{t,m,j,1}\\ \vdots\\ \bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m+\eta_{t,m,j,p}\end{bmatrix}[\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m+\eta_{t,m,j,1},\dots,\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m+\eta_{t,m,j,p}]^\top(\boldsymbol{X}_{\text{batch}}^{+})^\top-\boldsymbol{X}_{\text{batch}}^{+}(\boldsymbol{X}_{\text{batch}}^{+})^\top\\
=&\frac{1}{MT_t}\sum_{m=1}^{M}\sum_{j=1}^{T_t}\boldsymbol{X}_{\text{batch}}^{+}\begin{bmatrix}(\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m+\eta_{t,m,j,1})^2 & \cdots & (\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m+\eta_{t,m,j,1})(\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m+\eta_{t,m,j,p})\\ \cdots & \cdots & \cdots\\ (\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m+\eta_{t,m,j,p})(\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m+\eta_{t,m,j,1}) & \cdots & (\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m+\eta_{t,m,j,p})^2\end{bmatrix}(\boldsymbol{X}_{\text{batch}}^{+})^\top\\
&-\boldsymbol{X}_{\text{batch}}^{+}(\boldsymbol{X}_{\text{batch}}^{+})^\top\\
=&\frac{1}{MT_t}\sum_{m=1}^{M}\sum_{j=1}^{T_t}\boldsymbol{X}_{\text{batch}}^{+}\left(\begin{bmatrix}(\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m)^2 & \cdots & \bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\\ \cdots & \cdots & \cdots\\ \bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m & \cdots & (\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m)^2\end{bmatrix}\right.\\
&+\begin{bmatrix}2\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\eta_{t,m,j,1} & \cdots & \bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\eta_{t,m,j,p}+\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\eta_{t,m,j,1}\\ \cdots & \cdots & \cdots\\ \bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\eta_{t,m,j,p}+\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\eta_{t,m,j,1} & \cdots & 2\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\eta_{t,m,j,p}\end{bmatrix}\\
&\left.+\begin{bmatrix}(\eta_{t,m,j,1})^2 & \cdots & \eta_{t,m,j,1}\eta_{t,m,j,p}\\ \cdots & \cdots & \cdots\\ \eta_{t,m,j,1}\eta_{t,m,j,p} & \cdots & (\eta_{t,m,j,p})^2\end{bmatrix}\right)(\boldsymbol{X}_{\text{batch}}^{+})^\top-\boldsymbol{X}_{\text{batch}}^{+}(\boldsymbol{X}_{\text{batch}}^{+})^\top.
\end{aligned}\tag{3}$$

Then, taking the expectation on $\boldsymbol{Z}_t$, we have

$$\mathbb{E}[\boldsymbol{Z}_t]=\frac{1}{MT_t}\sum_{m=1}^{M}\sum_{j=1}^{T_t}\boldsymbol{X}_{\text{batch}}^{+}\left(\begin{bmatrix}(\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m)^2 & \cdots & \bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\\ \cdots & \cdots & \cdots\\ \bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m & \cdots & (\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m)^2\end{bmatrix}+\boldsymbol{I}_d\right)(\boldsymbol{X}_{\text{batch}}^{+})^\top-\boldsymbol{X}_{\text{batch}}^{+}(\boldsymbol{X}_{\text{batch}}^{+})^\top$$

$$
\begin{aligned}
=&\frac{1}{MT_t}\sum_{m=1}^{M}\sum_{j=1}^{T_t}\boldsymbol{X}_{\text{batch}}^{+}\begin{bmatrix}(\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m)^2 & \cdots & \bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\\ \cdots & \cdots & \cdots\\ \bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m & \cdots & (\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m)^2\end{bmatrix}(\boldsymbol{X}_{\text{batch}}^{+})^\top\\
=&\frac{1}{MT_t}\sum_{m=1}^{M}\sum_{j=1}^{T_t}\boldsymbol{X}_{\text{batch}}^{+}\begin{bmatrix}\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\\ \vdots\\ \bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\end{bmatrix}[\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m,\dots,\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m]^\top(\boldsymbol{X}_{\text{batch}}^{+})^\top\\
=&\frac{1}{MT_t}\sum_{m=1}^{M}\sum_{j=1}^{T_t}\boldsymbol{X}_{\text{batch}}^{+}\boldsymbol{X}_{\text{batch}}\boldsymbol{\theta}_m\boldsymbol{\theta}_m^\top\boldsymbol{X}_{\text{batch}}^\top(\boldsymbol{X}_{\text{batch}}^{+})^\top\\
=&\frac{1}{MT_t}\sum_{m=1}^{M}\sum_{j=1}^{T_t}\boldsymbol{\theta}_m\boldsymbol{\theta}_m^\top\\
=&\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\theta}_m\boldsymbol{\theta}_m^\top.
\end{aligned}
$$

□

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

For any phase $t>0$, define events

$$
\mathcal{E}_t:=\left\{\|\boldsymbol{Z}_t-\mathbb{E}[\boldsymbol{Z}_t]\|\le\frac{96\left\|\boldsymbol{X}_{\text{batch}}^{+}\right\|^2 pL_xL_\theta\log\left(\frac{16p}{\delta_t}\right)}{\sqrt{MT_t}}\log\left(\frac{16pMT_t}{\delta_t}\right)\right\},
$$

and

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

Lemma C.3 (Concentration of $\boldsymbol{Z}_t$). *It holds that*

$$
\Pr\left[\mathcal{E}\right]\ge\frac{\delta}{2}.
$$

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

$$
\begin{aligned}
\boldsymbol{Z}_t-\mathbb{E}[\boldsymbol{Z}_t]=&\frac{1}{MT_t}\sum_{m=1}^{M}\sum_{j=1}^{T_t}\boldsymbol{X}_{\text{batch}}^{+}\Bigg(\begin{bmatrix}2\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\eta_{t,m,j,1} & \cdots & \bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\eta_{t,m,j,p}+\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\eta_{t,m,j,1}\\ \cdots & \cdots & \cdots\\ \bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\eta_{t,m,j,p}+\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\eta_{t,m,j,1} & \cdots & 2\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\eta_{t,m,j,p}\end{bmatrix}\\
&-\mathbb{E}\begin{bmatrix}2\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\eta_{t,m,j,1} & \cdots & \bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\eta_{t,m,j,p}+\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\eta_{t,m,j,1}\\ \cdots & \cdots & \cdots\\ \bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\eta_{t,m,j,p}+\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\eta_{t,m,j,1} & \cdots & 2\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\eta_{t,m,j,p}\end{bmatrix}\\
&+\begin{bmatrix}(\eta_{t,m,j,1})^2 & \cdots & \eta_{t,m,j,1}\eta_{t,m,j,p}\\ \cdots & \cdots & \cdots\\ \eta_{t,m,j,1}\eta_{t,m,j,p} & \cdots & (\eta_{t,m,j,p})^2\end{bmatrix}-\mathbb{E}\begin{bmatrix}(\eta_{t,m,j,1})^2 & \cdots & \eta_{t,m,j,1}\eta_{t,m,j,p}\\ \cdots & \cdots & \cdots\\ \eta_{t,m,j,1}\eta_{t,m,j,p} & \cdots & (\eta_{t,m,j,p})^2\end{bmatrix}\Bigg)(\boldsymbol{X}_{\text{batch}}^{+})^\top.
\end{aligned}
$$

Define the following matrices:

$$
\begin{aligned}
\boldsymbol{A}_{t,m,j}:=&\frac{1}{MT_t}\begin{bmatrix}2\bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\eta_{t,m,j,1} & \cdots & \bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\eta_{t,m,j,p}+\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\eta_{t,m,j,1}\\ \cdots & \cdots & \cdots\\ \bar{\boldsymbol{x}}_1^\top\boldsymbol{\theta}_m\eta_{t,m,j,p}+\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\eta_{t,m,j,1} & \cdots & 2\bar{\boldsymbol{x}}_p^\top\boldsymbol{\theta}_m\eta_{t,m,j,p}\end{bmatrix},\\
\boldsymbol{A}_t:=&\sum_{m=1}^{M}\sum_{j=1}^{T_t}\boldsymbol{A}_{t,m,j},
\end{aligned}
$$

$$\boldsymbol{C}_{t,m,j} := \frac{1}{MT_t} \begin{bmatrix} (\eta_{t,m,j,1})^2 & \cdots & \eta_{t,m,j,1}\eta_{t,m,j,p} \\ \cdots & \cdots & \cdots \\ \eta_{t,m,j,1}\eta_{t,m,j,p} & \cdots & (\eta_{t,m,j,p})^2 \end{bmatrix},$$

$$\boldsymbol{C}_t := \sum_{m=1}^{M} \sum_{j=1}^{T_t} \boldsymbol{C}_{t,m,j}.$$

Then, we can write $\boldsymbol{Z}_t - \mathbb{E}[\boldsymbol{Z}_t]$ as

$$\boldsymbol{Z}_t - \mathbb{E}[\boldsymbol{Z}_t] = \boldsymbol{X}_{\text{batch}}^{+} \left(\boldsymbol{A}_t - \mathbb{E}[\boldsymbol{A}_t] + \boldsymbol{C}_t - \mathbb{E}[\boldsymbol{C}_t] \right) (\boldsymbol{X}_{\text{batch}}^{+})^{\top},$$

and thus,

$$\|\boldsymbol{Z}_t - \mathbb{E}[\boldsymbol{Z}_t]\| \leq \left\|\boldsymbol{X}_{\text{batch}}^{+}\right\|^2 \left(\|\boldsymbol{A}_t - \mathbb{E}[\boldsymbol{A}_t]\| + \|\boldsymbol{C}_t - \mathbb{E}[\boldsymbol{C}_t]\| \right). \tag{4}$$

Next, we analyze $\|\boldsymbol{A}_t - \mathbb{E}[\boldsymbol{A}_t]\|$ and $\|\boldsymbol{C}_t - \mathbb{E}[\boldsymbol{C}_t]\|$. 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 $\tilde{\eta}_{t,m,j,i} = \eta_{t,m,j,i} \mathbb{1}\{|\eta_{t,m,j,i}| \leq R\}$ denote the truncated noise. Then, we define the following truncated matrices:

$$\begin{aligned}
\tilde{\boldsymbol{A}}_{t,m,j} &:= \frac{1}{MT_t} \begin{bmatrix} 2\bar{\boldsymbol{x}}_1^{\top}\boldsymbol{\theta}_m \tilde{\eta}_{t,m,j,1} & \cdots & \bar{\boldsymbol{x}}_1^{\top}\boldsymbol{\theta}_m \tilde{\eta}_{t,m,j,p} + \bar{\boldsymbol{x}}_p^{\top}\boldsymbol{\theta}_m \tilde{\eta}_{t,m,j,1} \\ \cdots & \cdots & \cdots \\ \bar{\boldsymbol{x}}_1^{\top}\boldsymbol{\theta}_m \tilde{\eta}_{t,m,j,p} + \bar{\boldsymbol{x}}_p^{\top}\boldsymbol{\theta}_m \tilde{\eta}_{t,m,j,1} & \cdots & 2\bar{\boldsymbol{x}}_p^{\top}\boldsymbol{\theta}_m \tilde{\eta}_{t,m,j,p} \end{bmatrix} \\
\tilde{\boldsymbol{A}}_t &:= \sum_{m=1}^{M} \sum_{j=1}^{T_t} \tilde{\boldsymbol{A}}_{t,m,j}, \\
\tilde{\boldsymbol{C}}_{t,m,j} &:= \frac{1}{MT_t} \begin{bmatrix} (\tilde{\eta}_{t,m,j,1})^2 & \cdots & \tilde{\eta}_{t,m,j,1}\tilde{\eta}_{t,m,j,p} \\ \cdots & \cdots & \cdots \\ \tilde{\eta}_{t,m,j,1}\tilde{\eta}_{t,m,j,p} & \cdots & (\tilde{\eta}_{t,m,j,p})^2 \end{bmatrix} \\
\tilde{\boldsymbol{C}}_t &:= \sum_{m=1}^{M} \sum_{j=1}^{T_t} \tilde{\boldsymbol{C}}_{t,m,j}
\end{aligned} \tag{5}$$

First, we bound $\|\boldsymbol{A}_t - \mathbb{E}[\boldsymbol{A}_t]\|$. Since for any $t > 0$, $m \in [M]$, $j \in [T_t]$ and $i \in [p]$, $|\tilde{\eta}_{t,m,j,i}| \leq R$ and $|\bar{\boldsymbol{x}}_i^{\top}\boldsymbol{\theta}_m| \leq L_x L_\theta$, we have $\|\tilde{\boldsymbol{A}}_{t,m,j}\| \leq \frac{1}{MT_t} \cdot 2pL_xL_\theta R$.

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 $1 - 2p\exp(-\frac{R^2}{2})$, $\|\boldsymbol{A}_{t,m,j}\| \leq \frac{1}{MT_t} \cdot 2pL_xL_\theta R$.

Then, we have

$$\begin{aligned}
\left\|\mathbb{E}[\boldsymbol{A}_{t,m,j}] - \mathbb{E}[\tilde{\boldsymbol{A}}_{t,m,j}]\right\| \leq & \left\|\mathbb{E}\left[\boldsymbol{A}_{t,m,j} \cdot \mathbb{1}\left\{\|\boldsymbol{A}_{t,m,j}\| \geq \frac{2pL_xL_\theta R}{MT_t}\right\}\right]\right\| \\
\leq & \mathbb{E}\left[\|\boldsymbol{A}_{t,m,j}\| \cdot \mathbb{1}\left\{\|\boldsymbol{A}_{t,m,j}\| \geq \frac{2pL_xL_\theta R}{MT_t}\right\}\right] \\
= & \mathbb{E}\left[\frac{2pL_xL_\theta R}{MT_t} \cdot \mathbb{1}\left\{\|\boldsymbol{A}_{t,m,j}\| \geq \frac{2pL_xL_\theta R}{MT_t}\right\}\right] \\
& + \left[\left(\|\boldsymbol{A}_{t,m,j}\| - \frac{2pL_xL_\theta R}{MT_t}\right) \cdot \mathbb{1}\left\{\|\boldsymbol{A}_{t,m,j}\| \geq \frac{2pL_xL_\theta R}{MT_t}\right\}\right] \\
= & \frac{2pL_xL_\theta R}{MT_t} \cdot \Pr\left[\|\boldsymbol{A}_{t,m,j}\| \geq \frac{2pL_xL_\theta R}{MT_t}\right] + \int_0^{\infty} \Pr\left[\|\boldsymbol{A}_{t,m,j}\| - \frac{2pL_xL_\theta R}{MT_t} > x\right] dx
\end{aligned}$$

$$
\begin{aligned}
\leq&\frac{2pL_xL_\theta R}{MT_t}\cdot 2p\cdot\exp\left(-\frac{R^2}{2}\right)+\frac{2pL_xL_\theta}{MT_t}\int_R^\infty \Pr\left[\|\boldsymbol{A}_{t,m,j}\|>\frac{2pL_xL_\theta y}{MT_t}\right]dy\\
\leq&\frac{2pL_xL_\theta R}{MT_t}\cdot 2p\cdot\exp\left(-\frac{R^2}{2}\right)+\frac{2pL_xL_\theta}{MT_t}\int_R^\infty 2p\exp\left(-\frac{y^2}{2}\right)dy\\
\leq&\frac{2pL_xL_\theta R}{MT_t}\cdot 2p\cdot\exp\left(-\frac{R^2}{2}\right)+\frac{2pL_xL_\theta}{MT_t}\cdot 2p\cdot\frac{1}{R}\cdot\exp\left(-\frac{R^2}{2}\right)\\
=&\frac{2pL_xL_\theta}{MT_t}\cdot 2p\cdot\left(R+\frac{1}{R}\right)\exp\left(-\frac{R^2}{2}\right).
\end{aligned}
$$

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

$$
\begin{aligned}
\|\boldsymbol{A}_t-\mathbb{E}[\boldsymbol{A}_t]\|\leq&\frac{4\cdot 2pL_xL_\theta\sqrt{2\log\left(\frac{2pMT_t}{\delta'}\right)}\log\left(\frac{2p}{\delta'}\right)}{\sqrt{MT_t}}+\frac{4\cdot 2pL_xL_\theta\sqrt{2\log\left(\frac{2pMT_t}{\delta'}\right)}\log\left(\frac{2p}{\delta'}\right)}{MT_t}\\
\leq&\frac{8\cdot 2pL_xL_\theta\sqrt{2\log\left(\frac{2pMT_t}{\delta'}\right)}\log\left(\frac{2p}{\delta'}\right)}{\sqrt{MT_t}}.
\end{aligned}\tag{6}
$$

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

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 $1-2p\exp(-\frac{R^2}{2})$, $\|\boldsymbol{C}_{t,m,j}\|\leq\frac{1}{MT_t}\cdot pR^2$. Then, we have

$$
\begin{aligned}
\left\|\mathbb{E}[\boldsymbol{C}_{t,m,j}]-\mathbb{E}[\tilde{\boldsymbol{C}}_{t,m,j}]\right\|\leq&\left\|\mathbb{E}\left[\boldsymbol{C}_{t,m,j}\cdot\mathbb{1}\left\{\|\boldsymbol{C}_{t,m,j}\|\geq\frac{pR^2}{MT_t}\right\}\right]\right\|\\
\leq&\mathbb{E}\left[\|\boldsymbol{C}_{t,m,j}\|\cdot\mathbb{1}\left\{\|\boldsymbol{C}_{t,m,j}\|\geq\frac{pR^2}{MT_t}\right\}\right]\\
=&\mathbb{E}\left[\frac{pR^2}{MT_t}\cdot\mathbb{1}\left\{\|\boldsymbol{C}_{t,m,j}\|\geq\frac{pR^2}{MT_t}\right\}\right]+\left[\left(\|\boldsymbol{C}_{t,m,j}\|-\frac{pR^2}{MT_t}\right)\cdot\mathbb{1}\left\{\|\boldsymbol{C}_{t,m,j}\|\geq\frac{pR^2}{MT_t}\right\}\right]\\
=&\frac{pR^2}{MT_t}\cdot\Pr\left[\|\boldsymbol{C}_{t,m,j}\|\geq\frac{pR^2}{MT_t}\right]+\int_0^\infty\Pr\left[\|\boldsymbol{C}_{t,m,j}\|-\frac{pR^2}{MT_t}>x\right]dx\\
\leq&\frac{pR^2}{MT_t}\cdot 2p\cdot\exp\left(-\frac{R^2}{2}\right)+\frac{2p}{MT_t}\int_R^\infty\boldsymbol{y}\cdot\Pr\left[\|\boldsymbol{C}_{t,m,j}\|>\frac{dy^2}{MT_t}\right]dy\\
\leq&\frac{pR^2}{MT_t}\cdot 2p\cdot\exp\left(-\frac{R^2}{2}\right)+\frac{2p}{MT_t}\int_R^\infty\boldsymbol{y}\cdot 2p\exp\left(-\frac{y^2}{2}\right)dy\\
\leq&\frac{pR^2}{MT_t}\cdot 2p\cdot\exp\left(-\frac{R^2}{2}\right)+\frac{2p}{MT_t}\cdot 2p\cdot\exp\left(-\frac{R^2}{2}\right)\\
=&\frac{p}{MT_t}\cdot 2p\cdot\left(R^2+2\right)\exp\left(-\frac{R^2}{2}\right).
\end{aligned}
$$

Using the truncated matrix Bernstein inequality (Lemma E.2) with $n=MT_t$, $R=\sqrt{2\log\left(\frac{2pMT_t}{\delta'}\right)}$, $n\Pr[\|\boldsymbol{C}_{t,m,j}\|\geq\frac{1}{MT_t}\cdot pR^2]\leq\delta'$, $U=\frac{p\cdot 2\log\left(\frac{2pMT_t}{\delta'}\right)}{MT_t}$, $\sigma^2=\frac{32p}{MT_t}$, $\tau=\frac{4\cdot p\cdot 2\log\left(\frac{2pMT_t}{\delta'}\right)\log\left(\frac{2p}{\delta'}\right)}{\sqrt{MT_t}}+\frac{4\cdot p\cdot 2\log\left(\frac{2pMT_t}{\delta'}\right)\log\left(\frac{2p}{\delta'}\right)}{MT_t}$ and $\Delta=$

$\frac{p\cdot 2\cdot 2\log\left(\frac{2pMT_t}{\delta'}\right)}{MT_t}\cdot\frac{\delta'}{MT_t}$, we have that with probability at least $1-2\delta'$,

$$
\begin{aligned}
\left\| \boldsymbol{C}_t - \mathbb{E}\left[\boldsymbol{C}_t\right] \right\| \leq & \frac{4\cdot 2p\log\left(\frac{2pMT_t}{\delta'}\right)\log\left(\frac{2p}{\delta'}\right)}{\sqrt{MT_t}} + \frac{4\cdot 2p\log\left(\frac{2pMT_t}{\delta'}\right)\log\left(\frac{2p}{\delta'}\right)}{MT_t} \\
\leq & \frac{8\cdot 2p\log\left(\frac{2pMT_t}{\delta'}\right)\log\left(\frac{2p}{\delta'}\right)}{\sqrt{MT_t}}
\end{aligned} \tag{7}
$$

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

$$
\begin{aligned}
\left\| \boldsymbol{Z}_t - \mathbb{E}[\boldsymbol{Z}_t] \right\| \leq & \left\| \boldsymbol{X}^{+}_{\text{batch}} \right\|^2 \left(\left\| \boldsymbol{A}_t - \mathbb{E}\left[\boldsymbol{A}_t\right] \right\| + \left\| \boldsymbol{C}_t - \mathbb{E}\left[\boldsymbol{C}_t\right] \right\| \right) \\
\leq & \left\| \boldsymbol{X}^{+}_{\text{batch}} \right\|^2 \left(\frac{8\cdot 2pL_xL_\theta\sqrt{2\log\left(\frac{2pMT_t}{\delta'}\right)}\log\left(\frac{2p}{\delta'}\right)}{\sqrt{MT_t}} + \frac{8\cdot 2p\log\left(\frac{2pMT_t}{\delta'}\right)\log\left(\frac{2p}{\delta'}\right)}{\sqrt{MT_t}} \right) \\
\leq & \frac{96\left\| \boldsymbol{X}^{+}_{\text{batch}} \right\|^2 pL_xL_\theta\log\left(\frac{2p}{\delta'}\right)}{\sqrt{MT_t}}\log\left(\frac{2pMT_t}{\delta'}\right).
\end{aligned}
$$

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

$$
\left\| \boldsymbol{Z}_t - \mathbb{E}[\boldsymbol{Z}_t] \right\| \leq \frac{96\left\| \boldsymbol{X}^{+}_{\text{batch}} \right\|^2 pL_xL_\theta\log\left(\frac{16p}{\delta_t}\right)}{\sqrt{MT_t}}\log\left(\frac{16pMT_t}{\delta_t}\right),
$$

which implies that $\Pr\left[\mathcal{E}_t\right] \geq 1-\frac{\delta_t}{2}$.

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

$$
\begin{aligned}
\Pr\left[\mathcal{E}\right] \geq & 1 - \sum_{t=1}^{\infty}\Pr\left[\bar{\mathcal{E}}_t\right] \\
\geq & 1 - \sum_{t=1}^{\infty}\frac{\delta_t}{2} \\
= & 1 - \sum_{t=1}^{\infty}\frac{\delta}{4t^2} \\
\geq & 1 - \frac{\delta}{2}.
\end{aligned}
$$

□

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

For any matrix $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ with $m\geq n$, let $\boldsymbol{A}_\perp$ denote the orthogonal complement matrix of $\boldsymbol{A}$, where the columns of $\boldsymbol{A}_\perp$ are the orthogonal complement of those of $\boldsymbol{A}$. Then, it holds that $\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{A}_\perp\boldsymbol{A}_\perp^\top = \boldsymbol{I}_m$, where $\boldsymbol{I}_m$ is the $m\times m$ identity matrix.

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

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

$$
\left\| \hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B} \right\| \leq \frac{192\left\| \boldsymbol{X}^{+}_{\text{batch}} \right\|^2 kpL_xL_\theta\log\left(\frac{16p}{\delta_t}\right)}{\sqrt{MT_t}}\log\left(\frac{16pMT_t}{\delta_t}\right).
$$

Furthermore, for any phase $t > 0$, if

$$T_t = \Bigg\lceil \frac{68 \cdot 192^2 \cdot 8^2 \left(1+\zeta\right)^3 (\rho^E)^2 k^4 L_x^4 L_\theta^2 L_w^2}{c_0^2 M} \cdot \max\left\{2^{2t},\ \frac{L_x^4}{\omega^2}\right\} \cdot \log^2\left(\frac{16p}{\delta_t}\right)$$
$$\log^2\left(\frac{192 \cdot 16 \cdot 8 \left(1+\zeta\right)^{\frac{3}{2}} \rho^E k^2 p L_x^2 L_\theta L_w}{c_0} \cdot \max\left\{2^t,\ \frac{L_x^2}{\omega}\right\} \cdot \frac{1}{\delta_t} \cdot \log\left(\frac{16p}{\delta_t}\right)\right) \Bigg\rceil, \tag{8}$$

then

$$\left\| \hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B} \right\| \leq \min\left\{\frac{1}{8kL_xL_w \cdot 2^t \sqrt{1+\zeta}},\ \frac{\omega}{6L_x^2}\right\}.$$

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

$$\begin{aligned}
\left\| \hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B} \right\| \leq & \frac{\|\boldsymbol{Z}_t - \mathbb{E}[\boldsymbol{Z}_t]\|}{\sigma_k(\mathbb{E}[\boldsymbol{Z}_t]) - \sigma_{k+1}(\mathbb{E}[\boldsymbol{Z}_t]) - \|\boldsymbol{Z}_t - \mathbb{E}[\boldsymbol{Z}_t]\|} \\
\leq & \frac{2k}{c_0} \|\boldsymbol{Z}_t - \mathbb{E}[\boldsymbol{Z}_t]\| \\
\overset{\text{(a)}}{\leq} & \frac{192 \left\|\boldsymbol{X}_{\text{batch}}^+\right\|^2 kpL_xL_\theta \log\left(\frac{16p}{\delta_t}\right)}{c_0\sqrt{MT_t}} \log\left(\frac{16pMT_t}{\delta_t}\right), .
\end{aligned}$$

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

Using Lemma E.3 with $A = \frac{192\left\|\boldsymbol{X}_{\text{batch}}^+\right\|^2 kpL_xL_\theta}{c_0} \log\left(\frac{16p}{\delta_t}\right)$, $B = \frac{16p}{\delta_t}$ and $\kappa = \min\{\frac{1}{8kL_xL_w \cdot 2^t\sqrt{1+\zeta}}, \frac{\omega}{6L_x^2}\}$, we have that if

$$MT_t \geq 68 \left(\frac{192 \left\|\boldsymbol{X}_{\text{batch}}^+\right\|^2 kpL_xL_\theta}{c_0} \log\left(\frac{16p}{\delta_t}\right)\right)^2 \cdot \max\left\{\left(8kL_xL_w \cdot 2^t\sqrt{1+\zeta}\right)^2,\ \frac{6^2L_x^4}{\omega^2}\right\} \cdot$$
$$\log^2\left(\frac{192 \left\|\boldsymbol{X}_{\text{batch}}^+\right\|^2 kpL_xL_\theta}{c_0} \log\left(\frac{16p}{\delta_t}\right) \cdot \frac{16p}{\delta_t} \cdot \max\left\{8kL_xL_w \cdot 2^t\sqrt{1+\zeta},\ \frac{6L_x^2}{\omega}\right\}\right),$$

then $\left\| \hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B} \right\| \leq \min\left\{\frac{1}{8kL_xL_w \cdot 2^t\sqrt{1+\zeta}},\ \frac{\omega}{6L_x^2}\right\}$.

According to Lemma C.1, we have $\|\boldsymbol{X}_{\text{batch}}^+\| \leq \sqrt{\frac{(1+\zeta)\rho^E}{p}}$.

Then, further enlarging MT_t, we have that if

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

then

$$\left\| \hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B} \right\| \leq \min\left\{\frac{1}{8kL_xL_w \cdot 2^t \sqrt{1+\zeta}},\ \frac{\omega}{6L_x^2}\right\}.$$

□

C.3. Elimination with Low-dimensional Representations

For clarity of notation, we also add subscript t to the notations in subroutine `EliLowRep` to denote the quantities generated in phase t. Specifically, we use the notations $\hat{\boldsymbol{B}}_t$, $\hat{\mathcal{X}}_{t,m}$, $\boldsymbol{\lambda}^G_{t,m}$, $\rho^G_{t,m}$, $N_{t,m}$, $\{\boldsymbol{z}_{t,m,i}\}_{i\in[N_{t,m}]}$, $\{r_{t,m,i}\}_{i\in[N_{t,m}]}$, $\hat{\boldsymbol{w}}_{t,m}$ and $\hat{\boldsymbol{\theta}}_{t,m}$ to denote the corresponding quantities used in `EliLowRep` in phase t.

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

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

$$\boldsymbol{\lambda}^*_m := \operatorname*{argmin}_{\boldsymbol{\lambda}\in\triangle_{\mathcal{X}}} \max_{\boldsymbol{x}\in\mathcal{X}\setminus\{\boldsymbol{x}^*_m\}} \frac{\|\boldsymbol{B}^\top\boldsymbol{x}^*_m - \boldsymbol{B}^\top\boldsymbol{x}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i)\boldsymbol{B}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{B}\right)^{-1}}}{\left((\boldsymbol{x}^*_m-\boldsymbol{x})^\top\boldsymbol{\theta}_m\right)^2}.$$

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

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

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

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

$$\begin{aligned}\sum_{i=1}^n \lambda^*_m(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\hat{\boldsymbol{B}}_t =&\hat{\boldsymbol{B}}_t^\top\boldsymbol{A}_m\hat{\boldsymbol{B}}_t\\ =&\hat{\boldsymbol{B}}_t^\top\left(\boldsymbol{B}\boldsymbol{B}^\top+\boldsymbol{B}_\perp\boldsymbol{B}_\perp^\top\right)\boldsymbol{A}_m\left(\boldsymbol{B}\boldsymbol{B}^\top+\boldsymbol{B}_\perp\boldsymbol{B}_\perp^\top\right)\hat{\boldsymbol{B}}_t\\ =&\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{A}_m\boldsymbol{B}\boldsymbol{B}^\top\hat{\boldsymbol{B}}_t + \hat{\boldsymbol{B}}_t^\top\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{A}_m\boldsymbol{B}_\perp\boldsymbol{B}_\perp^\top\hat{\boldsymbol{B}}_t\\ &+\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}_\perp\boldsymbol{B}_\perp^\top\boldsymbol{A}_m\boldsymbol{B}\boldsymbol{B}^\top\hat{\boldsymbol{B}}_t + \hat{\boldsymbol{B}}_t^\top\boldsymbol{B}_\perp\boldsymbol{B}_\perp^\top\boldsymbol{A}_m\boldsymbol{B}_\perp\boldsymbol{B}_\perp^\top\hat{\boldsymbol{B}}_t.\end{aligned}$$

Hence, we have

$$\begin{aligned}\sigma_{\min}\left(\sum_{i=1}^n \lambda^*_m(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\hat{\boldsymbol{B}}_t\right) \ge&\sigma_{\min}\left(\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{A}_m\boldsymbol{B}\boldsymbol{B}^\top\hat{\boldsymbol{B}}_t\right) - \sigma_{\max}\left(\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{A}_m\boldsymbol{B}_\perp\boldsymbol{B}_\perp^\top\hat{\boldsymbol{B}}_t\right)\\ &-\sigma_{\max}\left(\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}_\perp\boldsymbol{B}_\perp^\top\boldsymbol{A}_m\boldsymbol{B}\boldsymbol{B}^\top\hat{\boldsymbol{B}}_t\right) - \sigma_{\max}\left(\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}_\perp\boldsymbol{B}_\perp^\top\boldsymbol{A}_m\boldsymbol{B}_\perp\boldsymbol{B}_\perp^\top\hat{\boldsymbol{B}}_t\right)\\ \ge&\sigma_{\min}\left(\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}\right)\sigma_{\min}\left(\boldsymbol{B}^\top\boldsymbol{A}_m\boldsymbol{B}\right)\sigma_{\min}\left(\boldsymbol{B}^\top\hat{\boldsymbol{B}}_t\right) - \left\|\boldsymbol{B}_\perp^\top\hat{\boldsymbol{B}}_t\right\|\|\boldsymbol{A}_m\|\\ &-\left\|\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}_\perp\right\|\|\boldsymbol{A}_m\| - \left\|\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}_\perp\right\|\|\boldsymbol{A}_m\|\\ \ge&\sigma^2_{\min}\left(\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}\right)\sigma_{\min}\left(\boldsymbol{B}^\top\boldsymbol{A}_m\boldsymbol{B}\right) - 3\left\|\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}_\perp\right\|L_x^2\\ \overset{\text{(a)}}{\ge}&\left(1-\left\|\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}_\perp\right\|^2\right)\omega - 3\left\|\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}_\perp\right\|L_x^2,\end{aligned}$$

where inequality (a) uses the fact that $\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}\boldsymbol{B}^\top\hat{\boldsymbol{B}}_t + \hat{\boldsymbol{B}}_t^\top\boldsymbol{B}_\perp\boldsymbol{B}_\perp^\top\hat{\boldsymbol{B}}_t = \hat{\boldsymbol{B}}_t^\top(B\boldsymbol{B}^\top+\boldsymbol{B}_\perp\boldsymbol{B}_\perp^\top)\hat{\boldsymbol{B}}_t = \hat{\boldsymbol{B}}_t^\top\hat{\boldsymbol{B}}_t = \boldsymbol{I}_k$, and thus, $\sigma^2_{\min}(\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}) = 1-\|\hat{\boldsymbol{B}}_t^\top\boldsymbol{B}_\perp\|^2$.

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

$$\begin{aligned}\sigma_{\min}\left(\sum_{i=1}^n \lambda^*_m(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\hat{\boldsymbol{B}}_t\right) \ge&\left(1-\frac{\omega^2}{36L_x^4}\right)\omega - \frac{\omega}{2}\\ =&\frac{\omega}{2} - \frac{\omega^3}{36L_x^4}\end{aligned}$$

$$>0,$$

where the last inequality is due to $\omega \leq L_x^2 < \sqrt{18} L_x^2$. $\square$

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

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

$$\rho^G_{t,m} := \min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{y} \in \mathcal{Y}(\hat{\mathcal{X}}_{t,m})} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}}.$$

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

$$\rho^G_{t,m} \leq 4k.$$

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 $\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}$,

$$\begin{aligned}
\max_{\boldsymbol{y} \in \mathcal{Y}(\hat{\mathcal{X}}_{t,m})} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \leq & \max_{\boldsymbol{y} \in \mathcal{Y}(\mathcal{X})} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
= & \|\hat{\boldsymbol{B}}_t^\top (\boldsymbol{x}'_1 - \boldsymbol{x}'_2)\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
\leq & \left(\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}'_1\|_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} + \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}'_2\|_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \right)^2 \\
\leq & 2\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}'_1\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} + 2\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}'_2\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
\leq & 4 \max_{\boldsymbol{x} \in \mathcal{X}} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}},
\end{aligned}$$

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

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

$$\min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{x} \in \mathcal{X}} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} = k.$$

Therefore, we have

$$\begin{aligned}
4k = & 4 \min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{x} \in \mathcal{X}} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
= & 4 \max_{\boldsymbol{x} \in \mathcal{X}} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}\|^2_{\left(\sum_{i=1}^n \lambda'(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
\geq & \max_{\boldsymbol{y} \in \mathcal{Y}(\hat{\mathcal{X}}_{t,m})} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda'(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
\geq & \min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{y} \in \mathcal{Y}(\hat{\mathcal{X}}_{t,m})} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
= & \rho^G_{t,m},
\end{aligned}$$

where $\boldsymbol{\lambda}' := \operatorname{argmin}_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{x} \in \mathcal{X}} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}}$. $\square$

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

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 $\boldsymbol{z}_{t,m,j}$ for task m, during the execution of `EliLowRep` in phase t (Line 5 in Algorithm 3).

For any phase $t > 0$, define events

$$\mathcal{F}_t := \Bigg\{ \boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t \right)^{-1} \sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \xi_{t,m,j} \\ \leq \left\| \hat{\boldsymbol{B}}_t^\top \boldsymbol{y} \right\|_{\left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t \right)^{-1}} \sqrt{2 \log \left(\frac{4n^2 M}{\delta_t} \right)},\ \forall m \in [M],\ \forall \boldsymbol{y} \in \mathcal{Y}(\hat{\mathcal{X}}_{t,m}) \Bigg\}, \tag{9}$$

and

$$\mathcal{F} := \cap_{t=1}^{\infty} \mathcal{F}_t.$$

Lemma C.7 (Concentration of the Variance Term)**.** *It holds that*

$$\Pr[\mathcal{F}] \geq 1 - \frac{\delta}{2}.$$

Proof of Lemma C.7. Let $\boldsymbol{\Sigma}_{t,m} := \sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t$. Then, we can write

$$\boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t \right)^{-1} \sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \xi_{t,m,j} = \sum_{j=1}^{N_{t,m}} \boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \boldsymbol{\Sigma}_{t,m}^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \xi_{t,m,j}.$$

For any phase $t > 0$, task $m \in [M]$ and arm $j \in [N_{t,m}]$, $\hat{\boldsymbol{B}}_t$, $\boldsymbol{\Sigma}_{t,m}$ and $\{\boldsymbol{z}_{t,m,j}\}_{j=1}^{N_{t,m}}$ 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 $j \in [N_{t,m}]$, $\boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \boldsymbol{\Sigma}_{t,m}^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \xi_{t,m,j}$ is $(\boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \boldsymbol{\Sigma}_{t,m}^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j})$-sub-Gaussian.

Using Hoeffding's inequality and taking a union bound over all $m \in [M]$ and $\boldsymbol{y} \in \mathcal{Y}(\hat{\mathcal{X}}_{t,m})$, we have that with probability at least $1 - \frac{\delta_t}{2}$,

$$\begin{aligned}
&\sum_{j=1}^{N_{t,m}} \boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \boldsymbol{\Sigma}_{t,m}^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \xi_{t,m,j} \\
\leq & \sqrt{2 \sum_{j=1}^{N_{t,m}} \left(\boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \boldsymbol{\Sigma}_{t,m}^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \right)^2 \cdot \log \left(\frac{4n^2 M}{\delta_t} \right)} \\
= & \sqrt{2 \sum_{j=1}^{N_{t,m}} \boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \boldsymbol{\Sigma}_{t,m}^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t \boldsymbol{\Sigma}_{t,m}^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{y} \cdot \log \left(\frac{4n^2 M}{\delta_t} \right)} \\
= & \sqrt{2 \boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \boldsymbol{\Sigma}_{t,m}^{-1} \left(\hat{\boldsymbol{B}}_t^\top \sum_{j=1}^{N_{t,m}} \boldsymbol{z}_{t,m,j} \cdot \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t \right) \boldsymbol{\Sigma}_{t,m}^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{y} \cdot \log \left(\frac{4n^2 M}{\delta_t} \right)} \\
= & \sqrt{2 \boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \boldsymbol{\Sigma}_{t,m}^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{y} \cdot \log \left(\frac{4n^2 M}{\delta_t} \right)} \\
= & \left\| \hat{\boldsymbol{B}}_t^\top \boldsymbol{y} \right\|_{\boldsymbol{\Sigma}_{t,m}^{-1}} \sqrt{2 \log \left(\frac{4n^2 M}{\delta_t} \right)},
\end{aligned}$$

which implies that

$$\Pr[\mathcal{F}_t] \geq 1 - \frac{\delta_t}{2}.$$

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

$$\begin{aligned}
\Pr\left[\mathcal{F}\right] \geq & 1 - \sum_{t=1}^{\infty} \Pr\left[\bar{\mathcal{F}}_t\right] \\
\geq & 1 - \sum_{t=1}^{\infty} \frac{\delta_t}{2} \\
= & 1 - \sum_{t=1}^{\infty} \frac{\delta}{4t^2} \\
\geq & 1 - \frac{\delta}{2}.
\end{aligned}$$

□

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

$$\left| \boldsymbol{y}^\top \left(\hat{\boldsymbol{\theta}}_{t,m} - \boldsymbol{\theta}_m \right) \right| \leq \frac{1}{2^t}.$$

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

$$\begin{aligned}
\boldsymbol{y}^\top \left(\hat{\boldsymbol{\theta}}_{t,m} - \boldsymbol{\theta}_m \right) = & \boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,m} - \boldsymbol{y}^\top \left(\hat{\boldsymbol{B}}_t \hat{\boldsymbol{B}}_t^\top + \hat{\boldsymbol{B}}_{t,\perp} \hat{\boldsymbol{B}}_{t,\perp}^\top \right) \boldsymbol{\theta}_m \\
= & \boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \left(\hat{\boldsymbol{w}}_{t,m} - \hat{\boldsymbol{B}}_t^\top \boldsymbol{\theta}_m \right) - \boldsymbol{y}^\top \hat{\boldsymbol{B}}_{t,\perp} \hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{\theta}_m. \qquad (10)
\end{aligned}$$

Here, $\hat{\boldsymbol{w}}_{t,m}$ can be written as

$$\begin{aligned}
\hat{\boldsymbol{w}}_{t,m} = & \left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t \right)^{-1} \sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot r_{t,m,j} \\
= & \left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t \right)^{-1} \sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \left(\boldsymbol{z}_{t,m,j}{}^\top \boldsymbol{\theta}_m + \xi_{t,m,j} \right) \\
= & \left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t \right)^{-1} \sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \left(\boldsymbol{z}_{t,m,j}{}^\top \left(\hat{\boldsymbol{B}}_t \hat{\boldsymbol{B}}_t^\top + \hat{\boldsymbol{B}}_{t,\perp} \hat{\boldsymbol{B}}_{t,\perp}^\top \right) \boldsymbol{\theta}_m + \xi_{t,m,j} \right) \\
= & \hat{\boldsymbol{B}}_t^\top \boldsymbol{\theta}_m + \left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t \right)^{-1} \sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_{t,\perp} \hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{\theta}_m \\
& + \left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t \right)^{-1} \sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \xi_{t,m,j}. \qquad (11)
\end{aligned}$$

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

$$\boldsymbol{y}^\top \left(\hat{\boldsymbol{\theta}}_{t,m} - \boldsymbol{\theta}_m \right) = \underbrace{\boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t \right)^{-1} \sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_{t,\perp} \hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B} \boldsymbol{w}_m}_{\text{Bias}}$$

$$+\underbrace{\boldsymbol{y}^\top \hat{\boldsymbol{B}}_t \left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t\right)^{-1} \sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \xi_{t,m,j}}_{\text{Variance}} - \underbrace{\boldsymbol{y}^\top \hat{\boldsymbol{B}}_{t,\perp} \hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B} \boldsymbol{w}_m}_{\text{Estimation error of } \hat{\boldsymbol{B}}_t}.$$

Taking the absolute value on both sides, and using the Cauchy–Schwarz inequality and definition of event $\mathcal{F}$ (Eq. (9)), we have

$$\begin{aligned}
&\left|\boldsymbol{y}^\top \hat{\boldsymbol{\theta}}_{t,m} - \boldsymbol{y}^\top \boldsymbol{\theta}_m\right| \\
\leq & \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\right\|_{\left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \cdot \left\|\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \cdot \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_{t,\perp} \hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B} \boldsymbol{w}_m\right\|_{\left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
& + \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\right\|_{\left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \sqrt{2 \log\left(\frac{4n^2 M}{\delta_t}\right)} + \left|\boldsymbol{y}^\top \hat{\boldsymbol{B}}_{t,\perp} \hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B} \boldsymbol{w}_m\right| \\
\overset{\text{(a)}}{\leq} & \frac{\sqrt{1+\zeta} \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\right\|_{\left(\sum_{i=1}^n \lambda_{t,m}^G(\boldsymbol{x}_i) \cdot \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}}}{\sqrt{N_{t,m}}} \cdot L_x L_w \left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| \cdot \sum_{j=1}^{N_{t,m}} \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j}\right\|_{\left(\sum_{j=1}^{N_{t,m}} \hat{\boldsymbol{B}}_t^\top \boldsymbol{z}_{t,m,j} \boldsymbol{z}_{t,m,j}{}^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
& + \frac{\sqrt{1+\zeta} \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\right\|_{\left(\sum_{i=1}^n \lambda_{t,m}^G(\boldsymbol{x}_i) \cdot \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}}}{\sqrt{N_{t,m}}} \cdot \sqrt{2 \log\left(\frac{4n^2 M}{\delta_t}\right)} + 2 L_x L_w \left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| \\
\overset{\text{(b)}}{\leq} & \frac{\sqrt{1+\zeta} \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\right\|_{\left(\sum_{i=1}^n \lambda_{t,m}^G(\boldsymbol{x}_i) \cdot \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}}}{\sqrt{N_{t,m}}} \cdot L_x L_w \left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| \cdot \sqrt{k N_{t,m}} \\
& + \frac{\sqrt{1+\zeta} \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\right\|_{\left(\sum_{i=1}^n \lambda_{t,m}^G(\boldsymbol{x}_i) \cdot \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}}}{\sqrt{N_{t,m}}} \cdot \sqrt{2 \log\left(\frac{4n^2 M}{\delta_t}\right)} + 2 L_x L_w \left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| \\
\leq & \sqrt{1+\zeta} \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\right\|_{\left(\sum_{i=1}^n \lambda_{t,m}^G(\boldsymbol{x}_i) \cdot \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \cdot L_x L_w \left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| \cdot \sqrt{k} \\
& + \frac{\sqrt{1+\zeta} \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\right\|_{\left(\sum_{i=1}^n \lambda_{t,m}^G(\boldsymbol{x}_i) \cdot \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}}}{\sqrt{N_{t,m}}} \cdot \sqrt{2 \log\left(\frac{4n^2 M}{\delta_t}\right)} + 2 L_x L_w \left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| \\
\leq & \sqrt{(1+\zeta) \cdot k \cdot \rho_{t,m}^G} \cdot L_x L_w \left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| + \frac{\sqrt{(1+\zeta) \cdot \rho_{t,m}^G \cdot 2 \log\left(\frac{4n^2 M}{\delta_t}\right)}}{\sqrt{N_{t,m}}} + 2 L_x L_w \left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| \\
\overset{\text{(c)}}{\leq} & \sqrt{(1+\zeta) \cdot 4k^2} \cdot L_x L_w \left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| + \frac{\sqrt{(1+\zeta) \cdot \rho_{t,m}^G \cdot 2 \log\left(\frac{4n^2 M}{\delta_t}\right)}}{\sqrt{N_{t,m}}} + 2 L_x L_w \left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| \\
\overset{\text{(d)}}{\leq} & \sqrt{(1+\zeta) \cdot 4k^2} \cdot L_x L_w \cdot \frac{1}{8k L_x L_w \cdot 2^t \sqrt{1+\zeta}} + \frac{1}{4 \cdot 2^t} + 2 L_x L_w \cdot \frac{1}{8k L_x L_w \cdot 2^t \sqrt{1+\zeta}} \\
\leq & \frac{1}{4 \cdot 2^t} + \frac{1}{4 \cdot 2^t} + \frac{1}{4 \cdot 2^t} \\
\leq & \frac{1}{2^t}.
\end{aligned}$$

Here inequality (a) is due to the guarantee of rounding procedure `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 $N_{t,m} := \max\{\lceil 32 \cdot 2^{2t}(1+\zeta)\rho_{t,m}^G \log(\frac{4n^2 M}{\delta_t})\rceil,\ \frac{180k}{\zeta^2}\}$. □

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

and arm $\boldsymbol{x}$ in task m. For any phase $t>0$ and task $m \in [M]$, let $\mathcal{Z}_{t,m} := \{\boldsymbol{x} \in \mathcal{X} : \Delta_m(\boldsymbol{x}) \leq 4 \cdot 2^{-t}\}$.

Lemma C.9. *Suppose that event $\mathcal{E} \cap \mathcal{F}$ holds. For any phase $t>0$ and task $m \in [M]$,*

$$\boldsymbol{x}^*_m \in \hat{\mathcal{X}}_{t,m},$$

and for any phase $t \geq 2$ and task $m \in [M]$,

$$\hat{\mathcal{X}}_{t,m} \subseteq \mathcal{Z}_{t,m}.$$

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

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

$$(\boldsymbol{x}' - \boldsymbol{x}^*_m)^\top \hat{\boldsymbol{\theta}}_{t,m} > 2^{-t}.$$

Then, we have

$$\begin{aligned}
(\boldsymbol{x}' - \boldsymbol{x}^*_m)^\top \boldsymbol{\theta}_m =& (\boldsymbol{x}' - \boldsymbol{x}^*_m)^\top \hat{\boldsymbol{\theta}}_{t,m} - (\boldsymbol{x}' - \boldsymbol{x}^*_m)^\top \left(\hat{\boldsymbol{\theta}}_{t,m} - \boldsymbol{\theta}_m\right) \\
\geq& (\boldsymbol{x}' - \boldsymbol{x}^*_m)^\top \hat{\boldsymbol{\theta}}_{t,m} - 2^{-t} \\
>& 2^{-t} - 2^{-t} \\
=& 0,
\end{aligned}$$

which contradicts the definition of $\boldsymbol{x}^*_m$. Thus, we obtain that $\boldsymbol{x}^*_m \in \hat{\mathcal{X}}_{t,m}$ for any phase $t>0$ and task $m \in [M]$.

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

Suppose that there exists some phase t, some task m and some $\boldsymbol{x} \in \hat{\mathcal{X}}_{t,m}$ such that $\Delta_m(\boldsymbol{x}) > 4 \cdot 2^{-t}$. Then, in phase $t-1 \geq 1$, we have

$$\begin{aligned}
(\boldsymbol{x}^*_m - \boldsymbol{x})^\top \hat{\boldsymbol{\theta}}_{t-1,m} =& (\boldsymbol{x}^*_m - \boldsymbol{x})^\top \boldsymbol{\theta}_m - (\boldsymbol{x}^*_m - \boldsymbol{x})^\top \left(\boldsymbol{\theta}_m - \hat{\boldsymbol{\theta}}_{t-1,m}\right) \\
\geq& (\boldsymbol{x}^*_m - \boldsymbol{x})^\top \boldsymbol{\theta}_m - 2^{-(t-1)} \\
>& 4 \cdot 2^{-t} - 2^{-(t-1)} \\
=& 2^{-(t-1)},
\end{aligned}$$

which implies that x should have been eliminated from $\hat{\mathcal{X}}_{t,m}$ 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

$$\boldsymbol{\lambda}^*_m := \operatorname*{argmin}_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{x} \in \mathcal{X} \setminus \{\boldsymbol{x}^*_m\}} \frac{\|\boldsymbol{B}^\top \boldsymbol{x}^*_m - \boldsymbol{B}^\top \boldsymbol{x}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}}}{\left((\boldsymbol{x}^*_m - \boldsymbol{x})^\top \boldsymbol{\theta}_m\right)^2},$$

and

$$\rho^*_m := \min_{\boldsymbol{\lambda} \in \triangle_{\mathcal{X}}} \max_{\boldsymbol{x} \in \mathcal{X} \setminus \{\boldsymbol{x}^*_m\}} \frac{\|\boldsymbol{B}^\top \boldsymbol{x}^*_m - \boldsymbol{B}^\top \boldsymbol{x}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}}}{\left((\boldsymbol{x}^*_m - \boldsymbol{x})^\top \boldsymbol{\theta}_m\right)^2}.$$

$\boldsymbol{\lambda}^*_m$ and ρ^*_m are the optimal solution and the optimal value of the G-optimal design optimization with true feature extractor $\boldsymbol{B}$, respectively.

Lemma C.10. *Suppose that event $\mathcal{E} \cap \mathcal{F}$ holds. For any task $m \in [M]$ and $\boldsymbol{y} \in \mathbb{R}^d$,*

$$\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \le \|\boldsymbol{B}^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\boldsymbol{B}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}} + \frac{11L_x^4}{k\omega^2 \cdot 2^t}.$$

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

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

$$\begin{aligned}
&\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t \\
=&\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i + \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i\right)\cdot\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i + \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i\right)^\top \\
=&\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\Bigg(\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right)\cdot\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right)^\top + \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right)\cdot\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i\right)^\top \\
&+\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i\right)\cdot\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right)^\top + \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i\right)\cdot\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i\right)^\top\Bigg) \\
=&\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right)\cdot\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right)^\top + \sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\Bigg(\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right)\cdot\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i\right)^\top \\
&+\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i\right)\cdot\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right)^\top + \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i\right)\cdot\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i\right)^\top\Bigg).
\end{aligned}$$

Let $\boldsymbol{P}_t := \sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i)\cdot(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i)^\top$. Let $\boldsymbol{Q}_t := \sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)((\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i)\cdot(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i)^\top + (\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i)\cdot(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i)^\top + (\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i)\cdot(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{x}_i)^\top)$. Then, we have $\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t = \boldsymbol{P}_t + \boldsymbol{Q}_t$.

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

$$\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1} = \boldsymbol{P}_t^{-1} - (\boldsymbol{P}_t + \boldsymbol{Q}_t)^{-1}\boldsymbol{Q}_t\boldsymbol{P}_t^{-1}$$

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

$$\begin{aligned}
\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} &= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\right)^\top\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}\hat{\boldsymbol{B}}_t^\top \boldsymbol{y} \\
&= \underbrace{\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\right)^\top \boldsymbol{P}_t^{-1}\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}}_{\text{Term 1}} - \underbrace{\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\right)^\top(\boldsymbol{P}_t + \boldsymbol{Q}_t)^{-1}\boldsymbol{Q}_t\boldsymbol{P}_t^{-1}\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}}_{\text{Term 2}}. \qquad (12)
\end{aligned}$$

From Lemma C.4, we have

$$\left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| \le \min\left\{\frac{1}{8k\cdot 2^t\sqrt{1+\zeta}},\, \frac{\omega}{6L_x^2}\right\} \le \min\left\{\frac{1}{8k\cdot 2^t},\, \frac{\omega}{6L_x^2}\right\}.$$

Since $\boldsymbol{B}^\top \hat{\boldsymbol{B}}_t\hat{\boldsymbol{B}}_t^\top \boldsymbol{B} + \boldsymbol{B}^\top \hat{\boldsymbol{B}}_{t,\perp}\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B} = \boldsymbol{B}^\top(\hat{\boldsymbol{B}}_t\hat{\boldsymbol{B}}_t^\top + \hat{\boldsymbol{B}}_{t,\perp}\hat{\boldsymbol{B}}_{t,\perp}^\top)\boldsymbol{B} = \boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}_k$, we have $\sigma_{\min}^2(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}) = 1 - \|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\|^2$.

Thus, we have

$$\sigma_{\min}(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}) = \sqrt{1 - \left\| \hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B} \right\|^2} \geq \sqrt{1 - \min\left\{ \frac{1}{64k^2 \cdot 2^{2t}},\, \frac{\omega^2}{36L_x^4} \right\}} > 0,$$

which implies that $\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}$ is invertible.

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

$$\begin{aligned}
\text{Term 1} &= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\right)^\top \boldsymbol{P}_t^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{y} \\
&= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y} + \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)^\top \boldsymbol{P}_t^{-1} \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y} + \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right) \\
&= \underbrace{\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y}\right)^\top \boldsymbol{P}_t^{-1} \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y}\right)}_{\text{Term 1-1}} + \underbrace{\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y}\right)^\top \boldsymbol{P}_t^{-1} \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)}_{\text{Term 1-2}} \\
&\quad + \underbrace{\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)^\top \boldsymbol{P}_t^{-1} \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y}\right)}_{\text{Term 1-3}} + \underbrace{\left(\hat{\boldsymbol{B}}_t^\top B_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)^\top \boldsymbol{P}_t^{-1} \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)}_{\text{Term 1-4}}.
\end{aligned}$$

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

First, we have

$$\begin{aligned}
\text{Term 1-1} &= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y}\right)^\top \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right) \cdot \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right)^\top\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y} \\
&= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y}\right)^\top \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B} \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right) \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^\top\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y} \\
&= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y}\right)^\top \left(\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1}\right)^\top \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1} \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y} \\
&= \left(\boldsymbol{B}^\top \boldsymbol{y}\right)^\top \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top \boldsymbol{y} \\
&= \left\|\boldsymbol{B}^\top \boldsymbol{y}\right\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}}.
\end{aligned}$$

We note that since $\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \hat{\boldsymbol{B}}_t + \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \hat{\boldsymbol{B}}_t = \hat{\boldsymbol{B}}_t^\top (B\boldsymbol{B}^\top + \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top) \hat{\boldsymbol{B}}_t = \hat{\boldsymbol{B}}_t^\top \hat{\boldsymbol{B}}_t = \boldsymbol{I}_k$, $\sigma_{\min}^2\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right) = 1 - \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp\right\|^2$. In addition, $\left\|\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1}\right\| = \frac{1}{\sigma_{\min}(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B})} = \frac{1}{\sqrt{1 - \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp\|^2}}$.

Then, second, we have

$$\begin{aligned}
\text{Term 1-2} &= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y}\right)^\top \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right) \cdot \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right)^\top\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y} \\
&= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y}\right)^\top \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B} \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right) \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^\top\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y} \\
&= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y}\right)^\top \left(\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1}\right)^\top \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1} \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\boldsymbol{B}^\top \boldsymbol{y}\right)^\top \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1} \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y} \\
&\le 2L_x \cdot \frac{1}{\omega} \cdot \frac{1}{\sqrt{1 - \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp\right\|^2}} \cdot \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp\right\| \cdot 2L_x \\
&\le 4L_x^2 \cdot \frac{1}{\omega} \cdot \frac{1}{\sqrt{1 - \left(\frac{1}{8k \cdot 2^t}\right)^2}} \cdot \frac{1}{8k \cdot 2^t} \\
&\le 4L_x^2 \cdot \frac{1}{\omega} \cdot \frac{1}{\sqrt{1 - \frac{3}{4}}} \cdot \frac{1}{8k \cdot 2^t} \\
&= \frac{L_x^2}{k\omega \cdot 2^t}.
\end{aligned}
$$

Third, we have

$$
\begin{aligned}
\text{Term 1-3} &= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)^\top \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right) \cdot \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right)^\top\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y} \\
&= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)^\top \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B} \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right) \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^\top\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y} \\
&= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)^\top \left(\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1}\right)^\top \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1} \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{y} \\
&= \left(\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)^\top \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top \boldsymbol{y}, \\
&\le \frac{1}{\sqrt{1 - \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp\right\|^2}} \cdot \left\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp\right\| \cdot 2L_x \cdot \frac{1}{\omega} \cdot 2L_x \\
&\le 4L_x^2 \cdot \frac{1}{\omega} \cdot \frac{1}{\sqrt{1 - \left(\frac{1}{8k \cdot 2^t}\right)^2}} \cdot \frac{1}{8k \cdot 2^t} \\
&\le 4L_x^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{aligned}
$$

Finally, we have

$$
\begin{aligned}
\text{Term 1-4} &= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)^\top \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right) \cdot \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_i\right)^\top\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y} \\
&= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)^\top \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B} \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right) \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^\top\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y} \\
&= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)^\top \left(\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1}\right)^\top \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1} \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y} \\
&= \left(\left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y}\right)^\top \left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1} \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}\right)^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top \boldsymbol{y},
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left(\frac{1}{\sqrt{1 - \left\| \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \right\|^2}} \cdot \left\| \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \right\| \cdot 2L_x \right)^2 \cdot \frac{1}{\omega} \\
&\leq \left(2L_x \cdot \frac{1}{\sqrt{1 - \left(\frac{1}{8k \cdot 2^t}\right)^2}} \cdot \frac{1}{8k \cdot 2^t} \right)^2 \cdot \frac{1}{\omega} \\
&\leq \left(2L_x \cdot \frac{1}{\sqrt{1 - \frac{3}{4}}} \cdot \frac{1}{8k \cdot 2^t} \right)^2 \cdot \frac{1}{\omega} \\
&= \frac{L_x^2}{4k^2 \omega \cdot 2^{2t}}.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\text{Term 1} &= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{y} \right)^\top \boldsymbol{P}_t^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{y} \\
&\leq \left\| \boldsymbol{B}^\top \boldsymbol{y} \right\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}} + \frac{2L_x^2}{k\omega \cdot 2^t} + \frac{L_x^2}{4k^2 \omega \cdot 2^{2t}} \\
&\leq \left\| \boldsymbol{B}^\top \boldsymbol{y} \right\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}} + \frac{3L_x^2}{k\omega \cdot 2^t}.
\end{aligned}
\tag{13}
$$

Next, we investigate Term 2. In order to bound Term 2, we first bound the minimum singular value of $\boldsymbol{P}_t$ and the maximum singular value of $\boldsymbol{Q}_t$.

Since $\boldsymbol{P}_t = \hat{\boldsymbol{B}}_t^\top \boldsymbol{B} (\sum_{i=1}^n \lambda^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B})(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B})^\top$, we have

$$
\begin{aligned}
\sigma_{\min}(\boldsymbol{P}_t) &\geq \sigma^2_{\min}(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}) \cdot \omega \\
&= \left(1 - \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp\|^2\right) \omega \\
&\geq \left(1 - \frac{1}{8^2 k^2 \cdot 2^{2t}}\right) \omega \\
&\geq \frac{3}{4} \omega.
\end{aligned}
$$

Since $\boldsymbol{Q}_t = \hat{\boldsymbol{B}}_t^\top \boldsymbol{B} (\sum_{i=1}^n \lambda^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}_\perp)(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp)^\top + \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp (\sum_{i=1}^n \lambda^*(\boldsymbol{x}_i) \boldsymbol{B}_\perp^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B})(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B})^\top + \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp (\sum_{i=1}^n \lambda^*(\boldsymbol{x}_i) \boldsymbol{B}_\perp^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}_\perp)(\hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp)^\top$, we have

$$
\begin{aligned}
\sigma_{\max}(\boldsymbol{Q}_t) &\leq 3L_x^2 \left\| \hat{\boldsymbol{B}}_t^\top \boldsymbol{B}_\perp \right\| \\
&\leq \min \left\{ \frac{3L_x^2}{8k \cdot 2^t}, \, \frac{\omega}{2} \right\}.
\end{aligned}
$$

Then, we can bound Term 2 as

$$
\begin{aligned}
\text{Term 2} &= \left(\hat{\boldsymbol{B}}_t^\top \boldsymbol{y} \right)^\top (\boldsymbol{P}_t + \boldsymbol{Q}_t)^{-1} \boldsymbol{Q}_t \boldsymbol{P}_t^{-1} \hat{\boldsymbol{B}}_t^\top \boldsymbol{y} \\
&\leq \left\| \hat{\boldsymbol{B}}_t^\top \boldsymbol{y} \right\|^2 \cdot \left\| (\boldsymbol{P}_t + \boldsymbol{Q}_t)^{-1} \right\| \cdot \|\boldsymbol{Q}_t\| \cdot \left\| \boldsymbol{P}_t^{-1} \right\| \\
&\leq \frac{4L_x^2 \cdot \sigma_{\max}(\boldsymbol{Q}_t)}{\sigma_{\min}(\boldsymbol{P}_t + \boldsymbol{Q}_t) \cdot \sigma_{\min}(\boldsymbol{P}_t)} \\
&\leq \frac{4L_x^2 \cdot \sigma_{\max}(\boldsymbol{Q}_t)}{(\sigma_{\min}(\boldsymbol{P}_t) - \sigma_{\max}(\boldsymbol{Q}_t)) \cdot \sigma_{\min}(\boldsymbol{P}_t)}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{4L_x^2 \cdot \frac{3L_x^2}{8k \cdot 2^t}}{\left(\frac{3}{4}\omega - \frac{1}{2}\omega\right) \cdot \frac{3}{4}\omega} \\
&= \frac{8L_x^4}{k\omega^2 \cdot 2^t}. \qquad (14)
\end{aligned}
$$

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

$$
\begin{aligned}
\|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} &\leq \left\|\boldsymbol{B}^\top \boldsymbol{y}\right\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}} + \frac{3L_x^2}{k\omega \cdot 2^t} + \frac{8L_x^4}{k\omega^2 \cdot 2^t}. \\
&\leq \left\|\boldsymbol{B}^\top \boldsymbol{y}\right\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i) \boldsymbol{B}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}} + \frac{11L_x^4}{k\omega^2 \cdot 2^t}.
\end{aligned}
$$

□

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 $|\hat{\mathcal{X}}_{t,m}| = 1$. Let $t_* = \max_{m \in [M]} t_m^*$ denote the total number of phases used. For any task $m \in [M]$, let $\Delta_{m,\min} := \min_{\boldsymbol{x} \in \mathcal{X} \setminus \{\boldsymbol{x}_m^*\}} (\boldsymbol{x}_m^* - \boldsymbol{x})^\top \boldsymbol{\theta}_m$ denote the minimum reward gap for task m. Let $\Delta_{\min} := \min_{m \in [M]} \Delta_{\min,m}$ 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 $\boldsymbol{x}_m^*$ will never be eliminated. (ii) $t_m^* \leq \lceil \log(\frac{4}{\Delta_{m,\min}}) \rceil + 1$, and thus, $t_* \leq \lceil \log(\frac{4}{\Delta_{\min}}) \rceil + 1$. Therefore, after at most $\lceil \log(\frac{4}{\Delta_{\min}}) \rceil + 1$ phases, algorithm DouExpDes will return the optimal arms $\boldsymbol{x}_m^*$ 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 $\tilde{O}(\frac{Mk}{\Delta_{\min}^2} \log(\delta^{-1}) + (\rho^E)^2 dk^4 L_x^2 L_w^2 D \log^4(\delta^{-1}))$.

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

$$
\begin{aligned}
&\sum_{t=1}^{t^*} pMT_t + \sum_{m=1}^{M} \sum_{t=1}^{t_m^*} N_{t,m} \\
=& \sum_{t=1}^{t^*} p \cdot O\Bigg((1+\zeta)^3 (\rho^E)^2 k^4 L_x^2 L_\theta^2 \max\left\{2^{2t}, \frac{L_x^4}{\omega^2}\right\} \log^2\left(\frac{p}{\delta_t}\right) \cdot \\
&\qquad \log^2\left((1+\zeta)\, \rho^E kpL_x L_\theta \max\left\{2^t, \frac{L_x}{\omega}\right\} \frac{1}{\delta_t} \log\left(\frac{p}{\delta_t}\right)\right)\Bigg) \\
&+ \sum_{m=1}^{M} \sum_{t=1}^{t_m^*} O\left(2^{2t}(1+\zeta)\rho_{t,m}^G \log\left(\frac{n^2 M}{\delta_t}\right) + \frac{k}{\zeta^2}\right) \\
=& \sum_{t=1}^{O(\log(\Delta_{\min}^{-1}))} O\Bigg((\rho^E)^2 k^4 dL_x^2 L_\theta^2 \max\left\{2^{2t}, \frac{L_x^4}{\omega^2}\right\} \log^2\left(\frac{d\log(\Delta_{\min}^{-1})}{\delta}\right) \cdot \\
&\qquad \log^2\left(\rho^E kdL_x L_\theta \max\left\{\Delta_{\min}^{-1}, \frac{L_x}{\omega}\right\} \frac{\log(\Delta_{\min}^{-1})}{\delta} \log\left(\frac{d\log(\Delta_{\min}^{-1})}{\delta}\right)\right)\Bigg) \\
&+ \sum_{m=1}^{M} \sum_{t=1}^{O(\log(\Delta_{m,\min}^{-1}))} O\left(2^{2t}\rho_{t,m}^G \log\left(\frac{n^2 M \log(\Delta_{m,\min}^{-1})}{\delta}\right) + k\right) \qquad (15)
\end{aligned}
$$

$$
\begin{aligned}
=&O\Bigg((\rho^E)^2 k^4 d L_x^2 L_\theta^2 \max\left\{\Delta_{\min}^{-2},\ \frac{L_x^4 \log(\Delta_{\min}^{-1})}{\omega^2}\right\} \log^2\left(\frac{d\log(\Delta_{\min}^{-1})}{\delta}\right) \cdot \\
&\qquad \log^2\left(\rho^E k d L_x L_\theta \max\left\{\Delta_{\min}^{-1},\ \frac{L_x}{\omega}\right\} \frac{\log(\Delta_{\min}^{-1})}{\delta} \log\left(\frac{d\log(\Delta_{\min}^{-1})}{\delta}\right)\right)\Bigg) \\
&+ O\left(Mk\Delta_{\min}^{-2} \log\left(\frac{n^2 M \log(\Delta_{\min}^{-1})}{\delta}\right)\right).
\end{aligned}
$$

□

Next, we prove that the sample complexity of algorithm `DouExpDes` is bounded by $\tilde{O}(\sum_{m=1}^{M} \min_{\boldsymbol{\lambda}\in\triangle_{\mathcal{X}}} \max_{\boldsymbol{x}\in\mathcal{X}\setminus\{\boldsymbol{x}_m^*\}} \frac{\|\boldsymbol{B}^\top(\boldsymbol{x}_m^*-\boldsymbol{x})\|^2_{\boldsymbol{A}(\boldsymbol{\lambda})^{-1}}}{((\boldsymbol{x}_m^*-\boldsymbol{x})^\top\boldsymbol{\theta}_m)^2} \log(\delta^{-1}) + (\rho^E)^2 d k^4 L_x^2 L_w^2 D \log^4(\delta^{-1}))$.

From Eq. (15), we have that with probability $1-\delta$, the number of samples used by algorithm `DouExpDes` is bounded by

$$
\tilde{O}\Bigg(\sum_{t=1}^{\log(\Delta_{\min}^{-1})} (\rho^E)^2 k^4 d L_x^2 L_\theta^2 \max\left\{2^{2t},\ \frac{L_x^4}{\omega^2}\right\} + \sum_{m=1}^{M} \sum_{t=1}^{\log(\Delta_{m,\min}^{-1})} 2^{2t} \rho^G_{t,m} + Mk \Bigg). \tag{16}
$$

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

$$
\begin{aligned}
\left(2^t\right)^2 \rho^G_{t,m} =& \left(2^t\right)^2 \min_{\boldsymbol{\lambda}\in\triangle_{\mathcal{X}}} \max_{\boldsymbol{y}\in\mathcal{Y}(\hat{\mathcal{X}}_{t,m})} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
\leq& \left(2^t\right)^2 \max_{\boldsymbol{y}\in\mathcal{Y}(\hat{\mathcal{X}}_{t,m})} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
\overset{\text{(a)}}{\leq}& \left(2^t\right)^2 \max_{\boldsymbol{y}\in\mathcal{Y}(\mathcal{Z}_{t,m})} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
\overset{\text{(b)}}{\leq}& 4\left(2^t\right)^2 \max_{\boldsymbol{y}\in\mathcal{Y}_m^*(\mathcal{Z}_{t,m})} \|\hat{\boldsymbol{B}}_t^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\hat{\boldsymbol{B}}_t^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \hat{\boldsymbol{B}}_t\right)^{-1}} \\
\overset{\text{(c)}}{\leq}& 4\left(2^t\right)^2 \left(\max_{\boldsymbol{y}\in\mathcal{Y}_m^*(\mathcal{Z}_{t,m})} \|\boldsymbol{B}^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\boldsymbol{B}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}} + \frac{11L_x^4}{\omega^2 k\cdot 2^t} \right) \\
=& 4\left(\frac{16\max_{\boldsymbol{y}\in\mathcal{Y}_m^*(\mathcal{Z}_{t,m})} \|\boldsymbol{B}^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\boldsymbol{B}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}}}{\left(4\cdot 2^{-t}\right)^2} + \frac{11L_x^4\cdot 2^t}{\omega^2 k} \right) \\
\overset{\text{(d)}}{\leq}& 4\left(16 \max_{\boldsymbol{y}\in\mathcal{Y}_m^*(\mathcal{Z}_{t,m})} \frac{\|\boldsymbol{B}^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\boldsymbol{B}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}}}{\left(\boldsymbol{y}^\top\boldsymbol{\theta}_m\right)^2} + \frac{11L_x^4\cdot 2^t}{\omega^2 k} \right) \\
\leq& 4\left(16 \max_{\boldsymbol{y}\in\mathcal{Y}_m^*(\mathcal{X})} \frac{\|\boldsymbol{B}^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda_m^*(\boldsymbol{x}_i)\boldsymbol{B}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}}}{\left(\boldsymbol{y}^\top\boldsymbol{\theta}_m\right)^2} + \frac{11L_x^4\cdot 2^t}{\omega^2 k} \right) \\
\overset{\text{(e)}}{=}& 4\left(16 \min_{\boldsymbol{\lambda}\in\triangle_{\mathcal{X}}} \max_{\boldsymbol{y}\in\mathcal{Y}_m^*(\mathcal{X})} \frac{\|\boldsymbol{B}^\top \boldsymbol{y}\|^2_{\left(\sum_{i=1}^n \lambda(\boldsymbol{x}_i)\boldsymbol{B}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{B}\right)^{-1}}}{\left(\boldsymbol{y}^\top\boldsymbol{\theta}_m\right)^2} + \frac{11L_x^4\cdot 2^t}{\omega^2 k} \right).
\end{aligned} \tag{17}
$$

Here inequality (a) is due to $\hat{\mathcal{X}}_{t,m}\subseteq\mathcal{Z}_{t,m}$ (from Lemma C.9). Inequality (b) uses the fact that for any $\boldsymbol{y}=\boldsymbol{x}_i-\boldsymbol{x}_j\in\mathcal{Y}(\mathcal{Z}_{t,m})$, we can write $\boldsymbol{y}=(\boldsymbol{x}_m^*-\boldsymbol{x}_j)-(\boldsymbol{x}_m^*-\boldsymbol{x}_i)$, and the triangle inequality. Inequality (c) follows from Lemma C.10, and inequality (d) is due to that for any $\boldsymbol{y}\in\mathcal{Y}_m^*(\mathcal{Z}_{t,m})$, $\boldsymbol{y}^\top\boldsymbol{\theta}_m\leq 4\cdot 2^{-t}$ (from the definition of $\mathcal{Z}_{t,m}$). Equality (e) comes from the definition of $\boldsymbol{\lambda}_m^*$.

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

$$
\begin{aligned}
& O\left(\sum_{m=1}^{M}\sum_{t=1}^{\log(\Delta_{m,\min}^{-1})} 2^{2t}\rho_{t,m}^{G}\log\left(\frac{n^2M\log(\Delta_{m,\min}^{-1})}{\delta}\right) + Mk\log(\Delta_{\min}^{-1}) + \sum_{t=1}^{\log(\Delta_{\min}^{-1})}(\rho^E)^2k^4dL_x^2L_\theta^2\max\left\{2^{2t}, \frac{L_x^4}{\omega^2}\right\}L\right) \\
=& O\Bigg(\sum_{m=1}^{M}\sum_{t=2}^{\log(\Delta_{m,\min}^{-1})}\left(\min_{\boldsymbol{\lambda}\in\triangle_{\mathcal{X}}}\max_{\boldsymbol{y}\in\mathcal{Y}_m^*(\mathcal{X})}\frac{\|\boldsymbol{B}^\top\boldsymbol{y}\|^2_{(\sum_{i=1}^n\lambda(\boldsymbol{x}_i)\boldsymbol{B}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{B})^{-1}}}{(\boldsymbol{y}^\top\boldsymbol{\theta}_m)^2} + \frac{L_x^4\cdot 2^t}{\omega^2 k}\right)\cdot\log\left(\frac{n^2M\log(\Delta_{m,\min}^{-1})}{\delta}\right) \\
& + \sum_{m=1}^{M}\rho_{1,m}^*\log\left(\frac{n^2M\log(\Delta_{m,\min}^{-1})}{\delta}\right) + Mk\log(\Delta_{\min}^{-1}) + \sum_{t=1}^{\log(\Delta_{\min}^{-1})}(\rho^E)^2k^4dL_x^2L_\theta^2\max\left\{2^{2t}, \frac{L_x^4}{\omega^2}\right\}\cdot L\Bigg) \\
\overset{(a)}{=}& O\Bigg(\sum_{m=1}^{M}\min_{\boldsymbol{\lambda}\in\triangle_{\mathcal{X}}}\max_{\boldsymbol{y}\in\mathcal{Y}_m^*(\mathcal{X})}\frac{\|\boldsymbol{B}^\top\boldsymbol{y}\|^2_{(\sum_{i=1}^n\lambda(\boldsymbol{x}_i)\boldsymbol{B}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{B})^{-1}}}{(\boldsymbol{y}^\top\boldsymbol{\theta}_m)^2}\cdot\log\left(\frac{n^2M\log(\Delta_{m,\min}^{-1})}{\delta}\right)\cdot\log(\Delta_{m,\min}^{-1}) \\
& + \frac{ML_x^4}{\omega^2k\cdot\Delta_{\min}}\cdot\log\left(\frac{n^2M\log(\Delta_{\min}^{-1})}{\delta}\right) + Mk\cdot\log\left(\frac{n^2M\log(\Delta_{\min}^{-1})}{\delta}\right)\cdot\log(\Delta_{\min}^{-1}) \\
& \quad + (\rho^E)^2k^4dL_x^2L_\theta^2\max\left\{\Delta_{\min}^{-2}, \frac{L_x^4\cdot\log(\Delta_{\min}^{-1})}{\omega^2}\right\}\cdot L\Bigg),
\end{aligned}
$$

where 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

$$
\tilde{O}\left(\sum_{m=1}^{M}\min_{\boldsymbol{\lambda}\in\triangle_{\mathcal{X}}}\max_{\boldsymbol{y}\in\mathcal{Y}_m^*(\mathcal{X})}\frac{\|\boldsymbol{B}^\top\boldsymbol{y}\|^2_{(\sum_{i=1}^n\lambda(\boldsymbol{x}_i)\boldsymbol{B}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{B})^{-1}}}{(\boldsymbol{y}^\top\boldsymbol{\theta}_m)^2}\log\left(\frac{1}{\delta}\right) + (\rho^E)^2k^4dL_x^2L_\theta^2\max\left\{\Delta_{\min}^{-2}, \frac{L_x^4}{\omega^2}\right\}\log^4\left(\frac{1}{\delta}\right)\right).
$$

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 $\lambda_{\mathcal{D}}^E$ and $\rho_{\mathcal{D}}^E$ as the optimal solution and the optimal value of the following E-optimal design optimization:

$$
\min_{\boldsymbol{\lambda}\in\triangle_{\mathcal{A}}}\left\|\left(\sum_{a\in\mathcal{A}}\lambda(a)\mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\top\right]\right)^{-1}\right\|. \tag{18}
$$

Lemma D.1. *It holds that*

$$
\rho_{\mathcal{D}}^E \le \frac{1}{\nu}.
$$

Proof of Lemma D.1. The optimization in Eq. (18) is equivalent to maximize the minimum singular value of the matrix $\sum_{a\in\mathcal{A}}\lambda(a)\mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\top\right]$.

Thus, $\lambda_{\mathcal{D}}^E$ is the optimal solution of the following optimization:

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

Using Assumption 3.3, we have

$$
\sigma_{\min}\left(\sum_{a\in\mathcal{A}}\lambda_{\mathcal{D}}^E\mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\top\right]\right) \ge \nu.
$$

Then, we have

$$\begin{aligned}\rho^E_{\mathcal{D}} =& \left\| \left(\sum_{a \in \mathcal{A}} \lambda^E_{\mathcal{D}} \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \right)^{-1} \right\| \\ =& \frac{1}{\sigma_{\min} \left(\sum_{a \in \mathcal{A}} \lambda^E_{\mathcal{D}} \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \right)} \\ \leq& \frac{1}{\nu}.\end{aligned}$$

□

Define event

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

Lemma D.2. *It holds that*

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

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

$$T_0 = \left\lceil \frac{32^2 (1+\zeta)^2 L_\phi^4}{\nu^2} \log^2 \left(\frac{20 d |\mathcal{A}|}{\delta} \right) \right\rceil,$$

we have that for any $a \in \mathcal{A}$,

$$\left\| \mathbb{E}_{s \sim \hat{\mathcal{D}}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] - \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \right\| \leq \frac{\nu}{4(1+\zeta)}.$$

Proof of Lemma D.2. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\|\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\top\| \leq L_\phi^2$. Then, using the matrix Bernstern inequality (Lemma E.2) and a union bound over $a \in \mathcal{A}$, we have that with probability $1 - \frac{\delta}{5}$, for any $a \in \mathcal{A}$,

$$\begin{aligned}\left\| \mathbb{E}_{s \sim \hat{\mathcal{D}}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] - \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \right\| \leq& 4 L_\phi^2 \sqrt{\frac{\log \left(\frac{10 \cdot 2 d |\mathcal{A}|}{\delta} \right)}{T_0}} + \frac{4 L_\phi^2 \log \left(\frac{10 \cdot 2 d |\mathcal{A}|}{\delta} \right)}{T_0} \\ \leq& \frac{8 L_\phi^2 \log \left(\frac{20 d |\mathcal{A}|}{\delta} \right)}{\sqrt{T_0}}.\end{aligned}$$

If $T_0 \geq 32^2 (1+\zeta)^2 \nu^{-2} L_\phi^4 \log^2 \left(\frac{20 d |\mathcal{A}|}{\delta} \right)$, we have

$$\left\| \mathbb{E}_{s \sim \hat{\mathcal{D}}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] - \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \right\| \leq \frac{\nu}{4(1+\zeta)},$$

which completes the proof. □

Define event

$$\begin{aligned}\mathcal{L} := \Bigg\{ & \left\| \sum_{i=1}^{p} \boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i) \boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)^\top - \sum_{i=1}^{p} \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s, \bar{a}_i) \boldsymbol{\phi}(s, \bar{a}_i)^\top \right] \right\| \leq 8 L_\phi^2 \sqrt{p} \log \left(\frac{40 d M T}{\delta} \right), \\ & \forall m \in [M],\ \forall j \in [T],\ \forall \ell \in \{1,2\} \Bigg\}.\end{aligned}$$

Lemma D.3. *It holds that*

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

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

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

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

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

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

Proof of Lemma D.3. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\|\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\top\| \leq L_\phi^2$. Then, using the matrix Bernstern 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 $\ell \in \{1,2\}$,

$$\begin{aligned}\left\| \sum_{i=1}^{p} \boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)\boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)^\top - \sum_{i=1}^{p} \mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right] \right\| &\leq 4L_\phi^2 \sqrt{p \log\left(\frac{10 \cdot 4dMT}{\delta}\right)} + 4L_\phi^2 \log\left(\frac{10 \cdot 4dMT}{\delta}\right) \\ &\leq 8L_\phi^2 \sqrt{p} \log\left(\frac{40dMT}{\delta}\right).\end{aligned}$$

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,

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

which completes the proof. $\square$

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

$$\boldsymbol{\Phi}^{(\ell)}_{m,j} = \begin{bmatrix} \boldsymbol{\phi}(s^{(\ell)}_{m,j,1}, \bar{a}_1)^\top \\ \cdots \\ \boldsymbol{\phi}(s^{(\ell)}_{m,j,p}, \bar{a}_p)^\top \end{bmatrix},$$

and

$$(\boldsymbol{\Phi}^{(\ell)}_{m,j})^+ = ((\boldsymbol{\Phi}^{(\ell)}_{m,j})^\top \boldsymbol{\Phi}^{(\ell)}_{m,j})^{-1} (\boldsymbol{\Phi}^{(\ell)}_{m,j})^\top.$$

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

$$\left\| (\boldsymbol{\Phi}^{(\ell)}_{m,j})^+ \right\| \leq 2\sqrt{\frac{(1+\zeta)}{p\nu}}.$$

Proof of Lemma D.4. We first assume that $(\mathbf{\Phi}^{(\ell)}_{m,j})^\top \mathbf{\Phi}^{(\ell)}_{m,j}$ is invertible. In our later analysis, we will prove that as long as T_0 and p are large enough, $(\mathbf{\Phi}^{(\ell)}_{m,j})^\top \mathbf{\Phi}^{(\ell)}_{m,j}$ is invertible.

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

$$\begin{aligned}
\left\|(\mathbf{\Phi}^{(\ell)}_{m,j})^{+}\right\| =& \left\|((\mathbf{\Phi}^{(\ell)}_{m,j})^\top \mathbf{\Phi}^{(\ell)}_{m,j})^{-1}(\mathbf{\Phi}^{(\ell)}_{m,j})^\top\right\| \\
=& \sqrt{\left\|((\mathbf{\Phi}^{(\ell)}_{m,j})^\top \mathbf{\Phi}^{(\ell)}_{m,j})^{-1}(\mathbf{\Phi}^{(\ell)}_{m,j})^\top \mathbf{\Phi}^{(\ell)}_{m,j}((\mathbf{\Phi}^{(\ell)}_{m,j})^\top \mathbf{\Phi}^{(\ell)}_{m,j})^{-1}\right\|} \\
=& \sqrt{\left\|((\mathbf{\Phi}^{(\ell)}_{m,j})^\top \mathbf{\Phi}^{(\ell)}_{m,j})^{-1}\right\|} \\
=& \frac{1}{\sqrt{\sigma_{\min}\left((\mathbf{\Phi}^{(\ell)}_{m,j})^\top \mathbf{\Phi}^{(\ell)}_{m,j}\right)}}. \qquad (20)
\end{aligned}$$

In addition, we have

$$\begin{aligned}
&\sigma_{\min}\left((\mathbf{\Phi}^{(\ell)}_{m,j})^\top \mathbf{\Phi}^{(\ell)}_{m,j}\right) \\
=&\sigma_{\min}\left(\sum_{i=1}^{p} \boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)\boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)^\top\right) \\
=&\sigma_{\min}\left(\sum_{i=1}^{p} \mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right] + \sum_{i=1}^{p} \boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)\boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)^\top - \sum_{i=1}^{p} \mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right]\right) \\
\geq&\sigma_{\min}\left(\sum_{i=1}^{p} \mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right]\right) - \left\|\sum_{i=1}^{p} \boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)\boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)^\top - \sum_{i=1}^{p} \mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right]\right\| \\
=&\sigma_{\min}\left(\sum_{i=1}^{p} \mathbb{E}_{s\sim\hat{\mathcal{D}}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right] + \sum_{i=1}^{p} \mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right] - \sum_{i=1}^{p} \mathbb{E}_{s\sim\hat{\mathcal{D}}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right]\right) \\
&- \left\|\sum_{i=1}^{p} \boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)\boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)^\top - \sum_{i=1}^{p} \mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right]\right\| \\
\geq&\sigma_{\min}\left(\sum_{i=1}^{p} \mathbb{E}_{s\sim\hat{\mathcal{D}}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right]\right) - \left\|\sum_{i=1}^{p} \mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right] - \sum_{i=1}^{p} \mathbb{E}_{s\sim\hat{\mathcal{D}}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right]\right\| \\
&- \left\|\sum_{i=1}^{p} \boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)\boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)^\top - \sum_{i=1}^{p} \mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right]\right\| \\
\geq&\sigma_{\min}\left(\sum_{i=1}^{p} \mathbb{E}_{s\sim\hat{\mathcal{D}}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right]\right) - \sum_{i=1}^{p} \left\|\mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right] - \mathbb{E}_{s\sim\hat{\mathcal{D}}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right]\right\| \\
&- \left\|\sum_{i=1}^{p} \boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)\boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)^\top - \sum_{i=1}^{p} \mathbb{E}_{s\sim\mathcal{D}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right]\right\| \\
\geq&\sigma_{\min}\left(\sum_{i=1}^{p} \mathbb{E}_{s\sim\hat{\mathcal{D}}}\left[\boldsymbol{\phi}(s, \bar{a}_i)\boldsymbol{\phi}(s, \bar{a}_i)^\top\right]\right) - \frac{p\nu}{4(1+\zeta)} - \frac{p\nu}{4(1+\zeta)}, \qquad (21)
\end{aligned}$$

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

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

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

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

which implies that

$$\begin{aligned}
&\sigma_{\min} \left(\sum_{i=1}^{p} \mathbb{E}_{s \sim \hat{\mathcal{D}}} \left[\boldsymbol{\phi}(s, \bar{a}_i) \boldsymbol{\phi}(s, \bar{a}_i)^\top \right] \right) \\
\geq & \frac{p}{1+\zeta} \sigma_{\min} \left(\sum_{a \in \mathcal{A}} \lambda^E(a) \mathbb{E}_{s \sim \hat{\mathcal{D}}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \right) \\
\geq & \frac{p}{1+\zeta} \sigma_{\min} \Bigg(\sum_{a \in \mathcal{A}} \lambda^E(a) \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \\
& + \sum_{a \in \mathcal{A}} \lambda^E(a) \mathbb{E}_{s \sim \hat{\mathcal{D}}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] - \sum_{a \in \mathcal{A}} \lambda^E(a) \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \Bigg) \\
\geq & \frac{p}{1+\zeta} \Bigg(\sigma_{\min} \left(\sum_{a \in \mathcal{A}} \lambda^E(a) \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \right) \\
& - \left\| \sum_{a \in \mathcal{A}} \lambda^E(a) \mathbb{E}_{s \sim \hat{\mathcal{D}}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] - \sum_{a \in \mathcal{A}} \lambda^E(a) \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \right\| \Bigg) \\
\geq & \frac{p}{1+\zeta} \left(\frac{1}{\rho^E_{\mathcal{D}}} - \sum_{a \in \mathcal{A}} \lambda^E(a) \left\| \mathbb{E}_{s \sim \hat{\mathcal{D}}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] - \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \right\| \right) \\
\overset{\text{(a)}}{\geq} & \frac{p}{1+\zeta} \left(\nu - \frac{\nu}{4(1+\zeta)} \right) \\
\geq & \frac{3p\nu}{4(1+\zeta)},
\end{aligned} \tag{22}$$

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

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

$$\begin{aligned}
\sigma_{\min} \left((\boldsymbol{\Phi}^{(\ell)}_{m,j})^\top \boldsymbol{\Phi}^{(\ell)}_{m,j} \right) \geq & \frac{3p\nu}{4(1+\zeta)} - \frac{p\nu}{4(1+\zeta)} - \frac{p\nu}{4(1+\zeta)} \\
= & \frac{p\nu}{4(1+\zeta)}.
\end{aligned} \tag{23}$$

Equations (21) and (23) show that if T_0 and p are large enough to satisfy that $\left\| \mathbb{E}_{s \sim \hat{\mathcal{D}}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] - \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s,a) \boldsymbol{\phi}(s,a)^\top \right] \right\| \leq \frac{\nu}{4(1+\zeta)}$ for any $a \in \mathcal{A}$ and $\left\| \sum_{i=1}^{p} \boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i) \boldsymbol{\phi}(s^{(\ell)}_{m,j,i}, \bar{a}_i)^\top - \sum_{i=1}^{p} \mathbb{E}_{s \sim \mathcal{D}} \left[\boldsymbol{\phi}(s, \bar{a}_i) \boldsymbol{\phi}(s, \bar{a}_i)^\top \right] \right\| \leq \frac{p\nu}{4(1+\zeta)}$ for any $m \in [M]$, $j \in [T]$ and $\ell \in \{1,2\}$, respectively, then we have that $(\boldsymbol{\Phi}^{(\ell)}_{m,j})^\top \boldsymbol{\Phi}^{(\ell)}_{m,j}$ is invertible.

Continuing with Eq. (20), we have

$$\left\| (\boldsymbol{\Phi}^{(\ell)}_{m,j})^+ \right\| \leq 2 \sqrt{\frac{(1+\zeta)}{p\nu}}.$$

□

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 $s^{(\ell)}_{m,j,i}$ and $\eta^{(\ell)}_{m,j,i}$ denote the random context and noise of the ℓ-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 $\boldsymbol{\alpha}^{(\ell)}_{m,j} \leftarrow [\alpha^{(\ell)}_{m,j,1}, \dots, \alpha^{(\ell)}_{m,j,p}]^\top$, and then, $\tilde{\boldsymbol{\theta}}^{(\ell)}_{m,j} = (\boldsymbol{\Phi}^{(\ell)}_{m,j})^+ \boldsymbol{\alpha}^{(\ell)}_{m,j}$. Recall that $\boldsymbol{Z} = \frac{1}{MT}\sum_{m=1}^M \sum_{j=1}^T \tilde{\boldsymbol{\theta}}^{(1)}_{m,j}(\tilde{\boldsymbol{\theta}}^{(2)}_{m,j})^\top$.

Lemma D.5 (Expectation of $\boldsymbol{Z}$). *It holds that*

$$\mathbb{E}\left[\boldsymbol{Z}\right] = \frac{1}{M}\sum_{m=1}^M \boldsymbol{\theta}_m \boldsymbol{\theta}_m^\top.$$

Proof of Lemma D.5. $\boldsymbol{Z}$ can be written as

$$\begin{aligned}
\boldsymbol{Z} =& \frac{1}{MT}\sum_{m=1}^M \sum_{j=1}^T \tilde{\boldsymbol{\theta}}^{(1)}_{m,j}(\tilde{\boldsymbol{\theta}}^{(2)}_{m,j})^\top \\
=& \frac{1}{MT}\sum_{m=1}^M \sum_{j=1}^T (\boldsymbol{\Phi}^{(1)}_{m,j})^+ \begin{bmatrix} \alpha^{(1)}_{m,j,1} \\ \vdots \\ \alpha^{(1)}_{m,j,p} \end{bmatrix} \left[\alpha^{(2)}_{m,j,1}, \dots, \alpha^{(2)}_{m,j,p}\right] ((\boldsymbol{\Phi}^{(2)}_{m,j})^+)^\top \\
=& \frac{1}{MT}\sum_{m=1}^M \sum_{j=1}^T (\boldsymbol{\Phi}^{(1)}_{m,j})^+ \Bigg(\begin{bmatrix} \left(\boldsymbol{\phi}(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m\right)\left(\boldsymbol{\phi}(s^{(2)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m\right) \dots & \left(\boldsymbol{\phi}(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m\right)\left(\boldsymbol{\phi}(s^{(2)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m\right) \\ \dots \quad\quad \dots & \dots \\ \left(\boldsymbol{\phi}(s^{(1)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m\right)\left(\boldsymbol{\phi}(s^{(2)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m\right) \dots & \left(\boldsymbol{\phi}(s^{(1)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m\right)\left(\boldsymbol{\phi}(s^{(2)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m\right) \end{bmatrix} \\
&+ \begin{bmatrix} \boldsymbol{\phi}(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m \cdot \eta^{(2)}_{m,j,1} + \eta^{(1)}_{m,j,1} \cdot \boldsymbol{\phi}(s^{(2)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m \dots & \boldsymbol{\phi}(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m \cdot \eta^{(2)}_{m,j,p} + \eta^{(1)}_{m,j,1} \cdot \boldsymbol{\phi}(s^{(2)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m \\ \dots \quad\quad \dots & \dots \\ \boldsymbol{\phi}(s^{(1)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m \cdot \eta^{(2)}_{m,j,1} + \eta^{(1)}_{m,j,p} \cdot \boldsymbol{\phi}(s^{(2)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m \dots & \boldsymbol{\phi}(s^{(1)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m \cdot \eta^{(2)}_{m,j,p} + \eta^{(1)}_{m,j,p} \cdot \boldsymbol{\phi}(s^{(2)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m \end{bmatrix} \\
&+ \begin{bmatrix} \eta^{(1)}_{m,j,1} \cdot \eta^{(2)}_{m,j,1} & \dots & \eta^{(1)}_{m,j,1} \cdot \eta^{(2)}_{m,j,p} \\ \dots & \dots & \dots \\ \eta^{(1)}_{m,j,p} \cdot \eta^{(2)}_{m,j,1} & \dots & \eta^{(1)}_{m,j,p} \cdot \eta^{(2)}_{m,j,p} \end{bmatrix} \Bigg) ((\boldsymbol{\Phi}^{(2)}_{m,j})^+)^\top.
\end{aligned} \tag{24}$$

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 $\boldsymbol{Z}$, we obtain

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Z}] =& \frac{1}{MT}\sum_{m=1}^M \sum_{j=1}^T \mathbb{E}\Bigg[(\boldsymbol{\Phi}^{(1)}_{m,j})^+ \cdot \\
& \begin{bmatrix} \left(\boldsymbol{\phi}(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m\right)\left(\boldsymbol{\phi}(s^{(2)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m\right) \dots & \left(\boldsymbol{\phi}(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m\right)\left(\boldsymbol{\phi}(s^{(2)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m\right) \\ \dots \quad\quad \dots & \dots \\ \left(\boldsymbol{\phi}(s^{(1)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m\right)\left(\boldsymbol{\phi}(s^{(2)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m\right) \dots & \left(\boldsymbol{\phi}(s^{(1)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m\right)\left(\boldsymbol{\phi}(s^{(2)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m\right) \end{bmatrix} ((\boldsymbol{\Phi}^{(2)}_{m,j})^+)^\top \Bigg] \\
=& \frac{1}{MT}\sum_{m=1}^M \sum_{j=1}^T \mathbb{E}\Bigg[((\boldsymbol{\Phi}^{(1)}_{m,j})^\top \boldsymbol{\Phi}^{(1)}_{m,j})^{-1} (\boldsymbol{\Phi}^{(1)}_{m,j})^\top \begin{bmatrix} \boldsymbol{\phi}(s^{(1)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m \\ \vdots \\ \boldsymbol{\phi}(s^{(1)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m \end{bmatrix} \left[\boldsymbol{\phi}(s^{(2)}_{m,j,1}, \bar{a}_1)^\top \boldsymbol{\theta}_m, \dots, \boldsymbol{\phi}(s^{(2)}_{m,j,p}, \bar{a}_p)^\top \boldsymbol{\theta}_m \right] \cdot \\
& \boldsymbol{\Phi}^{(2)}_{m,j} ((\boldsymbol{\Phi}^{(2)}_{m,j})^\top \boldsymbol{\Phi}^{(2)}_{m,j})^{-1} \Bigg] \\
=& \frac{1}{MT}\sum_{m=1}^M \sum_{j=1}^T \mathbb{E}\left[((\boldsymbol{\Phi}^{(1)}_{m,j})^\top \boldsymbol{\Phi}^{(1)}_{m,j})^{-1} (\boldsymbol{\Phi}^{(1)}_{m,j})^\top \cdot \boldsymbol{\Phi}^{(1)}_{m,j} \boldsymbol{\theta}_m (\boldsymbol{\theta}_m)^\top (\boldsymbol{\Phi}^{(2)}_{m,j})^\top \cdot \boldsymbol{\Phi}^{(2)}_{m,j} ((\boldsymbol{\Phi}^{(2)}_{m,j})^\top \boldsymbol{\Phi}^{(2)}_{m,j})^{-1} \right] \\
=& \frac{1}{MT}\sum_{m=1}^M \sum_{j=1}^T \boldsymbol{\theta}_m (\boldsymbol{\theta}_m)^\top
\end{aligned}$$

$$=\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\theta}_m\boldsymbol{\theta}_m^\top.$$

□

Define event

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

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

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

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

$$\boldsymbol{D}_{m,j} := \frac{1}{MT}(\boldsymbol{\Phi}^{(1)}_{m,j})^{+}\cdot \begin{bmatrix} \left(\boldsymbol{\phi}(s^{(1)}_{m,j,1},\bar{a}_1)^\top\boldsymbol{\theta}_m\right)\left(\boldsymbol{\phi}(s^{(2)}_{m,j,1},\bar{a}_1)^\top\boldsymbol{\theta}_m\right) & \dots & \left(\boldsymbol{\phi}(s^{(1)}_{m,j,1},\bar{a}_1)^\top\boldsymbol{\theta}_m\right)\left(\boldsymbol{\phi}(s^{(2)}_{m,j,p},\bar{a}_p)^\top\boldsymbol{\theta}_m\right) \\ \dots & \dots & \dots \\ \left(\boldsymbol{\phi}(s^{(1)}_{m,j,p},\bar{a}_p)^\top\boldsymbol{\theta}_m\right)\left(\boldsymbol{\phi}(s^{(2)}_{m,j,1},\bar{a}_1)^\top\boldsymbol{\theta}_m\right) & \dots & \left(\boldsymbol{\phi}(s^{(1)}_{m,j,p},\bar{a}_p)^\top\boldsymbol{\theta}_m\right)\left(\boldsymbol{\phi}(s^{(2)}_{m,j,p},\bar{a}_p)^\top\boldsymbol{\theta}_m\right) \end{bmatrix} ((\boldsymbol{\Phi}^{(2)}_{m,j})^{+})^\top$$

$$\boldsymbol{D} := \sum_{m=1}^{M}\sum_{j=1}^{T}\boldsymbol{D}_{m,j}$$

$$\boldsymbol{E}_{m,j} := \frac{1}{MT}(\boldsymbol{\Phi}^{(1)}_{m,j})^{+}\cdot \begin{bmatrix} \boldsymbol{\phi}(s^{(1)}_{m,j,1},\bar{a}_1)^\top\boldsymbol{\theta}_m\cdot\eta^{(2)}_{m,j,1}+\eta^{(1)}_{m,j,1}\cdot\boldsymbol{\phi}(s^{(2)}_{m,j,1},\bar{a}_1)^\top\boldsymbol{\theta}_m & \dots & \boldsymbol{\phi}(s^{(1)}_{m,j,1},\bar{a}_1)^\top\boldsymbol{\theta}_m\cdot\eta^{(2)}_{m,j,p}+\eta^{(1)}_{m,j,1}\cdot\boldsymbol{\phi}(s^{(2)}_{m,j,p},\bar{a}_p)^\top\boldsymbol{\theta}_m \\ \dots & \dots & \dots \\ \boldsymbol{\phi}(s^{(1)}_{m,j,p},\bar{a}_p)^\top\boldsymbol{\theta}_m\cdot\eta^{(2)}_{m,j,1}+\eta^{(1)}_{m,j,p}\cdot\boldsymbol{\phi}(s^{(2)}_{m,j,1},\bar{a}_1)^\top\boldsymbol{\theta}_m & \dots & \boldsymbol{\phi}(s^{(1)}_{m,j,p},\bar{a}_p)^\top\boldsymbol{\theta}_m\cdot\eta^{(2)}_{m,j,p}+\eta^{(1)}_{m,j,p}\cdot\boldsymbol{\phi}(s^{(2)}_{m,j,p},\bar{a}_p)^\top\boldsymbol{\theta}_m \end{bmatrix}\cdot ((\boldsymbol{\Phi}^{(2)}_{m,j})^{+})^\top$$

$$\boldsymbol{E} := \sum_{m=1}^{M}\sum_{j=1}^{T}\boldsymbol{E}_{m,j}$$

$$\boldsymbol{F}_{m,j} := \frac{1}{MT}(\boldsymbol{\Phi}^{(1)}_{m,j})^{+} \begin{bmatrix} \eta^{(1)}_{m,j,1}\cdot\eta^{(2)}_{m,j,1} & \dots & \eta^{(1)}_{m,j,1}\cdot\eta^{(2)}_{m,j,p} \\ \dots & \dots & \dots \\ \eta^{(1)}_{m,j,p}\cdot\eta^{(2)}_{m,j,1} & \dots & \eta^{(1)}_{m,j,p}\cdot\eta^{(2)}_{m,j,p} \end{bmatrix} ((\boldsymbol{\Phi}^{(2)}_{m,j})^{+})^\top$$

$$\boldsymbol{F} := \sum_{m=1}^{M}\sum_{j=1}^{T}\boldsymbol{F}_{m,j}$$

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

$$\|\boldsymbol{Z}-\mathbb{E}[\boldsymbol{Z}]\| \leq \|\boldsymbol{D}-\mathbb{E}[\boldsymbol{D}]\|+\|\boldsymbol{E}-\mathbb{E}[\boldsymbol{E}]\|+\|\boldsymbol{F}-\mathbb{E}[\boldsymbol{F}]\|. \tag{25}$$

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 $\tilde{\eta}^{(\ell)}_{m,j,i}=\eta^{(\ell)}_{m,j,i}\mathbb{1}\{|\eta^{(\ell)}_{m,j,i}|\leq R\}$ denote the truncated noise. Furthermore, we define the following matrices with truncated noises:

$$\tilde{\boldsymbol{E}}_{m,j} := \frac{1}{MT}(\boldsymbol{\Phi}^{(1)}_{m,j})^{+}\cdot$$

$$\begin{bmatrix} \phi(s^{(1)}_{m,j,1},\bar{a}_1)^\top \theta_m \cdot \tilde{\eta}^{(2)}_{m,j,1} + \tilde{\eta}^{(1)}_{m,j,1} \cdot \phi(s^{(2)}_{m,j,1},\bar{a}_1)^\top \theta_m \dots \phi(s^{(1)}_{m,j,1},\bar{a}_1)^\top \theta_m \cdot \tilde{\eta}^{(2)}_{m,j,p} + \tilde{\eta}^{(1)}_{m,j,1} \cdot \phi(s^{(2)}_{m,j,p},\bar{a}_p)^\top \theta_m \\ \dots \qquad \dots \qquad \dots \\ \phi(s^{(1)}_{m,j,p},\bar{a}_p)^\top \theta_m \cdot \tilde{\eta}^{(2)}_{m,j,1} + \tilde{\eta}^{(1)}_{m,j,p} \cdot \phi(s^{(2)}_{m,j,1},\bar{a}_1)^\top \theta_m \dots \phi(s^{(1)}_{m,j,p},\bar{a}_p)^\top \theta_m \cdot \tilde{\eta}^{(2)}_{m,j,p} + \tilde{\eta}^{(1)}_{m,j,p} \cdot \phi(s^{(2)}_{m,j,p},\bar{a}_p)^\top \theta_m \end{bmatrix}.$$

$$((\Phi^{(2)}_{m,j})^+)^\top$$

$$\tilde{E} := \sum_{m=1}^{M} \sum_{j=1}^{T} \tilde{E}_{m,j}$$

$$\tilde{F}_{m,j} := \frac{1}{MT} (\Phi^{(1)}_{m,j})^+ \begin{bmatrix} \tilde{\eta}^{(1)}_{m,j,1} \cdot \tilde{\eta}^{(2)}_{m,j,1} & \dots & \tilde{\eta}^{(1)}_{m,j,1} \cdot \tilde{\eta}^{(2)}_{m,j,p} \\ \dots & \dots & \dots \\ \tilde{\eta}^{(1)}_{m,j,p} \cdot \tilde{\eta}^{(2)}_{m,j,1} & \dots & \tilde{\eta}^{(1)}_{m,j,p} \cdot \tilde{\eta}^{(2)}_{m,j,p} \end{bmatrix} ((\Phi^{(2)}_{m,j})^+)^\top$$

$$\tilde{F} := \sum_{m=1}^{M} \sum_{j=1}^{T} \tilde{F}_{m,j}$$

Recall that from Lemma D.4, we have that for any $m \in [M]$, $j \in [T]$ and $\ell \in \{1,2\}$, $\|(\Phi^{(\ell)}_{m,j})^+\| \leq 2\sqrt{\frac{(1+\zeta)}{p\nu}}$. Let $B_\Phi := 2\sqrt{\frac{(1+\zeta)}{p\nu}}$.

We first analyze $\|D - \mathbb{E}[D]\|$. Since $|\phi(s^{(\ell)}_{m,j,i},\bar{a}_i)^\top \theta_m| \leq L_\phi L_\theta$ for any $m \in [M]$, $j \in [T]$, $i \in [p]$ and $\ell \in \{1,2\}$, we have that $\|D_{m,j}\| \leq \frac{1}{MT} \cdot pL_\phi L_\theta B^2_\Phi$ and $\|\sum_{m=1}^M \sum_{j=1}^T \mathbb{E}[D^2_{m,j}]\| \leq MT \cdot \frac{1}{M^2T^2} \cdot p^2 L^2_\phi L^2_\theta B^4_\Phi = \frac{1}{MT} \cdot p^2 L^2_\phi L^2_\theta B^4_\Phi$ 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 $1-\delta'$,

$$\begin{aligned} \|D - \mathbb{E}[D]\| \leq & 4\sqrt{\frac{p^2 L^2_\phi L^2_\theta B^4_\Phi \log\left(\frac{2d}{\delta'}\right)}{MT}} + \frac{4pL_\phi L_\theta B^2_\Phi \log\left(\frac{2d}{\delta'}\right)}{MT} \\ \leq & \frac{8 \cdot 4pL_\phi L_\theta B^2_\Phi \log\left(\frac{2d}{\delta'}\right)}{\sqrt{MT}}. \end{aligned} \tag{26}$$

Next, we bound $\|E - \mathbb{E}[E]\|$. Since $|\phi(s^{(\ell)}_{m,j,i},\bar{a}_i)^\top \theta_m| \leq L_\phi L_\theta$ and $|\tilde{\eta}^{(\ell)}_{m,j,i}| \leq R$ for any $m \in [M]$, $j \in [T]$, $i \in [p]$ and $\ell \in \{1,2\}$, we have that $\|\tilde{E}_{m,j}\| \leq \frac{1}{MT} \cdot 2pRL_\phi L_\theta B^2_\Phi$ and $\left\|\sum_{m=1}^M \sum_{j=1}^T \mathbb{E}[\tilde{E}^2_{m,j}]\right\| \leq \frac{1}{MT} \cdot 4p^2R^2 L^2_\phi L^2_\theta B^4_\Phi$ for any $m \in [M]$ and $j \in [T]$.

Since $\eta^{(\ell)}_{m,j,i}$ 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 $1 - 4p\exp(-\frac{R^2}{2})$, $|\eta^{(\ell)}_{m,j,i}| \leq R$ for all $i \in [p]$ and $\ell \in \{1,2\}$, and thus, $\|E_{m,j}\| \leq \frac{1}{MT} \cdot 2pRL_\phi L_\theta B^2_\Phi$. Then, we have

$$\begin{aligned} & \left\|\mathbb{E}[E_{m,j}] - \mathbb{E}[\tilde{E}_{m,j}]\right\| \\ \leq & \left\|\mathbb{E}\left[E_{m,j} \cdot \mathbb{1}\left\{\|E_{m,j}\| \geq \frac{2pRL_\phi L_\theta B^2_\Phi}{MT}\right\}\right]\right\| \\ \leq & \mathbb{E}\left[\|E_{m,j}\| \cdot \mathbb{1}\left\{\|E_{m,j}\| \geq \frac{2pRL_\phi L_\theta B^2_\Phi}{MT}\right\}\right] \\ = & \mathbb{E}\left[\frac{2pRL_\phi L_\theta B^2_\Phi}{MT} \cdot \mathbb{1}\left\{\|E_{m,j}\| \geq \frac{2pRL_\phi L_\theta B^2_\Phi}{MT}\right\}\right] + \mathbb{E}\left[\left(\|E_{m,j}\| - \frac{2pRL_\phi L_\theta B^2_\Phi}{MT}\right) \cdot \mathbb{1}\left\{\|E_{m,j}\| \geq \frac{2pRL_\phi L_\theta B^2_\Phi}{MT}\right\}\right] \\ = & \frac{2pRL_\phi L_\theta B^2_\Phi}{MT} \cdot \Pr\left[\|E_{m,j}\| \geq \frac{2pRL_\phi L_\theta B^2_\Phi}{MT}\right] + \int_0^\infty \Pr\left[\|E_{m,j}\| - \frac{2pRL_\phi L_\theta B^2_\Phi}{MT} > x\right] dx \end{aligned}$$

$$
\begin{aligned}
&\le \frac{2pRL_\phi L_\theta B_\Phi^2}{MT} \cdot 4p \cdot \exp\left(-\frac{R^2}{2}\right) + \frac{2pL_\phi L_\theta B_\Phi^2}{MT} \int_R^\infty \Pr\left[\|\boldsymbol{E}_{m,j}\| > \frac{2pL_\phi L_\theta B_\Phi^2 y}{MT}\right] dy \\
&\le \frac{2pRL_\phi L_\theta B_\Phi^2}{MT} \cdot 4p \cdot \exp\left(-\frac{R^2}{2}\right) + \frac{2pL_\phi L_\theta B_\Phi^2}{MT} \int_R^\infty 4p \exp\left(-\frac{y^2}{2}\right) dy \\
&\le \frac{2pRL_\phi L_\theta B_\Phi^2}{MT} \cdot 4p \cdot \exp\left(-\frac{R^2}{2}\right) + \frac{2pL_\phi L_\theta B_\Phi^2}{MT} \cdot 4p \cdot \frac{1}{R} \cdot \exp\left(-\frac{R^2}{2}\right) \\
&= \frac{2pL_\phi L_\theta B_\Phi^2}{MT} \cdot 4p \cdot \left(R + \frac{1}{R}\right) \exp\left(-\frac{R^2}{2}\right).
\end{aligned}
$$

Using the truncated matrix Bernstein inequality (Lemma E.2) with $n = MT$, $R = \sqrt{2\log\left(\frac{4pMT}{\delta'}\right)}$, $U = \frac{2pL_\phi L_\theta B_\Phi^2 \sqrt{2\log\left(\frac{4pMT}{\delta'}\right)}}{MT}$, $\sigma^2 = \frac{(2pL_\phi L_\theta B_\Phi^2 \sqrt{2\log\left(\frac{4pMT}{\delta'}\right)})^2}{MT}$, $\tau = 4\sqrt{\frac{(2pL_\phi L_\theta B_\Phi^2 \sqrt{2\log\left(\frac{4pMT}{\delta'}\right)})^2 \cdot \log\left(\frac{2d}{\delta'}\right)}{MT}} + \frac{4 \cdot 2pL_\phi L_\theta B_\Phi^2 \sqrt{2\log\left(\frac{4pMT}{\delta'}\right)} \cdot \log\left(\frac{2d}{\delta'}\right)}{MT}$ and $\Delta = \frac{2pL_\phi L_\theta B_\Phi^2 \cdot 2\sqrt{2\log\left(\frac{4pMT}{\delta'}\right)}}{MT} \cdot \frac{\delta'}{MT}$, we have that with probability at least $1 - 2\delta'$,

$$
\|\boldsymbol{E} - \mathbb{E}[\boldsymbol{E}]\| \le \frac{8 \cdot 2pL_\phi L_\theta B_\Phi^2 \sqrt{2\log\left(\frac{4pMT}{\delta'}\right)} \cdot \log\left(\frac{2d}{\delta'}\right)}{\sqrt{MT}}. \tag{27}
$$

Now we investigate $\|\boldsymbol{F} - \mathbb{E}[\boldsymbol{F}]\|$. Since $|\tilde{\eta}_{m,j,i}^{(\ell)}| \le R$ for any $m \in [M]$, $j \in [T]$, $i \in [p]$ and $\ell \in \{1, 2\}$, we have that $\|\tilde{\boldsymbol{F}}_{m,j}\| \le \frac{1}{MT} \cdot pR^2 B_\Phi^2$ and $\left\|\sum_{m=1}^M \sum_{j=1}^T \mathbb{E}\left[\tilde{\boldsymbol{F}}_{m,j}^2\right]\right\| \le \frac{1}{MT} \cdot p^2 R^4 B_\Phi^4$.

Recall that for any $m \in [M]$ and $j \in [T]$, with probability at least $1 - 4p\exp(-\frac{R^2}{2})$, $|\eta_{m,j,i}^{(\ell)}| \le R$ for all $i \in [p]$ and $\ell \in \{1, 2\}$, and thus, $\|\boldsymbol{F}_{m,j}\| \le \frac{1}{MT} \cdot pB_\Phi^2 R^2$. Then, we have

$$
\begin{aligned}
\left\|\mathbb{E}[\boldsymbol{F}_{m,j}] - \mathbb{E}[\tilde{\boldsymbol{F}}_{m,j}]\right\| \le& \left\|\mathbb{E}\left[\boldsymbol{F}_{m,j} \cdot \mathbb{1}\left\{\|\boldsymbol{F}_{m,j}\| \ge \frac{pB_\Phi^2 R^2}{MT}\right\}\right]\right\| \\
\le& \mathbb{E}\left[\|\boldsymbol{F}_{m,j}\| \cdot \mathbb{1}\left\{\|\boldsymbol{F}_{m,j}\| \ge \frac{pB_\Phi^2 R^2}{MT}\right\}\right] \\
=& \mathbb{E}\left[\frac{pB_\Phi^2 R^2}{MT} \cdot \mathbb{1}\left\{\|\boldsymbol{F}_{m,j}\| \ge \frac{pB_\Phi^2 R^2}{MT}\right\}\right] + \left[\left(\|\boldsymbol{F}_{m,j}\| - \frac{pB_\Phi^2 R^2}{MT}\right) \cdot \mathbb{1}\left\{\|\boldsymbol{F}_{m,j}\| \ge \frac{pB_\Phi^2 R^2}{MT}\right\}\right] \\
=& \frac{pB_\Phi^2 R^2}{MT} \cdot \Pr\left[\|\boldsymbol{F}_{m,j}\| \ge \frac{pB_\Phi^2 R^2}{MT}\right] + \int_0^\infty \Pr\left[\|\boldsymbol{F}_{m,j}\| - \frac{pB_\Phi^2 R^2}{MT} > x\right] dx \\
\le& \frac{pB_\Phi^2 R^2}{MT} \cdot 4p \cdot \exp\left(-\frac{R^2}{2}\right) + \frac{2pB_\Phi^2}{MT} \int_R^\infty \boldsymbol{y} \cdot \Pr\left[\|\boldsymbol{F}_{m,j}\| > \frac{pB_\Phi^2 y^2}{MT}\right] dy \\
\le& \frac{pB_\Phi^2 R^2}{MT} \cdot 4p \cdot \exp\left(-\frac{R^2}{2}\right) + \frac{2pB_\Phi^2}{MT} \int_R^\infty \boldsymbol{y} \cdot 4p \exp\left(-\frac{y^2}{2}\right) dy \\
\le& \frac{pB_\Phi^2 R^2}{MT} \cdot 4p \cdot \exp\left(-\frac{R^2}{2}\right) + \frac{2pB_\Phi^2}{MT} \cdot 4p \cdot \exp\left(-\frac{R^2}{2}\right) \\
=& \frac{pB_\Phi^2}{MT} \cdot 4p \cdot \left(R^2 + 2\right) \exp\left(-\frac{R^2}{2}\right).
\end{aligned}
$$

Using the truncated matrix Bernstein inequality (Lemma E.2) with $n = MT$, $R = \sqrt{2\log\left(\frac{4pMT}{\delta'}\right)}$, $U = \frac{pB_\Phi^2 \cdot 2\log\left(\frac{4pMT}{\delta'}\right)}{MT}$, $\sigma^2 = \frac{(pB_\Phi^2 \cdot 2\log\left(\frac{4pMT}{\delta'}\right))^2}{MT}$, $\tau = 4\sqrt{\frac{(pB_\Phi^2 \cdot 2\log\left(\frac{4pMT}{\delta'}\right))^2 \cdot \log\left(\frac{2d}{\delta'}\right)}{MT}} + \frac{4 \cdot pB_\Phi^2 \cdot 2\log\left(\frac{4pMT}{\delta'}\right) \cdot \log\left(\frac{2d}{\delta'}\right)}{MT}$ and $\Delta = \frac{pB_\Phi^2 \cdot 2 \cdot 2\log\left(\frac{4pMT}{\delta'}\right)}{MT} \cdot \frac{\delta'}{MT}$, we have that with probability at least $1 - 2\delta'$,

$$
\|\boldsymbol{F} - \mathbb{E}\left[\boldsymbol{F}\right]\| \le \frac{8 \cdot pB_\Phi^2 \cdot 2\log\left(\frac{4pMT}{\delta'}\right) \cdot \log\left(\frac{2d}{\delta'}\right)}{\sqrt{MT}}. \tag{28}
$$

Plugging Eqs. (26)-(28) into Eq. (25), we have that with probability at least $1-5\delta'$,

$$\begin{aligned}\|\boldsymbol{Z}-\mathbb{E}[\boldsymbol{Z}]\| &\leq \|\boldsymbol{D}-\mathbb{E}\left[\boldsymbol{D}\right]\|+\|\boldsymbol{E}-\mathbb{E}\left[\boldsymbol{E}\right]\|+\|\boldsymbol{F}-\mathbb{E}\left[\boldsymbol{F}\right]\| \\ &\leq \frac{64pL_\phi L_\theta B_\Phi^2 \log\left(\frac{4pMT}{\delta'}\right)\log\left(\frac{2d}{\delta'}\right)}{\sqrt{MT}}.\end{aligned}$$

Let $\delta'=\frac{\delta}{25}$. Recall that $B_\Phi := 2\sqrt{\frac{(1+\zeta)}{p\nu}}$. Then, we obtain that with probability at least $1-\frac{\delta}{5}$,

$$\|\boldsymbol{Z}-\mathbb{E}[\boldsymbol{Z}]\| \leq \frac{256(1+\zeta)L_\phi L_\theta \log\left(\frac{50d}{\delta}\right)}{\nu\sqrt{MT}}\log\left(\frac{100pMT}{\delta}\right),$$

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 $\sigma_{\min}(\frac{1}{M}\sum_{m=1}^M \boldsymbol{w}_m\boldsymbol{w}_m^\top) = \sigma_{\min}(\frac{1}{M}\sum_{m=1}^M \boldsymbol{\theta}_m\boldsymbol{\theta}_m^\top)\geq \frac{c_0}{k}$.

Lemma D.7 (Concentration of $\hat{\boldsymbol{B}}$)**.** *Suppose that event $\mathcal{G}$ holds. Then,*

$$\left\|\hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B}\right\| \leq \frac{2048(1+\zeta)kL_\phi L_\theta\log\left(\frac{50d}{\delta}\right)}{c_0\nu\sqrt{MT}}\log\left(\frac{135(1+\zeta)dL_\phi MT}{\nu\delta}\right).$$

Furthermore, if

$$T=\left\lceil \frac{68\cdot 2048^2\cdot 96^2(1+\zeta)^2k^4L_\phi^4L_\theta^2L_w^2}{c_0^2\nu^2\varepsilon^2M}\log^6\left(\frac{2048\cdot 135\cdot 96\cdot 50\cdot 5(1+\zeta)^2k^2d^2L_\phi^3L_\theta L_w N}{c_0\nu^2\delta^3\varepsilon}\right)\right\rceil, \tag{29}$$

we have

$$\left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| \leq \frac{\varepsilon}{96k\log\left(\frac{5N}{\delta}\right)L_\phi L_w}.$$

Proof of Lemma D.7. First, we have that $\sigma_k(\mathbb{E}[\boldsymbol{Z}])-\sigma_{k+1}(\mathbb{E}[\boldsymbol{Z}])=\sigma_{\min}(\frac{1}{M}\sum_{m=1}^M\boldsymbol{\theta}_m\boldsymbol{\theta}_m^\top)\geq\frac{c_0}{k}$. Let $p:=\lceil 32^2(1+\zeta)^2\nu^{-2}L_\phi^4\log^2\left(\frac{40dMT}{\delta}\right)\rceil$. Then, using the Davis-Kahan $\sin\theta$ Theorem (Bhatia, 2013) and letting T_t be large enough to satisfy that $\|\boldsymbol{Z}-\mathbb{E}[\boldsymbol{Z}]\|\leq\frac{c_0}{2k}$, we have

$$\begin{aligned}\left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| &\leq \frac{\|\boldsymbol{Z}-\mathbb{E}[\boldsymbol{Z}]\|}{\sigma_k(\mathbb{E}[\boldsymbol{Z}])-\sigma_{k+1}(\mathbb{E}[\boldsymbol{Z}])-\|\boldsymbol{Z}-\mathbb{E}[\boldsymbol{Z}]\|} \\ &\leq \frac{2k}{c_0}\|\boldsymbol{Z}-\mathbb{E}[\boldsymbol{Z}]\| \\ &\leq \frac{512(1+\zeta)kL_\phi L_\theta\log\left(\frac{50d}{\delta}\right)}{c_0\nu\sqrt{MT}}\log\left(\frac{100pMT}{\delta}\right) \\ &\leq \frac{512(1+\zeta)kL_\phi L_\theta\log\left(\frac{50d}{\delta}\right)}{c_0\nu\sqrt{MT}}\log\left(\frac{100MT}{\delta}\cdot\frac{2\cdot 32^2(1+\zeta)^2L_\phi^4}{\nu^2}\log^2\left(\frac{40dMT}{\delta}\right)\right) \\ &\leq \frac{512(1+\zeta)kL_\phi L_\theta\log\left(\frac{50d}{\delta}\right)}{c_0\nu\sqrt{MT}}\log\left(\frac{2\cdot 100\cdot 32^2\cdot 40^2(1+\zeta)^2d^2L_\phi^4M^3T^3}{\nu^2\delta^3}\right) \\ &\leq \frac{2048(1+\zeta)kL_\phi L_\theta\log\left(\frac{50d}{\delta}\right)}{c_0\nu\sqrt{MT}}\log\left(\frac{135(1+\zeta)dL_\phi MT}{\nu\delta}\right).\end{aligned}$$

Using Lemma E.3 with $A=2048(1+\zeta)kc_0^{-1}\nu^{-1}L_\phi L_\theta\log\left(\frac{50d}{\delta}\right)$, $B=\frac{135(1+\zeta)dL_\phi}{\nu\delta}$ and $\kappa=\frac{\varepsilon}{96k\log\left(\frac{5N}{\delta}\right)L_\phi L_w}$, we have that if

$$MT\geq\frac{68\cdot 2048^2\cdot 96^2(1+\zeta)^2k^4L_\phi^4L_\theta^2L_w^2}{c_0^2\nu^2\varepsilon^2}.$$

$$\log^2\left(\frac{50d}{\delta}\right)\log^2\left(\frac{5N}{\delta}\right)\log^2\left(\frac{2048\cdot 135\cdot 96(1+\zeta)^2k^2dL_\phi^3L_\theta L_w}{c_0\nu^2\delta\varepsilon}\log\left(\frac{50d}{\delta}\right)\log\left(\frac{5N}{\delta}\right)\right),$$

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

Further enlarging MT, if

$$MT\geq\frac{68\cdot 2048^2\cdot 96^2(1+\zeta)^2k^4L_\phi^4L_\theta^2L_w^2}{c_0^2\nu^2\varepsilon^2}\log^6\left(\frac{2048\cdot 135\cdot 96\cdot 50\cdot 5(1+\zeta)^2k^2d^2L_\phi^3L_\theta L_wN}{c_0\nu^2\delta^3\varepsilon}\right),$$

then

$$\left\|\hat{\boldsymbol{B}}_{t,\perp}^\top\boldsymbol{B}\right\|\leq\frac{\varepsilon}{96k\log\left(\frac{5N}{\delta}\right)L_\phi L_w}.$$

□

D.3. Estimation with Low-dimensional Representations

Lemma D.8. *In subroutine* `EstLowRep` *(Algorithm 6), for any $m\in[M]$ and $t>0$, we have*

$$\log\left(\frac{\det\left(\gamma I+\sum_{\tau=1}^t\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\hat{\boldsymbol{B}}\right)}{\det\left(\gamma I\right)}\right)\leq k\log\left(1+\frac{t}{\gamma k}\right).$$

Proof of Lemma D.8. This proof uses a similar idea as Lemma 11 in (Abbasi-Yadkori et al., 2011).

It holds that

$$\begin{aligned}
&\log\left(\frac{\det\left(\gamma I+\sum_{\tau=1}^t\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\hat{\boldsymbol{B}}\right)}{\det\left(\gamma I\right)}\right)\\
\leq&\log\left(\frac{\left(\frac{\mathrm{Trace}\left(\gamma I+\sum_{\tau=1}^t\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\hat{\boldsymbol{B}}\right)}{k}\right)^k}{\gamma^k}\right)\\
=&k\log\left(\frac{\mathrm{Trace}\left(\gamma I\right)+\sum_{\tau=1}^t\mathrm{Trace}\left(\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\hat{\boldsymbol{B}}\right)}{\gamma k}\right)\\
=&k\log\left(\frac{\gamma k+\sum_{\tau=1}^t\left\|\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\right\|^2}{\gamma k}\right)\\
\leq&k\log\left(1+\frac{t}{\gamma k}\right).
\end{aligned}$$

□

Lemma D.9. *In subroutine* `EstLowRep` *(Algorithm 6), for any $m\in[M]$ and $t\geq 0$, we have*

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

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 $\boldsymbol{\Sigma}_{m,t+1} \succeq \boldsymbol{\Sigma}_{m,t}$, we have $\boldsymbol{\Sigma}_{m,t}^{-1} \succeq \boldsymbol{\Sigma}_{m,t+1}^{-1}$. Hence, for any $m \in [M]$, $t \geq 0$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have

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

which implies that

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

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

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

□

In subroutine `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

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

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 δ' be a confidence parameter which will be chosen later. Since $\hat{\boldsymbol{B}}$ 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 $t > 0$,

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

where inequality (a) uses Lemma D.8.

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

Define event

$$\mathcal{J} := \left\{ \sum_{t=1}^{N} \mathbb{E}_{s \sim \mathcal{D}} \left[\max_{a \in \mathcal{A}} \left\| \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a) \right\|_{\boldsymbol{\Sigma}_{t-1}^{-1}} \right] \leq \right.$$

$$\frac{1}{4}\left(2\sqrt{\log\left(\frac{5}{\delta}\right)}+\sqrt{4\log\left(\frac{5}{\delta}\right)+4\left(\sum_{t=1}^{N}\max_{a\in\mathcal{A}}\left\|\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_t,a)\right\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}+2\log\left(\frac{5}{\delta}\right)\right)}\right)^2\Bigg\}.$$

Lemma D.11. *It holds that*

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

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

Lemma D.12. *Suppose that event $\mathcal{K}\cap\mathcal{L}\cap\mathcal{G}\cap\mathcal{H}\cap\mathcal{J}$ holds. For any task $m\in[M]$, we have*

$$\mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}}\left|\boldsymbol{\phi}(s,a)^\top\left(\hat{\boldsymbol{\theta}}_{m,N}-\boldsymbol{\theta}_m\right)\right|\right]\leq\left(2\sqrt{\frac{2k\log\left(1+\frac{N}{\gamma k}\right)}{N}}+\frac{8\log\left(\frac{5}{\delta}\right)}{N}\right)\cdot$$
$$\left(\left\|\hat{\boldsymbol{B}}_\perp^\top\boldsymbol{B}\right\|\sqrt{Nk}+\sqrt{k\log\left(1+\frac{N}{\gamma k}\right)+2\log\left(\frac{5}{\delta}\right)}+\sqrt{\gamma}\right)+\left\|\hat{\boldsymbol{B}}_\perp^\top\boldsymbol{B}\right\|.$$

Furthermore, if

$$N=\left\lceil\frac{4^2\cdot 26^4\cdot 24^2\cdot 2\left(k^2+k\gamma L_\theta^2\right)\log^4\left(\frac{240(k+\sqrt{k\gamma}L_\theta)}{\varepsilon\delta}\right)}{\varepsilon^2}\right\rceil,$$

then

$$\mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}}\left|\boldsymbol{\phi}(s,a)^\top\left(\hat{\boldsymbol{\theta}}_{m,N}-\boldsymbol{\theta}_m\right)\right|\right]\leq\frac{\varepsilon}{2}$$

Proof of Lemma D.12. For any task $m\in[M]$ and $t\in[N]$,

$$\begin{aligned}
\hat{\boldsymbol{w}}_{m,t}=&\boldsymbol{\Sigma}_{m,t}^{-1}\sum_{\tau=1}^{t}\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})r_{m,\tau}\\
=&\left(\gamma I+\sum_{\tau=1}^{t}\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\hat{\boldsymbol{B}}\right)^{-1}\sum_{\tau=1}^{t}\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\left(\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\boldsymbol{\theta}_m+\xi_{m,j}\right)\\
=&\left(\gamma I+\sum_{\tau=1}^{t}\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\hat{\boldsymbol{B}}\right)^{-1}\sum_{\tau=1}^{t}\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\cdot\\
&\left(\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top\boldsymbol{\theta}_m+\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\hat{\boldsymbol{B}}_\perp\hat{\boldsymbol{B}}_\perp^\top\boldsymbol{\theta}_m+\xi_{m,j}\right)\\
&+\gamma\left(\gamma I+\sum_{\tau=1}^{t}\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\hat{\boldsymbol{B}}\right)^{-1}\hat{\boldsymbol{B}}^\top\boldsymbol{\theta}_m\\
&-\gamma\left(\gamma I+\sum_{\tau=1}^{t}\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\hat{\boldsymbol{B}}\right)^{-1}\hat{\boldsymbol{B}}^\top\boldsymbol{\theta}_m\\
=&\hat{\boldsymbol{B}}^\top\boldsymbol{\theta}_m+\left(\gamma I+\sum_{\tau=1}^{t}\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\hat{\boldsymbol{B}}\right)^{-1}\cdot\\
&\sum_{\tau=1}^{t}\hat{\boldsymbol{B}}^\top\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^\top\hat{\boldsymbol{B}}_\perp\hat{\boldsymbol{B}}_\perp^\top\boldsymbol{B}\boldsymbol{w}_m
\end{aligned}$$

$$+ \left(\gamma I + \sum_{\tau=1}^{t} \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau})^\top \hat{\boldsymbol{B}} \right)^{-1} \sum_{\tau=1}^{t} \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \xi_{m,j}$$
$$- \gamma \left(\gamma I + \sum_{\tau=1}^{t} \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau})^\top \hat{\boldsymbol{B}} \right)^{-1} \hat{\boldsymbol{B}}^\top \boldsymbol{\theta}_m.$$

Hence, for any task $m \in [M]$, $t \in [N]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
\boldsymbol{\phi}(s,a)^\top \left(\hat{\boldsymbol{\theta}}_{m,t} - \boldsymbol{\theta}_m \right) =& \boldsymbol{\phi}(s,a)^\top \hat{\boldsymbol{B}} \hat{\boldsymbol{w}}_{m,t} - \boldsymbol{\phi}(s,a)^\top \left(\hat{\boldsymbol{B}} \hat{\boldsymbol{B}}^\top + \hat{\boldsymbol{B}}_\perp \hat{\boldsymbol{B}}_\perp^\top \right) \boldsymbol{\theta}_m \\
=& \boldsymbol{\phi}(s,a)^\top \hat{\boldsymbol{B}} \left(\hat{\boldsymbol{w}}_{m,t} - \hat{\boldsymbol{B}}^\top \boldsymbol{\theta}_m \right) - \boldsymbol{\phi}(s,a)^\top \hat{\boldsymbol{B}}_\perp \hat{\boldsymbol{B}}_\perp^\top \boldsymbol{\theta}_m \\
=& \boldsymbol{\phi}(s,a)^\top \hat{\boldsymbol{B}} \left(\gamma I + \sum_{\tau=1}^{t} \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau})^\top \hat{\boldsymbol{B}} \right)^{-1} \cdot \\
& \sum_{\tau=1}^{t} \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau})^\top \hat{\boldsymbol{B}}_\perp \hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B} \boldsymbol{w}_m \\
& + \boldsymbol{\phi}(s,a)^\top \hat{\boldsymbol{B}} \left(\gamma I + \sum_{\tau=1}^{t} \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau})^\top \hat{\boldsymbol{B}} \right)^{-1} \sum_{\tau=1}^{t} \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \xi_{m,j} \\
& - \gamma \boldsymbol{\phi}(s,a)^\top \hat{\boldsymbol{B}} \left(\gamma I + \sum_{\tau=1}^{t} \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau})^\top \hat{\boldsymbol{B}} \right)^{-1} \hat{\boldsymbol{B}}^\top \boldsymbol{\theta}_m - \boldsymbol{\phi}(s,a)^\top \hat{\boldsymbol{B}}_\perp \hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B} \boldsymbol{w}_m.
\end{aligned}$$

For any $m \in [M]$, let $\boldsymbol{\Sigma}_{m,0} := \gamma I$. For any $m \in [M]$ and $t \geq 1$, let $\boldsymbol{\Sigma}_{m,t} := \gamma I + \sum_{\tau=1}^{t} \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau})^\top \hat{\boldsymbol{B}}$.

Taking the absolute value on both sides and using the Cauchy–Schwarz inequality, we obtain that for any $m \in [M]$, $t \in [N]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
& \left| \boldsymbol{\phi}(s,a)^\top \left(\hat{\boldsymbol{\theta}}_{m,t} - \boldsymbol{\theta}_m \right) \right| \\
\leq & \left\| \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a) \right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \left\| \sum_{\tau=1}^{t} \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau})^\top \hat{\boldsymbol{B}}_\perp \hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B} \boldsymbol{w}_m \right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \\
& + \left\| \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a) \right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \left\| \sum_{\tau=1}^{t} \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \xi_{m,j} \right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \\
& + \gamma \left\| \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a) \right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \left\| \hat{\boldsymbol{B}}^\top \boldsymbol{\theta}_m \right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \\
& + \left| \boldsymbol{\phi}(s,a)^\top \hat{\boldsymbol{B}}_\perp \hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B} \boldsymbol{w}_m \right| \\
\overset{\text{(a)}}{\leq} & \left\| \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a) \right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \sum_{\tau=1}^{t} \left| \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau})^\top \hat{\boldsymbol{B}}_\perp \hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B} \boldsymbol{w}_m \right| \cdot \left\| \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \\
& + \left\| \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a) \right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \sqrt{k \log\left(1 + \frac{t}{\gamma k}\right) + 2 \log\left(\frac{5}{\delta}\right)} \\
& + \gamma \left\| \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a) \right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \cdot \frac{1}{\sqrt{\gamma}} \cdot \left\| \hat{\boldsymbol{B}}^\top \boldsymbol{\theta}_m \right\| + \left\| \hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B} \right\| L_\phi L_w \\
\leq & \left\| \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a) \right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \cdot \left\| \hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B} \right\| L_\phi L_w \cdot \sum_{\tau=1}^{t} \left\| \hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_{m,\tau}, a_{m,\tau}) \right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}}
\end{aligned}$$

$$
\begin{aligned}
&+ \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \sqrt{k \log\left(1+\frac{t}{\gamma k}\right) + 2\log\left(\frac{5}{\delta}\right)} \\
&+ \sqrt{\gamma} L_\theta \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} + \left\|\hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B}\right\| L_\phi L_w \\
\overset{\text{(b)}}{\le} & \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \cdot \left\|\hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B}\right\| L_\phi L_w \cdot \sqrt{tk} \\
&+ \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \sqrt{k \log\left(1+\frac{t}{\gamma k}\right) + 2\log\left(\frac{5}{\delta}\right)} \\
&+ \sqrt{\gamma} L_\theta \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} + \left\|\hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B}\right\| L_\phi L_w \\
= & \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}_{m,t}^{-1}} \left(\left\|\hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B}\right\| L_\phi L_w \cdot \sqrt{tk} + \sqrt{k \log\left(1+\frac{t}{\gamma k}\right) + 2\log\left(\frac{5}{\delta}\right)} + \sqrt{\gamma} L_\theta \right) \\
&+ \left\|\hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B}\right\| L_\phi L_w,
\end{aligned}
$$

where inequality (a) uses the triangle inequality and the definition of event $\mathcal{H}$, and inequality (b) is due to Lemma E.6.

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

$$
\begin{aligned}
\mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}} \left|\boldsymbol{\phi}(s,a)^\top \left(\hat{\boldsymbol{\theta}}_{m,N} - \boldsymbol{\theta}_m\right)\right|\right] \le & \mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}} \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}_N^{-1}}\right] \cdot \\
& \left(\left\|\hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B}\right\| L_\phi L_w \cdot \sqrt{Nk} + \sqrt{k \log\left(1+\frac{N}{\gamma k}\right) + 2\log\left(\frac{5}{\delta}\right)} + \sqrt{\gamma} L_\theta \right) \\
& + \left\|\hat{\boldsymbol{B}}_\perp^\top \boldsymbol{B}\right\| L_\phi L_w.
\end{aligned} \tag{30}
$$

According to Lemma D.9, $\mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}} \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}_t^{-1}}\right]$ is non-increasing with respect to t. Hence, we have

$$
\begin{aligned}
\mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}} \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}_N^{-1}}\right] \le & \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}} \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}_t^{-1}}\right] \\
\le & \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}} \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}\right] \\
\overset{\text{(a)}}{\le} & \frac{1}{4N}\left(2\sqrt{\log\left(\frac{5}{\delta}\right)}\right. \\
& \left. + \sqrt{4\log\left(\frac{5}{\delta}\right) + 4\left(\sum_{t=1}^{N} \max_{a\in\mathcal{A}} \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_t,a)\right\|_{\boldsymbol{\Sigma}_{t-1}^{-1}} + 2\log\left(\frac{5}{\delta}\right)\right)}\right)^2 \\
= & \frac{1}{4N}\left(2\sqrt{\log\left(\frac{5}{\delta}\right)} + \sqrt{4\log\left(\frac{5}{\delta}\right) + 4\left(\sum_{t=1}^{N} \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_t,a_t)\right\|_{\boldsymbol{\Sigma}_{t-1}^{-1}} + 2\log\left(\frac{5}{\delta}\right)\right)}\right)^2,
\end{aligned} \tag{31}
$$

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

In addition, we have

$$
\sum_{t=1}^{N} \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_t,a_t)\right\|_{\boldsymbol{\Sigma}_{t-1}^{-1}} \le \sqrt{N \cdot \sum_{t=1}^{N} \left\|\hat{\boldsymbol{B}}^\top \boldsymbol{\phi}(s_t,a_t)\right\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}^2}
$$

$$\overset{\text{(a)}}{\leq}\sqrt{2N\log\left(\frac{\det\left(\gamma I+\sum_{\tau=1}^{N}\hat{\boldsymbol{B}}^{\top}\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})\boldsymbol{\phi}(s_{m,\tau},a_{m,\tau})^{\top}\hat{\boldsymbol{B}}\right)}{\det\left(\gamma I\right)}\right)}$$

$$\overset{\text{(b)}}{\leq}\sqrt{2Nk\log\left(1+\frac{N}{\gamma k}\right)}, \tag{32}$$

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

Combining Eqs. (31) and (32), we have

$$\begin{aligned}
\mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}}\left\|\hat{\boldsymbol{B}}^{\top}\boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}_N^{-1}}\right] \leq & \frac{1}{4N}\left(2\sqrt{\log\left(\frac{5}{\delta}\right)}+\sqrt{4\log\left(\frac{5}{\delta}\right)+4\left(\sqrt{2Nk\log\left(1+\frac{N}{\gamma k}\right)}+2\log\left(\frac{5}{\delta}\right)\right)}\right)^2 \\
\overset{\text{(a)}}{\leq} & \frac{1}{2N}\left(4\log\left(\frac{5}{\delta}\right)+4\log\left(\frac{5}{\delta}\right)+4\left(\sqrt{2Nk\log\left(1+\frac{N}{\gamma k}\right)}+2\log\left(\frac{5}{\delta}\right)\right)\right) \\
= & \frac{1}{N}\left(2\sqrt{2Nk\log\left(1+\frac{N}{\gamma k}\right)}+8\log\left(\frac{5}{\delta}\right)\right) \\
= & 2\sqrt{\frac{2k\log\left(1+\frac{N}{\gamma k}\right)}{N}}+\frac{8\log\left(\frac{5}{\delta}\right)}{N},
\end{aligned} \tag{33}$$

where inequality (a) uses the Cauchy–Schwarz inequality.

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

$$\mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}}\left|\boldsymbol{\phi}(s,a)^{\top}\left(\hat{\boldsymbol{\theta}}_{m,N}-\boldsymbol{\theta}_m\right)\right|\right] \tag{34}$$

$$\begin{aligned}
\leq & \left(2\sqrt{\frac{2k\log\left(1+\frac{N}{\gamma k}\right)}{N}}+\frac{8\log\left(\frac{5}{\delta}\right)}{N}\right)\cdot \\
& \left(\left\|\hat{\boldsymbol{B}}_{\perp}^{\top}\boldsymbol{B}\right\|L_{\phi}L_w\sqrt{Nk}+\sqrt{k\log\left(1+\frac{N}{\gamma k}\right)+2\log\left(\frac{5}{\delta}\right)}+\sqrt{\gamma}L_{\theta}\right)+\left\|\hat{\boldsymbol{B}}_{\perp}^{\top}\boldsymbol{B}\right\|L_{\phi}L_w \\
\leq & \frac{12\sqrt{k}\log\left(\frac{5N}{\delta}\right)}{\sqrt{N}}\left(\left\|\hat{\boldsymbol{B}}_{\perp}^{\top}\boldsymbol{B}\right\|L_{\phi}L_w\sqrt{Nk}+2\sqrt{k}\log\left(\frac{5N}{\delta}\right)+\sqrt{\gamma}L_{\theta}\right)+\left\|\hat{\boldsymbol{B}}_{\perp}^{\top}\boldsymbol{B}\right\|L_{\phi}L_w \\
\leq & \frac{\left(24k+12\sqrt{k\gamma}L_{\theta}\right)\log^2\left(\frac{5N}{\delta}\right)}{\sqrt{N}}+24k\log\left(\frac{5N}{\delta}\right)\left\|\hat{\boldsymbol{B}}_{\perp}^{\top}\boldsymbol{B}\right\|L_{\phi}L_w.
\end{aligned} \tag{35}$$

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

$$N \geq \frac{26^4\left(24k+12\sqrt{k\gamma}L_{\theta}\right)^2\log^4\left(\frac{2\cdot 5(24k+12\sqrt{k\gamma}L_{\theta})}{\varepsilon\delta}\right)}{\left(\frac{\varepsilon}{4}\right)^2},$$

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

Further enlarging N, if

$$N \geq \frac{4^2\cdot 26^4\cdot 24^2\cdot 2\left(k^2+k\gamma L_{\theta}^2\right)\log^4\left(\frac{240(k+\sqrt{k\gamma}L_{\theta})}{\varepsilon\delta}\right)}{\varepsilon^2}, \tag{36}$$

then

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

According to Lemma D.7, we have $\left\|\hat{\boldsymbol{B}}_{t,\perp}^\top \boldsymbol{B}\right\| \leq \frac{\varepsilon}{96k\log\left(\frac{5N}{\delta}\right)L_\phi L_w}$.

Thus, setting N as the value in Eq. (36), and continuing with Eq. (35), we have

$$\mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}}\left|\boldsymbol{\phi}(s,a)^\top\left(\hat{\boldsymbol{\theta}}_{m,N} - \boldsymbol{\theta}_m\right)\right|\right] \leq \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}.$$

□

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 $Pr[\mathcal{K}\cap\mathcal{L}\cap\mathcal{G}\cap\mathcal{H}\cap\mathcal{J}] \geq 1-\delta$. Suppose that event $\mathcal{K}\cap\mathcal{L}\cap\mathcal{G}\cap\mathcal{H}\cap\mathcal{J}$ holds.

First, we uses 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 $m \in [M]$,

$$\mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}}\left|\boldsymbol{\phi}(s,a)^\top\left(\hat{\boldsymbol{\theta}}_{m,N} - \boldsymbol{\theta}_m\right)\right|\right] \leq \frac{\varepsilon}{2}.$$

For any $m \in [M]$ and $s \in \mathcal{S}$, let $\beta_m(s) := \max_{a\in\mathcal{A}}|\boldsymbol{\phi}(s,a)^\top(\hat{\boldsymbol{\theta}}_{m,N} - \boldsymbol{\theta}_m)|$ and $\pi^*_m(s) := \operatorname{argmax}_{a\in\mathcal{A}}\boldsymbol{\phi}(s,a)^\top\boldsymbol{\theta}_m$.

For any $m \in [M]$ and $s \in \mathcal{S}$, we have

$$\begin{aligned}
\boldsymbol{\phi}(s,\hat{\pi}_m(s))^\top\boldsymbol{\theta}_m \geq& \boldsymbol{\phi}(s,\hat{\pi}_m(s))^\top\hat{\boldsymbol{\theta}}_{m,N} - \beta_m(s) \\
\overset{\text{(a)}}{\geq}& \boldsymbol{\phi}(s,\pi^*_m(s))^\top\hat{\boldsymbol{\theta}}_{m,N} - \beta_m(s) \\
\geq& \boldsymbol{\phi}(s,\pi^*_m(s))^\top\boldsymbol{\theta}_m - 2\beta_m(s),
\end{aligned}$$

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

Rearranging the above equation and taking the expectation of s on both sides, we have

$$\mathbb{E}_{s\sim\mathcal{D}}\left[\max_{a\in\mathcal{A}}\left(\boldsymbol{\phi}(s,a) - \boldsymbol{\phi}(s,\hat{\pi}_m(s))\right)^\top\boldsymbol{\theta}_m\right] \leq 2\mathbb{E}_{s\sim\mathcal{D}}[\beta_m(s)] \leq \varepsilon.$$

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

$$\begin{aligned}
& T_0 + 2MTp + MN \\
=& O\Bigg(\frac{L_\phi^4}{\nu^2}\log^2\left(\frac{d|\mathcal{A}|}{\delta}\right) + \frac{k^4L_\phi^4L_\theta^2L_w^2}{\nu^2\varepsilon^2}\log^6\left(\frac{kdL_\phi L_\theta L_w N}{\nu\delta\varepsilon}\right)\cdot\frac{L_\phi^4}{\nu^2}\log^2\left(\frac{dMT}{\delta}\right) \\
& + M\cdot\frac{\left(k^2 + k\gamma L_\theta^2\right)\log^4\left(\frac{k+\sqrt{k\gamma}L_\theta}{\varepsilon\delta}\right)}{\varepsilon^2}\Bigg) \\
=& O\Bigg(\frac{k^4L_\phi^4L_\theta^2L_w^2}{\nu^2\varepsilon^2}\log^6\left(\frac{|\mathcal{A}|kdL_\phi L_\theta L_w N}{\nu\delta\varepsilon}\right)\cdot\frac{L_\phi^4}{\nu^2}\log^2\left(\frac{dMT}{\delta}\right)
\end{aligned}$$

$$+ M \cdot \frac{\left(k^2 + k\gamma L_\theta^2\right)\log^4\left(\frac{k+\sqrt{k\gamma}L_\theta}{\varepsilon\delta}\right)}{\varepsilon^2}\Bigg)$$

$$=\tilde{O}\left(\frac{k^4 L_\phi^8 L_\theta^2 L_w^2}{\nu^4\varepsilon^2} + \frac{M\left(k^2 + k\gamma L_\theta^2\right)}{\varepsilon^2}\right).$$

□

E. Technical Tools

In this section, we provide some useful technical tools.

Lemma E.1 (Matrix Bernstern Inequality - Average, Lemma 31 in (Tripuraneni et al., 2021)). *Consider a truncation level $U > 0$. If $\{\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n\}$ is a sequence of $d_1 \times d_2$ independent random matrices and $\boldsymbol{Z}'_i = \boldsymbol{Z}_i \cdot \mathbb{1}\{\|\boldsymbol{Z}_i\| \le U\}$ for any $i \in [n]$, then*

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

where $\Delta \ge \|\mathbb{E}[\boldsymbol{Z}_i] - \mathbb{E}[\boldsymbol{Z}'_i]\|$ for any $i \in [n]$.

In addition, for $t \ge \Delta$, we have

$$\Pr\left[\left\|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{Z}'_i - \mathbb{E}[\boldsymbol{Z}'_i])\right\| \ge t - \Delta\right] \le (d_1 + d_2)\exp\left(-\frac{n^2(t-\Delta)^2}{2\sigma^2 + \frac{2Un(t-\Delta)}{3}}\right),$$

where

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

Lemma 31 in (Tripuraneni et al., 2021) gives a truncated matrix Bernstern 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 Bernstern inequality for general random matrices).

Lemma E.2 (Matrix Bernstern Inequality - Summation). *Consider a truncation level $U > 0$. If $\{\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n\}$ is a sequence of $d_1 \times d_2$ independent random matrices, and $\boldsymbol{Z}'_i = \boldsymbol{Z}_i \cdot \mathbb{1}\{\|\boldsymbol{Z}_i\| \le U\}$ and $\Delta \ge \|\mathbb{E}[\boldsymbol{Z}_i] - \mathbb{E}[\boldsymbol{Z}'_i]\|$ for any $i \in [n]$, then for $\tau \ge 2n\Delta$,*

$$\Pr\left[\left\|\sum_{i=1}^n (\boldsymbol{Z}_i - \mathbb{E}[\boldsymbol{Z}_i])\right\| \ge \tau\right] \le (d_1 + d_2)\exp\left(-\frac{1}{4}\cdot\frac{\tau^2}{2\sigma^2 + \frac{U\tau}{3}}\right) + n\Pr\left[\|\boldsymbol{Z}_i\| \ge U\right],$$

where

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

Furthermore, we have

$$\Pr\left[\left\|\sum_{i=1}^n (\boldsymbol{Z}_i - \mathbb{E}[\boldsymbol{Z}_i])\right\| \ge 4\sqrt{\sigma^2\log\left(\frac{d_1+d_2}{\delta}\right)} + 4U\log\left(\frac{d_1+d_2}{\delta}\right)\right] \le \delta + n\Pr\left[\|\boldsymbol{Z}_i\| \ge U\right].$$

Proof of Lemma E.2. Using Lemma E.1 and defining $\tau := nt$, we have that for $\tau > n\Delta$,

$$\Pr\left[\left\|\sum_{i=1}^{n}\left(\boldsymbol{Z}_i - \mathbb{E}[\boldsymbol{Z}_i]\right)\right\| \geq \tau\right] \leq (d_1+d_2)\exp\left(-\frac{(\tau-n\Delta)^2}{2\sigma^2 + \frac{2U(\tau-n\Delta)}{3}}\right) + n\Pr\left[\|\boldsymbol{Z}_i\| \geq U\right].$$

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

$$\begin{aligned}
\Pr\left[\left\|\sum_{i=1}^{n}\left(\boldsymbol{Z}_i - \mathbb{E}[\boldsymbol{Z}_i]\right)\right\| \geq \tau\right] \leq& (d_1+d_2)\exp\left(-\frac{\left(\frac{1}{2}\tau\right)^2}{2\sigma^2 + \frac{2U\left(\frac{1}{2}\tau\right)}{3}}\right) + n\Pr\left[\|\boldsymbol{Z}_i\| \geq U\right] \\
\leq& (d_1+d_2)\exp\left(-\frac{1}{4}\cdot\frac{\tau^2}{2\sigma^2 + \frac{U\tau}{3}}\right) + n\Pr\left[\|\boldsymbol{Z}_i\| \geq U\right]. \qquad (37)
\end{aligned}$$

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

$$\begin{aligned}
&\Pr\left[\left\|\sum_{i=1}^{n}\left(\boldsymbol{Z}_i - \mathbb{E}[\boldsymbol{Z}_i]\right)\right\| \geq 4\sqrt{\sigma^2\log\left(\frac{d_1+d_2}{\delta}\right)} + 4U\log\left(\frac{d_1+d_2}{\delta}\right)\right] \\
\leq& (d_1+d_2)\exp\left(-\frac{1}{4}\cdot\frac{16\sigma^2\log\left(\frac{d_1+d_2}{\delta}\right) + 16U^2\log^2\left(\frac{d_1+d_2}{\delta}\right) + 32U\log\left(\frac{d_1+d_2}{\delta}\right)\sqrt{\sigma^2\log\left(\frac{d_1+d_2}{\delta}\right)}}{2\sigma^2 + \frac{1}{3}\left(4U\sqrt{\sigma^2\log\left(\frac{d_1+d_2}{\delta}\right)} + 4U^2\log\left(\frac{d_1+d_2}{\delta}\right)\right)}\right) \\
&+ n\Pr\left[\|\boldsymbol{Z}_i\| \geq U\right] \\
\leq& (d_1+d_2)\exp\left(-\frac{1}{4}\cdot 4\log\left(\frac{d_1+d_2}{\delta}\right)\right) + n\Pr\left[\|\boldsymbol{Z}_i\| \geq U\right] \\
=& \delta + n\Pr\left[\|\boldsymbol{Z}_i\| \geq U\right].
\end{aligned}$$

□

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

$$\begin{aligned}
\frac{A}{\sqrt{T}}\log(BT) =& \frac{A\kappa}{\sqrt{68}A\log\left(\frac{AB}{\kappa}\right)}\log\left(\frac{68A^2B\log^2\left(\frac{AB}{\kappa}\right)}{\kappa^2}\right) \\
=& \frac{\kappa}{\sqrt{68}\log\left(\frac{AB}{\kappa}\right)}\left(\log(68) + \log\left(\frac{A^2B}{\kappa^2}\right) + \log\left(\log^2\left(\frac{AB}{\kappa}\right)\right)\right) \\
\leq& \frac{\kappa}{\sqrt{68}\log\left(\frac{AB}{\kappa}\right)}\left(\log(68) + 2\log\left(\frac{AB}{\kappa}\right) + 2\log\left(\frac{AB}{\kappa}\right)\right) \\
\leq& \frac{\kappa}{\sqrt{68}\log\left(\frac{AB}{\kappa}\right)}\left(\log(68)\log\left(\frac{AB}{\kappa}\right) + 4\log\left(\frac{AB}{\kappa}\right)\right) \\
\leq& \kappa.
\end{aligned}$$

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

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

□

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*

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

then

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

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

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

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

Lemma E.5. *For any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^k$, we have*

$$\sum_{j=1}^{n} \|\boldsymbol{x}_j\|_{\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}} \leq \sqrt{nk}.$$

Proof of Lemma E.5. It holds that

$$\begin{aligned}
\sum_{j=1}^{n} \|\boldsymbol{x}_j\|_{\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}} &= \sum_{j=1}^{n} \sqrt{\boldsymbol{x}_j^\top \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \boldsymbol{x}_j} \\
&\leq \sqrt{n \cdot \sum_{j=1}^{n} \boldsymbol{x}_j^\top \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \boldsymbol{x}_j} \\
&\leq \sqrt{n \cdot \sum_{j=1}^{n} \text{Trace}\left(\boldsymbol{x}_j^\top \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \boldsymbol{x}_j\right)} \\
&= \sqrt{n \cdot \sum_{j=1}^{n} \text{Trace}\left(\boldsymbol{x}_j \boldsymbol{x}_j^\top \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}\right)} \\
&= \sqrt{n \cdot \text{Trace}\left(\sum_{j=1}^{n} \boldsymbol{x}_j \boldsymbol{x}_j^\top \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}\right)} \\
&= \sqrt{n \cdot \text{Trace}\left(\boldsymbol{I}_k\right)} \\
&= \sqrt{nk}
\end{aligned}$$

□

Lemma E.6. *For any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^k$ and $\gamma > 0$, we have*

$$\sum_{j=1}^{n} \|\boldsymbol{x}_j\|_{\left(\gamma I + \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}} \leq \sqrt{nk}.$$

Proof of Lemma E.6. It holds that

$$\begin{aligned}
\sum_{j=1}^{n} \|\boldsymbol{x}_j\|_{\left(\gamma I + \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}} &= \sum_{j=1}^{n} \sqrt{\boldsymbol{x}_j^\top \left(\gamma I + \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \boldsymbol{x}_j} \\
&\leq \sqrt{n \cdot \sum_{j=1}^{n} \boldsymbol{x}_j^\top \left(\gamma I + \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \boldsymbol{x}_j} \\
&= \sqrt{n \cdot \sum_{j=1}^{n} \text{Trace}\left(\boldsymbol{x}_j^\top \left(\gamma I + \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \boldsymbol{x}_j\right)} \\
&= \sqrt{n \cdot \sum_{j=1}^{n} \text{Trace}\left(\boldsymbol{x}_j \boldsymbol{x}_j^\top \left(\gamma I + \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}\right)} \\
&= \sqrt{n \cdot \text{Trace}\left(\sum_{j=1}^{n} \boldsymbol{x}_j \boldsymbol{x}_j^\top \left(\gamma I + \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}\right)}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leq}\sqrt{n\cdot\left(\mathrm{Trace}\left(\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^\top\left(\gamma I+\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\right)+\mathrm{Trace}\left(\gamma\left(\gamma I+\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\right)\right)}\\
&=\sqrt{n\cdot\mathrm{Trace}\left(\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^\top\left(\gamma I+\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}+\gamma\left(\gamma I+\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\right)}\\
&=\sqrt{n\cdot\mathrm{Trace}\left(\left(\gamma I+\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^\top\right)\left(\gamma I+\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\right)}\\
&=\sqrt{n\cdot\mathrm{Trace}\left(\boldsymbol{I}_k\right)}\\
&=\sqrt{nk},
\end{aligned}
$$

where inequality (a) is due to that $\left(\gamma I+\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)$ is a positive definite matrix. □

Lemma E.7 (Self-normalized Concentration for Martingales, Theorem 1 in (Abbasi-Yadkori et al., 2011))**.** *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration such that for any $t\geq 1$, the selected action $\boldsymbol{X}_t\in\mathbb{R}^k$ is $\mathcal{F}_{t-1}$-measurable, the noise $\eta_t\in\mathbb{R}$ is $\mathcal{F}_t$-measurable, and conditioning on $\mathcal{F}_{t-1}$, η_t is zero-mean and R-sub-Gaussian. Let $\boldsymbol{V}_0\in\mathbb{R}^{k\times k}$ be a positive definite matrix and let $\boldsymbol{V}_t=\sum_{i=1}^{t}\boldsymbol{X}_i\boldsymbol{X}_i^\top$ for any $t\geq 1$. Then, for any $\delta>0$, with probability at least $1-\delta$, for all $t\geq 1$,*

$$
\left\|\sum_{i=1}^{t}\boldsymbol{X}_i\cdot\eta_i\right\|^2_{(\boldsymbol{V}_0+\boldsymbol{V}_t)^{-1}}\leq 2R^2\log\left(\frac{\det(\boldsymbol{V}_t)^{\frac{1}{2}}}{\det(\boldsymbol{V}_0)^{\frac{1}{2}}\cdot\delta}\right).
$$

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

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

Lemma E.9 (Elliptical Potential Lemma, Lemma 11 in (Abbasi-Yadkori et al., 2011))**.** *Let $\{\boldsymbol{X}_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^k$. Let $\boldsymbol{V}_0$ be a $k\times k$ positive definite matrix and let $\boldsymbol{V}_t=\boldsymbol{V}_0+\sum_{i=1}^{t}\boldsymbol{X}_i\boldsymbol{X}_i^\top$ such that for any $t\geq 1$, $\|\boldsymbol{X}_t\|^2_{\boldsymbol{V}_{t-1}^{-1}}\leq 1$. Then, we have that*

$$
\sum_{t=1}^{n}\|\boldsymbol{X}_t\|^2_{\boldsymbol{V}_{t-1}^{-1}}\leq 2\log\frac{\det(\boldsymbol{V}_n)}{\det(\boldsymbol{V}_0)}.
$$

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

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

we have that for any integer $n\geq 1$,

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

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