Provable Active Multi-Task Representation Learning

Jiabin Lin, Shana Moothedath, *Senior Member, IEEE*, and Tuan Anh Le

***Abstract*—Multi-task representation learning is an emerging machine learning paradigm that integrates data from multiple sources, harnessing task similarities to enhance overall model performance. The application of multi-task learning to real-world settings is hindered due to data scarcity, along with challenges related to scalability and computational resources. To address these challenges, we develop a fast and sample-efficient approach for multi-task active learning with linear representation when the amount of data from source tasks and target tasks is limited. By leveraging the techniques from active learning, we propose an adaptive sampling-based alternating projected gradient descent (GD) and minimization algorithm that iteratively estimates the relevance of each source task to the target task and samples from each source task based on the estimated relevance. We present the convergence guarantees and the sample and time complexities of our algorithm. We evaluated the effectiveness of our algorithm using experiments and compared it with four benchmark algorithms using synthetic and real-world MNIST-C and MovieLens-100K datasets.**

***Index Terms*—Multi-task representation learning, transfer learning, alternating gradient descent and minimization algorithm, active learning.**

I. Introduction

MULTI-TASK representation learning has emerged as a promising machine learning (ML) approach for simultaneously learning multiple related models by integrating data from various sources. The approach leverages shared structures between tasks to improve the performance of each individual task by collaboratively training similar but different tasks to overcome a scarcity of data for any one task. This paradigm has been used with great success in the natural language processing domains GPT-2 [1], GPT-3 [2], Bert [3], as well as the vision domains CLIP [4]. As noted in [1], despite the notable advances, existing learning systems require hundreds to thousands of examples to effectively induce functions that generalize well. With current approaches, this implies that multi-task training may need just as many effective training pairs to realize its potential.

Most of the existing work on multi-task representation learning often assumes an unlimited number of samples for source tasks and a limited number of samples for the target task [5], [6]. In practical applications such as medical imaging, drug discovery, fraud detection, and natural language processing in low-resource languages, data availability is limited, restricting the application of existing ML approaches due to poor sample efficiency. It may be challenging to continue scaling the creation of datasets to the extent that might be necessary using current techniques. This motivates exploring new approaches for multi-task learning, specifically to develop provable methods that are fast and sample-efficient. Additionally, as noted in [6], not all tasks equally contribute to learning a representation. For instance, modern datasets like CIFAR-10, ImageNet, and the CLIP dataset were created using a list of search terms and a variety of different sources like search engines, news websites, and Wikipedia [4], [7], [8]. Further, it is often unclear which tasks will best maximize performance on the target task.

In this paper, we introduce a novel active (adaptive) multi-task learning framework and an associated algorithm with guarantees. Our goal is to learn a multi-task linear representation while prioritizing the relevance of the source tasks to generalize to a specific target task. Our approach involves using alternating gradient descent (GD) and minimization estimator to estimate the unknown parameters. Additionally, we utilize adaptive sampling to incorporate data samples from more relevant tasks into the learning process, which is beneficial to generalize the task to a target with fewer samples. In our approach, the tasks only share their parameter estimates rather than the raw data itself.

Our Contributions. We present our main contributions below.

(i) Active Low-Rank Representation Learning (A-LRRL) algorithm. We propose an alternating gradient descent and minimization approach to provide a provable solution to the *active* multi-task representation learning problem. Our proposed A-LRRL algorithm is fast, federated (tasks do not share raw data but only the parameter estimates), and sample-efficient for learning the common low-dimensional representation. Our algorithm adopts an iterative approach: it first learns the representation and then leverages this learned representation to estimate the unknown relevance parameter. Using the relevance estimate, we design an adaptive sampling strategy that selectively samples source task data, optimizing the learning process. Both the time and the sample complexity of our solution depend only logarithmically on $1/\epsilon$ for an ϵ guarantee. We summarize the sample complexity in Table I.

Received 26 February 2025; revised 10 September 2025 and 10 October 2025; accepted 15 October 2025. Date of publication 24 October 2025; date of current version 23 December 2025. This work was supported in part by NSF-CAREER under Grant 2440455 and in part by NSF-ECCS under Grant 2213069. The associate editor coordinating the review of this article and approving it for publication was Thomas Oberlin. *(Corresponding author: Shana Moothedath.)*

Jiabin Lin and Shana Moothedath are with the Department of Electrical Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: jiabin@iastate.edu; mshana@iastate.edu).

Tuan Anh Le is with the Department of Computer Science, Iowa State University, Ames, IA 50011 USA (e-mail: tuanle@iastate.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TSP.2025.3623098, provided by the authors.

Digital Object Identifier 10.1109/TSP.2025.3623098

TABLE I
TABLE COMPARING SAMPLE COMPLEXITIES OF EXISTING WORK VERSUS OURS

	Sample Complexity Per Source Task (All Epochs)	Sample Complexity Across All Source Tasks (All Epochs)	Sample Complexity for the Target Task
Collins et al., 2021 [11]	$\widetilde{O}\left(k^3 + d\frac{k^2}{M}\right)$	$\widetilde{O}\left(Mk^3 + dk^2\right)$	Did not study transferability to target task
Chen et al., 2022 [6]	$\times$	$\widetilde{O}\left((d+M)\frac{k}{M}s_\Gamma^\star\epsilon^{-1} + dk^2\sqrt{M}\epsilon^{-\frac{1}{2}}\frac{R}{\underline{\sigma}^3}\right)$	$\widetilde{O}\left(\frac{k}{M}\epsilon^{-1} + \min\{\frac{\sqrt{R}}{\underline{\sigma}^2 k}, \sqrt{(d+M)k}\}\sqrt{\frac{s_\Gamma^\star}{M}}\epsilon^{-\frac{1}{2}}\right)$
Our approach	$\widetilde{O}\left(\max\{\log d, \log M, k\}s_\Gamma^\star\epsilon^{-2}\right)$	$\widetilde{O}\left((d+k)k^2 s_\Gamma^\star\epsilon^{-2}\right)$	$\widetilde{O}\left(\max\left\{\frac{k}{M}\epsilon^{-1}, \max\{\log M, k, \log d\}\frac{s_\Gamma^\star}{k^3}\epsilon^{-2}\right\}\right)$

Note: R represents the upper bound of $\|w_m^\star\|_2$, while $\underline{\sigma}$ denotes the lower bound of $\sigma_{\min}(W^\star)$ as stated in Chen et al., 2022 [6].

(ii) Convergence guarantees. We present the convergence guarantee of excess risk for the proposed approach along with sample and time complexities. Our results show that the number of target samples scales with the rank of the low-dimensional feature space and log of the input dimension and number of tasks to achieve $\epsilon-$accuracy in the excess risk for generalizing to the target task. We show that the sample complexity for the source tasks scales according to the sparsity of the relevance parameter. Hence the sample complexity of the proposed approach improves by a factor of the number of tasks compared to the naive uniform sampling approach.

(iii) Numerical performance. We compared our algorithm with four benchmark approaches via simulations. We conducted experiments by varying the number of tasks, problem dimension, and rank of the feature matrix. We performed experiments on synthetic and real-world MNIST-C [9] and Movielens-100K [10] datasets. The proposed approach consistently outperformed the benchmark algorithms in all cases, validating its effectiveness.

II. NOTATIONS AND PROBLEM SETTING

Notations. We denote the set containing the first n positive integers as $[n]$, which is defined as $\{1, 2, \ldots, n\}$. The ℓ_2 norm of a vector x is represented by $||x||$, while the spectral norm and the Frobenius norm of a matrix A are denoted by $||A||$ and $||A||_F$, respectively. The transpose operation for matrices and vectors is indicated by $\top$, and $|x|$ refers to the element-wise absolute value of the vector x. The identity matrix of size $n \times n$ is denoted as I_n, often abbreviated as I, and e_k denotes the k-th canonical basis vector, i.e., the k-th column of I_n. We define the n_m i.i.d. samples from the m-th source task as an input matrix $X_m \in \mathbb{R}^{n_m \times d}$, with the corresponding output vector $Y_m \in \mathbb{R}^{n_m}$ and a noise vector $Z_m \in \mathbb{R}^{n_m}$. The vectors $\{w_m\}_{m\in[M]}$, is assembled into the matrix $W \in \mathbb{R}^{k\times M}$. For basis matrices B_1 and B_2, we define Subspace Distance (SD) as $\mathrm{SD}(B_1, B_2) := \|(I - B_1 B_1^\top)B_2\|$.

Problem Setting. Consider M source tasks and one target task, referred to as the $(M+1)$-th task. Every task $m \in [M+1]$ is associated with a distinct joint distribution μ_m over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \in \mathbb{R}^d$ represents the input space and $\mathcal{Y} \in \mathbb{R}$ represents the output space. For each task $m \in [M+1]$, we are given n_m data samples $(x_{m,1}, y_{m,1}), \cdots, (x_{m,n_m}, y_{m,n_m})$, which are i.i.d. and sampled from the distribution μ_m. The goal of multi-task learning is to simultaneously produce predictive models for all M source tasks, with the aim of finding common property among these tasks. We consider the existence of an underlying representation function $\phi^\star := \mathcal{X} \to \mathcal{Z}$, which transforms inputs into a feature space $\mathcal{Z} \in \mathbb{R}^k$ with $k \ll d$, within a specified set of functions Φ such as linear functions. Furthermore, we consider a linear transformation from the feature space to the output space, represented by the vector $w_m^\star \in \mathbb{R}^k$. Specifically, we assume that a sample (x, y) from μ_m for any task $m \in [M+1]$ can be represented as $y = \phi^\star(x)^\top w_m^\star + z_m$, where z_m is a noise.

We consider a data-scarce regime. Typically, the number of data samples for the target task is even fewer than that of the source task. This setting aligns with our main objective of active representation learning under scarce data, in which we have a limited amount of data available for the source task but have even less access to the target task data. Define $\mathcal{L}_{M+1}(\phi, w) := \mathbb{E}_{(x,y)\sim\mu_{M+1}}[(\langle\phi(x), w\rangle - y)^2]$. The main objective is to use as few total samples from the source task as possible to learn a representation and linear predictor ϕ, w_{M+1} that effectively minimizes the excess risk on the target task, defined as

$$\mathrm{ER}_{M+1}(\phi, w) = \mathcal{L}_{M+1}(\phi, w) - \mathcal{L}_{M+1}(\phi^\star, w_{M+1}^\star). \quad (1)$$

We focus on the linear representation function class, studied in [5], [6], [12], [13]. We have the low-dimensional assumption.

Assumption II.1 (Low-dimension linear representation): $\Phi = \{x \to B^\top x | B \in \mathbb{R}^{d\times k}\}$. We assume that the true underlying representation is given by an orthonormal matrix $B^\star \in \mathbb{R}^{d\times k}$.

The low-dimensional assumption captures the relatedness between the tasks and is used in many works on representation learning, including [5], [6], [12], [14], [15], [16], [17]. Under Assumption II.1, we can rewrite y as $y = x^\top B^\star w_m^\star + z_m = x^\top \theta_m^\star + z_m$. This model aligns with a two-layer linear neural network: a shared representation $B^\star \in \mathbb{R}^{d\times k}$ (hidden layer) maps the inputs from a high-dimensional space into a lower-dimension subspace, and each task m utilizes a task-specific vector $w_m^\star \in \mathbb{R}^k$ (output layer) to make predictions. This is the standard "representation + task prediction" structure widely used in multi-task and transfer settings, rather than a simple single-layer linear regression. It is crucial to note that the decomposition into $(B^\star, w_m^\star)$ is not unique. In fact, the pair $(B^\star Q, Q^{-1} w_m^\star)$ is the same as $(B^\star, w_m^\star)$ for any invertible matrix $Q \in \mathbb{R}^{k\times k}$. Thus, the ground-truth factors cannot be explicitly determined, and only their product holds significance.

Let $\Theta^\star := [\theta_1^\star, \ldots, \theta_M^\star] \in \mathbb{R}^{d\times M}$ be a rank-k matrix, where $k \ll \min\{d, M\}$. The reduced (rank k) SVD is given by $\Theta^\star := [\theta_1^\star, \ldots, \theta_M^\star] \overset{SVD}{=} B^\star \Sigma^\star V^{\star\top} = B^\star W^\star$, where $B^\star \in \mathbb{R}^{d\times k}$ and $V^{\star\top} \in \mathbb{R}^{M\times k}$ are matrices with orthonormal columns, and Σ is an $k \times k$ diagonal matrix with singular values. We let

$W^\star := \Sigma V^\star$. We use $\sigma^\star_{\max}$ and $\sigma^\star_{\min}$ to denote the maximum and minimum singular values of Σ, and condition number $\kappa := \sigma^\star_{\max}/\sigma^\star_{\min}$. Given that $W^\star$ has full row rank, it follows that $\text{span}\{w^\star_1, \cdots, w^\star_M\} = \mathbb{R}^k$. Thus, any k-dimensional vector can be expressed as a linear combination of $\{w^\star_m\}_{m=1}^M$.

Inspired by [6], in our model, the relevance of the task is a crucial factor. We consider a setting where the goal is to learn a *specific* target task, rather than a *generic* target task as in [5], [12]. Since $\sigma_{\min}(W^\star) > 0$, the coefficient $w^\star_{M+1}$ can be considered a linear combination of the coefficients $\{w^\star_m\}_{m\in[M]}$. Therefore, we assume that $\nu^\star \in \mathbb{R}^M$, such that $W^\star \nu^\star = w^\star_{M+1}$, where a larger value of $|\nu^\star(m)|$ (the m-th element of the vector $\nu^\star$) indicates a stronger relevance for the source task m for the target task. We prioritize samples from source tasks with the highest relevance. In this paper, we aim to learn the low-dimensional representation and the relevance parameter $\nu^\star$ to expedite collaborative learning among the source tasks and facilitate generalization to a target task.

Assumption II.2: We assume $x_{m,n}$ follows an i.i.d. standard Gaussian distribution and noise variables z_m follow i.i.d. Gaussian distribution with zero mean and σ^2 variance.

We work in the random design linear regression setting, and in this context, Assumption II.2 is standard [5], [6], [12], [13].

Assumption II.3 (Incoherence): We assume that $\|w^\star_m\|^2 \leqslant \mu^2 \frac{k}{M} {\sigma^\star_{\max}}^2$ for a constant $\mu \geqslant 1$.

Recovering the feature matrix is impossible without any structural assumption. Notice that y_ms are not global functions of $\Theta^\star$, i.e., no $y_{m,n}$ is a function of the entire matrix $\Theta^\star$. We thus need an assumption that enables correct interpolation across the different columns. The incoherence (w.r.t. the canonical basis) assumption on the right singular vectors suffices for this purpose. It was first introduced in [18] and used recently in representation learning [11], [12], [19].

III. Related Work

Multi-task representation learning has been extensively explored, with roots traced back to seminal works such as [20], [21], [22]. Works such as [5], [11], [12], [19], [23] focus on learning a representation function that is applicable to *any* potential target task, assuming the existence of a shared low-dimensional linear representation across all tasks. More recently, [6], [24] proposed adaptive representation learning methods tailored to a *specific* target task, operating under a similar setting as in [5]. [24] improved the sample complexity on [6] under a high-dimension input assumption. There also exists works on empirical multi-task representation learning [25], [26].

While representation learning has achieved tremendous success, there remain challenges in providing theoretical guarantees. Most of the existing theoretical studies adopt a convex relaxation of the original non-convex problem and rely on the assumption that an optimal solution to the non-convex problem is known for their theoretical analysis [5], [12], [27]. The primary focus of these works is to demonstrate the dimensionality-reducing benefits of representation learning by showing that the number of target samples exceeds only $O(k)$, where k is assumed to be small. Our work complements these results by showing how to provably and efficiently learn the representation in the linear case. The work most closely related to ours is [6]. Our work extends and complements Chen et al. in two ways: (1) The estimation approach in [6] utilizes [5], which assumes an optimal solution to the non-convex estimation problem is available. In this work, we present a novel adaptive sampling-based alternating gradient descent and minimization-based estimator to solve the non-convex representation problem with generalization guarantees to a target task. (2) [5], [6] considered that the number of source task samples must exceed the problem dimension d. We relax this assumption in our approach, and our guarantees hold for setting where the number of data samples is fewer than the feature dimension. Our approach is thus viable for many practical applications with large problem sizes and fewer data samples. This is specifically true in image-related learning as validated through our simulations. This work extends our preliminary study [28], which considered the simplified setting where the relevance parameter $\nu^\star$ was assumed to be known, reducing the problem to learning $\Theta^\star$. When the relevance parameter is unknown, which is the focus of this paper, the error in estimating $\nu^\star$ affects the estimation of $\Theta^\star$ in the next epoch, leading to a temporal error propagation. This requires novel algorithmic techniques and proof approaches to derive guarantees and ensure convergence.

Matrix learning is another related line of work in the low-rank matrix learning literature [11], [29]. Our approach builds on the alternating gradient descent (GD) minimization algorithm introduced in [11], [29], which is designed for recovering a low-rank matrix from compressed signals. The focus of these works is learning the low-dimensional linear representation. [11] provided guarantees on linear convergence relative to the initialization error. However, it does not offer guarantees for the initialization error itself. Further, the matrix learning analysis in [11], [29], [30] considered a non-noisy setting where the observed signals are not affected by noise. Additionally, these works focus on learning a low-dimensional representation using a non-adaptive data sampling of the source samples and not on the generalization of the target task and quantifying the excess risk for the target task. Our work focuses on *active* representation learning and generalizability to target tasks via adaptive sampling. Through theoretical analysis and numerical simulations, we showcase how the adaptive sampling approach enhances generalizability over uniform sampling.

Multi-task learning for sequential decision-making has been studied in the context of bandit learning and reinforcement learning (RL). Multi-task learning in RL domains is studied in many works, including [31], [32], [33], [34]. [33] demonstrated that representation learning has the potential to enhance the rate of the approximate value iteration algorithm. [34] proved that representation learning can reduce the sample complexity of imitation learning. Multi-task bandit learning is studied in many works, including [14], [15], [16], [35]. [14], [16] considered a convex relaxation-based approach to estimate the unknown parameter matrix, while [15] proposed an optimism in the face of uncertainty approach. [35] proposed an alternating GD-minimization approach with guarantees.

IV. Active Low-Rank Representation Learning (A-LRRL) via Alternating GD and Minimization

Our objective is to acquire a low-dimensional linear representation and task relevance estimation from the training samples (source tasks) through an adaptive sampling approach, allowing the utilization of more data from source tasks that are more relevant to the target task rather than a uniform sampling approach. The rationale is that by incorporating more samples from pertinent tasks, we can accelerate the learning process. To this end, our algorithm draws $\propto (\nu^\star(m))^2$ i.i.d. samples from the corresponding offline data for each source task $m \in [M]$. We partition the learning horizon into Γ epochs. Let $\{X_m^{(i)}, Y_m^{(i)}\}_{m=1}^M$ denote the source tasks data in epoch i. Using the source task data, in each epoch $i \in [\Gamma]$, we minimize

$$f_i(\widehat{B}^{(i)}, \widehat{W}^{(i)}) = \sum_{m=1}^{M} \sum_{n=1}^{n_m^i} \|y_{m,n} - x_{m,n}^\top \widehat{B}^{(i)} \widehat{w}_m^{(i)}\|^2, \quad (2)$$

where $\widehat{B}^{(i)} \in \mathbb{R}^{d\times k}$ and $\widehat{W}^{(i)} \in \mathbb{R}^{k\times M}$. Subsequently, we use the estimated parameter $\widehat{B}^{(i)}$ along with the sample for the target task to further optimize the cost function

$$\widehat{w}_{M+1}^{(i)} = \arg\min_w \|{X_{M+1}^{(i)}}^\top \widehat{B}^{(i)} w - Y_{M+1}^{(i)}\|^2. \quad (3)$$

Equation (3) via a least-squares solution yields the estimated parameter $\widehat{w}_{M+1}^{(i)}$ for the target task. Finally using the estimates $\widehat{W}^{(i)}, \widehat{w}_{M+1}^{(i)}$, we now solve the constrained least-squares problem to get the minimum-norm (unique) solution

$$\hat{\nu}_{i+1} = \arg\min_\nu \|\nu\|_2^2 \quad \text{such that} \quad \widehat{W}^{(i)}\nu = \widehat{w}_{M+1}^{(i)}. \quad (4)$$

Using the relevance estimate $\hat{\nu}_{i+1}$, in the next epoch, we sample the source task data such that we utilize more samples from tasks that are more relevant to the specific target task. This observation is motivated by [6] that demonstrated the benefit of adaptive relevance-based sampling over uniform sampling. In our theoretical analysis, we show that the estimate of the relevance parameter obtained at the end of each epoch $|\hat{\nu}_i(m)|$ is $\sqrt{\epsilon_i}$-close to the true parameter $|\nu^\star(m)|$.

Now, we will elaborate on our approach for solving the cost function in Eq. (2). Eq. (2) is non-convex and hence it requires careful initialization. Thus, in the first epoch, we perform a spectral initialization [29], [36]. The initialization process starts by extracting the top k singular vector from

$$\widehat{\Theta}_{0,full} = \left[\left(\frac{1}{n_1} {X_1^{(1)}}^\top Y_1^{(1)}\right), \cdots, \left(\frac{1}{n_M} {X_M^{(1)}}^\top Y_M^{(1)}\right)\right] = \sum_{m=1}^{M} \frac{1}{n_m^1} \sum_{n=1}^{n_m^1} x_{m,n} y_{m,n} e_m^\top,$$

where $X_m^{(1)}$ is the feature matrix obtained by concatenating the feature vectors of task m. The expected value of the $m-$th task represents $B^\star w_m^\star$ with $\mathbb{E}[\widehat{\Theta}_{0,full}] = B^\star W^\star$. However, the large magnitude of the sum of independent sub-exponential random variables restricts the ability to determine a bound for the $\|\widehat{\Theta}_{0,full} - B^\star W^\star\|$ within the desired sample complexity.

Algorithm 1 Spectral Initialization

1: **Input:** $\{X_{m,00}^{(1)}, Y_{m,00}^{(1)}\}_{m=1}^M$
2: **Parameters:** multiplier for α in init step, $\tilde{C} = 9\kappa^2\mu^2$
3: Use $Y_m^{(1)} \equiv Y_{m,00}^{(1)}$, $X_m^{(1)} \equiv X_{m,00}^{(1)}$, set $\alpha = \frac{\tilde{C}}{\sum_{m=1}^M n_m^1} \sum_{n=1}^{n_m^1} y_{m,n}^2$
4: $y_{m,trunc}(\alpha) := Y_m^{(1)} \circ \mathbb{1}_{\{|Y_m^{(1)}| \leqslant \sqrt{\alpha}\}}$
5: $\widehat{\Theta}_0 := \sum_{m=1}^M \frac{1}{n_m^1} {X_m^{(1)}}^\top y_{m,trunc}(\alpha) e_m^\top$
6: Set $\widehat{B}^{(0)} \leftarrow$ top-k-singular-vectors of $\widehat{\Theta}_0$

To tackle this, we use the truncation method introduced in [36], starting with the top k singular vectors of

$$\widehat{\Theta}_0 = \sum_{m=1}^{M} \sum_{n=1}^{n_m^1} x_{m,n} y_{m,n} e_m^\top \mathbb{1}_{\{y_{m,n}^2 \leqslant \alpha\}},$$

where $\alpha = \frac{\tilde{C}}{\sum_{m=1}^M n_m^1} \sum_{m=1,n=1}^{M,n_m^1} y_{m,n}^2$, $\tilde{C} = 9\kappa^2\mu^2$, and $y_{m,trunc}(\alpha) := Y_m^{(1)} \circ \mathbb{1}_{\{|Y_m^{(1)}| \leqslant \sqrt{\alpha}\}}$. Using Singular Value Decomposition (SVD), we obtain the top k singular vectors from $\widehat{\Theta}_0$ and set as our initial estimate $\widehat{B}^{(0)}$. This method effectively filters out large values while maintaining the remaining values and serves as a reliable initial step in accurately estimating parameters.

After the initialization phase, we perform an alternating GD and minimization step to estimate $\widehat{B}^{(i)}$ and $\widehat{W}^{(i)}$ by minimizing Eq. (2). Each iteration consists of two stages: independently optimizing $\widehat{w}_m$ for each task via a least square minimization step, followed by a GD step to update $\widehat{B}$, utilizing the QR decomposition to obtain the updated matrix B^+, represented as $\widehat{B}^+ \overset{QR}{=} B^+ R^+$. Then, the estimate of $B^\star$ for the i^{th} epoch is set as the orthonormal B^+ obtained using the QR decomposition (step 17 in Algorithm 2). We now compute the estimated parameter $\widehat{w}_{M+1}^{(i)}$ by minimizing the cost function in Eq. (3) using the least squares estimator. Finally, we solve the minimum-norm least squares problem in Eq. (4) to estimate the relevance parameter. The estimate of the relevance parameter serves as the sampling parameter in the next epoch. We sample the source task data for the next epoch $\propto (\hat{\nu}_i(m))^2$, giving more weightage to the more relevant task.

Practical setting of parameters. In our algorithm, we set the parameters, GD step size η and the multiplier $\tilde{C}$ in the spectral initialization step. Our theorem states that $\eta = \frac{c}{\sigma_{\max}^{\star 2}}$ with $c \leqslant 0.5$. However, $\sigma_{\max}^\star$ is unknown. Given the initialization matrix $\widehat{\Theta}_0$ provides an approximation to $\Theta^\star$, we set $\sigma_{\max}^\star \approx \|\widehat{\Theta}_0\|$ and $\eta = \frac{c}{\|\widehat{\Theta}_0\|^2}$. Our analysis requires $\tilde{C} = 9\kappa^2\mu^2$, with κ and μ being functions of $\Theta^\star$ and hence unknown. Using the incoherence assumption, we can set $\kappa^2\mu^2$ by an estimate of its lower bound, $M \max_m \frac{\|\widehat{\theta}_m\|^2}{\|\widehat{\Theta}\|_F^2}$, with $\|\widehat{\theta}_m\|^2 = \frac{1}{n_m^1} \sum_{n=1}^{n_m^1} y_{m,n}^2$ and $\|\widehat{\Theta}\|_F^2 = \frac{1}{n_m^1} \sum_{m=1}^M \sum_{n=1}^{n_m^1} y_{m,n}^2$.

Remark IV.1: The truncated spectral initialization guarantees that our initial estimate achieves a small subspace distance with respect to $B^\star$ with high probability (Theorem 2.2 in [37]). Starting from a good initial estimate, the alternating projected

Algorithm 2: Active Low-Rank Representation Learning

Input: Representation function class Φ, multiplier for α in init step, $\tilde{C}$, GD step size, η, Number of GD iterations in the i^{th} epoch, $T^{(\text{i})}$, number of epochs Γ
Initialize $\hat{\nu}_1 = [\frac{1}{M}, \cdots, \frac{1}{M}]$ and $\epsilon_i = 2^{-i}$
for $i = 1, 2, \ldots, \Gamma$ **do**
  Set $n_m^i \propto \max\{\epsilon_i^{-1}, \hat{\nu}_i(m)^2 \epsilon_i^{-1}\}$
  For each task m, draw n_m^i i.i.d samples from the corresponding dataset $\{X_m^{(\text{i})}, Y_m^{(\text{i})}\}_{m=1}^M$
  Sample-split: Partition the measurements and measure matrices into $2T^{(\text{i})}+1$ equal-sized disjoint sets: one for initialization and $2T^{(\text{i})}$ sets each for the iterations in each epoch. Denote these by $\{X_{m,\tau}^{(\text{i})}, Y_{m,\tau}^{(\text{i})}\}_{m=1}^M$, $\tau = 00$ (only for epoch 1), $01, \cdots 2T^{(\text{i})}$.
  if $i = 1$ **then**
    Initialize $\widehat{B}^{(0)}$ using Spectral Initialization (Alg 1)
  end if
  AltGDmin iterations:
  Set $\widehat{B}_0 \leftarrow \widehat{B}^{(\text{i-1})}$
  for $\ell = 1$ to $T^{(\text{i})}$ **do**
    Let $\widehat{B} \leftarrow \widehat{B}_{\ell-1}$
    Update $\widehat{w}_{m,\ell}$, $\widehat{\theta}_{m,\ell}$: For $m \in [M]$, $\widehat{w}_{m,\ell} \leftarrow (X_{m,\tau}^{(\text{i})}\widehat{B})^{\dagger} Y_{m,\tau}^{(\text{i})}$ and $\widehat{\theta}_{m,\ell} \leftarrow \widehat{B}\widehat{w}_{m,\ell}$
    Gradient w.r.t $\widehat{B}$: With $Y_m^{(\text{i})} \equiv Y_{m,T^{(\text{i})}+\tau}^{(\text{i})}$, $X_m^{(\text{i})} \equiv X_{m,T^{(\text{i})}+\tau}^{(\text{i})}$, compute $\nabla_{\widehat{B}} f(\widehat{B}, \widehat{W}_\ell) = \sum_{m=1}^{M} \frac{1}{n_m^i} X_m^{(\text{i})\top}(X_m^{(\text{i})}\widehat{B}\widehat{w}_{m,\ell} - Y_m^{(\text{i})})\widehat{w}_{m,\ell}^{\top}$
    GD step: Set $\widehat{B}^{+} \leftarrow \widehat{B} - \eta \nabla_{\widehat{B}} f(\widehat{B}, \widehat{W}_\ell)$
    Projection: Compute $\widehat{B}^{+} \overset{QR}{=} B^{+}R^{+}$, set $\widehat{B}_\ell \leftarrow B^{+}$
  end for
  Set $\widehat{B}^{(\text{i})} \leftarrow \widehat{B}_{T^{(\text{i})}}$ and set $\widehat{W}^{(\text{i})} \leftarrow \widehat{W}_{T^{(\text{i})}}$
  Observe n_{M+1}^i samples $(X_{M+1}^{(\text{i})}, Y_{M+1}^{(\text{i})})$ for target task
  Compute $\widehat{w}_{M+1}^{(\text{i})} = \arg\min_w \|X_{M+1}^{(\text{i})\top}\widehat{B}^{(\text{i})}w - Y_{M+1}^{(\text{i})}\|^2$
  Estimate the relevance parameter as $\hat{\nu}_{i+1} = \arg\min_\nu \|\nu\|_2^2$ s.t. $\widehat{W}^{(\text{i})}\nu = \widehat{w}_{M+1}^{(\text{i})}$
end for

gradient descent with minimization achieves exponential decay in subspace distance (Theorem 2.3 in [37]). These guarantees provide the required control to finalize the proof of Lemma B.2, and demonstrate that the estimation of the relevance parameter achieves $\sqrt{\epsilon_i}$-accuracy with high probability.

V. Theoretical Analysis of A-LRRL Algorithm

In this section, we present the main result that provides guarantees for excess risk and sample complexities for the source and target tasks, and time complexity.

Theorem V.1: Assume Assumptions II.2 and II.3 hold. For $C > 1$, set the excess risk accuracy parameter $\epsilon \in (0, \frac{1}{6M}]$, $\eta = \frac{c}{\sigma_{\max}^{\star 2}}$ with a $c \leqslant 0.5$, and the number of GD iterations in the i^{th} epoch, $T^{(i)} = C\kappa^2 \log\left(\frac{\mu^4\kappa^2 k^{\frac{5}{2}}}{M\sqrt{\epsilon_i}}\right)$. If for all $i \in [\Gamma]$, and $m \in [M]$,

$$n_m^i \geqslant C \max\{\log d, \log M, k\} \max\left\{1, \frac{\mu^8\kappa^8 k^5 NSR}{M^2\epsilon_i}\right\}$$

$$\sum_{m=1}^{M} n_m^i \geqslant C\mu^2\kappa^4(d+k)k \max\left\{k\kappa^4, \kappa^2 NSR, \frac{\mu^8\kappa^{10}k^6 NSR}{M^2\epsilon_i}\right\},$$

and if the number of epochs

$$\Gamma \geqslant \log_2\left(\frac{2(\|\nu^\star\|_2^2 + M)M s_\Gamma^\star \sigma_{\max}^{\star 2}}{15\mu^6\kappa^8 k^5 \epsilon^2}\right),$$

then with probability at least $1 - \exp(-c(d+M)) - [\Gamma(T^{(\Gamma+1)} + \frac{1}{4}\log\epsilon_{\Gamma+1}) + 6]d^{-10}$, Algorithm 2 guarantees that at the end of the last epoch Γ,

$$\text{ER}(\widehat{B}^{(\Gamma)}, \widehat{w}_{M+1}^{(\Gamma)}) \leqslant \epsilon,$$

if, for each epoch $i \in [\Gamma]$, the number of target task samples n_{M+1}^i is at least

$$O\left(\max\left\{\frac{k + \log\frac{1}{\delta}}{M}\sigma_{\min}^{\star 2} NSR\epsilon^{-1}, \max\{\log M, k, \log d\}\max\{1, \frac{\mu^6\kappa^6 k^2 NSR}{M^2\epsilon_i}\}\right\}\right),$$

where $\|\nu^\star\|_{0,\gamma}^{\Gamma} = |\{m : |\nu^\star(m)| > \sqrt{\gamma}\epsilon_\Gamma\}|$, $s_\Gamma^\star = (1-\gamma)\|\nu^\star\|_{0,\gamma}^{\Gamma} + \gamma M$, $\epsilon_\Gamma = 2^{-\Gamma}$, and $NSR = \frac{\sigma^2 M}{\sigma_{\min}^{\star 2}}$.

Remark V.2: While our proof utilizes standard arguments from matrix learning literature, the primary focus of our work is fundamentally different. Our analysis focuses on the benefits of transferring the knowledge learned from source tasks to a target task, in addition to learning the subspace of the representation. Specifically, we alternate between (i) estimating the shared representation (which is the focus in matrix learning works; however, without adaptive sampling) and (ii) estimating the task relevance parameter for adaptive sampling. This leads to a temporal progression of the error, which necessitates a unique analysis and proof approach to quantify the error and ensure convergence. We derive an accurate estimation of $\widehat{\nu}$ (Lemma B.2), which directly impacts the significance of source tasks. This estimation allows us to substantially reduce the data requirements. Based on this, we further illustrate how the estimated $\widehat{\nu}$ can be utilized to derive an excess risk bound (Theorem V.1) with a reduced number of data samples.

Remark V.3: To clarify the benefits of active learning, we considered two boundary cases: Case (i): $\nu^\star$ is a 1-sparse vector, indicating that the target task relies only on a single source task. Case (ii): $\nu^\star$ is a scaled vector $\mathbf{1}$, where $\mathbf{1}$ represents the vector consisting of all ones. This relates to the uniform sampling scenario where all source tasks hold equal significance. For $\gamma = 0$, case (i) results in $s_\Gamma^\star = 1$, whereas case (ii) results in $s_\Gamma^\star = M$. As stated in Theorem V.1, in case (i), the number of epochs must satisfy $\Gamma \geqslant \log_2\left(\frac{2(\|\nu^\star\|_2^2+M)M\sigma_{\max}^{\star 2}}{15\mu^6\kappa^8 k^5\epsilon^2}\right)$, and in case (ii), the requirement becomes $\Gamma \geqslant \log_2\left(\frac{2(\|\nu^\star\|_2^2+M)M^2\sigma_{\max}^{\star 2}}{15\mu^6\kappa^8 k^5\epsilon^2}\right)$. This analysis indicates that uniform sampling (case (ii)) requires significantly more epochs, and thereby more source and target data, than the adaptive case (i). This result highlights that adaptive sampling achieves substantial benefits in settings with significant task relevance variability and fewer data samples.

Discussion on Theorem V.1: Theorem V.1 provides a high probability ϵ-guarantee for excess risk along with sample complexity requirements and number of epochs. The guarantee holds when (i) the per-source sample n_m^i and the total source sample $\sum_{m=1}^M n_m^i$ in each epoch satisfy the stated bounds, (ii) the target sample n_{M+1}^i meets its lower bound, and the number of epochs satisfies the provided condition. The theorem does not determine the distribution of samples among the source tasks. Rather, our algorithm uses an active allocation strategy by distributing samples based on estimated task relevance $\hat{\nu}$, while following the per-task and total sample requirements of Theorem V.1. Thus, the theorem specifies the theoretical feasibility conditions, while the algorithm provides a relevance-guided method for implementing such a feasible allocation.

Theorem V.1 shows that the sample complexity of the target tasks depends only on $k, \log(M), \log(d)$, rather than d, M, which is an advantage given $k \ll \min\{d, M\}$. Theorem V.1 shows that the number of epochs, thus the total number of source samples, depends on the task relevance denoted by $\nu^\star$ and the approximate sparsity of $\nu^\star$ denoted by $s^\star$, validating the effectiveness of the adaptive sampling approach over a uniform sampling approach (Remark V.3).

The guarantees for the excess risk and sample complexities are based on the estimation guarantee of the proposed estimator. By setting the number of GD iterations in i^{th} epoch as $T^{(i)} = C\kappa^2 \log(\frac{\mu^4\kappa^2 k^{\frac{5}{2}}}{M\sqrt{\epsilon_i}})$, at the end of epoch i, the alternating GD and minimization approach estimates the unknown rank-k feature matrix $\Theta^\star$ with high probability, if for each epoch i, the total source task samples $\sum_{m=1}^M n_m^i \geqslant C\mu^2\kappa^4(d+k)k\max\{1, \frac{\mu^8\kappa^{10}k^6 NSR}{M^2\epsilon_i}\}$ and the number of samples from each source task $n_m^i \geqslant C\max\{\log M, k\}\max\{1, \frac{\mu^8\kappa^8 k^5 NSR}{M^2\epsilon_i}\}$. Using the convergence guarantee for $\widehat{B}^{(i)}$ and $\widehat{W}^{(i)}$, we provide a guarantee to estimate the relevance parameter. In Lemma B.2, we show that under the $(\widehat{B}^{(i)}, \widehat{W}^{(i)})$ guarantee the estimate of the relevance parameter $|\hat{\nu}_i(m)|$ is $\sqrt{\epsilon_i}$ close to the true value $|\nu^\star(m)|$. Using this and after deriving some intermediate results using linear algebra and adopting some of the proof techniques of [6] for our proposed alternating GD and minimization algorithm, we provide the convergence guarantee for excess risk. We present the complete proof in Section V-B.

A. Time and Communication Complexities

To analyze the time complexity of a given epoch i, we first calculate the computation time for the initialization step. To calculate Θ_0, we need a time of order $\sum_{m=1}^M n_m^1 d$. The time complexity of the k-SVD step dMk times the number of iterations required. We notice that to obtain an initial estimate of the span of $B^\star$ that is δ_0-accurate, where $\delta_0 = \frac{c}{\kappa^2\sqrt{k}}$, it is sufficient to use an order $\log(\kappa k)$ number of iterations. Thus, since $n_m^1 \geqslant k$, the total complexity of the initialization phase is $O(d(\sum_{m=1}^M n_m^1 + Mk)\log(\kappa k)) = O(\sum_{m=1}^M n_m^1 d\log\kappa k)$. The time required for each gradient computation is $\sum_{m=1}^M n_m^i dk$. The QR decomposition process requires a time complexity of order dk^2. Additionally, the time required to update the columns of matrix W using the least squares method is $O(\sum_{m=1}^M n_m^i dk)$. The number of iterations of these steps for each epoch can be expressed as $T^{(i)} = O(\kappa^2\log(\frac{\mu^4\kappa^2 k^{\frac{5}{2}}}{M\sqrt{\epsilon_i}}))$. Upon finishing the alternating GD minimization iterations, in every epoch, we solve the least squared estimator to calculate $\widehat{w}_{M+1}^{(i)}$ and $\hat{\nu}_{i+1}$, with a complexity of $O(n_{M+1}^i dk + k^2 M)$. Thus, the overall time complexity is $O(\sum_{m=1}^M n_m^1 d\log(\kappa k) + \sum_{i=1}^\Gamma \max\{\sum_{m=1}^M n_m^i dk, dk^2\}T^{(i)} + \sum_{i=1}^\Gamma (n_{M+1}^i dk + k^2 M)) = O(\kappa^2 \sum_{i=1}^\Gamma \sum_{m=1}^M n_m^i dk \log\frac{\mu^4\kappa^2 k^{\frac{5}{2}}}{M\sqrt{\epsilon_i}} + k^2 M\Gamma) + \sum_{i=1}^\Gamma n_{M+1}^i dk$. The communication complexity for each task in each iteration is of the order of dk. Hence, the total is $O(dk\cdot\kappa\log\frac{1}{\epsilon})$.

B. Proof of Theorem V.1

Proof: From Lemma B.5 and Lemma B.4, using the union bound, we conclude that with probability at least $1 - 2\delta - [\Gamma(T^{(\Gamma+1)} + \frac{1}{4}\log\epsilon_{\Gamma+1}) + 4]d^{-10}$, the excess risk at the end of the last epoch Γ is bounded by

$$\text{ER}(\widehat{B}^{(\Gamma)}, \widehat{w}_{M+1}^{(\Gamma)}) \leqslant \frac{\sigma^2(2k + 3\log\frac{1}{\delta})}{1.8 n_{M+1}^\Gamma} + \frac{s_\Gamma^\star \sigma_{\max}^{\star^2} M^2 \epsilon_\Gamma^2}{75\beta\mu^6\kappa^8 k^5} \cdot \left(\sqrt{3}\sum_{m=1}^M n_m^\Gamma + \sqrt{12\log\frac{1}{\delta}\sum_{m=1}^M n_m^\Gamma} + \sqrt{12M}\log\frac{1}{\delta}\right).$$

Define $q_1 := \frac{s_\Gamma^\star \sigma_{\max}^{\star^2} M^2\epsilon_\Gamma^2}{75\beta\mu^6\kappa^8 k^5}$ and $q_2 := 10\epsilon_\Gamma^{-1}\beta(\|\nu^\star\|_2^2 + M)q_1$. Given $\Gamma \geqslant \log_2\left(\frac{2(\|\nu^\star\|_2^2 + M)M s_\Gamma^\star \sigma_{\max}^{\star^2}}{15\mu^6\kappa^8 k^5 \epsilon^2}\right)$, it follows that $2q_2\epsilon^{-1} \leqslant \frac{1}{3}$ when $\epsilon \leqslant \frac{1}{6M}$. To guarantee

$$q_1\left(\sum_{m=1}^M n_m^\Gamma + 2\sqrt{\log\frac{1}{\delta}}\sqrt{\sum_{m=1}^M n_m^\Gamma} + 2\sqrt{M}\log\frac{1}{\delta}\right) \leqslant \frac{1}{2}\epsilon,$$

by using Lemma B.6 and definition of q_2, we need to show

$$(1 - 2q_2\epsilon^{-1})\sum_{m=1}^M n_m^\Gamma - 4q_2\epsilon^{-1}\sqrt{\log\frac{1}{\delta}}\sqrt{\sum_{m=1}^M n_m^\Gamma} - 4q_2\epsilon^{-1}\sqrt{M}\log\frac{1}{\delta} \geqslant 0.$$

Using the quadratic formula for the inequality, we need to show

$$\sqrt{\sum_{m=1}^M n_m^\Gamma} \geqslant \frac{4q_2\epsilon^{-1}\sqrt{\log\frac{1}{\delta}} + \sqrt{16\left(q_2^2\epsilon^{-2} + (1-2q_2\epsilon^{-1})q_2\epsilon^{-1}\sqrt{M}\right)\log\frac{1}{\delta}}}{2(1-2q_2\epsilon^{-1})}.$$

By applying $\sqrt{x} + \sqrt{y} \leqslant \sqrt{2(x+y)}$, the inequality becomes

$$\sqrt{\sum_{m=1}^M n_m^\Gamma} \geqslant \sqrt{\frac{8\left(2q_2^2\epsilon^{-2} + (1-2q_2\epsilon^{-1})q_2\epsilon^{-1}\sqrt{M}\right)\log\frac{1}{\delta}}{(1-2q_2\epsilon^{-1})^2}}.$$

Utilizing the inequality $q_2\epsilon^{-1} \leqslant (1-2q_2\epsilon^{-1})^2$, we have

$$\sum_{m=1}^{M} n_m^{\Gamma} \geqslant 16q_2\epsilon^{-1}(1-\sqrt{M})\log\frac{1}{\delta} + 8\sqrt{M}\log\frac{1}{\delta}.$$

Since $M>1$ and by setting $\delta = d^{-10}$, the above inequality holds if

$$\sum_{m=1}^{M} n_m^{\Gamma} \geqslant 80\sqrt{M}\log d.$$

Given that $n_m^i \geqslant C\max\{\log d, \log M, k\}\max\left\{1, \frac{\mu^8\kappa^8k^5NSR}{M^2\epsilon_i}\right\}$, it is clearly that the above condition is inherently satisfied. By setting the target sample size as

$$n_{M+1}^{\Gamma} \geqslant \frac{(2k+3\log\frac{1}{\delta})\sigma_{\min}^{\star^2}NSR\epsilon^{-1}}{0.9M},$$

using $NSR = \frac{\sigma^2 M}{\sigma_{\min}^{\star^2}}$, we derive

$$\frac{\sigma^2(2k+3\log\frac{1}{\delta})}{1.8n_{M+1}^{\Gamma}} \leqslant \frac{1}{2}\epsilon.$$

Based on this analysis, we can conclude that with probability at least $1-[\Gamma(T^{(\Gamma+1)}+\frac{1}{4}\log\epsilon_{\Gamma+1})+6]d^{-10}$,

$$\mathrm{ER}(\widehat{B}^{(\Gamma)}, \widehat{w}_{M+1}^{(\Gamma)}) \leqslant \epsilon.$$

According to Theorem 2.2 in [37], if $\sum_{i=1}^{M} n_m^i \geqslant C\mu^2\kappa^6dk(k\kappa^2+NSR)$, then with probability at least $1-\exp(-c(d+M))$, $\mathrm{SD}(B^\star, \widehat{B}^{(0)}) \leqslant \delta_0 = \frac{c}{\sqrt{k}\kappa^2}$. Applying the union bound, we complete our proof. □

VI. Simulations

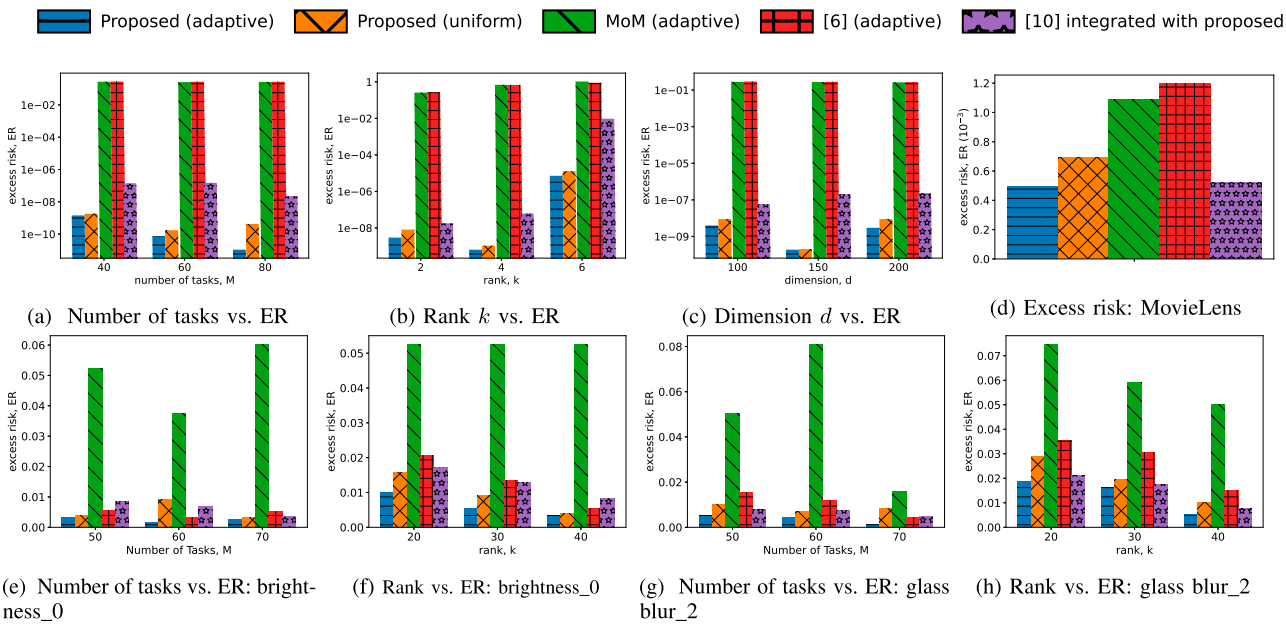
In this section, we present the numerical experiments that validate the effectiveness of our proposed algorithm on both synthetic, real-world MNIST-C, and MovieLens-100K datasets. While the proposed algorithm and guarantees are designed for linear representations, we conducted experiments on the MNIST-C dataset to evaluate the effectiveness of our approach on non-linear models. We performed a comparative analysis of our algorithm with four benchmark approaches: (i) the Method-of-Moments (MoM) estimator presented in [12], [14], (ii) the approach in Chen et al. [6], (iii) our proposed estimator via a uniform sampling approach, (iv) the approach in Collins et al. [11]. The algorithm in [11] does not consider transferability to new target task; for the sake of this analysis we augment their algorithm with the task relevance estimation approach proposed in the paper. We performed experiments on synthetic and MNIST-C datasets, varying the number of tasks M and the rank k. In experiments on synthetic data, we varied the dimension d in addition to the number of tasks M and the rank k. We noticed that the proposed algorithm consistently outperforms all four benchmark approaches, validating the benefit of our proposed approach.

A. Datasets

Synthetic data: In our experimental setup for the synthetic data, we defined the default setting parameters as $n_m^i = 50$, $d = 100$, $k = 2$, $M = 80$. Notice that $n_m^i < d$, which captures the data-scarce setting. In the experiments, we varied one of the parameters by keeping others fixed to the default setting. The entries of matrix $B^\star$ were randomly generated by orthonormalizing an i.i.d. standard Gaussian matrix. Similarly, the entries of matrix $W^\star$ for the source tasks were randomly generated according to an i.i.d. Gaussian distribution. The task relevance parameter $\nu^\star$ was generated by assigning 20% of tasks a weight of 2, 60% of tasks set to 6, and the remaining 20% tasks to 10. Using the generated $\nu^\star$ and $W^\star$, we construct $w_{M+1}^\star := W^\star\nu^\star$ for the target task. The matrices X_m were randomly generated using an i.i.d. standard Gaussian distribution. In addition, we utilized a noise model with a mean of zero and a variance of 10^{-6}. It is important to note that in our experiments when we change the rank, number of source tasks, or dimensions, the matrices $B^\star$ and $W^\star$, as well as the data, are generated based on the specific dimensional setting of the problem.

MNIST-C data: In our experiments for the MNIST-C data, we evaluated our proposed algorithms on the corrupted MNIST dataset (MNIST-C) used in [9], which consists of 16 unique types of corruption. Although the MNIST problem is typically framed as a classification task with cross-entropy loss, we reformulate it as a regression problem with ℓ_2 loss to align with the setting studied in this paper. To generate source and target tasks, each corrupted sub-dataset was partitioned into 10 tasks through the application of one-hot encoding to labels 0 through 9, resulting in 160 tasks, each identified as "corruption type + label." For each task, we converted the label into a binary format of $1/0$ based on the correspondence between the image and the label. Each task contained 28×28 dimensional 6000 images, which were normalized before processing. Experimental results are presented for two specific target tasks: brightness_0 and glass_blur_2. In our experiments, we defined the default parameter settings as $n_m^i = 100, d = 28^2 = 784, M = 50$, and $k = 40$. We varied the rank and the number of source tasks and evaluated the performance of our approach.

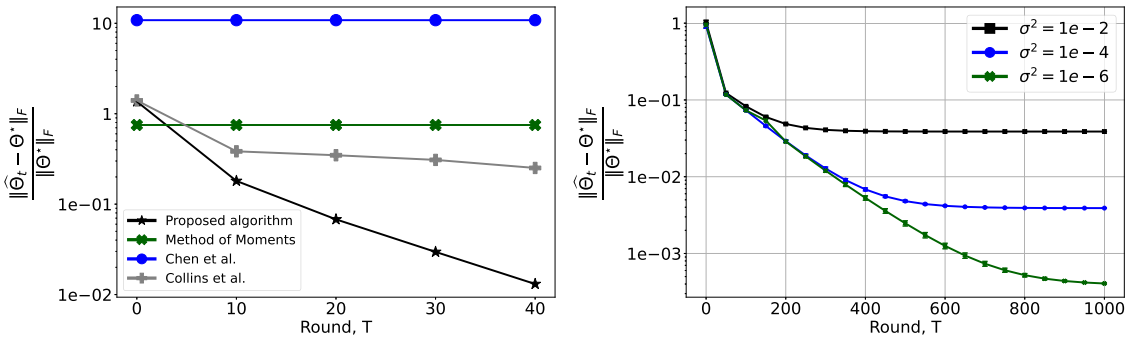
MovieLens-100K data: We evaluated our proposed method on the MovieLens-100K dataset [10], which contains 100,000 ratings from 943 users on 1,682 movies, with each user having rated at least 20 movies. To construct our multi-task learning framework, we first load the sparse rating matrix $\mathbf{R} \in \mathbb{R}^{943\times 1682}$. Since the original matrix is highly sparse (approximately 93.7% missing entries), we apply collaborative filtering using matrix factorization with 50 iterations, learning rate 0.01, and regularization parameter 0.1 to obtain a complete rating matrix. The completed ratings are then normalized to the range [0,1] by dividing by the maximum rating value of 5. To generate source and target tasks, we perform Non-negative Matrix Factorization (NMF) on the completed rating matrix to obtain user feature matrix $\mathbf{W} \in \mathbb{R}^{943\times d}$ and item feature matrix $\mathbf{H} \in \mathbb{R}^{1682\times d}$, where d represents the latent feature dimensionality. To generate M source tasks, we apply K-means clustering to the item feature matrix $\mathbf{H}$, where each cluster centroid serves



(a) Number of tasks vs. ER (b) Rank k vs. ER (c) Dimension d vs. ER (d) Excess risk: MovieLens

(e) Number of tasks vs. ER: brightness_0 (f) Rank vs. ER: brightness_0 (g) Number of tasks vs. ER: glass blur_2 (h) Rank vs. ER: glass blur_2

Fig. 1. **Synthetic data:** We considered 50 data samples for each source task and 30 data samples for the target task. We varied the number of tasks as $M = 40, 60, 80$, varied the rank of the $\Theta^\star$ as $k = 2, 4, 6$, and varied the dimension as $d = 100, 150, 200$. As shown in the plots (Fig. 1(a)–1(c)), our proposed approach with adaptive sampling outperforms the existing approaches. MNIST-C data: We considered 100 data samples for each source task and 50 data samples for the target task. We varied the number of tasks as $M = 50, 60, 70$, varied the rank of the $\Theta^\star$ as $k = 20, 30, 40$. The plot for MNIST-C data are presented in Fig. 1(e)–1(h). MovieLens-100K data: We consider 50 samples for each source task and 20 data samples for the target task. The plot for MovieLens-100K are presented in Fig. 1(d).



(a) Estimation error vs. GD iteration (b) Estimation error with varying noise variance

Fig. 2. We set the parameter as $d = 100$, $M = 80$, $k = 2$, noise variance $= 10^{-6}$.

as a task parameter. This approach ensures that similar movies are clustered as a task. Our experimental setup uses $d = 100$ latent dimensions, $M + 1 = 30$ total tasks (29 source tasks and 1 target task), rank $k = 10$, and $n_m^i = 50$ samples per source task, and 20 target task samples.

B. Results and Discussions

Estimation error plot. In Fig. 2(a), we present the plot for estimation error vs. GD iterations during the first epoch for the different algorithms. The MoM estimator is a noniterative method; hence, the estimation error is a single line. [6] considered a convex relaxed solution of the original non-convex problem via the projected gradient descent method to obtain the estimation. [11], on the other hand, does not provide an initialization guarantee, which affects the estimation again due to the non-convexity of the problem. It is difficult to obtain guarantees for the non-convex problem if the initialization error is not sufficiently small. The estimation error for the parameter matrix for the M tasks $\Theta^\star$ is very low in our proposed estimator and it outperforms all the benchmark approaches. Fig. 2(b) shows the estimation error of our proposed algorithm during the first epoch across different noise variances. This result clearly demonstrates that the estimation error decreases exponentially with each iteration. Furthermore, the error consistently decreases as the noise variance reduces, although it cannot be less than the noise variance. This validates the benefit of adaptive sampling for generalizing to target tasks.

Excess risk plots. Fig. 1 presents the plots for the excess risk. Fig. 1(a)–1(c) illustrate the excess risk for five algorithms as the number of source tasks M, rank k, and dimension d vary for synthetic data. Similarly, Fig. 1(e)–1(h) show the excess risk for the same algorithms while varying M and k for two target tasks

from the MNIST-C dataset. Fig. 1(d) displays the excess risk for all the algorithms on the MovieLens-100K dataset. We notice that our proposed approach outperforms the MoM estimator-based approach and the approach in Chen et al. This is because, as also noted in [6], during the iterative estimation process, the estimation error propagates from round to round due to unknown $\nu^\star$. Since the MoM estimator and the convex-relaxation approach in [6] have considerable errors in the estimation of $\Theta^\star$, it negatively affects the estimation of $\nu^\star$. Our adaptive sampling approach slightly outperforms the uniform sampling method. We note that the benefit of adaptive sampling is majorly in the sample complexity while ensuring no worse convergence error guarantee compared to uniform sampling. Our approach also outperforms a baseline approach in which we augmented the algorithm in [11] with the adaptive transfer learning approach presented in this paper. We note that, the approach in [11], do not consider transferability to a new task. This is expected due to the initialization error in [11]. Thus, the numerical experiments validate our theoretical findings and the effectiveness of our approach.

VII. Conclusion and Future Work

In this work, we introduced a novel active-representation learning algorithm using an adaptive sampling-based alternating GD and minimization approach. Our proposed algorithm is specifically designed for active multi-task representation learning by considering the *unknown* task relevance to enable adaptive sampling. Our proposed approach can handle data-scarce settings where the number of source data samples is fewer than the problem dimension. We have demonstrated the algorithm's convergence guarantee in estimating the unknown feature matrix and the unknown relevance parameter. Additionally, we have evaluated the effectiveness of our approach in comparison with benchmark algorithms. The results clearly show that our proposed algorithm outperforms the benchmark approaches, thus validating its advantage over existing methods. As part of our future work, we aim to extend our approach to handle non-i.i.d. data. Inspired by promising empirical results on nonlinear models, we plan to extend our approach to nonlinear representations in future work.

Appendix A
Preliminaries

Proposition A.1 (Theorem 2.8.1, [38]): Let $X_1, \cdots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $g \geqslant 0$, we have

$$\mathbb{P}\Big\{|\sum_{i=1}^{N} X_i \geqslant g|\Big\} \leqslant 2\exp\left[-c\min\left(\frac{g^2}{\sum_{i=1}^{N}||X_i||_{\psi_1}^2}, \frac{g}{\max_i ||X_i||_{\psi_1}^2}\right)\right],$$

where $c > 0$ is an absolute constant.

Proposition A.2 ([39]): If V is an $n \times n$ symmetrical matrix and if X is an arbitrary $n \times q$ real matrix, then

$$(V+XX^\top)^\dagger = V^\dagger - V^\dagger X(I+X^\top V^\dagger X)^{-1}X^\top V^\dagger + (X_\perp^\dagger)^\top X_\perp^\dagger,$$

where $X_\perp = (I - VV^\dagger)X$.

Lemma A.3: If for any $m \in [M+1]$, $n_m \geqslant C\max\{\log d, \log M\}$, then for each $m \in [M+1]$, with probability at least $1 - 2d^{-10}$, it holds that

$$0.9n_m I \preceq X_m^\top X_m \preceq 1.1n_m I,$$

where n_m denotes the number of rows in X_m.

Proof: Given that $X_m^\top X_m = \sum_{n=1}^{n_m} x_{m,n}x_{m,n}^\top$, where $x_{m,n}x_{m,n}^\top \succeq 0$ and $\lambda_{\max}(x_{m,n}x_{m,n}^\top) \leqslant 1$. Since

$$\lambda_{\min}\left(\sum_{n=1}^{n_m}\mathbb{E}\left[x_{m,n}x_{m,n}^\top\right]\right) = \lambda_{\max}\left(\sum_{n=1}^{n_m}\mathbb{E}\left[x_{m,n}x_{m,n}^\top\right]\right) = n_m,$$

by applying the Matrix Chernoff inequality, we have with probability at least $1 - d\exp(-\frac{\delta'^2 n_m}{2})$,

$$\lambda_{\min}\left(\sum_{n=1}^{n_m} x_{m,n}x_{m,n}^\top\right) \geqslant (1-\delta')n_m$$

and with probability at least $1 - d\exp(-\frac{\delta'^2 n_m}{3})$,

$$\lambda_{\max}\left(\sum_{n=1}^{n_m} x_{m,n}x_{m,n}^\top\right) \leqslant (1+\delta')n_m.$$

Applying union bound and setting $\delta' = 0.1$, with probability at least $1 - 2\exp(\log d + \log(M+1) - \frac{\min_{m\in[M+1]} n_m}{300})$, for all $m \in [M+1]$, we conclude that

$$0.9n_m I \preceq X_m^\top X_m \preceq 1.1n_m I.$$

To ensure probability at least $1 - 2d^{-10}$ guarantees for our lemma, it is necessary to set the bounds of n_m for all $m \in [M+1]$. These bounds must guarantee that the following probability is at least $1 - 2d^{-10}$: $1 - 2\exp(\log d + \log(M+1) - \frac{\min_{m\in[M+1]} n_m}{300})$, for all $m \in [M+1]$. This required that the exponential term be substantially smaller than or equal to d^{-10}. Given $M \geqslant 2$, we have $\log(M+1) \leqslant C\log M$, we obtain for all $m \in [M+1]$,

$$\log d + \log(M+1) - \frac{\min_{m\in[M+1]} n_m}{300} \leqslant -10\log d$$
$$\Rightarrow n_m \geqslant C\max\{\log d, \log M\}.$$

□

Appendix B
Supporting Results and Proofs

Lemma B.1: For any given epoch $i \in [\Gamma]$ and task $m \in [M]$, consider the estimates $(\widehat{B}^{(i)}, \widehat{W}^{(i)})$. Define $\Delta_i := B^\star W^\star - \widehat{B}^{(i)}\widehat{W}^{(i)}$. Then the following inequality holds

$$\Big|(B^\star w_m^\star)^\top\left(((B^\star W^\star)(B^\star W^\star)^\top)^\dagger - ((\widehat{B}^{(i)}\widehat{W}^{(i)})(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top)^\dagger\right)B^\star w_M^\star\Big| \leqslant \|w_m^\star\|_2\|w_{M+1}^\star\|_2\|(\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top)^\dagger\|_F\|\Delta_i\|_F \cdot \left(\|(\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top)^\dagger\|_F\|\Delta_i\|_F + 2\|(\widehat{W}^{(i)})^\dagger\|_F\right).$$

Proof: First, we analyze the term $((B^\star W^\star)(B^\star W^\star)^\top)^\dagger - ((\widehat{B}^{(i)}\widehat{W}^{(i)})(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top)^\dagger$.

$$((B^\star W^\star)(B^\star W^\star)^\top)^\dagger - ((\widehat{B}^{(i)}\widehat{W}^{(i)})(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top)^\dagger$$

$$
\begin{aligned}
&= \left((\widehat{B}^{(i)}\widehat{W}^{(i)} + \Delta_i)(\widehat{B}^{(i)}\widehat{W}^{(i)} + \Delta_i)^\top\right)^\dagger \\
&\quad - \left((\widehat{B}^{(i)}\widehat{W}^{(i)})(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top\right)^\dagger \\
&= \left((\widehat{B}^{(i)}\widehat{W}^{(i)})(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top \right. \\
&\quad \left. + \left(\Delta_i\Delta_i^\top + \Delta_i(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top + (\widehat{B}^{(i)}\widehat{W}^{(i)})\Delta_i^\top\right)\right)^\dagger \\
&\quad - \left((\widehat{B}^{(i)}\widehat{W}^{(i)})(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top\right)^\dagger. \qquad (5)
\end{aligned}
$$

Let $XX^\top := (\Delta_i\Delta_i^\top + \Delta_i(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top + (\widehat{B}^{(i)}\widehat{W}^{(i)})\Delta_i^\top)$ and $V := (\widehat{B}^{(i)}\widehat{W}^{(i)})(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top$. We can rewrite Eq. (5) as

$$
\begin{aligned}
&\left(((B^\star W^\star)(B^\star W^\star)^\top)^\dagger - V^\dagger\right) \\
&= ((V + \Delta_i\Delta_i^\top + \Delta_i(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top + (\widehat{B}^{(i)}\widehat{W}^{(i)})\Delta_i^\top)^\dagger - V^\dagger)
\end{aligned}
$$

We use Proposition A.2 in the rest of our proof. Let us denote SVD of $\widehat{B}^{(i)}\widehat{W}^{(i)}$ as $U\Sigma\bar{V}^\top$. We have

$$
\begin{aligned}
VV^\dagger &= U\Sigma^2U^\top(U\Sigma^2U^\top)^\dagger = UU^\top, \qquad (6)\\
X_\perp^\dagger B^\star &= ((I - VV^\dagger)X)^\dagger B^\star = ((I - UU^\top)X)^\dagger B^\star \\
&= (U_\perp U_\perp^\top X)^\dagger B^\star = X^\dagger(U_\perp U_\perp^\top)^\dagger B^\star \qquad (7)\\
&= X^\dagger U_\perp U_\perp^\top B^\star = 0, \qquad (8)
\end{aligned}
$$

where Eq. (7) is derived from $UU^\top + U_\perp U_\perp^\top = I$ for an orthonormal matrix U and $(AB)^\dagger = B^\dagger A^\dagger$. Eq. (8) is derived from $A^\dagger = A$ for any orthogonal projection matrix A, and that $B^\star$ lies in the column spaces as $B^\star W^\star$. Thus $X_\perp^{\dagger^\top} X_\perp^\dagger B^\star = 0$. By applying Proposition A.2 and taking into account the fact that $(I + X^\top V^\dagger X)^{-1} \preceq I$, and $X_\perp^{\dagger^\top} X_\perp^\dagger B^\star = 0$, we have

$$
\begin{aligned}
&|(B^\star w_m^\star)^\top \left(((B^\star W^\star)(B^\star W^\star)^\top)^\dagger - V^\dagger\right) B^\star w_{M+1}^\star| \\
&\leqslant \|\left(V^\dagger(\Delta_i\Delta_i^\top + \Delta_i(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top + (\widehat{B}^{(i)}\widehat{W}^{(i)})\Delta_i^\top)V^\dagger\right)\|_F \cdot \\
&\|B^\star w_m^\star\|_2 \cdot \|B^\star w_{M+1}^\star\|_2 \\
&\leqslant \|B^\star w_m^\star\|_2 \cdot \|V^\dagger\Delta_i\Delta_i^\top V^\dagger\|_F \cdot \|B^\star w_{M+1}^\star\|_2 \\
&\quad + \|B^\star w_m^\star\|_2\Big(\|V^\dagger\Delta_i(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top V^\dagger\|_F \\
&\quad + \|V^\dagger(\widehat{B}^{(i)}\widehat{W}^{(i)})\Delta_i^\top V^\dagger\|_F\Big) \cdot \|B^\star w_{M+1}^\star\|_2.
\end{aligned}
$$

Given that

$$
\begin{aligned}
&\|(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top((\widehat{B}^{(i)}\widehat{W}^{(i)})(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top)^\dagger\|_F \\
&= \|((\widehat{B}^{(i)}\widehat{W}^{(i)})(\widehat{B}^{(i)}\widehat{W}^{(i)})^\top)^\dagger(\widehat{B}^{(i)}\widehat{W}^{(i)})\|_F \\
&= \|((\widehat{B}^{(i)})^\top)^\dagger(\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top)^\dagger(\widehat{B}^{(i)})^\dagger(\widehat{B}^{(i)}\widehat{W}^{(i)})\|_F \\
&= \|\widehat{B}^{(i)}(\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top)^\dagger\widehat{W}^{(i)}\|_F = \|(\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top)^\dagger\widehat{W}^{(i)}\|_F \qquad (9)\\
&= \|((\widehat{W}^{(i)})^\top)^\dagger((\widehat{W}^{(i)})^\dagger\widehat{W}^{(i)})\|_F = \|((\widehat{W}^{(i)})^\dagger\widehat{W}^{(i)})^\top(\widehat{W}^{(i)})^\dagger\|_F \\
&= \|(\widehat{W}^{(i)})^\dagger\widehat{W}^{(i)}(\widehat{W}^{(i)})^\dagger\|_F = \|(\widehat{W}^{(i)})^\dagger\|_F, \qquad (10)
\end{aligned}
$$

where Eq. (9) follows since $\widehat{B}^{(i)}$ is a unitary matrix and $(AB)^\dagger = B^\dagger A^\dagger$. We then use $\|A\|_F = \|A^\top\|_F$ and $(A^\dagger A)^\top = A^\dagger A$. Eq. (10) is from $A^\dagger AA^\dagger = A^\dagger$. Therefore, the final bound can be expressed as

$$
\begin{aligned}
&\left|(B^\star w_m^\star)^\top \left(((B^\star W^\star)(B^\star W^\star)^\top)^\dagger - V^\dagger\right) B^\star w_{M+1}^\star\right| \\
&\leqslant \|B^\star w_m^\star\|_2\|V^\dagger\|_F^2\|\Delta_i\|_F^2\|B^\star w_{M+1}^\star\|_2 \\
&\quad + \|B^\star w_m^\star\|_2\|V^\dagger\|_F\|\Delta_i\|_F\|(\widehat{W}^{(i)})^\dagger\|_F\|B^\star w_{M+1}^\star\|_2 \\
&\quad + \|B^\star w_m^\star\|_2\|(\widehat{W}^{(i)})^\dagger\|_F\|\Delta_i\|_F\|V^\dagger\|_F\|B^\star w_{M+1}^\star\|_2 \\
&\leqslant \|w_m^\star\|_2\,\|(\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top)^\dagger\|_F^2\|\Delta_i\|_F^2\,\left\|w_{M+1}^\star\right\|_2 \\
&\quad + 2\|w_m^\star\|_2\|(\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top)^\dagger\|_F\|\Delta_i\|_F\|(\widehat{W}^{(i)})^\dagger\|_F\left\|w_{M+1}^\star\right\|_2 \\
&= \|w_m^\star\|_2\,\|w_{M+1}^\star\|_2\|(\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top)^\dagger\|_F\|\Delta_i\|_F \\
&\quad \cdot \left(\|(\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top)^\dagger\|_F\|\Delta_i\|_F + 2\|(\widehat{W}^{(i)})^\dagger\|_F\right).
\end{aligned}
$$

Thus, we complete the proof. □

Using Lemma B.1, we now present the result for estimation guarantee of the relevance parameter.

Lemma B.2: Assume Assumptions II.2 and II.3 hold. Let $\mathrm{SD}(B^\star, \widehat{B}^{(0)}) \leqslant \delta_0 = \frac{c}{\sqrt{k}\kappa^2}$, $\eta \leqslant \frac{0.5}{\sigma_{\max}^{\star 2}}$, and the number of GD iterations in the i^{th} epoch, $T^{(i)} = C\kappa^2\log\left(\frac{\mu^4\kappa^2k^{\frac{5}{2}}}{M\sqrt{\epsilon_i}}\right)$. If

$$
\begin{aligned}
n_m^i &\geqslant C\max\{\log M, k\}\max\left\{1, \frac{\mu^8\kappa^8k^5NSR}{M^2\epsilon_i}\right\} \\
\sum_{i=1}^{M} n_m^i &\geqslant C\mu^2\kappa^4(d+k)k\max\left\{1, \frac{\mu^8\kappa^{10}k^6NSR}{M^2\epsilon_i}\right\} \\
n_{M+1}^i &\geqslant C\max\{\log M, k, \log d\}\max\left\{1, \frac{\mu^6\kappa^6k^2NSR}{M^2\epsilon_i}\right\}
\end{aligned}
$$

then with probability at least $1 - [i(T^{(i+1)} + \frac{1}{4}\log\epsilon_{i+1}) + 4]d^{-10}$,

$$
|\nu^\star(m)| - \sqrt{\epsilon_i} \leqslant |\hat{\nu}_{i+1}(m)| \leqslant |\nu^\star(m)| + \sqrt{\epsilon_i}.
$$

and

$$
|\hat{\nu}_{i+1}(m)| = \begin{cases} \left[\frac{1}{2}|\nu^\star(m)|, \frac{3}{2}|\nu^\star(m)|\right], & \text{if} |\nu^\star(m)| \geqslant 2\sqrt{\epsilon_i} \\ \left[0, 3\sqrt{\epsilon_i}\right], & \text{if} |\nu^\star(m)| \leqslant 2\sqrt{\epsilon_i}. \end{cases}
$$

Proof: Consider any epoch i and its corresponding estimated representation $\widehat{B}^{(i)}$. Using the least squared method, we obtain

$$
\begin{aligned}
\widehat{w}_{M+1}^{(i)} &= \arg\min_w \|X_{M+1}^{(i)}\widehat{B}^{(i)}w - Y_{M+1}\|^2 \\
&= ((X_{M+1}^{(i)}\widehat{B}^{(i)})^\top(X_{M+1}^{(i)}\widehat{B}^{(i)}))^{-1}(X_{M+1}^{(i)}\widehat{B}^{(i)})^\top Y_{M+1}^{(i)} \\
&= (\widehat{B}^{(i)^\top}X_{M+1}^{(i)^\top}X_{M+1}^{(i)}\widehat{B}^{(i)})^{-1}\widehat{B}^{(i)^\top}X_{M+1}^{(i)^\top}X_{M+1}^{(i)}B^\star w_{M+1}^\star \\
&\quad + (\widehat{B}^{(i)^\top}X_{M+1}^{(i)^\top}X_{M+1}^{(i)}\widehat{B}^{(i)})^{-1}\widehat{B}^{(i)^\top}X_{M+1}^{(i)^\top}Z_{M+1}^{(i)}. \qquad (11)
\end{aligned}
$$

In Eq. (11) we use $Y_{M+1}^{(i)} = (X_{M+1}^{(i)}B^\star w_{M+1}^\star + Z_{M+1}^{(i)})$. Using Lemma E.8 in [6] with $M_1 = \widehat{W}^{(i)}$ and $M_2 = \widehat{w}_{M+1}^{(i)}$, we have

$$
\begin{aligned}
|\hat{\nu}_{i+1}(m)| &= |\widehat{w}_m^{(i)^\top}(\widehat{W}^{(i)}\widehat{W}^{(i)^\top})^\dagger\widehat{w}_{M+1}^{(i)}| \\
&= |(\widehat{B}^{(i)}\widehat{w}_m^{(i)})^\top V^\dagger(\widehat{B}^{(i)}\widehat{w}_{M+1}^{(i)})| \qquad (12)
\end{aligned}
$$

Now replacing $\widehat{w}_{M+1}^{(i)}$ in Eq. (12) using Eq. (11) and $Q := \widehat{B}^{(i)^\top}X_{M+1}^{(i)^\top}X_{M+1}^{(i)}\widehat{B}^{(i)}$ and by adding and substracting terms, we rewrite Eq. (12) as $|\hat{\nu}_{i+1}(m)|$

$$
\begin{aligned}
&= |(\widehat{B}^{(i)}\widehat{w}_m^{(i)})^\top V^\dagger\widehat{B}^{(i)}Q^{-1}\widehat{B}^{(i)^\top}X_{M+1}^{(i)^\top}X_{M+1}^{(i)} \\
&\quad \widehat{B}^{(i)}\widehat{B}^{(i)^\top}B^\star w_{M+1}^\star
\end{aligned}
$$

$$
\begin{aligned}
&+(\widehat{B}^{(i)}\widehat{w}_m^{(i)})^\top V^\dagger \widehat{B}^{(i)} Q^{-1}\widehat{B}^{(i)^\top} X_{M+1}^{(i)^\top} X_{M+1}^{(i)}\\
&(I-\widehat{B}^{(i)}\widehat{B}^{(i)^\top})B^\star w_{M+1}^\star\\
&+\widehat{w}_m^{(i)^\top}(\widehat{W}^{(i)}\widehat{W}^{(i)^\top})^\dagger Q^{-1}\widehat{B}^{(i)^\top} X_{M+1}^{(i)^\top} Z_{M+1}^{(i)}|\\
=&|(B^\star w_m^\star)^\top V^\dagger B^\star w_{M+1}^\star + \underbrace{(\widehat{B}^{(i)}\widehat{w}_m^{(i)} - B^\star w_m^\star)^\top V^\dagger B^\star w_{M+1}^\star}_{\text{Term 1}}\\
&+\underbrace{(\widehat{B}^{(i)}\widehat{w}_m^{(i)})^\top V^\dagger(\widehat{B}^{(i)}\widehat{B}^{(i)^\top}-I)B^\star w_{M+1}^\star}_{\text{Term 2}}\\
&+(\widehat{B}^{(i)}\widehat{w}_m^{(i)})^\top V^\dagger \widehat{B}^{(i)} Q^{-1}\widehat{B}^{(i)^\top} X_{M+1}^{(i)^\top}\\
&\underbrace{X_{M+1}^{(i)}(I-\widehat{B}^{(i)}\widehat{B}^{(i)^\top})B^\star w_{M+1}^\star}_{\text{Term 3}}\\
&+\underbrace{\widehat{w}_m^{(i)^\top}(\widehat{W}^{(i)}\widehat{W}^{(i)^\top})^\dagger Q^{-1}\widehat{B}^{(i)^\top} X_{M+1}^{(i)^\top} Z_{M+1}^{(i)}}_{\text{Term 4}}|
\end{aligned}
$$

Using the triangular inequality, we have

$$
\begin{aligned}
|\hat{\nu}_{i+1}(m)| \leqslant\ & |(B^\star w_m^\star)^\top V^\dagger B^\star w_{M+1}^\star|\\
&+|\text{Term 1}|+|\text{Term 2}|+|\text{Term 3}|+|\text{Term 4}|
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
|\hat{\nu}_{i+1}(m)| \geqslant\ & |((B^\star w_m^\star)^\top V^\dagger B^\star w_{M+1}^\star|\\
&-(|\text{Term 1}|+|\text{Term 2}|+|\text{Term 3}|+|\text{Term 4}|)
\end{aligned}
$$

From Lemma E.8 in [6], $\nu^\star(m) = w_m^{\star^\top}(W^\star W^{\star^\top})^{-1}w_{M+1}^\star = (B^\star w_m^\star)^\top((B^\star W^\star)(B^\star W^\star)^\top)^\dagger(B^\star w_{M+1}^\star)$. Applying the Cauchy–Schwarz inequality and Lemma B.1, we have

$$
\begin{aligned}
&|\hat{\nu}_{i+1}(m)| - |\nu^\star(m)|\\
&\leqslant \left|(B^\star w_m^\star)^\top\left(V^\dagger - ((B^\star W^\star)(B^\star W^\star)^\top)^\dagger\right)B^\star w_{M+1}^\star\right|\\
&\quad +|\text{Term 1}|+|\text{Term 2}|+|\text{Term 3}|+|\text{Term 4}|
\end{aligned}
$$

We bound Term 1, Term 2, Term 3, and Term 4 as

$$
\begin{aligned}
|\text{Term 1}| &\leqslant \|w_{M+1}^\star\|_2\|(\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top)^\dagger\|_F\|\widehat{\theta}_m^{(i)} - \theta_m^\star\|_2\\
|\text{Term 2}| &\leqslant \|\widehat{w}_m^{(i)}\|_2\|(\widehat{W}^{(i)}\widehat{W}^{(i)\top})^\dagger\|_F\|\\
&\qquad (\widehat{B}^{(i)}\widehat{B}^{(i)^\top}-I)B^\star\|\|w_{M+1}^\star\|_2\\
|\text{Term 3}| &\leqslant \|\widehat{w}_m^{(i)}\|_2\|(\widehat{W}^{(i)}\widehat{W}^{(i)\top})^\dagger\\
&\qquad \|_F\|\widehat{B}^{(i)}Q^{-1}\widehat{B}^{(i)^\top}X_{M+1}^{(i)^\top}X_{M+1}^{(i)}\|_F\\
&\qquad \cdot\|(I-\widehat{B}^{(i)}\widehat{B}^{(i)^\top})B^\star\|\|w_{M+1}^\star\|_2\\
|\text{Term 4}| &\leqslant \|(\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top)^\dagger\|_F\|Q^{-1}\|\|\widehat{w}_m^{(i)}\|_2\|\\
&\qquad \widehat{B}^{(i)^\top}X_{M+1}^{(i)^\top}Z_{M+1}^{(i)}\|_2
\end{aligned}
$$

We now substitute the above bounds and the bound in Lemma B.1 to obtain the bound for $|\hat{\nu}_{i+1}(m)| - |\nu^\star(m)|$. Similarly we obtain the upper bound for $|\nu^\star(m)| - |\hat{\nu}_{i+1}(m)|$. Our goal now is to obtain bounds for each of the terms.

Given matrix $\widehat{W}^{(i)} \in \mathbb{R}^{k\times M}$, utilizing the SVD, we derive

$$
\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top = \widehat{U}\widehat{\Sigma}\widehat{V}^\top\widehat{V}\widehat{\Sigma}^\top\widehat{U}^\top = \widehat{U}(\widehat{\Sigma}\widehat{\Sigma}^\top)\widehat{U}^\top.
$$

Thus $\|(\widehat{W}^{(i)}(\widehat{W}^{(i)})^\top)^\dagger\|_F$

$$
= \|\widehat{U}(\widehat{\Sigma}\widehat{\Sigma}^\top)^\dagger\widehat{U}^\top\|_F = \sqrt{\sum_{i=1}^k\left(\frac{1}{\sigma_i^2}\right)^2} \leqslant \frac{\sqrt{k}}{\sigma_{\min}^2(\widehat{W}^{(i)})}
$$

and

$$
\|(\widehat{W}^{(i)})^\dagger\|_F = \sqrt{\sum_{i=1}^k\frac{1}{\sigma_i^2}} \leqslant \frac{\sqrt{k}}{\sigma_{\min}(\widehat{W}^{(i)})}.
$$

According to Lemma B.9 in [37], with probability at least $1 - M\exp(k - cn_{M+1}^i)$, $\|Q^{-1}\| \leqslant \frac{1}{0.7n_{M+1}^i}$. By combining this with Lemma A.3 and applying the union bound, with probability at least $1 - M\exp(k - cn_{M+1}^i) - 2d^{-10}$,

$$
\begin{aligned}
&\|\widehat{B}^{(i)}Q^{-1}\widehat{B}^{(i)^\top}X_{M+1}^{(i)^\top}X_{M+1}^{(i)}\|_F\\
&\leqslant \|\widehat{B}^{(i)}Q^{-1}\widehat{B}^{(i)^\top}\|_F\|X_{M+1}^{(i)^\top}X_{M+1}^{(i)}\|_2\\
&\leqslant \|\widehat{B}^{(i)}\|_2\|Q^{-1}\widehat{B}^{(i)^\top}\|_F\|X_{M+1}^{(i)^\top}X_{M+1}^{(i)}\|_2\\
&\leqslant \|\widehat{B}^{(i)}\|_2^2\|Q^{-1}\|_2\|X_{M+1}^{(i)^\top}X_{M+1}^{(i)}\|_2\\
&\leqslant \frac{1}{0.7n_{M+1}^i}1.1n_{M+1}^i \leqslant 1.6
\end{aligned}
$$

According to Lemma 3.5 in [37], assume $\mathrm{SD}(\widehat{B}^{(i)}, B^\star) \leqslant \delta^{(i)}$, if $\delta^{(i)} \leqslant \frac{c}{\sqrt{k}\kappa^2}$, and if $n_m^i \geqslant C\max\{\log M, k, \max\{\log M, k\}\frac{NSR}{\delta^{(i)^2}k}\}$, then with probability at least $1 - d^{-10}$,

$$
\begin{aligned}
&\|\widehat{\theta}_m^{(i)} - \theta_m^\star\|_2 \leqslant 1.4\mu\delta^{(i)}\sqrt{\frac{k}{M}}\sigma_{\max}^\star,\\
&\|\widehat{\Theta}^{(i)} - \Theta^\star\|_F \leqslant 1.4\mu\delta^{(i)}\sqrt{k}\sigma_{\max}^\star,\\
&\sigma_{\min}(\widehat{W}^{(i)}) \geqslant 0.8\sigma_{\min}^\star,\\
&\|\widehat{w}_m^{(i)}\| \leqslant 1.1\mu\sqrt{\frac{k}{M}}\sigma_{\max}^\star.
\end{aligned}
$$

Based on Assumption II.3, we have $\|w_m^\star\|^2 \leqslant \mu^2\frac{k}{M}\sigma_{\max}^{\star^2}$. In order to determine the upper bound for $\widehat{B}^{(i)^\top}X_{M+1}^{(i)^\top}Z_{M+1}^{(i)}$, let's consider a fixed $z \in \mathcal{S}_k$. we analyze $z^\top\widehat{B}^{(i)^\top}X_{M+1}^{(i)^\top}Z_{M+1}^{(i)} = \sum_{j=1}^{n_{M+1}^i}(\widehat{B}^{(i)}z)^\top x_{M+1,j}^{(i)}Z_{M+1,j}^{(i)}$, resulting in $\mathbb{E}[(\widehat{B}^{(i)}z)^\top x_{M+1,j}^{(i)}Z_{M+1,j}^{(i)}] = 0$ and

$$
\begin{aligned}
&\mathrm{Var}((\widehat{B}^{(i)}z)^\top x_{M+1,j}^{(i)})\\
&= \mathbb{E}[(\widehat{B}^{(i)}z)^\top x_{M+1,j}^{(i)}]^2 - (\mathbb{E}[(\widehat{B}^{(i)}z)^\top x_{M+1,j}^{(i)}])^2\\
&= \mathbb{E}[(\widehat{B}^{(i)}z)^\top x_{M+1,j}^{(i)}]^2 = \mathbb{E}[z^\top\widehat{B}^{(i)^\top}x_{M+1,j}^{(i)}x_{M+1,j}^{(i)^\top}\widehat{B}^{(i)}z]\\
&= z^\top\widehat{B}^{(i)^\top}\mathbb{E}[x_{M+1,j}^{(i)}x_{M+1,j}^{(i)^\top}]\widehat{B}^{(i)}z = 1.
\end{aligned}
$$

Given $Z_{M+1,j}^{(i)} \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$, we have $\mathrm{Var}(Z_{M+1,j}^{(i)}) = \sigma^2$. Thus, $z^\top\widehat{B}^{(i)^\top}X_{M+1}^{(i)^\top}Z_{M+1}^{(i)}$ is a sum of n_{M+1}^i subexponential random variables with parameter $K_j \leqslant C\sigma$. We apply the subexponential Bernstein inequality in Proposition A.1 by setting $g = \epsilon_1 n_{M+1}^i$. To apply the inequality, we show that

$$
\frac{g^2}{\sum_{j=1}^{n_{M+1}^i}K_j^2} \geqslant \frac{\epsilon_1^2 n_{M+1}^{i^2}}{Cn_{M+1}^i\sigma^2} = \frac{c\epsilon_1^2 n_{M+1}^i}{\sigma^2}
$$

$$\frac{g}{\max_j K_j} \geqslant \frac{\epsilon_1 n_{M+1}^i}{C\sigma} = \frac{c\epsilon_1 n_{M+1}^i}{\sigma}.$$

For $\epsilon_1 \leqslant \sigma$, the first term above is smaller. Therefore, for a fixed $z \in \mathcal{S}_k$, with probability at least $1 - \exp(-\frac{c\epsilon_1^2 n_{M+1}^i}{\sigma^2})$, $z^\top \widehat{B}^{(i)\top} X_{M+1}^{(i)\top} Z_{M+1}^{(i)} \leqslant \epsilon_1 n_{M+1}^i$. Using epsilon-net over all z adds a factor of $\exp(k)$. Thus, setting $\epsilon_1 = \mu\frac{k^2}{\sqrt{M}}\sigma^\star_{\max}\delta^{(i)}$, with probability at least $1 - \exp(k - \frac{c\mu^2 k^4 \sigma^{\star 2}_{\max}\delta^{(i)2} n_{M+1}^i}{M\sigma^2})$, we have $\|\widehat{B}^{(i)\top} X_{M+1}^{(i)\top} Z_{M+1}^{(i)}\| \leqslant \epsilon_1 n_{M+1}^i \leqslant \mu\frac{k^2}{\sqrt{M}}\sigma^\star_{\max}\delta^{(i)} n_{M+1}^i$. By combining the aforementioned results and the union bound, we can determine with probability at least $1 - 2d^{-10} - M\exp(k - cn_{M+1}^i) - \exp(k - \frac{c\mu^2 k\sigma^{\star 2}_{\max}\delta^{(i)2} n_{M+1}^i}{M\sigma^2}) - i(T^{(i+1)} + \frac{1}{4}\log\epsilon_{i+1})d^{-10}$,

$$\begin{aligned}
&|\hat{\nu}_{i+1}(m)| - |\nu^\star(m)| \\
&\leqslant 4.8\mu^4\frac{k^3}{M}\kappa^4\delta^{(i)2} + 5.5\mu^3\frac{k^2\sqrt{k}}{M}\kappa^3\delta^{(i)} + 2.2\mu^2\frac{k\sqrt{k}}{M}\kappa^2\delta^{(i)} \\
&\quad + 1.8\mu^2\frac{k\sqrt{k}}{M}\kappa^2\delta^{(i)} + 2.75\mu^2\frac{k\sqrt{k}}{M}\kappa^2\delta^{(i)} + 2.5\mu^2\frac{k^3}{M}\kappa^2\delta^{(i)} \\
&\leqslant 20\mu^4\frac{k^3}{M}\kappa^4\delta^{(i)} = \sqrt{\epsilon_i}. \qquad (13)
\end{aligned}$$

Eq. (13) follows by setting $T^{(i)} = C\kappa^2\log\left(\frac{\mu^4\kappa^2 k^{\frac{5}{2}}}{M\sqrt{\epsilon_i}}\right)$ and applying Theorem 2.3 in [37]. Similarly, we get

$$|\nu^\star(m)| - |\hat{\nu}_{i+1}(m)| \leqslant \sqrt{\epsilon_i}. \qquad (14)$$

From Eq. (13) and Eq. (14), we can show with probability at least $1 - 2d^{-10} - M\exp(k - cn_{M+1}^i) - \exp(k - \frac{c\mu^2 k\sigma^{\star 2}_{\max}\delta^{(i)2} n_{M+1}^i}{M\sigma^2}) - i(T^{(i+1)} + \frac{1}{4}\log\epsilon_{i+1})d^{-10}$,

$$|\nu^\star(m)| - \sqrt{\epsilon_i} \leqslant |\hat{\nu}_{i+1}(m)| \leqslant |\nu^\star(m)| + \sqrt{\epsilon_i}.$$

Hence, when considering $|\nu^\star(m)| \geqslant 2\sqrt{\epsilon_i}$, we can conclude

$$\begin{aligned}
|\hat{\nu}_{i+1}(m)| &\leqslant |\nu^\star(m)| + \sqrt{\epsilon_i} \leqslant \frac{3}{2}|\nu^\star(m)| \\
|\hat{\nu}_{i+1}(m)| &\geqslant |\nu^\star(m)| - \sqrt{\epsilon_i} \geqslant \frac{1}{2}|\nu^\star(m)|.
\end{aligned}$$

When we consider $|\nu^\star(m)| \leqslant 2\sqrt{\epsilon_i}$, we can conclude

$$\begin{aligned}
|\hat{\nu}_{i+1}(m)| &\leqslant |\nu^\star(m)| + \sqrt{\epsilon_i} \leqslant 3\sqrt{\epsilon_i} \\
|\hat{\nu}_{i+1}(m)| &\geqslant 0
\end{aligned}$$

Note that we have a probability of $1 - 2d^{-10} - M\exp(k - cn_{M+1}^i) - \exp(k - \frac{c\mu^2 k\sigma^{\star 2}_{\max}\delta^{(i)2} n_{M+1}^i}{M\sigma^2}) - i(T^{(i+1)} + \frac{1}{4}\log\epsilon_{i+1})\,d^{-10}$. To ensure a probability guarantee of at least $1 - [i(T^{(i+1)} + \frac{1}{4}\log\epsilon_{i+1}) + 4]d^{-10}$ for our lemma, it is required to set the bound for n_{M+1}^i such that the exponential terms are less than or equal to d^{-10}. We obtain

$$\log M + k - cn_{M+1}^i \leqslant -10\log d$$

This implies

$$n_{M+1}^i \geqslant C\max\{\log M, \log d, k\}$$

Also,

$$k - \frac{c\mu^2 k^4 \sigma^{\star 2}_{\max}\delta^{(i)2} n_{M+1}^i}{M\sigma^2} \leqslant -10\log d$$

This implies

$$n_{M+1}^i \geqslant C\frac{NSR}{\mu^2\kappa^2 k^4\delta^{(i)2}}\max\{k, \log d\}.$$

Combining these results with Lemma A.3 and Lemma B.9, Lemma 3.5, Theorem 2.3 in [37], we have the complexity as

$$\begin{aligned}
n_m^i &\geqslant C\max\{\log M, k\}\max\left\{1, \frac{\mu^8\kappa^8 k^5 NSR}{M^2\epsilon_i}\right\} \\
\sum_{i=1}^M n_m^i &\geqslant C\mu^2\kappa^4(d+k)k\max\left\{1, \frac{\mu^8\kappa^{10}k^6 NSR}{M^2\epsilon_i}\right\} \\
n_{M+1}^i &\geqslant C\max\{\log M, k, \log d\}\max\left\{1, \frac{\mu^6\kappa^6 k^2 NSR}{M^2\epsilon_i}\right\}.
\end{aligned}$$

This completes the proof. □

Define $P_A := A(A^\top A)^\dagger A^\top$ and $P_A^\perp = I - P_A$. We bound $\|P^\perp_{X_{M+1}\widehat{B}_T} X_{M+1} B^\star \widetilde{W}^{\star(i)}\|_F^2$ below.

Lemma B.3: Assume Assumptions II.2 and II.3 hold. If

$$\begin{aligned}
n_m^i &\geqslant C\max\{\log d, \log M\} \\
n_{M+1}^i &\geqslant C\max\{\log d, \log M\},
\end{aligned}$$

then with probability at least $1 - \delta - 2d^{-10}$,

$$\begin{aligned}
&\|P^\perp_{X_{M+1}^{(i)}\widehat{B}^{(i)}} X_{M+1}^{(i)} B^\star \widetilde{W}^{\star(i)}\|_F^2 \\
&\leqslant \frac{1.1 n_{M+1}^i}{0.9}\|B^\star W^\star - \widehat{B}^{(i)}\widehat{W}^{(i)}\|_F^2 \\
&\cdot\left(\sqrt{3}\sum_{m=1}^M n_m^i + \sqrt{12\log\frac{1}{\delta}}\sum_{m=1}^M n_m^i + \sqrt{12M\log^2\frac{1}{\delta}}\right)
\end{aligned}$$

where $\widetilde{W}^{\star(i)} = W^\star\sqrt{\mathrm{diag}([n_1^i, n_2^i, \cdots, n_M^i])}$.

Proof: Given two matrices A_1 and A_2 with the same number of columns that satisfy $A_1^\top A_1 \succeq A_2^\top A_2$, for any two matrices B and B' with compatible dimensions, from Lemma A.7 from [5], we have the following inequality

$$\|P^\perp_{A_1 B} A_1 B'\|_F^2 \geqslant \|P^\perp_{A_2 B} A_2 B'\|_F^2.$$

Using the above result and Lemma A.3, if the number of samples is at least $C\max\{\log d, \log M\}$, then with probability at least $1 - 2d^{-10}$, the following inequalities hold.

$$\begin{aligned}
&\|P^\perp_{X_{M+1}^{(i)}\widehat{B}^{(i)}} X_{M+1}^{(i)} B^\star \widetilde{W}^{\star(i)}\|_F^2 \\
&= \sum_{m=1}^M n_m^i \|P^\perp_{X_{M+1}^{(i)}\widehat{B}^{(i)}} X_{M+1}^{(i)} B^\star w_m^\star\|_2^2 \\
&\leqslant \sum_{m=1}^M 1.1 n_{M+1}^i n_m^i \|P^\perp_{I\widehat{B}^{(i)}} I B^\star w_m^\star\|_2^2 \\
&\leqslant \sum_{m=1}^M \frac{1.1 n_{M+1}^i n_m^i}{0.9 n_m^i}\|P^\perp_{X_m^{(i)}\widehat{B}^{(i)}} X_m^{(i)} B^\star w_m^\star\|_2^2 \\
&= \frac{1.1 n_{M+1}^i}{0.9}\sum_{m=1}^M \|P^\perp_{X_m^{(i)}\widehat{B}^{(i)}} X_m^{(i)} B^\star w_m^\star\|_2^2. \qquad (15)
\end{aligned}$$

And we have

$$\sum_{m=1}^{M} \|P^{\perp}_{X_m^{(i)}\widehat{B}^{(i)}} X_m^{(i)} B^\star w_m^\star\|_2^2$$
$$\leqslant \sum_{m=1}^{M} \| - P^{\perp}_{X_m^{(i)}\widehat{B}^{(i)}} X_m^{(i)} B^\star w_m^\star\|_2^2 + \sum_{m=1}^{M} \|P_{X_m^{(i)}\widehat{B}^{(i)}} Z_m^{(i)}\|_2^2$$
$$= \sum_{m=1}^{M} \| - P^{\perp}_{X_m^{(i)}\widehat{B}^{(i)}} X_m^{(i)} B^\star w_m^\star + P_{X_m^{(i)}\widehat{B}^{(i)}} Z_m^{(i)}\|_2^2 \quad (16)$$
$$= \sum_{m=1}^{M} \|P_{X_m^{(i)}\widehat{B}^{(i)}} (X_m^{(i)} B^\star w_m^\star + Z_m^{(i)}) - X_m^{(i)} B^\star w_m^\star\|_2^2$$
$$= \sum_{m=1}^{M} \|X_m^{(i)} (\widehat{B}^{(i)} \widehat{w}_m^{(i)} - B^\star w_m^\star)\|_2^2. \quad (17)$$

Eq. (16) follows since the cross term is 0. Eq. (17) is derived from $P_{X_m^{(i)}\widehat{B}^{(i)}} Y_m^{(i)} = (X_m^{(i)}\widehat{B}^{(i)})\left((X_m^{(i)}\widehat{B}^{(i)})^\top (X_m^{(i)}\widehat{B}^{(i)})\right)^{-1} (X_m^{(i)}\widehat{B}^{(i)})^\top Y_m^{(i)} = X_m^{(i)}\ \widehat{B}^{(i)}\ \widehat{w}_m^{(i)}$. Given

$$\|X_m^{(i)} (\widehat{B}^{(i)} \widehat{w}_m^{(i)} - B^\star w_m^\star)\|_2^2$$
$$= (\widehat{B}^{(i)} \widehat{w}_m^{(i)} - B^\star w_m^\star)^\top {X_m^{(i)}}^\top X_m^{(i)} (\widehat{B}^{(i)} \widehat{w}_m^{(i)} - B^\star w_m^\star)$$
$$= \sum_{n=1}^{n_m^i} (\widehat{B}^{(i)} \widehat{w}_m^{(i)} - B^\star w_m^\star)^\top x_{m,n}^{(i)} {x_{m,n}^{(i)}}^\top (\widehat{B}^{(i)} \widehat{w}_m^{(i)} - B^\star w_m^\star),$$

it follows that $\frac{\|X_m^{(i)}(\widehat{B}^{(i)}\widehat{w}_m^{(i)} - B^\star w_m^\star)\|_2^2}{\|(\widehat{B}^{(i)}\widehat{w}_m^{(i)} - B^\star w_m^\star)\|_2^2} \sim \chi^2(n_m^i)$. Applying Lemma 1 in [40], with probability at least $1-\delta$,

$$\frac{1}{\|(\widehat{B}^{(i)} \widehat{w}_m^{(i)} - B^\star w_m^\star)\|_2^2} \|X_m^{(i)} (\widehat{B}^{(i)} \widehat{w}_m^{(i)} - B^\star w_m^\star)\|_2^2$$
$$\leqslant n_m^i + 2\sqrt{n_m^i \log\frac{1}{\delta}} + 2\log\frac{1}{\delta}.$$

Substituting in Eq. (15) and using the union bound, with probability at least $1 - \delta - 2d^{-10}$,

$$\|P^{\perp}_{X_{M+1}^{(i)}\widehat{B}^{(i)}} X_{M+1}^{(i)} B^\star \widetilde{W}^{\star(i)}\|_F^2 \leqslant \frac{1.1 n_{M+1}^i}{0.9}$$
$$\cdot \underbrace{\sum_{m=1}^{M} \|(\widehat{B}^{(i)} \widehat{w}_m^{(i)} - B^\star w_m^\star)\|_2^2 \left(n_m^i + 2\sqrt{n_m^i \log\frac{1}{\delta}} + 2\log\frac{1}{\delta}\right)}_{\text{Term}}. \quad (18)$$

Our goal now is to bound Term in Eq. 18. Using Cauchy–Schwarz inequality, we have Eq. (19).

$$\text{Term} \leqslant \|B^\star W^\star - \widehat{B}^{(i)} \widehat{W}^{(i)}\|_F^2 \sqrt{\sum_{m=1}^{M} \left(n_m^i + 2\sqrt{n_m^i \log\frac{1}{\delta}} + 2\log\frac{1}{\delta}\right)^2}. \quad (19)$$

Using the fact that $(a+b+c)^2 \leqslant 3(a^2+b^2+c^2)$ and $\sum_i a_i^2 \leqslant (\sum_i a_i)^2$, we can rewrite the square root term in Eq. (19) as

$$\sqrt{\sum_{m=1}^{M} \left(n_m^i + 2\sqrt{n_m^i \log\frac{1}{\delta}} + 2\log\frac{1}{\delta}\right)^2}$$
$$\leqslant \sqrt{3\sum_{m=1}^{M} \left(n_m^{i^2} + 4n_m^i \log\frac{1}{\delta} + 4\log^2\frac{1}{\delta}\right)} \quad (20)$$
$$\leqslant \sqrt{3\sum_{m=1}^{M} n_m^{i^2}} + \sqrt{12\log\frac{1}{\delta}\sum_{m=1}^{M} n_m^i} + \sqrt{12\sum_{m=1}^{M}\log^2\frac{1}{\delta}} \quad (21)$$
$$\leqslant \sqrt{3\left(\sum_{m=1}^{M} n_m^i\right)^2} + \sqrt{12\log\frac{1}{\delta}\sum_{m=1}^{M} n_m^i} + \sqrt{12M\log^2\frac{1}{\delta}}$$
$$= \sqrt{3}\sum_{m=1}^{M} n_m^i + \sqrt{12\log\frac{1}{\delta}\sum_{m=1}^{M} n_m^i} + \sqrt{12M\log^2\frac{1}{\delta}}. \quad (22)$$

Eq. (21) is derived from $\sqrt{a+b+c} \leqslant \sqrt{a} + \sqrt{b} + \sqrt{c}$. Putting the pieces together by substituting in Eq. 18, we have

$$\|P^{\perp}_{X_{M+1}^{(i)}\widehat{B}^{(i)}} X_{M+1}^{(i)} B^\star \widetilde{W}^{\star(i)}\|_F^2$$
$$\leqslant \frac{1.1 n_{M+1}^i}{0.9} \|B^\star W^\star - \widehat{B}^{(i)}\widehat{W}^{(i)}\|_F^2$$
$$\cdot \left(\sqrt{3}\sum_{m=1}^{M} n_m^i + \sqrt{12\log\frac{1}{\delta}\sum_{m=1}^{M} n_m^i} + \sqrt{12M\log^2\frac{1}{\delta}}\right).$$

This completes the proof. □

Lemma B.4: Assume Assumptions II.2 and II.3 hold. Let $\text{SD}(B^\star, \widehat{B}^{(0)}) \leqslant \delta_0 = \frac{c}{\sqrt{k}\kappa^2}$, $\eta \leqslant \frac{0.5}{\sigma_{\max}^{\star 2}}$ and the number of GD iterations in the i^{th} epoch, $T^{(i)} = C\kappa^2 \log\left(\frac{\mu^4\kappa^2 k^{\frac{5}{2}}}{M\sqrt{\epsilon_i}}\right)$. If

$$n_m^i \geqslant C\max\{\log d, \log M, k\} \max\left\{1, \frac{\mu^8\kappa^8 k^5 NSR}{M^2\epsilon_i}\right\}$$
$$\sum_{i=1}^{M} n_m^i \geqslant C\mu^2\kappa^4(d+k)k \max\left\{1, \frac{\mu^8\kappa^{10}k^6 NSR}{M^2\epsilon_i}\right\}$$
$$n_{M+1}^i \geqslant C\max\{\log d, \log M\},$$

then after epoch i, with probability at least $1 - 2\delta - [i(T^{(i+1)} + \frac{1}{4}\log\epsilon_{i+1}) + 3]d^{-10}$,

$$\text{ER}(\widehat{B}^{(i)}, \widehat{w}_{M+1}^{(i)}) \leqslant \frac{\sigma^2(2k + 3\log\frac{1}{\delta})}{1.8 n_{M+1}^i} + \frac{\sigma_{\max}^{\star 2} M^2 \epsilon_i}{300\mu^6\kappa^8 k^5}$$
$$\cdot \left(\sqrt{3}\sum_{m=1}^{M} n_m^i + \sqrt{12\log\frac{1}{\delta}\sum_{m=1}^{M} n_m^i} + \sqrt{12M\log^2\frac{1}{\delta}}\right)\|\widetilde{\nu}^{\star(i)}\|_2^2,$$

where $\|\widetilde{\nu}^{\star(i)}\|_2^2 = \sum_{m=1}^{M} \frac{\nu^\star(m)^2}{n_m^i}$.

Proof: From the definition of $\text{ER}(\widehat{B}^{(i)}, \widehat{w}_{M+1}^{(i)})$, we have

$$\text{ER}(\widehat{B}^{(i)}, \widehat{w}_{M+1}^{(i)}) = \frac{1}{2}\mathbb{E}\left[\left(x_{M+1,n}^{(i)\top}(\widehat{B}^{(i)}\widehat{w}_{M+1}^{(i)} - B^\star w_{M+1}^\star)\right)^2\right]$$
$$= \frac{1}{2}(\widehat{B}^{(i)}\widehat{w}_{M+1}^{(i)} - B^\star w_{M+1}^\star)^\top (\widehat{B}^{(i)}\widehat{w}_{M+1}^{(i)} - B^\star w_{M+1}^\star) \quad (23)$$
$$\leqslant \frac{1}{1.8 n_{M+1}^i} \|X_{M+1}^{(i)}(\widehat{B}^{(i)}\widehat{w}_{M+1}^{(i)} - B^\star w_{M+1}^\star)\|^2. \quad (24)$$

Eq. (23) follows from $\mathbb{E}\left[x^{(i)}_{M+1,n}x^{(i)\top}_{M+1,n}\right]=I$ and Eq. (24) is derived from Lemma A.3. We substitute the least square estimator solution $\widehat{w}^{(i)}_{M+1}=((X^{(i)}_{M+1}\widehat{B}^{(i)})^\top(X^{(i)}_{M+1}\widehat{B}^{(i)}))^{-1}(X^{(i)}_{M+1}\widehat{B}^{(i)})^\top Y^{(i)}_{M+1}$ in Eq. (24) and $P_A:=A(A^\top A)^\dagger A^\top$, $P_A^\perp=I-P_A$, we get

$$\begin{aligned}&\|X^{(i)}_{M+1}(\widehat{B}^{(i)}\widehat{w}^{(i)}_{M+1}-B^\star w^\star_{M+1})\|^2\\&=\|P_{X^{(i)}_{M+1}\widehat{B}^{(i)}}(X^{(i)}_{M+1}B^\star w^\star_{M+1}+Z^{(i)}_{M+1})-X^{(i)}_{M+1}B^\star w^\star_{M+1}\|^2\\&=\|P_{X^{(i)}_{M+1}\widehat{B}^{(i)}}Z^{(i)}_{M+1}\|^2+\|P^\perp_{X^{(i)}_{M+1}\widehat{B}^{(i)}}X^{(i)}_{M+1}B^\star w^\star_{M+1}\|^2 \quad (25)\\&=\|P_{X^{(i)}_{M+1}\widehat{B}^{(i)}}Z^{(i)}_{M+1}\|^2+\|P^\perp_{X^{(i)}_{M+1}\widehat{B}^{(i)}}X^{(i)}_{M+1}B^\star\widetilde{W}^{\star(i)}\widetilde{\nu}^{\star(i)}\|^2 \quad (26)\\&\leqslant\|P_{X^{(i)}_{M+1}\widehat{B}^{(i)}}Z^{(i)}_{M+1}\|^2+\|P^\perp_{X^{(i)}_{M+1}\widehat{B}^{(i)}}X^{(i)}_{M+1}\\&B^\star\widetilde{W}^{\star(i)}\|_F^2\|\widetilde{\nu}^{\star(i)}\|_2^2\end{aligned}$$

where $\widetilde{W}^{\star(i)}=W^\star\sqrt{\mathrm{diag}([n^i_1,n^i_2,\cdots,n^i_M])}$ and $\widetilde{\nu}^{\star(i)}(m)=\frac{\nu^\star(m)}{\sqrt{n^i_m}}$. Eq. (25) is derived from $P^\perp_{X^{(i)}_{M+1}\widehat{B}^{(i)}}{}^\top P_{X^{(i)}_{M+1}\widehat{B}^{(i)}}=0$. Eq. (26) is derived from $w^\star_{M+1}=\widetilde{W}^{\star(i)}\widetilde{\nu}^{\star(i)}$. Given that $Z^{(i)}_{M+1}$ follows i.i.d. Gaussian distribution with a zero mean and variance σ^2, it follows that $\frac{1}{\sigma^2}\|P_{X^{(i)}_{M+1}\widehat{B}^{(i)}}Z^{(i)}_{M+1}\|^2\sim\chi^2(k)$. Applying the Chernoff bound for chi-square distribution, we have with probability at least $1-\delta$,

$$\|P_{X^{(i)}_{M+1}\widehat{B}^{(i)}}Z^{(i)}_{M+1}\|^2\leqslant\sigma^2\left(2k+3\log\frac{1}{\delta}\right).$$

To bound $\|P^\perp_{X^{(i)}_{M+1}\widehat{B}^{(i)}}X^{(i)}_{M+1}B^\star\widetilde{W}^{\star(i)}\|_F^2$ we use Lemma B.3, with probability at least $1-\delta-2d^{-10}$,

$$\begin{aligned}&\|P^\perp_{X^{(i)}_{M+1}\widehat{B}^{(i)}}X^{(i)}_{M+1}B^\star\widetilde{W}^{\star(i)}\|_F^2\\&\leqslant\frac{1.1n^i_{M+1}}{0.9}\|B^\star W^\star-\widehat{B}^{(i)}\widehat{W}^{(i)}\|_F^2\cdot\\&\left(\sqrt{3}\sum_{m=1}^M n^i_m+\sqrt{12\log\frac{1}{\delta}\sum_{m=1}^M n^i_m}+\sqrt{12M\log^2\frac{1}{\delta}}\right).\end{aligned}$$

Assume $\mathrm{SD}(\widehat{B}^{(i)},B^\star)\leqslant\delta^{(i)}$. Using Lemma 3.5 in [37], if $\delta^{(i)}\leqslant\frac{c}{\sqrt{k}\kappa^2}$, and $n^i_m\geqslant C\max\{\log M,k,\max\{\log M,k\}\frac{NSR}{\delta^{(i)2}k}\}$, then with probability at least $1-d^{-10}$,

$$\|B^\star W^\star-\widehat{B}^{(i)}\widehat{W}^{(i)}\|_F\leqslant 1.4\mu\delta^{(i)}\sqrt{k}\sigma^\star_{\max}.$$

Set $T^{(i)}=C\kappa^2\log\left(\frac{\mu^4\kappa^2k^{\frac{5}{2}}}{M\sqrt{\epsilon_i}}\right)$. Then, from Theorem 2.3 in [37], if $\mathrm{SD}(B^\star,\widehat{B}^{(0)})\leqslant\delta_0=\frac{c}{\sqrt{k}\kappa^2}$, $\eta\leqslant\frac{0.5}{\sigma^{\star2}_{\max}}$, and

$$\begin{aligned}n^i_m&\geqslant C\max\{\log M,k\}\max\left\{1,\frac{\mu^8\kappa^8k^5NSR}{M^2\epsilon_i}\right\}\\\sum_{i=1}^M n^i_m&\geqslant C\mu^2\kappa^4(d+k)k\max\left\{1,\frac{\mu^8\kappa^{10}k^6NSR}{M^2\epsilon_i}\right\},\end{aligned}$$

with probability at least $1-i(T^{(i+1)}+\frac{1}{4}\log\epsilon_{i+1})d^{-10}$,

$$\delta^{(i)}=\frac{M\sqrt{\epsilon_i}}{20\mu^4k^3\kappa^4}.$$

Following that, by combining these results and applying the union bound, we derive that with probability at least $1-2\delta-[i(T^{(i+1)}+\frac{1}{4}\log\epsilon_{i+1})+3]d^{-10}$,

$$\mathrm{ER}(\widehat{B}^{(i)},\widehat{w}^{(i)}_{M+1})\leqslant\frac{\sigma^2(2k+3\log\frac{1}{\delta})}{1.8n^i_{M+1}}+\frac{\sigma^{\star2}_{\max}M^2\epsilon_i}{300\mu^6\kappa^8k^5}\cdot\left(\sqrt{3}\sum_{m=1}^M n^i_m+\sqrt{12\log\frac{1}{\delta}\sum_{m=1}^M n^i_m}+\sqrt{12M\log^2\frac{1}{\delta}}\right)\|\widetilde{\nu}^{\star(i)}\|_2^2.$$

□

To obtain the bound for excess risk, we now obtain the bound for $\|\widetilde{\nu}^{\star(i)}\|_2^2$ in the lemma below and substitute in Lemma B.4.

Lemma B.5: Assume all conditions in Lemma B.2. After epoch i, with probability at least $1-[i(T^{(i+1)}+\frac{1}{4}\log\epsilon_{i+1})+4]d^{-10}$,

$$\|\widetilde{\nu}^{\star(i)}\|_2^2=\sum_{m=1}^M\frac{\nu^\star(m)^2}{n^i_m}\leqslant\frac{4\epsilon_i}{\beta}s^\star_i,$$

where $\|\nu^\star\|^i_{0,\gamma}=|\{m:|\nu^\star(m)|>\sqrt{\gamma}\epsilon_i\}|$, $s^\star_i=(1-\gamma)\|\nu^\star\|^i_{0,\gamma}+\gamma M$.

Proof: Recall that $n^i_m=\max\{\beta\epsilon_i^{-1},\beta\hat{\nu}_i(m)^2\epsilon_i^{-1}\}$, where $\beta>1$. From Lemma B.2, we have for any $\gamma\in[0,1]$, with probability at least $1-[i(T^{(i+1)}+\frac{1}{4}\log\epsilon_{i+1})+4]d^{-10}$, we have

$$\begin{aligned}\sum_{m=1}^M\frac{\nu^\star(m)^2}{n^i_m}&=\sum_{m=1}^M\frac{\nu^\star(m)^2}{n^i_m}\mathbb{1}\{|\nu^\star(m)|\leqslant\sqrt{\gamma}\epsilon_i\}\\&+\sum_{m=1}^M\frac{\nu^\star(m)^2}{n^i_m}\mathbb{1}\{\sqrt{\gamma}\epsilon_i\leqslant|\nu^\star(m)|\leqslant2\sqrt{\epsilon_{i-1}}\}\\&+\sum_{m=1}^M\frac{\nu^\star(m)^2}{n^i_m}\mathbb{1}\{|\nu^\star(m)|\geqslant2\sqrt{\epsilon_{i-1}}\}\\&\leqslant\sum_{m=1}^M\frac{\gamma\epsilon_i^2}{\beta\epsilon_i^{-1}}\mathbb{1}\{|\nu^\star(m)|\leqslant\sqrt{\gamma}\epsilon_i\}\\&+\sum_{m=1}^M\frac{8\epsilon_i}{\beta\epsilon_i^{-1}}\mathbb{1}\{\sqrt{\gamma}\epsilon_i\leqslant|\nu^\star(m)|\leqslant2\sqrt{\epsilon_{i-1}}\}\\&+\sum_{m=1}^M\frac{4\hat{\nu}_i(m)^2}{\beta\hat{\nu}_i(m)^2\epsilon_i^{-1}}\mathbb{1}\{|\nu^\star(m)|\geqslant2\sqrt{\epsilon_{i-1}}\}\\&\leqslant\frac{4\epsilon_i}{\beta}\sum_{m=1}^M(\gamma\mathbb{1}\{|\nu^\star(m)|\leqslant\sqrt{\gamma}\epsilon_i\}+\mathbb{1}\{|\nu^\star(m)|\geqslant\sqrt{\gamma}\epsilon_i\})\\&\leqslant\frac{4\epsilon_i}{\beta}\left(\gamma(M-\|\nu^\star\|^i_{0,\gamma})+\|\nu^\star\|^i_{0,\gamma}\right) \quad (27)\\&=\frac{4\epsilon_i}{\beta}\left((1-\gamma)\|\nu^\star\|^i_{0,\gamma}+\gamma M\right)=\frac{4\epsilon_i}{\beta}s^\star_i \quad (28)\end{aligned}$$

Eq. (27) is obtained from $\|\nu^\star\|^i_{0,\gamma}=|\{m:|\nu^\star(m)|>\sqrt{\gamma}\epsilon_i\}|$ and Eq. (28) is derived from $s^\star_i=(1-\gamma)\|\nu^\star\|^i_{0,\gamma}+\gamma M$. □

In the lemma below we bound the number of source task samples required for each epoch.

Lemma B.6: Let $\mathrm{SD}(B^\star, \widehat{B}^{(0)}) \leqslant \delta_0 = \frac{c}{\sqrt{k}\kappa^2}$, $\eta \leqslant \frac{0.5}{\sigma_{\max}^{\star 2}}$, and the number of GD iterations in the i^{th} epoch, $T^{(i)} = C\kappa^2 \log\left(\frac{\mu^4\kappa^2 k^{\frac{5}{2}}}{M\sqrt{\epsilon_i}}\right)$. If

$$n_m^i \geqslant C \max\{\log M, k\} \max\left\{1, \frac{\mu^8\kappa^8 k^5 NSR}{M^2\epsilon_i}\right\}$$

$$\sum_{i=1}^{M} n_m^i \geqslant C\mu^2\kappa^4(d+k)k \max\left\{1, \frac{\mu^8\kappa^{10}k^6 NSR}{M^2\epsilon_i}\right\}$$

$$n_{M+1}^i \geqslant C \max\{\log M, k, \log d\} \max\left\{1, \frac{\mu^6\kappa^6 k^2 NSR}{M^2\epsilon_i}\right\},$$

then with probability at least $1 - [i(T^{(i+1)} + \frac{1}{4}\log\epsilon_{i+1}) + 4]d^{-10}$, the total number of source samples for any epoch i is

$$10\beta\epsilon_i^{-1}(\|\nu^\star\|_2^2 + M).$$

Proof: Consider any given epoch i. We know $n_m^i = \max\left\{\beta\epsilon_i^{-1}, \beta\hat{\nu}_i(m)^2\epsilon_i^{-1}\right\}$. Thus, we can write

$$\begin{aligned}
&\sum_{m=1}^{M} n_m^i = \sum_{m=1}^{M} \max\left\{\beta\epsilon_i^{-1}, \beta\hat{\nu}_i(m)^2\epsilon_i^{-1}\right\} \\
&\leqslant \sum_{m=1}^{M} \beta\epsilon_i^{-1} + \sum_{m=1}^{M} \beta\hat{\nu}_i(m)^2\epsilon_i^{-1}\mathbb{1}\{|\nu^\star(m)| > 2\sqrt{\epsilon_{i-1}}\} \\
&+ \sum_{m=1}^{M} \beta\hat{\nu}_i(m)^2\epsilon_i^{-1}\mathbb{1}\{|\nu^\star(m)| \leqslant 2\sqrt{\epsilon_{i-1}}\} \qquad (29)
\end{aligned}$$

Using Lemma B.2 we can rewrite Eq. (29) as

$$\begin{aligned}
&\leqslant M\beta\epsilon_i^{-1} + \frac{9}{4}\sum_{m=1}^{M} \beta\nu^\star(m)^2\epsilon_i^{-1}\mathbb{1}\{|\nu^\star(m)| > 2\sqrt{\epsilon_{i-1}}\} \\
&+ \sum_{m=1}^{M} 9\beta\mathbb{1}\{|\nu^\star(m)| \leqslant 2\sqrt{\epsilon_{i-1}}\} \qquad (30) \\
&\leqslant M\beta\epsilon_i^{-1} + \frac{9}{4}\sum_{m=1}^{M} \beta\nu^\star(m)^2\epsilon_i^{-1} + \sum_{m=1}^{M} 9\beta \\
&= M\beta\epsilon_i^{-1} + \frac{9}{4}\beta\epsilon_i^{-1}\|\nu^\star\|_2^2 + 9M\beta \\
&= \beta\epsilon_i^{-1}(M + \frac{9}{4}\|\nu^\star\|_2^2 + 9M\epsilon_i) \\
&\leqslant 10\beta\epsilon_i^{-1}(\|\nu^\star\|_2^2 + M),
\end{aligned}$$

This completes the proof. □

REFERENCES

[1] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[2] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[4] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[5] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei, "Few-shot learning via learning the representation, provably," in *Proc. Int. Conf. Learn. Representations*, 2021.

[6] Y. Chen, K. Jamieson, and S. Du, "Active multi-task representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 3271–3298.

[7] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Technical report, Toronto University, 2009.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Piscataway, NJ, USA: IEEE Press, 2009, pp. 248–255.

[9] N. Mu and J. Gilmer, "MNIST-C: A robustness benchmark for computer vision," 2019, *arXiv:1906.02337*.

[10] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1999, pp. 230–237.

[11] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2089–2099.

[12] N. Tripuraneni, C. Jin, and M. Jordan, "Provable meta-learning of linear representations," in *Proc. Int. Conf. Mach. Learn.,* PMLR, 2021, pp. 10434–10443.

[13] L. Cella and M. Pontil, "Multi-task and meta-learning with sparse linear bandits," in *Uncertainty in Artificial Intelligence*, PMLR, 2021, pp. 1692–1702.

[14] J. Yang, W. Hu, J. D. Lee, and S. S. Du, "Impact of representation learning in linear bandits," 2020, *arXiv:2010.06531*.

[15] J. Hu, X. Chen, C. Jin, L. Li, and L. Wang, "Near-optimal representation learning for linear bandits and linear RL," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4349–4358.

[16] L. Cella, K. Lounici, G. Pacreau, and M. Pontil, "Multi-task representation learning with stochastic linear bandits," in *Proc. Int. Conf. Artif. Intell. Statist.,* PMLR, 2023, pp. 4822–4847.

[17] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," 2022, *arXiv:2202.10054*.

[18] E. Candes and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, no. 6, pp. 111–119, 2012.

[19] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh, "Sample efficient linear meta-learning by alternating minimization," 2021, *arXiv:2105.08306*.

[20] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, pp. 41–75, Jul. 1997.

[21] S. Thrun and L. Pratt, "Learning to learn: Introduction and overview," in *Learning to Learn*, New York, NY, USA: Springer, 1998, pp. 3–17.

[22] J. Baxter, "A model of inductive bias learning," *J. Artif. Intell. Res.*, vol. 12, pp. 149–198, Mar. 2000.

[23] Z. Xu and A. Tewari, "Representation learning beyond linear prediction functions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 4792–4804.

[24] Y. Wang, Y. Chen, K. Jamieson, and S. S. Du, "Improved active multi-task representation learning via lasso," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 35548–35578.

[25] X. Yao, Y. Zheng, X. Yang, and Z. Yang, "NLP from scratch without large-scale pretraining: A simple and efficient framework," in *Proc. Int. Conf. Mach. Learn.,* PMLR, 2022, pp. 25438–25451.

[26] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3712–3722.

[27] P. Knight and R. Duan, "Multi-task learning with summary statistics," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[28] J. Lin and S. Moothedath, "Fast and sample-efficient relevance-based multi-task representation learning," *IEEE Contr. Syst. Lett.*, vol. 8, pp. 1397–1402, 2024.

[29] S. Nayer and N. Vaswani, "Fast and sample-efficient federated low rank matrix recovery from column-wise linear and quadratic projections," *IEEE Trans. Inf. Theory*, vol. 69, no. 2, pp. 1177–1202, Feb. 2023.

[30] N. Vaswani, "Efficient federated low rank matrix recovery via alternating GD and minimization: A simple proof," *IEEE Trans. Inf. Theory*, vol. 70, no. 7, pp. 5162–5167, Jul. 2024.

[31] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *J. Mach. Learn. Res.*, vol. 10, no. 7, 2009, pp. 1633–1685.
[32] E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Actor-Mimic: Deep multitask and transfer reinforcement learning," 2015, *arXiv:1511.06342*.
[33] C. D'Eramo, D. Tateo, A. Bonarini, M. Restelli, and J. Peters, "Sharing knowledge in multi-task deep reinforcement learning," 2024, *arXiv:2401.09561*.
[34] S. Arora, S. Du, S. Kakade, Y. Luo, and N. Saunshi, "Provable representation learning for imitation learning via bi-level optimization," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 367–376.
[35] J. Lin, S. Moothedath, and N. Vaswani, "Fast and sample efficient multi-task representation learning in stochastic contextual bandits," in *Proc. Int. Conf. Mach. Learn.*, 2024.
[36] Y. Chen and E. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.
[37] A. P. Singh and N. Vaswani, "Noisy low rank column-wise sensing," 2024, *arXiv:2409.08384*.
[38] R. Vershynin, "High-dimensional probability: An introduction with applications in data science," Cambridge, MA, USA: Cambridge Univ. Press, vol. 47, 2018.
[39] P. Kovanic, "On the pseudoinverse of a sum of symmetric matrices with applications to estimation," *Kybernetika*, vol. 15, no. 5, pp. 341–348, 1979.
[40] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, vol. 28, pp. 1302–1338, Oct. 2000.

Jiabin Lin received the B.S. degree in computer science and communication engineering from Guangxi University of Science and Technology, Liuzhou, China, in 2019, and the M.S. degree in electrical engineering from the University of South Florida, Tampa, FL, USA, in 2021. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA. His research interests include bandit learning, multitask learning, and representation learning.

Shana Moothedath (Member, IEEE) received the Ph.D. degree in electrical engineering from the Indian Institute of Technology Bombay (IITB), in 2018. She is a Harpole-Pentair Assistant Professor of electrical and computer engineering with Iowa State University, Ames, IA, USA. She was a Postdoctoral Scholar in electrical and computer engineering with the University of Washington, Seattle, till 2021. Her research focuses on distributed decision making, control and security of dynamical systems, and signal processing. She received the NSF CAREER Award in 2025, the Best Research Thesis Award at IITB in 2019, and selected as a MIT-EECS Rising Star in 2019.

Tuan Le received the B.S. degree in computer science and mathematics (double major) from DePauw University, Greencastle, IN. He is currently working toward the M.S. degree in computer science with Iowa State University, Ames, IA, USA. His research interests lie in federated learning, representation learning, and trustworthy and safe artificial intelligence.