Fast and Sample-Efficient Relevance-Based Multi-Task Representation Learning

Jiabin Lin and Shana Moothedath *Member, IEEE*

***Abstract*— This letter explores an approach for task-relevant multi-task representation learning when the amount of data is limited for both source tasks and target tasks. Specifically, we consider a low-dimensional setting where the goal is to sample source task data based on their relevance so as to utilize task-relevant information effectively. We present a novel learning algorithm based on an alternating projected gradient descent (GD) and minimization estimator. We present the convergence guarantee of our algorithm, excess risk, and the sample complexity of our approach. We evaluated the effectiveness of our algorithm via numerical experiments and compared it empirically against three benchmark approaches.**

***Index Terms*— Representation learning, Multi-task learning, Meta learning, Alternating gradient descent**

I. Introduction

Representation learning is an emerging problem for learning in a data-scarce environment, where one first learns a feature extractor or representation, e.g., the last layer of a convolutional neural network, from different but related source tasks, and then uses a predictor on top of this representation in the target task [1]. This process involves uncovering features that capture essential characteristics and patterns within the data, allowing for more effective and efficient learning across various tasks. Representation learning plays a key role in enhancing the capabilities of machine learning models, particularly in scenarios with limited data, facilitating improved generalization and adaptability across diverse tasks.

Multi-task representation learning is one method that assumes all tasks are supported by a common representation. The fundamental approach to this learning strategy involves using the source samples to identify the optimal representation, which is subsequently used to train the linear predictor for a target task. 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 [1], [2]. Nonetheless, source tasks frequently have a limited number of samples as well. Often, in real-world applications like medical image analysis, it is difficult to have a substantial dataset, and the samples are limited. Moreover, not all source tasks contribute equally to learning representation in many applications. Therefore, it is crucial to prioritize relevant tasks during the training rather than assigning them uniform weight in multi-task learning.

This paper develops a framework for *task-relevant* multi-task representation learning to determine an optimal representation using limited samples from source tasks. Our goal is to prioritize the relevance of the source task while sampling the training (source) data rather than a uniform sampling approach. This situation happens in many practical applications, including data-driven control for robotics and autonomous driving [3]. For instance, in robotic systems [4], [5], where the model simultaneously learns representations for various control tasks, such as navigation, manipulation, and object recognition. This approach enables the system to leverage shared knowledge across tasks, improving efficiency and adaptability in diverse and complex environments.

Related Work: Multi-task representation learning has been extensively explored, starting with seminal works such as [4], [6], [7]. There have been many recent works on provable uniform multi-task representation learning under various assumptions. [1], [8]–[11] focus on learning a representation function for *any* potential target task under the assumption of the existence of a shared low-dimensional linear representation across all tasks. Recently, [2], [12] developed an adaptive representation learning for a specific target task, under a similar setting as in [1]. [12] improves the sample complexity on [2] under a high dimension input assumption. The primary distinctions between our approach and existing works, [1], [2], [8], [12], is in our consideration of a data-scarce regime, where the availability of source data is also limited, and that we propose an estimation algorithm with guarantees for solving the problem.

Contributions: In this letter, we propose a task-relevant representation learning algorithm based on an alternating gradient descent and minimization approach. With respect to the existing works [1], [8], [12] and the closely related work [2], our work differs in two key aspects. (i) We consider a data-scarce regime where the number of source data samples is limited, unlike in [2], which assumes unlimited availability of source data. Data scarcity is a prevalent challenge in learning, rendering our approach well-suited for practical settings such as medical imaging applications, where data samples are limited. (ii) [1], [2], [8], [12] assumed the availability of the optimal solution to the estimation problem. This is not feasible since the rank-constrained estimation problem (Eq.(2)) is a non-convex problem. We propose a novel estimator and establish the convergence of the proposed algorithm and sample complexity.

This work was supported in part by the U.S. National Science Foundation Grant 2213069.

J. Lin and S. Moothedath are with the Department of Electrical Engineering, Iowa State University, Ames, IA 50011, USA `jiabin@iastate.edu, mshana@iastate.edu`

We empirically validated our approach outperforms the state-of-the-art techniques consistently.

II. Problem Formulation and Notations

Problem Formulation: Consider M source tasks and a single 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 source task $m \in [M]$, 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 come from the distribution μ_m. The goal of multitask 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.

In this letter, we deal with a data-scarce regime, both for the source and the target task, i.e., $n_m < d$. We consider limited data for both source and target task, denoted as $\{(x_{m,1}, y_{m,1}), \cdots, (x_{m,n_m}, y_{m,n_m})\}_{m \in [M+1]}$ which is drawn i.i.d. from the task distributions μ_m for $m \in [M+1]$. The number of data samples for the target task is even fewer than that of the source task, i.e., $n_{M+1} \lll \{n_1, \ldots, n_M\}$. This setting aligns with our main objective of 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. 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

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

where $\mathcal{L}_{M+1}(\phi, w) = \mathbb{E}_{(x,y)\sim\mu_{M+1}}[(\langle \phi(x), w \rangle - y)^2]$.

We focus on the linear representation function class, which is studied in [1], [2], [8], [13]. We have the assumption below.

Assumption II.1 (Low-dimension linear representation)**.** $\Phi = \{x \to B^\top x | B \in \mathbb{R}^{d\times k}\}$. *We denote the true underlying representation function as $B^\star$.*

Inspired by [2], in our model, task relevance is a crucial factor. That is, we consider a setting where the goal is to learn a representation of a *specific* target task rather than a *generic* target task as in [1], [8]. Notice that, by Assumption II.1, $\Theta^\star := [\theta_1^\star, \ldots, \theta_M^\star] = B^\star W^\star$ is a rank-k matrix, where $W^\star \in \mathbb{R}^{k\times M}$ and $k \ll \min\{d, M\}$. Given that $\sigma_{\min}(W^\star) > 0$, the coefficient $w_{M+1}^\star$ can be considered a linear combination of the coefficients $\{w_m^\star\}_{m\in[M]}$. Therefore, we make the assumption that $\nu^\star \in \mathbb{R}^M$, such that $W^\star \nu^\star = w_{M+1}^\star$, where a larger value of $|\nu^\star(m)|$ indicates a stronger connection between the source task m and the target task. Based on the information provided by $\nu^\star$, we give priority to samples from source tasks that have the highest relevance.

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 max-norm is expressed as $\|A\|_{\max} = \max_{i,j} |A_{i,j}|$. 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 symbolized by 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 $Z_m \in \mathbb{R}^{n_m}$. Furthermore, the collection of vectors $\{w_m\}_{m\in[M]}$, where w_m is associated with the m-th source task, is assembled into the matrix $W \in \mathbb{R}^{k\times M}$. The notation $a \gtrsim b$ means that approximately $a \geqslant Cb$, $C > 1$.

Let $\Theta^\star := B^\star W^\star \overset{\text{SVD}}{=} B^\star \Sigma V^\star$ denote its reduced (rank k) SVD, i.e., $B^\star$ and ${V^\star}^\top$ are matrices with orthonormal columns *(basis matrices)*, $B^\star$ is $d \times k$, $V^\star$ is $k \times M$, and Σ is an $k \times k$ diagonal matrix with non-negative entries (singular values). We let $W^\star := \Sigma V^\star$. We use $\sigma_{\max}^\star$ and $\sigma_{\min}^\star$ to denote the maximum and minimum singular values of Σ, and we define its condition number as $\kappa := \sigma_{\max}^\star / \sigma_{\min}^\star$. We have the following standard assumptions.

Assumption II.2. *(Gaussian design and noise) We assume $x_{m,n}$ follows an i.i.d. standard Gaussian distribution. Moreover, the additive noise variables z_m follow i.i.d. Gaussian distribution with a zero mean and variance σ^2.*

Assumption II.3 (Incoherence of right singular vectors)**.** *We assume that $\|w_m^\star\|^2 \leqslant \mu^2 \frac{k}{M} {\sigma_{\max}^\star}^2$ for a constant $\mu \geqslant 1$.*

III. Proposed Algorithm and Analysis: Task Relevant Representation Learning via AltGDmin

Our objective is to learn a low-dimensional linear representation from the training samples (source tasks) through an task-relevance based 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 as in [1], [8]. The rationale is that by incorporating more samples from pertinent tasks, we can accelerate the learning process. To this end, our algorithm starts by drawing $\propto (\nu^\star(n_m))^2$ i.i.d. samples from the corresponding offline data for each source task $m \in [M]$. Following that, we use these samples in all source tasks to minimize the cost function

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

Subsequently, we use the estimated parameter $\widehat{B}$ and the sample of the target task to further optimize the cost function

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

Using least-squares, Eq. 3 estimates the parameter $\widehat{w}_{M+1}$ for the target task. We will elaborate on our approach for solving Eq. (2). Our approach utilizes the recently introduced alternating gradient descent and minimization (AltGDmin)

Algorithm 1: Active Representation Learning Algorithm

```
**Input:** Confidence $\delta$, representation function class $\Phi$, relevance parameter $\nu^\star$, source-task sampling budget $N \gg M(\frac{k}{\sqrt{M^3}}((d-k)+\log(\frac{1}{\delta}))$, multiplier for $\alpha$ in init step, $\tilde{C}$, GD step size $\eta$, number of GD iterations $T$
Initialize the lower bound $\underline{N} = \frac{k}{\sqrt{M^3}}((d-k)+\log(\frac{1}{\delta}))$ and number of samples $n_m = \max\{(N - M\underline{N})\frac{(\nu^\star(m))^2}{\|\nu^\star\|_2^2}, \underline{N}\}$
For each task $m$, draw $n_m$ i.i.d samples from the corresponding offline dataset denoted as $\{X_m, Y_m\}_{m=1}^M$
Set $\alpha = \frac{\tilde{C}}{NM}\sum_{m=1,n=1}^{M,n_m} y_{m,n}^2$
$y_{m,trunc}(\alpha) := Y_m \circ \mathbb{1}_{\{|Y_m| \leqslant \sqrt{\alpha}\}}$
$\widehat{\Theta}_0 := \sum_{m=1}^M \frac{1}{n_m} X_m^\top y_{m,trunc}(\alpha) e_m^\top$
Set $\widehat{B}_0 \leftarrow$ top-$k$-singular-vectors of $\widehat{\Theta}_0$
**GDmin iterations:**
**for** $t = 1$ to $T$ **do**
    Let $\widehat{B} \leftarrow \widehat{B}_{t-1}$
    **Update** $\widehat{w}_m$**,** $\widehat{\theta}_m$**:** For each $m \in [M]$, set $(\widehat{w}_m)_t \leftarrow (X_m\widehat{B})^\dagger Y_m$ and set $(\widehat{\theta}_m)_t \leftarrow \widehat{B}(\widehat{w}_m)_t$
    **Gradient w.r.t** $\widehat{B}$**:** Compute $\nabla_{\widehat{B}} f(\widehat{B}, \widehat{W}_t) = \sum_{m=1}^M X_m^\top (X_m\widehat{B}(\widehat{w}_m)_t - Y_m)(\widehat{w}_m)_t^\top$
    **GD step:** Set $\widehat{B}^+ \leftarrow \widehat{B} - \frac{\eta}{N/M}\nabla_{\widehat{B}} f(\widehat{B}, \widehat{W}_t)$
    **Projection step:** Compute $\widehat{B}^+ \overset{QR}{=} B^+R^+$
    Set $\widehat{B}_t \leftarrow B^+$
**end for**
Compute $\widehat{w}_{M+1} = \arg\min_w \|X_{M+1}^\top \widehat{B}_T w - Y_{M+1}\|^2$
Return $\widehat{B}_T$, $\widehat{w}_{M+1}$
```

algorithm [14], [15] for matrix learning. The main distinctions lie in our consideration of a noisy setting, where the observed signal contains noise, which is the common observation model studied in multi-task learning [1], [2], [8]. Further, we consider a task-relevant sampling technique as in [2] rather than uniform sampling, which is highly beneficial for generalizing to a target task as also demonstrated in the simulations (Fig. 1). Additionally, the goal of matrix learning works [14], [15] is to estimate an unknown low-rank matrix (under non-noisy settings) and there is no focus on generalizing to a target task and quantifying the excess risk.

Recall that $n_m < d$ and rank $k \ll d$. Due to the non-convex cost function $f(\widehat{B}, \widehat{W})$ with respect to the unknowns $\{\widehat{B}, \widehat{W}\}$ the AltGDmin algorithm [14] starts with a careful initialization, referred to as spectral initialization. We extract the top k singular vector from

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

where X_m represents the feature matrix obtained by concatenating the feature vectors associated with 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 presents a significant challenge that restricts the ability to determine a bound for the $\|\widehat{\Theta}_{0,full} - B^\star W^\star\|$ within the desired sample complexity. Consequently, a strategic approach is necessary to effectively handle this challenge. In order to tackle this issue, we use the truncation method introduced in [16], carefully starting with the top k singular vectors of

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

where $\alpha = \frac{\tilde{C}}{NM}\sum_{m=1,n=1}^{M,n_m} y_{m,n}^2$, $\tilde{C} = 9\kappa^2\mu^2$, and $y_{m,trunc}(\alpha) := Y_m \circ \mathbb{1}_{\{|Y_m| \leqslant \sqrt{\alpha}\}}$. Using Singular Value Decomposition (SVD), we derive the top k singular vectors from $\widehat{\Theta}_0$ to obtain initial estimate $\widehat{B}_0$. This method 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 minimize the cost function (2). In each iteration, we independently optimize $\widehat{w}_m$ 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 $\widehat{B}^+ \overset{QR}{=} B^+R^+$. Using the estimated parameter matrix $\widehat{B}$ obtained from the source tasks, we compute the estimated parameter $\widehat{w}_{M+1}$ by minimizing the cost function (3) using the least squares estimator.

Below, we present the excess risk bound for Algorithm 1.

Theorem III.1. *Consider Assumptions II.2 and II.3 hold. For any $\epsilon > 0$, success probabilities $\delta, \delta' \in [0,1]$, $C > 1$, let $\sigma^2 \leqslant \min\left\{\frac{c\|\theta_m^\star\|^2}{k^3\kappa^6}, \frac{\epsilon^2\|\theta_m^\star\|^2}{c^2\kappa^2}\right\}$, $\eta = \frac{0.4}{{\sigma_{\max}^\star}^2}$, and $T = C\kappa^2\log\frac{1}{\epsilon}$. If $n_m \geqslant C\max(\log d, \log M, k)\log\frac{1}{\epsilon}$, then with probability $O(1 - \delta - d^{-10} - de^{-\frac{\delta'^2 n_{M+1}}{3\|x_{M+1,n}\|^2}}$, the output of Algorithm 1 guarantees that $\mathrm{ER}(\widehat{B}_T, \widehat{w}_{M+1}) \leqslant \epsilon$ whenever the total sampling budget from all sources N is at least*

$$O\left(\min\left\{\frac{(1+\delta')}{(1-\delta')^2}k\|\nu^\star\|_2^2 s^\star\epsilon\log\frac{1}{\delta}, (d+M)k(k^2+\log\frac{1}{\epsilon})\right\}\right)$$

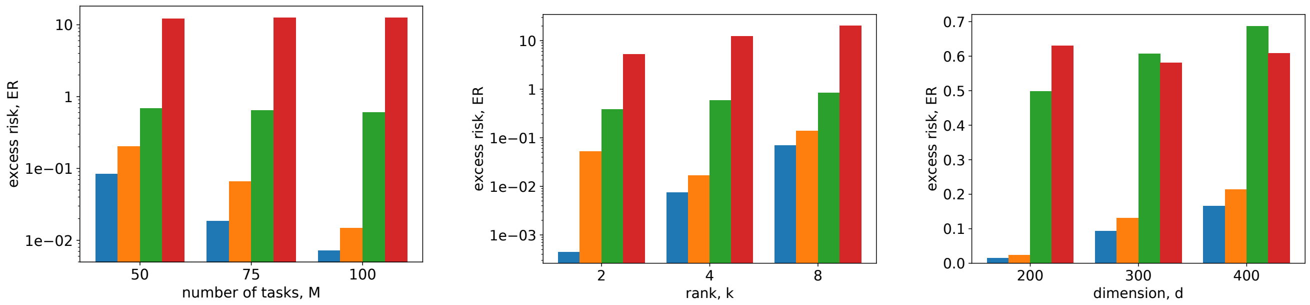
and the number of target samples n_{M+1} is at least

$$O\left(\frac{\sigma^2(k+\log\frac{1}{\delta})}{(1-\delta')}\epsilon^{-1}\right)$$

where $s^\star = (1-\gamma)\|\nu\|_{0,\gamma} + \gamma M$, $\|\nu\|_{0,\gamma} := \left|\left\{m : |\nu_m| > \sqrt{\gamma\frac{\|\nu^\star\|_2^2}{N}}\right\}\right|$ for $\gamma \in [0,1]$.

Proof. Proof is provided in Appendix II. □

Remark III.2. *The probability of the guarantee increases as the number of target samples n_{M+1} increases and the number of target samples scales only with $k \ll d$. Theorem III.1 shows that the number of source samples required depends on the task relevance denoted by $s^\star$. Since $\sqrt{\frac{\|\nu^\star\|_2^2}{N}}$ is of the order of ϵ, for $\gamma \approx 1/M$, we have $\mathrm{ER}(\widehat{B}_T, \widehat{w}_{M+1}) \leqslant \epsilon$ by using only those source tasks with relevance $|\nu^\star(m)| \gtrsim \epsilon$. Let us consider two boundary cases: (i) $\nu^\star$ is a 1-sparse vector, i.e., the target task only depends on one source task, and (ii) $\nu^\star$ is a scaled vector $\mathbf{1}$ where $\mathbf{1}$ is a vector of all ones, i.e., all source tasks are equally relevant (uniform sampling). For $\gamma = 0$, (i) gives $s^\star = 1$ and (ii) gives $s^\star = M$. Thus, uniform sampling requires M times more source data samples than (i), validating the effectiveness of the task-relevance-based sampling. The result in [2] requires that the total sampling budget from all sources N is at*



(a) Number of tasks M vs. excess risk (ER)

(b) Rank k vs. excess risk (ER)

(c) Dimension d vs. excess risk (ER)

Fig. 1: 🟦 proposed algorithm (relevance sampling), 🟧 proposed algorithm (uniform sampling), 🟩 MoM (relevance), 🟥 Chen et al. (relevance). We considered 200 data samples for each source task and 100 data samples for the target task. We varied the number of tasks as $M = 50, 75, 100$, varied the rank of the $\Theta^\star$ as $k = 2, 4, 8$, and varied the dimension as $d = 200, 300, 400$. Based on the plots (Figures 1a, 1b, and 1c), our proposed approach with adaptive sampling (also even if we use uniform sampling) outperforms the existing approaches.

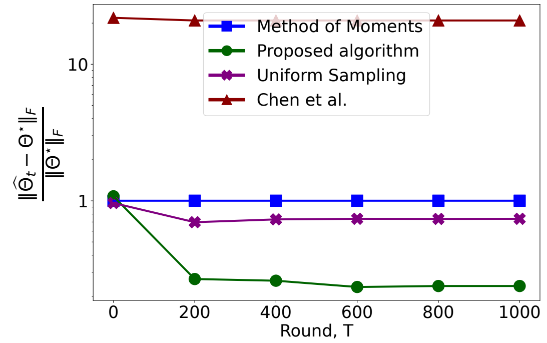


Fig. 2: Estimation error vs. GD iterations. We set $d = 300$, $M = 100$, $\tilde{C} = 3$, $k = 2$, noise variance $= 10^{-6}$.

least $O\left((kd + kM + \log(\frac{1}{\delta}))\sigma^2 s^\star \|\nu\|_2^2 \epsilon^{-2}\right)$ *and the number of target samples* n_{M+1} *is at least* $O\left(\sigma^2(k + \log(\frac{1}{\delta}))\epsilon^{-2}\right)$. *Further, the guarantees in [2] are under the assumption that an optimal solution to the non-convex cost function is available. Theorem III.1 presents the guarantees on the excess risk for the target task using an AltGDmin estimator.*

Remark III.3. *The proposed approach can be extended to multiple target task settings, say* K *tasks, each with a relevance parameter* $\nu_i^\star$ *where* $i \in \{M+1, \dots, M+K\}$. *The number of data samples for each source task will be determined by the target task with the highest relevance for it.*

IV. Simulations

We evaluated the effectiveness of our proposed algorithm compared to three benchmarks (i) our estimator with uniform sampling (to validate the effectiveness of task-relevant sampling), (ii) Method-of-Moments (MoM) estimator in [8], and (iii) Chen *et al.* in [2]. The MoM estimator computes the top k singular value decomposition on $\widehat{\Theta} = \frac{1}{NM}\sum_{m=1}^{M}\sum_{n=1}^{n_m} y_{m,n}^2 x_{m,n} x_{m,n}^\top$ to obtain the estimated matrix $\widehat{B}$. In our algorithm, we set GD step-size $\eta = 0.4/\|\widehat{\Theta}_0\|^2$ and GD iterations $T = 1000$. The entries of matrix $B^\star$ were randomly generated by orthonormalizing an i.i.d. standard Gaussian matrix, and 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 randomly and then used to calculate the parameter $w_{M+1}^\star$ for the target task. The matrices X_m were randomly generated using an i.i.d. standard Gaussian distribution. We used a noise model with a mean of zero and variance of 10^{-6}. All results are averaged over 100 independent trials.

Excess Risk Plots. The plots in Figure 1a, 1b, and 1c show the plots of the excess risk for the two algorithms by varying the number of tasks (M), the rank of $\Theta^\star$, k, and the dimension d. The results of our study show that as the number of tasks increases, the excess risk decreases for both algorithms, as expected. However, our algorithm consistently provides a significantly lower excess risk than the MoM-based algorithm and Chen et al.. We varied the rank k and the dimension d of the data and compared the performances of the algorithms. As shown in Figures 1b and 1c, our algorithm outperformed the other by a significant margin. Our algorithm consistently outperforms the MoM-based algorithm regarding accuracy, as demonstrated by its low excess risk for all experiments.

Estimation Error. In Figure 2, we present the plot for estimation error vs. GD iterations. The MoM estimator and Chen et al. [2] are noniterative methods; hence, the estimation error is a single line. We notice that, the estimation error for the parameter matrix for the M tasks $\Theta^\star$ is considerably less in our proposed estimator. We also notice that the estimation error with uniform sampling is lower than that of the adaptive sampling. However, the excess risk is lower for the adaptive sampling, as shown in Figure 1. This validates the benefit of adaptive sampling for generalizing to a target task.

V. Conclusion and Future Work

In this letter, we introduced a novel active-representation learning algorithm based on an alternating GD and minimization approach. The algorithm is specifically designed for active multi-task representation learning by considering the task relevance to enable adaptive sampling. We have demonstrated the algorithm's convergence and analyzed the sample complexity. Additionally, we have evaluated the effectiveness of our approach in comparison with three benchmark algorithms. As part of future work, we plan to study the unknown relevance setting and online learning approaches, including bandit learning and reinforcement learning.

Appendix I
Initial Results

We present initial lemmas and then prove our main theorem.

Lemma I.1. *For any $m \in [M+1]$, with probability at least $1 - 2de^{-\frac{\delta'^2 n_m}{3\|x_{m,n}\|^2}}$, it holds that $(1-\delta')n_m I \preceq X_m^\top X_m \preceq (1+\delta')n_m I$, 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 $\lambda_{\max}(x_{m,n}x_{m,n}^\top) \leqslant \|x_{m,n}\|^2$. Since

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

by applying the Matrix Chernoff inequality, we have with probability at least $1 - de^{-\frac{\delta'^2 n_m}{2\|x_{m,n}\|^2}}$, $\lambda_{\min}(\sum_{n=1}^{n_m} x_{m,n}x_{m,n}^\top) \geqslant (1-\delta')n_m$ and with probability at least $1 - de^{-\frac{\delta'^2 n_m}{3\|x_{m,n}\|^2}}$, $\lambda_{\max}(\sum_{n=1}^{n_m} x_{m,n}x_{m,n}^\top) \leqslant (1+\delta')n_m$. Applying union bound completes the proof. □

Define $P_A := A(A^\top A)^\dagger A^\top$ and $P_A^\perp = I - I_A$.

Lemma I.2. *Assume that Assumptions II.2 and II.3 hold and $\sigma^2 \leqslant \min\left\{\frac{c\|\theta_m^\star\|^2}{k^3\kappa^6}, \frac{\epsilon^2\|\theta_m^\star\|^2}{c^2\kappa^2}\right\}$. Set $\eta = \frac{0.4}{{\sigma_{\max}^\star}^2}$ and $T = C\kappa^2 \log\frac{1}{\epsilon}$. If $N \geqslant C\kappa^6\mu^2(d+M)k(\kappa^2k^2 + \log\frac{1}{\epsilon})$ and $n_m \geqslant C\max(\log d, \log M, k)\log\frac{1}{\epsilon}$, then with probability at least $O(1 - \delta - d^{-10} - de^{-\frac{\delta'^2 n_{M+1}}{3\|x_{M+1,n}\|^2}})$,*

$$\frac{1}{n_{M+1}}\|P^\perp_{X_{M+1}\widehat{B}_T} X_{M+1}B^\star\widetilde{W}^\star\|_F^2$$
$$\leqslant \frac{(1+\delta')}{(1-\delta')}\epsilon^2\mu^2 k{\sigma_{\max}^\star}^2\left(2N(d-k) + 3\log\frac{1}{\delta}\right)$$

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

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 [1], we have the following inequality

$$\|P^\perp_{A_1B}A_1B'\|_F^2 \geqslant \|P^\perp_{A_2B}A_2B'\|_F^2.$$

Using the above result and Lemma I.1, with probability at least $1 - 2de^{-\frac{\delta'^2 n_{M+1}}{3\|x_{M+1,n}\|^2}}$, the following inequalities hold.

$$\frac{1}{n_{M+1}}\|P^\perp_{X_{M+1}\widehat{B}_T}X_{M+1}B^\star\widetilde{W}^\star\|_F^2 \leqslant (1+\delta')\|P^\perp_{I\widehat{B}_T}IB^\star\widetilde{W}^\star\|_F^2$$
$$\leqslant \frac{(1+\delta')}{(1-\delta')}\sum_{m=1}^M \|P^\perp_{X_m\widehat{B}_T}X_mB^\star w_m^\star\|_2^2. \tag{4}$$

Using the definition of $P^\perp_{X_m\widehat{B}_T}X_mB^\star w_m^\star$, where $\widehat{B}_T$ is the estimate in the T-th GD iteration, we have

$$\sum_{m=1}^M \|P^\perp_{X_m\widehat{B}_T}X_mB^\star w_m^\star\|_2^2$$
$$= \sum_{m=1}^M \|X_m(B^\star w_m^\star - \widehat{B}_T(\widehat{w}_m)_T) - (X_m\widehat{B}_T)((X_m\widehat{B}_T)^\top$$
$$(X_m\widehat{B}_T))^{-1}(X_m\widehat{B}_T)^\top X_m(B^\star w_m^\star - \widehat{B}_T(\widehat{w}_m)_T)\|_2^2 \tag{5}$$
$$= \sum_{m=1}^M \|P^\perp_{X_m\widehat{B}_T}X_m(B^\star w_m^\star - \widehat{B}_T(\widehat{w}_m)_T)\|_2^2$$
$$\leqslant \left(\sum_{m=1}^M \|P^\perp_{X_m\widehat{B}_T}X_m\|_F^2\right)\cdot\left(\sum_{m=1}^M \|B^\star w_m^\star - \widehat{B}_T(\widehat{w}_m)_T\|_2^2\right) \tag{6}$$
$$= \|B^\star W^\star - \widehat{B}_T\widehat{W}_T\|_F^2 \sum_{m=1}^M \|P^\perp_{X_m\widehat{B}_T}X_m\|_F^2. \tag{7}$$

Eq. (5) is derived from adding and subtracting and by using $X_m\widehat{B}_T(\widehat{w}_m)_T - (X_m\widehat{B}_T)((X_m\widehat{B}_T)^\top(X_m\widehat{B}_T))^{-1}(X_m\widehat{B}_T)^\top X_m\widehat{B}_T(\widehat{w}_m)_T = X_m\widehat{B}_T(\widehat{w}_m)_T - X_m\widehat{B}_T(\widehat{w}_m)_T = 0$. Eq. (6) is derived from Cauchy-Schwarz inequality. Given that X_m follows i.i.d. standard Gaussian distribution, it follows that $\sum_{m=1}^M \|P^\perp_{X_m\widehat{B}_T}X_m\|_F^2 \sim \chi^2(\sum_{m=1}^M n_m(d-k))$. Applying the Chernoff bound for chi-square distribution, we have

$$\sum_{m=1}^M \|P^\perp_{X_m\widehat{B}_T}X_m\|_F^2$$
$$\leqslant \sum_{m=1}^M n_m(d-k) + 2\sqrt{\sum_{m=1}^M n_m(d-k)\log\frac{1}{\delta}} + 2\log\frac{1}{\delta},$$

with probability at least $1-\delta$. Using the inequality $\sqrt{ab} \leqslant \frac{a+b}{2}$, we can determine $2\sqrt{\sum_{m=1}^M n_m(d-k)\log\frac{1}{\delta}} \leqslant \sum_{m=1}^M n_m(d-k) + \log\frac{1}{\delta}$. Therefore, we conclude that with probability at least $1-\delta$, $\sum_{m=1}^M \|P^\perp_{X_m\widehat{B}_T}X_m\|_F^2 \leqslant 2\sum_{m=1}^M n_m(d-k) + 3\log\frac{1}{\delta}$. From Theorem 5.3 in [17], under the given assumptions and conditions, with probability at least $O(1-d^{-10})$, $\|\widehat{\theta}_{m,T} - \theta_m^\star\| \leqslant \epsilon\|\theta_m^\star\|$ for all $m \in [M]$. Then we have with probability at least $O(1-d^{-10})$,

$$\|\widehat{B}_T\widehat{W}_T - B^\star W^\star\|_F^2 \leqslant \sum_{m=1}^M \epsilon^2\|\theta_m^\star\|^2 \leqslant \epsilon^2\mu^2 k{\sigma_{\max}^\star}^2.$$

The above inequality uses the fact that $B^\star$ is a unitary matrix and Assumption II.3. Hence, by combining these results and using the union bound, we conclude that with probability at least $O(1 - \delta - d^{-10} - de^{-\frac{\delta'^2 n_{M+1}}{3\|x_{M+1,n}\|^2}})$, we have

$$\sum_{m=1}^M \|P^\perp_{X_m\widehat{B}_T}X_mB^\star w_m^\star\|_2^2$$
$$\leqslant \epsilon^2\mu^2k{\sigma_{\max}^\star}^2\left(2\sum_{m=1}^M n_m(d-k) + 3\log\frac{1}{\delta}\right).$$

Substituting in Eq. (4) completes the proof. □

Appendix II
Proof of Theorem III.1

From the definition of $\mathrm{ER}(\widehat{B}_T, \widehat{w}_{M+1})$, we have $\mathrm{ER}(\widehat{B}_T, \widehat{w}_{M+1})$

$$= \frac{1}{2}\mathbb{E}_{x_{M+1,n}\sim p_{M+1}}\left[\left(x_{M+1,n}^\top(\widehat{B}_T\widehat{w}_{M+1} - B^\star w_{M+1}^\star)\right)^2\right]$$

$$= (1/2)(\widehat{B}_T \widehat{w}_{M+1} - B^\star w^\star_{M+1})^\top (\widehat{B}_T \widehat{w}_{M+1} - B^\star w^\star_{M+1}) \tag{8}$$

$$\leqslant \frac{1}{2(1-\delta')n_{M+1}} \|X_{M+1}(\widehat{B}_T \widehat{w}_{M+1} - B^\star w^\star_{M+1})\|^2 \tag{9}$$

$$= \frac{1}{2(1-\delta')n_{M+1}} \|X_{M+1}\widehat{B}_T((X_{M+1}\widehat{B}_T)^\top (X_{M+1}\widehat{B}_T))^\dagger$$
$$(X_{M+1}\widehat{B}_T)^\top Y_{M+1} - X_{M+1}B^\star w^\star_{M+1}\|^2 \tag{10}$$

$$= \frac{1}{2(1-\delta')n_{M+1}} \|P_{X_{M+1}\widehat{B}_T}(X_{M+1}B^\star w^\star_{M+1} + Z_{M+1})$$
$$- X_{M+1}B^\star w^\star_{M+1}\|^2$$

$$= \frac{1}{2(1-\delta')n_{M+1}} \|P_{X_{M+1}\widehat{B}_T} Z_{M+1}\|^2$$
$$+ \frac{1}{2(1-\delta')n_{M+1}} \|P^\perp_{X_{M+1}\widehat{B}_T} X_{M+1}B^\star w^\star_{M+1}\|^2 \tag{11}$$

$$= \frac{1}{2(1-\delta')n_{M+1}} \|P_{X_{M+1}\widehat{B}_T} Z_{M+1}\|^2$$
$$+ \frac{1}{2(1-\delta')n_{M+1}} \|P^\perp_{X_{M+1}\widehat{B}_T} X_{M+1}B^\star \widetilde{W}^\star \widetilde{\nu}^\star\|^2 \tag{12}$$

$$\leqslant \frac{1}{2(1-\delta')n_{M+1}} \|P_{X_{M+1}\widehat{B}_T} Z_{M+1}\|^2$$
$$+ \frac{1}{2(1-\delta')n_{M+1}} \|P^\perp_{X_{M+1}\widehat{B}_T} X_{M+1}B^\star \widetilde{W}^\star\|_F^2 \|\widetilde{\nu}^\star\|_2^2$$

where $\widetilde{W}^\star = W^\star \sqrt{\mathrm{diag}([n_1, n_2, \cdots, n_M])}$ and $\widetilde{\nu}^\star(m) = \frac{\nu^\star(m)}{\sqrt{n_m}}$. Eq. (8) is derived from $\mathbb{E}\left[x_{M+1,n} x_{M+1,n}^\top\right] = I$. Eq. (9) is derived from Lemma I.1. Eq. (10) is derived from the least square estimator solution of the optimality of $\widehat{w}_{M+1}$. Eq. (11) is derived from $P^\perp_{X_{M+1}\widehat{B}_T}{}^\top P_{X_{M+1}\widehat{B}_T} = 0$. Eq. (12) is derived from $w^\star_{M+1} = \widetilde{W}^\star \widetilde{\nu}^\star$. Given that Z_{M+1} follows i.i.d. Gaussian distribution with a zero mean and variance σ^2, it follows that $\frac{1}{\sigma^2}\|P_{X_{M+1}\widehat{B}_T} Z_{M+1}\|^2 \sim \chi^2(k)$. Applying the Chernoff bound for chi-square distribution, we have with probability at least $1-\delta$, $\|P_{X_{M+1}\widehat{B}_T} Z_{M+1}\|^2 \leqslant \sigma^2(2k + 3\log\frac{1}{\delta})$. Following that, by combining the result obtained from Lemma I.2 along with applying the union bound, we derive that with probability at least $O(1 - \delta - d^{-10} - de^{-\frac{\delta'^2 n_{M+1}}{3\|x_{M+1,n}\|^2}}$,

$$\mathrm{ER}(\widehat{B}_T, \widehat{w}_{M+1}) \leqslant \frac{\sigma^2(2k + 3\log\frac{1}{\delta})}{2(1-\delta')n_{M+1}} + \frac{(1+\delta')}{2(1-\delta')^2}\mu^2 k {\sigma^\star_{\max}}^2$$
$$\epsilon^2 \left(2N(d-k) + 3\log\frac{1}{\delta}\right) \|\widetilde{\nu}^\star\|_2^2.$$

Our objective in the remaining analysis is to determine the upper bound of $\|\widetilde{\nu}^\star\|_2^2$. Define $\epsilon^{-2} = \frac{N}{\|\nu^\star\|_2^2}$. Using a technique similar to Theorem 3.2 in [2], for any $\gamma \in [0, 1]$,

$$\|\widetilde{\nu}^\star\|_2^2 \leqslant \frac{2\|\nu^\star\|_2^2}{N}((1-\gamma)\|\nu^\star\|_{0,\gamma} + \gamma M).$$

By combining these results, we obtain the upper bound as

$$\mathrm{ER}(\widehat{B}_T, \widehat{w}_{M+1}) \leqslant \frac{\sigma^2(2k + 3\log\frac{1}{\delta})}{2(1-\delta')n_{M+1}} + \frac{(1+\delta')}{(1-\delta')^2}\mu^2 k {\sigma^\star_{\max}}^2$$
$$\epsilon^2 \left(2(d-k) + \frac{3}{N}\log\frac{1}{\delta}\right) \|\nu^\star\|_2^2 s^\star.$$

For $0 < c < 1$, setting target sample size $n_{M+1} \geqslant \frac{\sigma^2(2k+3\log\frac{1}{\delta})}{2(1-c)(1-\delta')}\epsilon^{-1}$ ensures that

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

Define $t := \frac{(1+\delta')}{(1-\delta')^2}\mu^2 k {\sigma^\star_{\max}}^2 \|\nu^\star\|_2^2 s^\star$. For $C > 1$, setting source sample size $N \geqslant \frac{3C}{c} t\epsilon \log\frac{1}{\delta}$ results in

$$N \geqslant \frac{3C}{c} t\epsilon \log\frac{1}{\delta} = \frac{3\log\frac{1}{\delta}}{2(d-k)} C(\frac{2}{c}(d-k)t\epsilon)$$
$$\geqslant \frac{3\log\frac{1}{\delta}}{2(d-k)} \frac{\frac{2}{c}(d-k)t\epsilon}{1 - \frac{2}{c}(d-k)t\epsilon} = \frac{3t\log\frac{1}{\delta}}{c\epsilon^{-1} - 2(d-k)t} \tag{13}$$

where Eq. (13) is derived from the fact that there exists a constant $C > 1$ satisfying the inequality $\frac{x}{1-x} \leqslant Cx$ for $0 < x < 1$. Consequently, $(2(d-k) + \frac{3}{N}\log\frac{1}{\delta})t\epsilon^2 \leqslant c\epsilon$. Thus, $\mathrm{ER}(\widehat{B}_T, \widehat{w}_{M+1}) \leqslant \epsilon$ and completes the proof.

References

[1] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei, "Few-shot learning via learning the representation, provably," *arXiv preprint arXiv:2002.09434*, 2020.

[2] Y. Chen, K. Jamieson, and S. Du, "Active multi-task representation learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 3271–3298.

[3] A. Capone, A. Lederer, J. Umlauft, and S. Hirche, "Data selection for multi-task learning under dynamic constraints," *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 959–964, 2020.

[4] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, pp. 41–75, 1997.

[5] Y. Zhu, P. Thangeda, M. Ornik, and K. Hauser, "Few-shot adaptation for manipulating granular materials under domain shift," *arXiv preprint arXiv:2303.02893*, 2023.

[6] S. Thrun and L. Pratt, "Learning to learn: Introduction and overview," in *Learning to learn*. Springer, 1998, pp. 3–17.

[7] J. Baxter, "A model of inductive bias learning," *Journal of artificial intelligence research*, vol. 12, pp. 149–198, 2000.

[8] N. Tripuraneni, C. Jin, and M. Jordan, "Provable meta-learning of linear representations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 434–10 443.

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

[10] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *International conference on machine learning*, 2021, pp. 2089–2099.

[11] Z. Xu and A. Tewari, "Representation learning beyond linear prediction functions," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4792–4804, 2021.

[12] Y. Wang, Y. Chen, K. Jamieson, and S. S. Du, "Improved active multi-task representation learning via lasso," in *International Conference on Machine Learning*, 2023, pp. 35 548–35 578.

[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] S. Nayer and N. Vaswani, "Fast and sample-efficient federated low rank matrix recovery from column-wise linear and quadratic projections," *IEEE Transactions on Infomation Theory*, 2023.

[15] N. Vaswani, "Efficient federated low rank matrix recovery via alternating gd and minimization: A simple proof," *IEEE Transactions on Infomation Theory*, 2024 (to appear).

[16] Y. Chen and E. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[17] J. Lin, S. Moothedath, and N. Vaswani, "Fast and sample efficient multi-task representation learning in stochastic contextual bandits," in *International Conference on Machine Learning*, 2024.