Improved Active Multi-Task Representation Learning via Lasso

Yiping Wang [1] **Yifang Chen** [2] **Kevin Jamieson** [2] **Simon S. Du** [2]

Abstract

To leverage the copious amount of data from source tasks and overcome the scarcity of the target task samples, representation learning based on multi-task pretraining has become a standard approach in many applications. However, up until now, most existing works design a source task selection strategy from a purely empirical perspective. Recently, Chen et al. (2022) gave the first active multi-task representation learning (A-MTRL) algorithm which adaptively samples from source tasks and can provably reduce the total sample complexity using the L2-regularized-target-source-relevance parameter ν^2. But their work is theoretically suboptimal in terms of total source sample complexity and is less practical in some real-world scenarios where sparse training source task selection is desired. In this paper, we address both issues. Specifically, we show the strict dominance of the L1-regularized-relevance-based (ν^1-based) strategy by giving a lower bound for the ν^2-based strategy. When ν^1 is unknown, we propose a practical algorithm that uses the LASSO program to estimate ν^1. Our algorithm successfully recovers the optimal result in the known case. In addition to our sample complexity results, we also characterize the potential of our ν^1-based strategy in sample-cost-sensitive settings. Finally, we provide experiments on real-world computer vision datasets to illustrate the effectiveness of our proposed method.

1. Introduction

Deep learning has been successful because it can effectively learn a proper feature extractor that can map high-dimensional, highly structured inputs like natural images and natural language into a relatively low-dimensional representation. Recently, a big focus in deep learning has been on few-shot learning, where there is not enough data to learn a good representation and a prediction function from scratch. One solution is using multi-task learning, which uses data from other sources to help the few-shot target. This approach is based on the idea that different tasks can share a common representation. The process starts by training on a lot of source tasks to learn a simpler representation and then uses that pre-trained representation to train on a limited amount of target data.

Accessing a large amount of source data for multi-task representation learning (MTRL) may be easy, but processing and training on all that data can be costly. Therefore, it is important to find ways to minimize the number of samples, and perhaps the number of sources, needed from source tasks while still achieving the desired performance on the target task. Naturally, not all source tasks are equally important for learning the representation and maximizing performance on the target task. But to the best of our knowledge, most research in this area chooses which tasks to include in the training of the multi-task representation in an ad hoc way (Asai et al., 2022; Fifty et al., 2021; Yao et al., 2022; Zaiem et al., 2021; Zamir et al., 2018; Zhang et al., 2022b). Notable exceptions include (Chen et al., 2021; 2022) that study ways to improve training efficiency and reduce the cost of processing source data by prioritizing certain tasks during training with theoretical guarantees.

On the other hand, the significant empirical success of MTRL has motivated a number of theoretical studies (Du et al., 2021; Chen et al., 2022; Tripuraneni et al., 2021). In particular, (Du et al., 2021) and (Tripuraneni et al., 2021) provide generalization (excess risk) upper bounds on the estimation error of the target task for passive multi-task representation learning (P-MTRL). Here, *passive* means that samples are drawn from tasks according to some non-adaptive sampling strategy fixed before data is observed (e.g., an equal number of samples from each task). Tripuraneni et al. (2021) also proves a lower bound related to the quality of whole feature representations in P-MTRL.

In this paper, our main focus is to guarantee a specific level of accuracy on a target task while provably using the least amount of data from other related tasks. This is achieved

[1]College of Computer Science and Technology, Zhejiang University [2]Paul G. Allen School of Computer Science & Engineering, University of Washington. Correspondence to: Yiping Wang <yipingw@zju.edu.cn>.

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

through task-level active learning. Chen et al. (2022) is the first work to propose an active multi-task representation learning (A-MTRL) algorithm that can provably reduce the total number of samples from all the tasks compared to the passive learning version (P-MTRL) by estimating the relevance of each source task to the target task and sampling accordingly. However, this previous work has several limitations and leaves some questions open, in both theory and practical application. For example, they did not study the lower bounds of the excess risk on the target task for Multi-Task Transfer Learning. Furthermore, Chen et al. (2022) proposed an L_2 regularized source-to-target-task relevance quantity ν^2, but it is unclear whether this relevance score is the best criterion for the A-MTRL design compared to other possible relevance scores. As we will show later, their A-MTRL algorithm may not be optimal. In our work, we build on (Chen et al., 2022) by optimizing their upper bound of the excess risk and show that this yields an asymptotically optimal sampling strategy which corresponds to an L_1 regularized relevance quantity ν^1 and samples from this distribution accordingly. Moreover, we provide the first sampling-algorithm-dependent minimax lower bound of excess risk on the target task for both the A-MTRL in (Chen et al., 2022) and P-MTRL, which shows that our algorithm can strictly outperform these baselines even in the worst case.

In addition to the theoretical bounds, Chen et al. (2022) also has practical limitations. When there exist multiple sampling strategies that are seemingly equivalent under their framework, their algorithm tends to put a little weight on all tasks by nature of the L_2 regularized solution ν^2. This is sometimes undesirable in practice as will illustrate by two examples. First, setting up a sample-generating source can be more expensive than actually generating the samples. For example, in robotics, each source task can be considered as a specific real-world testing environment that can take weeks to set up, but then samples can be generated quickly and plentifully (Shi et al., 2021). Second, previous research assumes that the cost of samples is the same no matter how much data we need or for how long. However, in reality, subscribing to data from a single source for a long period of time can lead to a lower average label cost. Therefore, even with the same sample complexity from sources, choosing fewer source tasks can be more beneficial. We propose a general-purpose cost-sensitive A-MTRL strategy that addresses these scenarios and demonstrates the potential of our proposed L_1 regularized strategy in various cost-effective situations.

1.1. Our Contributions

We summarize our contributions as follows.

- We begin by proving that the sampling distribution over tasks using our proposed L_1 strategy defined in Def. 2.5 minimizes the target excess risk upper bound of (Chen et al., 2022). We then consider a class of strategies Lp-A-MTRL (A-MTRL with L_p strategy) and show that, when $T \gtrsim k^2$, for N_{tot} number of total source samples, L1-A-MTRL is strictly dominant over this class by proving that its estimation error decreases at least as fast as $\widetilde{\mathcal{O}}(\frac{k}{\sigma_k^2 N_{tot}})$ while the error of the L2-A-MTRL/P-MTRL strategies suffers algorithm-dependent minimax lower bound of at least $\widetilde{\Omega}(\frac{T}{k\sigma_k^2 N_{tot}})$. Here T is the number of source tasks, k is the dimension for the non-shared prediction function and σ_k characterizes the diversity of source tasks which will be specified later. These minimax lower bounds are novel to the MTRL literature. (Section 3.1, 3.2)

- While the L1-A-MTRL strategy provably has sample complexity benefits over other sampling strategies, it is not directly implementable in practice since it requires prior knowledge of ν^1 (i.e., those bounds only demonstrate the performance of the sampling distribution, not how to find it). Consequently, inspired by (Chen et al., 2022), we design a practical strategy that utilizes the Lasso and a low order number of samples to estimate the relevance vector ν^1, and then apply the L_1 strategy to sample source data using the estimated ν^1. We show that this practical algorithm achieves a sample complexity nearly as good as when ν^1 is known. The key technical innovation here is that when the regularization parameter is lower bounded, the Lasso solution can be close to the ground truth value. (Section 3.3)

- Going beyond these main results, we demonstrate that our L1-A-MTRL strategy can be extended to support many sample-cost-sensitive scenarios by levering its sparse source task selection properties. We formulate this setting as an optimization problem and formally characterize the benign cost function under which our L1-A-MTRL strategy is beneficial (Section 4)

- Finally, we empirically show the effectiveness of our algorithms. If we denote the practical algorithm of (Chen et al., 2022) by L2-A-MTRL, we show that our proposed L1-A-MTRL algorithm achieves 0.54% higher average accuracy on MNIST-C relative to L2-A-MTRL (92.6%), which confirms our theoretical results. We then restrict most of the data to be sampled from no more than 10 tasks, in order to mimic the sample-cost-sensitive setting with decreasing per-sample cost. Here we find L1-A-MTRL achieves 2.2% higher average accuracy relative to the uniform sampling (94.3%). (Section 5).

2. Preliminaries

In this section, We describe the relevant notations and the problem setup for further theoretical analysis.

2.1. Notation

Miscellaneous. Let $[T] := \{1, 2, ..., T\}$ denotes the set of source tasks and $n_{[T]} := \{n_1, n_2, ..., n_T\}$ denotes the number of samples dedicated to each task. Likewise, $n_{[T],i} := \{n_{1,i}, ..., n_{T,i}\}$ represents the data from each task at the i-stage for multi-stage learning procedure. If S is an index set, $|S|$ denotes the number of elements in S. We use $\|\cdot\|_p$ to denote the l_p norm of vectors and use $|\cdot|$ or $\|\cdot\|$ to denote the l_2 norm for convenience. Let $\text{subG}_d(\rho^2)$ be the d-dimensional sub-gaussian variables with variance ρ.

Singular Values. For $A \in \mathbb{R}^{m\times n}$, we denote by $\sigma_i(A)$ the i-th singular value of A, which satisfy $\sigma_1(A) \geq ... \geq \sigma_r(A) > 0$ with $r = rank(A)$. And we specify $\kappa(A)$ as the condition number of A, i.e., $\kappa(A) = \frac{\sigma_1(A)}{\sigma_r(A)}$ if $\sigma_r(A) > 0$.

Asymptotic comparison. We use the standard O, Ω, Θ notations to hide the universal constants, and further use $\widetilde{O}, \widetilde{\Omega}, \widetilde{\Theta}$ to hide logarithmic factors. We use $a \lesssim b$ or $b \gtrsim a$ to denote $a = O(b)$ and use $a \asymp b$ to denote $a = \Theta(b)$.

2.2. Problem Setup

Multi-Task. Let $t \in [T]$ be the index of the T source tasks and index $T+1$ denotes the target task. Each task $t \in [T+1]$ is associated with a joint distribution μ_t over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. In this paper we assume $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$.

Data Generation. Like in (Chen et al., 2022), we assume there exists an underlying representation function $\phi^* : \mathcal{X} \to \mathcal{R}$ which maps the input space $\mathcal{X}$ to a feature space $\mathcal{R} \in \mathbb{R}^k$ where $k \ll d$. And the representation functions are restricted to be in some function classes Φ, e.g., linear functions, convolutional networks, etc. We further assume that each t-th task for $t \in [T+1]$ follows a ground truth linear head w_t^* that maps the particular feature to the corresponding label. To be more specific, we assume the *i.i.d* sample $(x_t, y_t) \sim \mu_t$ satisfies

$$y_t = \phi^*(x_t)^\top w_t^* + z_t, \quad z_t \sim \mathcal{N}(0, \sigma_z^2) \tag{1}$$

where $x_t \sim p_t$ and x_t is independent to z_t. For convenience, we denote $X_t = [x_{t,1}, .., x_{t,n_t}]^\top \in \mathbb{R}^{n_t\times d}$ to be the input data matrix which contained n_t *i.i.d.* sampled data $(x_{t,1}, y_{t,1}), ..., (x_{t,n_t}, y_{t,n_t}) \sim \mu_t$ from the t-th task, and $Y_t = [y_{t,1}, ..., y_{t,n_t}]^\top \in \mathbb{R}^{n_t}$, $Z_t = [z_{t,1}, ..., z_{t,n_t}] \in \mathbb{R}^{n_t}$ to be the labels and noise terms aligned to the inputs. For convenience, we define $N_{tot} := \sum_{t=1}^T n_t$ to be the total sampling number from all the source tasks.

Transfer Learning Process. As in (Du et al., 2021), firstly we learn the representation map on the source tasks by solving the following optimization problem

$$\hat{\phi}, \hat{w}_1, ..., \hat{w}_T = \underset{\phi\in\Phi, w_1,...,w_T\in\mathbb{R}^k}{\arg\min} \frac{1}{T}\sum_{t=1}^T \frac{1}{n_t}\|Y_t - \phi(X_t)w_t\|_2^2 \tag{2}$$

Here we allow n_t to vary from different tasks rather than requiring uniform sampling and $\phi(X_t) := [\phi(x_{t,1}), ..., \phi(x_{t,n_t})]^\top \in \mathbb{R}^{n_t\times k}$. Then we retain the learned representation and apply it to the target task while training a specific linear head for this task:

$$\hat{w}_{T+1} = \underset{w_{T+1}}{\arg\min} \frac{1}{n_{T+1}}\|Y_{T+1} - \hat{\phi}(X_{T+1})w_{T+1}\|_2^2 \tag{3}$$

Goal. Our main goal is to bound the *excess risk* (ER) of our model on the target task with parameters $(\hat{\phi}(x), \hat{w}_{T+1})$ while minimizing the total cost of sampling data from the source tasks. Here, like in (Du et al., 2021; Chen et al., 2022), we define the population loss as $L_{T+1}(\hat{\phi}, \hat{w}_{T+1}) = \mathbb{E}_{(x,y)\sim\mu_{T+1}}[(y_{T+1} - \hat{\phi}(x_{T+1})^\top \hat{w}_{T+1})^2]$. Then from (1) we can define the excess risk:

$$\begin{aligned} &ER(\hat{\phi}, \hat{w}_{T+1}, \phi^*, w_{T+1}^*) \\ &= L_{T+1}(\hat{\phi}, \hat{w}_{T+1}) - L_{T+1}(\phi^*, w_{T+1}^*) \\ &= \mathbb{E}_{x\sim p_{T+1}}[(\hat{\phi}(x)^\top \hat{w}_t - \phi^*(x)^\top w_t^*)^2] \end{aligned} \tag{4}$$

It should be mentioned that in this paper, we consider the model performance under the worst circumstance, therefore we treat the ground truth parameters ϕ^*, w_{T+1}^* as the arguments of excess risk, which is different from that in the previous works (Du et al., 2021; Chen et al., 2022).

Linear Representation. Our theoretical study concentrates on the linear representation function class, which is widely used in many previous works (Du et al., 2021; Tripuraneni et al., 2020; 2021; Thekumparampil et al., 2021; Chen et al., 2022). Namely, we let $\Phi = \{x \mapsto B^\top x \mid B \in \mathbb{R}^{d\times k}\}$ and thus $\phi(X_t) = X_t B \in \mathbb{R}^{n_t\times k}$. Without loss of generality, we assume the ground truth representation map B^* is an orthonormal matrix, i.e., $B^* \in O_{d,k}$, which is also commonly used in the related works (Chen et al., 2022; Tripuraneni et al., 2021; Kumar et al., 2022).

Other assumptions. Assume $\mathbb{E}_{x_t\sim p_t}[x_t] = 0$, $\Sigma_t^* = \mathbb{E}_{x_t\sim p_t}[x_t x_t^\top]$ and $\hat{\Sigma}_t := \frac{1}{n_t}(X_t)^\top X_t$ for any $t \in [T+1]$. We have the following assumptions for the data distribution:

Assumption 2.1. (sub-gaussian input). There exists $\rho \geq 1$ such that $x_t \sim p_t$ is $\text{subG}_d(\rho^2)$ for all $t \in [T+1]$.

Assumption 2.2. (proper variance) For all $t \in [T+1]$, we have $\sigma_{\max}(\Sigma_t^*) = \Theta(1)$ and $\sigma_{\min}(\Sigma_t^*) = \Theta(1)$.

Variance conditions are common in the related works (Tripuraneni et al., 2021; Du et al., 2021; Chen et al., 2022) and

Assumption 2.2 is a generalization than identical variance assumption used in (Tripuraneni et al., 2021; Chen et al., 2022) which requires $\Sigma_1 = ... = \Sigma_{T+1} = I_d$. Specially, we only use the identical variance assumption in Section 3.3.

Assumption 2.3. (high dimension input and enough tasks) The parameters satisfy $d > T \geq k \geq 1$ and $d \gg k$.

Finally, we also need *diverse task* assumption mentioned in (Tripuraneni et al., 2021; Du et al., 2021; Chen et al., 2022). Denote $W^* := [w_1^*, ..., w_T^*] \in \mathbb{R}^{k\times T}$, then we assume:

Assumption 2.4. (diverse task) The matrix W^* satisfies $\sigma_{\min}(W^*) > 0$.

Assumption 2.4 claims that W^* has full row rank, so we can definitely find some $\nu \in \mathbb{R}^T$ such that $W^*\nu = w_{T+1}^*$, and thus (6) in Def. 2.5 is well-defined. It's a necessary assumption for learning reasonable features as proven by (Tripuraneni et al., 2021).

2.3. Scope of A-MTRL algorithms in this paper

Here we state the scope of the A-MTRL algorithm considered in this paper. Recall that in (Chen et al., 2022), the learner samples in proportional to $\frac{\hat{\nu}(t)^2}{\|\hat{\nu}\|_2^2}$ number of data from task t, where $\hat{\nu}$ is defined via the following solution:

$$\arg\min_{\nu} \|\nu\|_2 \qquad \text{s.t. } W^*\nu = w_{T+1}^* \tag{5}$$

Here instead of focusing on this L_2 regularization, we study the whole candidate set of source-target relevance terms and the corresponding sampling strategies. Formally, we generalize Definition 3.1 of (Chen et al., 2022) to propose:

Definition 2.5. (L_pN_q sampling strategy) Let $\nu(t)$ be the t-th element of vector $\nu \in \mathbb{R}^T$ and $\underline{N}$ be the minimum number of sampling data from every source task. The L_pN_q strategy is defined as taking $n_t = \max\{c'|\nu^p(t)|^q, \underline{N}\}$ for some constant $c' > 0$, where n_t is the number of samples drawn from from the t-th task, and

$$\nu^p = \arg\min_{\nu} \|\nu\|_p \qquad \text{s.t. } W^*\nu = w_{T+1}^*. \tag{6}$$

If $p = q$, we denote L_p as the abbreviation of L_pN_q. For example, if $\underline{N} = 0$, then the L_1 strategy corresponds to $n_t = \frac{N_{tot}}{\|\nu^1\|_1} \cdot |\nu^1(t)|$ and the L_2 strategy corresponds to $n_t = \frac{N_{tot}}{\|\nu^2\|_2^2}|\nu^2(t)|^2$, where N_{tot} is the total source sampling budget.

In the rest of the paper, we will focus on this L_pN_q sampling strategy set.

3. Main Results

3.1. Optimal Strategy L1-A-MTRL with Known ν

In this section, we aim to obtain the optimal sampling strategy that can achieve the required performance on the target task with the smallest possible number of samples from source tasks. Firstly, with linear representation assumption, we rewrite $ER(\hat{B}, \hat{w}_{T+1}, B^*, w_{T+1}^*)$ in (4) as follows:

$$\mathbb{E}_{x\sim p_{T+1}} \|x^\top(\hat{B}\hat{w}_{T+1} - B^* w_{T+1}^*)\|_2^2. \tag{7}$$

Then from the intermediate result of Theorem 3.2 in (Chen et al., 2022), we get the upper bound of excess risk for all A-MTRL methods:

Theorem 3.1. *(Chen et al., 2022) Fix a failure probability $\delta \in (0, 1)$. If Assumption 2.1, 2.2, 2.3, 2.4 hold, and the sample size in source and target tasks satisfy $n_t \gg \rho^4(d + \ln(\frac{T}{\delta}))$ for all $t \in [T]$ and $n_{T+1} \gg \rho^4(k + \ln(\frac{1}{\delta}))$, then with probability at least $1 - \delta$ we have:*

$$\begin{aligned} &ER(\hat{B}, \hat{w}_{T+1}, B^*, w_{T+1}^*) \\ &\lesssim \sigma^2(kd\ln(\frac{N_{tot}}{T}) + kT + \ln(\frac{1}{\delta}))\|\widetilde{\nu}\|_2^2 + \sigma^2\frac{(k + \ln(\frac{1}{\delta}))}{n_{T+1}} \end{aligned} \tag{8}$$

where $\nu \in \{\nu' \in \mathbb{R}^T | W^*\nu' = w_{T+1}^*\}$ and $\widetilde{\nu}(t) = \frac{\nu(t)}{\sqrt{n_t}}$.

The key idea behind Theorem 3.1 is as follows. (Du et al., 2021) provides the first upper bound for the MTRL problem. They consider sampling data evenly from each source task and demonstrated that following the transfer learning process (Eqn. 2, 3), the target task error can be controlled by the source-task training error $\mathcal{O}(\sigma^2(kd + kT)/N_{tot})$ and the target-task fine-tuning error $\mathcal{O}(\sigma^2 k/n_{T+1})$. However, in their proof, the ground truth linear head w_{T+1}^* is required to satisfy a distribution Q such that $\|E_{w\sim Q}[ww^\top]\| \leq O(\frac{1}{k})$. (Chen et al., 2022) go beyond this limitation by leveraging the equation $W^*\nu^* = \widetilde{W}^*\widetilde{\nu}^* = w_{T+1}^*$, where $\widetilde{w}_t^* = w_t^*\sqrt{n_t}$ and $\widetilde{\nu}^*(t) = \frac{\nu^*(t)}{\sqrt{n_t}}$. This idea introduces the source-target relevance vector $\nu^* \in R^T$ into the bound and results in Eqn. 8.

Inspired by Theorem 3.1, in order to minimize the excess risk bound with a fixed sampling quota N_{tot}, we need to find the optimal sampling strategy $n_{[T]} = \{n_1, ..., n_T\}$ by solving the following optimization problem:

$$\begin{aligned} \min_{\nu, n_{[T]}} \|\widetilde{\nu}\|_2^2 &= \sum_{t=1}^{T} \frac{(\nu(t))^2}{n_t} \\ \text{s.t. } W^*\nu &= w_{T+1}^* \\ \sum_{t=1}^{T} n_t &= N_{\text{tot}} \\ n_t &\geq \underline{N}, \quad \forall t \in [T] \end{aligned} \tag{9}$$

Here $\underline{N}(\gg \rho^4(d + \ln(\frac{T}{\delta})))$ is the minimum sampling number for every source task as in Theorem 3.1. In this section, we will transform (9) into a bi-level optimization problem and obtain the asymptotic optimal solutions of (9).

3.1.1. Optimal Strategy for Any Fixed ν

We first consider a fixed ν in (9) and find the optimal sampling strategy accordingly, and we get:

Lemma 3.2. *For any fixed ν such that $W^*\nu = w^*_{T+1}$, the optimal $n^*_{[T]}$ for minimizing $\|\widetilde{\nu}\|_2^2$ satisfies $n_t^* = \max\{c'|\nu(t)|, \underline{N}\}$ for every $t \in [T]$, where $c' > 0$ is some constant such that $\sum_{t=1}^T n_t^* = N_{tot}$.*

This lemma indicates an optimal sampling strategy under some fixed, arbitrary $\nu \in \{\nu' | W^*\nu' = w^*_{T+1}\}$. We can then apply Lemma 3.2 to the previous bound (8) and deduce the theoretical optimal bound on the sample complexity of the source tasks for any suitable ν. Here, for simplicity, we skip the trivial case where the model achieves sufficiently high accuracy with uniformly allocated sampling data $\underline{N}$ by requiring $\varepsilon^2 \ll \min(1, \sigma^2(kd+kT)\frac{\|\nu\|_1^2}{T\underline{N}})$. This condition guarantees that $N_{tot} \gg T\underline{N}$, and we get:

Corollary 3.3. *Assume Assumption 2.1, 2.2, 2.3, 2.4 hold and ν is fixed. Then the optimal sampling strategy $n_{[T]}$ satisfies $n_t = \max\{c'|\nu(t)|, \underline{N}\}$, $\forall t \in [T]$, and with probability at least $1-\delta$, the optimal A-MTRL algorithm satisfies $ER \le \varepsilon^2$ with $\varepsilon^2 \ll \min(1, \sigma^2(kd+kT)\frac{\|\nu\|_1^2}{T\underline{N}})$ whenever the total sampling budget from all source tasks N_{tot} is at least*

$$\widetilde{\mathcal{O}}(\sigma^2(kd+kT)\|\nu\|_1^2\varepsilon^{-2}) \tag{10}$$

and the number of target samples is at least $\widetilde{\mathcal{O}}(\sigma^2 k\varepsilon^{-2})$.

Discussion. To show the optimality of our bound, we compare this with the result in (Chen et al., 2022). Their known ν^2 (denoted as ν^* in their original paper) is equivalent to

$$\arg\min_{\nu} \|\nu\|_2 \quad \text{s.t. } W^*\nu = w^*_{T+1}.$$

Under the same setting but using this ν^2, with probability at least $1-\delta$, A-MTRL algorithm with sampling strategy $n_{[T]}$ such that $n_t = \max\{c''(\nu(t))^2, \underline{N}\}$, $\forall t \in [T]$ satisfies $ER \le \varepsilon^2$ with $\varepsilon \ll 1$ whenever N_{tot} is at least

$$\widetilde{\mathcal{O}}(\sigma^2(kd+kT)s^*\|\nu^2\|_2^2\varepsilon^{-2}) \tag{11}$$

and the required number of target samples is also $\widetilde{\mathcal{O}}(\sigma^2 k\varepsilon^{-2})$. Here $s^* = \min_{\gamma\in[0,1]}(1-\gamma)\|\nu^2\|_{0,\gamma} + \gamma T$ and $\|\nu^2\|_{0,\gamma} := |\{t : |\nu^2(t)| > \sqrt{\gamma\|\nu^2\|_2^2 N_{tot}^{-1}}\}|$. From Lemma 3.2 we know our strategy is better than the previous under given arbitrary ν setting, so we have $\|\nu\|_1 \lesssim \sqrt{s^*}\|\nu\|_2 \le \sqrt{T}\|\nu\|_2, \forall \nu \in \{\nu' | W^*\nu' = w^*_{T+1}\}$. In particular, we show the gap between $\|\nu\|_1$ and $\sqrt{s^*}\|\nu\|_2$ can be very large under some special cases as follows.

Example: Almost Sparse ν. Assume $T \gg 1$, $N_{tot} \gg \underline{N}T \ge T$, then we consider an extreme case where

$$\nu(t) = \begin{cases} \sqrt{1-\frac{1}{T-1}} & , t = 1 \\ \frac{1}{T-1} & , t = 2, ..., T \end{cases} \tag{12}$$

Then ν is approximately 1-sparse since $\frac{1}{T-1} \ll 1$, and we have $\|\nu\|_1 = \sqrt{1-\frac{1}{T-1}} + 1 < 2$, $\|\nu\|_2 = 1$. Let $\gamma_0 := \frac{N_{tot}}{(T-1)^2}$, it's easy to prove $s^* \ge \min\{\gamma_0, 1\} \times T \gg 1$. This result in $\sqrt{s^*}\|\nu\|_2 \gg \|\nu\|_1$ and A-MTRL in (Chen et al., 2022) requires a sample complexity that is $\min\{\frac{N_{tot}}{2(T-1)}, \frac{T}{2}\}$ times larger than our optimal sampling strategy.

3.1.2. Optimal ν in Candidate Set

Secondly, suppose we are able to access the whole set $\{\nu' | W^*\nu' = w^*_{T+1}\}$, now we aim to find the optimal ν from the candidate set for sampling. Once we find such a ν^*, we can utilize rules in Lemma 3.2 to obtain $n^*_{[T]}$ and apply all the results above. Here we focus on the case in (Chen et al., 2022) where ER bound $\varepsilon^2 \to 0$ and $N_{tot} \to +\infty$ and we deduce that L_1-minimization solution is the best choice.

Theorem 3.4. *Let $(\nu^1, n^1_{[T]})$ denotes the sampling parameters of L_1 strategy defined in Def. 2.5, i.e.,*

$$\begin{aligned} \nu^1 &= \arg\min_{\nu} \|\nu\|_1 \qquad \text{s.t. } W^*\nu = w^*_{T+1} \\ n_t^1 &= \max\{c'|\nu^1(t)|, \underline{N}\}, \qquad \forall t \in [T] \end{aligned} \tag{13}$$

Let $(\nu^*, n^*_{[T]})$ denote the optimal solution of (9). Then as $N_{tot} \to +\infty$ we have $\nu^1 \to \nu^*$, $n^1_{[T]} \to n^*_{[T]}$.

Theorem 3.4 shows that the L_1 strategy can correspond to the asymptotic optimal solution of (9). As a reference, Alg. 1 in (Chen et al., 2022) is equivalent to L_2 strategy, and we call these classes of methods **Lp-A-MTRL** (A-MTRL with L_p strategy) method with known ν^p for further discussion.

3.2. How Good Is L1-A-MTRL with Known ν? Comparison on the Worst Target Task

To show the effectiveness of the L_1 strategy with known ν^1, we analyze the performance of MTRL algorithms on a worst-case target task w^*_{T+1} that maximizes the excess risk. Firstly, for better comparison, we define the sampling-algorithm-dependent minimax lower bound of excess risk. Let $\Gamma(\sigma_k) = \{W \in \mathbb{R}^{k\times T} | \sigma_{\min}(W) \ge \sigma_k\}$ for any $\sigma_k > 0$, then we define:

Definition 3.5. (minimax ER lower bound) The mini-max lower bound of ER on the target task for Lp-A-MTRL method $\underline{ER}_{L_p}(\sigma_k)$ is defined as

$$\begin{aligned} &\inf_{(\hat{B}, \hat{w}_{T+1})} \sup_{(B^*, W^*, w^*_{T+1})} \mathbb{E}_{x\sim\mu_{T+1}} \|x^\top(\hat{B}\hat{w}_{T+1} - B^* w^*_{T+1})\|_2^2 \\ &= \inf_{(\hat{B}, \hat{W})} \sup_{(B^*, W^*, \nu^p)} \mathbb{E}_{x\sim\mu_{T+1}} \|x^\top(\hat{B}\hat{W}\nu^p - B^* W^* \nu^p)\|_2^2 \end{aligned} \tag{14}$$

where W^* varies on $\Gamma(\sigma_k)$ such that Assumption 2.4 holds and ν^p denotes the L_p-minimization solution of $W^*\nu = w^*_{T+1}$ like (6). Similar definitions hold for P-MTRL.

Remark 3.6. The left term of (14) denotes the case that we consider the average error of the best prediction model $(\hat{B}, \hat{w}_{T+1})$ on any target task facing any possible ground truth parameters (B^*, W^*, w^*_{T+1}). The equality of (14) holds because choosing $\hat{w}_{T+1}$ is equivalent to choosing any $\hat{W} \in \{W' | W'\nu^p = \hat{w}_{T+1}\}$, given the (W^*, w^*_{T+1})-dependent ν^p. Note that we consider the L_p strategy as Def. 2.5 which is determined by (ν^p, n_t), so once we choose some Lp-A-MTRL algorithm, (14) just depends on model parameters and σ_k.

With this definition, we show that with known ν^p, the ER on the worst target task for L_1-A-MTRL can reduce up to $\frac{T}{k}$ times of total sampling data from source tasks than that of L_2-A-MTRL(Chen et al., 2022) and P-MTRL.

Theorem 3.7. *Assume conditions in Theorem 3.1 hold, $\|w^*_{T+1}\| = \Theta(1)$ and ν^1, ν^2 defined in Def. 2.5 are known. Then for L1-A-MTRL, we claim ν^1 is at most k-sparse, i.e., $\|\nu^1\|_0 \leq k$. If $N_{tot} \gg T\underline{N}$ and $W^* \in \Gamma(\sigma_k)$, then with probability at least $1-\delta$, for ER defined in (7) we have* [1] *:*

$$ER_{L_1} \lesssim \sigma^2(kd\ln(\frac{N_{tot}}{T}) + kT + \ln(\frac{1}{\delta}))\frac{k}{\sigma_k^2 \cdot N_{tot}}$$

but for P-MTRL and L_2-A-MTRL, with probability at least $1-\delta$ we have :

$$\underline{ER}_{L_2}(\sigma_k), \underline{ER}_{passive}(\sigma_k) \gtrsim \sigma^2 \frac{dT}{\sigma_k^2 \cdot N_{tot}}$$

So when $T \gtrsim k^2$, L1-A-MTRL outperforms L2-A-MTRL and P-MTRL for the worst target task.

Discussion. In essence, the sparsity of ν^p causes the difference in model performance on the worst-case target task. For the upper bound of L1-A-MTRL, We show $\|\widetilde{\nu}^1\|_2^2 \lesssim k/(\sigma_k^2 \cdot N_{tot})$. And for the lower bound of L2-A-MTRL and P-MTRL, we utilize the fact that $\inf_{\hat{B},\hat{W}} \sup_{B^*,W^*} \|X_t(\hat{B}\hat{W} - B^*W^*)\|_2^2 \gtrsim \sigma^2 kd$ (up to logarithmic factors) and the result that when the row of $\widetilde{W}^*$ is well aligned with $\widetilde{\nu}^2$, then $\|(\hat{B} - B^*)\widetilde{W}^*\widetilde{\nu}^2\| \gtrsim \|(\hat{B} - B^*)\widetilde{W}^*\|_F\|\widetilde{\nu}^2\|$, where ν^2 can be chosen to satisfy $\|\widetilde{\nu}^2\| \gtrsim T/(k \cdot \sigma_k^2 \cdot N_{tot})$.

3.3. L1-A-MTRL Algorithm and Theory

In the previous sections, we showed the advantage of A-MTRL with the L_1 sampling strategy when ν^1 is given. However, in practice, ν^1 is unknown and needs to be estimated from W^* and w^*_{T+1}, which themselves need to be estimated with unknown representation B^* at the same time. In this section, we design a practical L1-A-MTRL algorithm shown in Algorithm 1 which estimates the model parameters $\hat{B}, \hat{W}, \hat{w}_{T+1}$ and relevance vector $\hat{\nu}^1$. Here in our algorithm setting, we let

$$\begin{aligned} \beta_1 &= 10^5 Tk^3 \cdot \frac{C_W^6}{\underline{\sigma}^6}(d + \ln(\frac{4T}{\delta})) \\ \beta_2 &= k(d + T + \ln(\frac{1}{\delta}))\|\hat{\nu}^1\|_1^2\varepsilon^{-2} + \beta_1 \end{aligned} \tag{15}$$

where C_W is defined in Assumption 3.8. β_1 and β_2 characterize the sample complexity required to explore at the first and second stage, respectively, and they are determined by $T\beta$ and N_{tot} defined in Theorem 3.10.

We want to highlight that unlike the L_2-minimization approach of (Chen et al., 2022), our L_1-minimization solution does not have a closed form solution which creates more challenges in controlling the estimation error between $\hat{\nu}^1$ and ν^1. To deal with this problem, we use the *Lasso Program* (Wainwright, 2019; Tibshirani, 1996) to estimate $\hat{\nu}^1$:

$$\hat{\nu}^1 \in \arg\min_{\nu \in \mathbb{R}^T} \{\frac{1}{2}\|\hat{w}_{T+1} - \hat{W}\nu\|_2^2 + \lambda_k\|\nu\|_1\} \tag{16}$$

where the regularization parameter λ_k is chosen by users. We prove that with proper λ_k, $\hat{\nu}^1$ will be sufficiently close to ν^1 in l_1 norm when the following assumptions hold.

Assumption 3.8. (bounded norm) There exists $C_W, R > 0$ s.t. $\sigma_{\max}(W^*) \leq C_W$ and $\|w^*_{T+1}\|_2 = \Theta(R)$.

Assumption 3.9. (identical covariance) we have: $\Sigma_t = \Sigma^* = I_d$ for all $t \in [T+1]$.

Assumption 3.8 implies $\forall t \in [T]$, $\|w_t^*\|_2 = \|W^*e_n\|_2 \leq C_W$, which is a very common condition in the previous work (Du et al., 2021; Tripuraneni et al., 2021; Chen et al., 2022). Assumption 3.9 is a stronger variance condition than Assumption 2.2, but it's also used in (Tripuraneni et al., 2021; Chen et al., 2022) and we only need it in this section. With these assumptions we are prepared to state our theoretical guarantee for our practical L1-A-MTRL algorithm:

Theorem 3.10. *Let Assumption 2.1, 2.3, 2.4, 3.8, 3.9 hold. Let $\gamma = \max\{2160k^{3/2}C_W^2/\underline{\sigma}, \sqrt{2160k^{3/2}C_W^3/\underline{\sigma}}\}$, where $\underline{\sigma} = \sigma_{\min}(W^*) > 0$. For L1-A-MTRL method, we set the regularization parameter by:*

$$\lambda_k = 45\frac{\sqrt{k}RC_W\underline{\sigma}}{\gamma}\max\{1, \frac{C_W}{\gamma}\} \tag{17}$$

Then to let $ER_{L_1} \leq \varepsilon^2$ where $\varepsilon^2 \ll \min(1, \sigma^2(kd + kT)\frac{\|\nu\|_1^2}{T\underline{N}})$ with probability $1-\delta$, the number of source samples N_{tot} is at most

$$\widetilde{\mathcal{O}}(\sigma^2(kd + kT)\|\nu^1\|_1^2\varepsilon^{-2} + T\beta) \tag{18}$$

[1] For the previous upper bound in Theorem 3.1, people estimate non-shared w^*_{T+1} by linear-probing on the target task so (8) contains target-related error term. However, under the "cheating" case in Theorem 3.7, knowing ν^p means we already have such information as long as n_t is large enough since $W^*\nu^p = w^*_{T+1}$. We want to emphasize that this known ν^p assumption is used for illustrating why L_1 strategy is better, but not for practical use.

Algorithm 1 L1-A-MTRL Method

Input: confidence δ, representation function class Φ, ER bound $\varepsilon \ll 1$, minimum singular value $\underline{\sigma}$
Initialize $\underline{N} = \beta_1/T$ with (15) and λ_k with (17),
Phase 1: Exploration ν
Draw $\underline{N}$ i.i.d samples from every source task datasets
Estimate $\hat{\phi}^1, \hat{W}^1$ and $\hat{w}^1_{T+1}$ with Eqn.(2), (3)
Estimate $\hat{\nu}^1$ by **Lasso Program** (16)
Set β_2 with Eqn. (15)
Phase 2: Sampling
Set $n_t^2 = \max\{\beta_2 |\hat{\nu}^1(t)| \cdot \|\hat{\nu}^1\|_1^{-1}, \underline{N}\}$.
Draw n_t i.i.d samples from t-th source task datasets
Estimate $\hat{\phi}^2, \hat{W}^2$ and $\hat{w}^2_{T+1}$ with Eqn.(2), (3)

where $\beta = \max\{\gamma^2 \frac{\sigma_z^2}{\underline{\sigma}^4}, \gamma^2 \frac{C_W^2}{\underline{\sigma}^4}\rho^4, \rho^4, \frac{\sigma_z^2}{\underline{\sigma}^2}\} \cdot (d + \ln(\frac{4T}{\delta}))$, *and target task sample complexity* n_{T+1} *is at most*

$$\widetilde{\mathcal{O}}(\sigma^2 k \varepsilon^{-2} + \alpha) \tag{19}$$

where $\alpha = \max\{\gamma^2 \frac{\sigma_z^2 C_W^2}{\underline{\sigma}^4 R^2}, \gamma^2 \frac{C_W^2}{\underline{\sigma}^4}\rho^4, \rho^4\} \cdot (k + \ln(\frac{4}{\delta}))$.

Discussion. Comparing to the known ν case in Corollary 3.3, in this unknown ν setting we find our algorithm only requires an additional ε-independent number of samples $T\beta$ for the sampling complexity from source tasks and α for that from target task to achieve the same performance. (Chen et al., 2022) have similar results, but their additional term β in their Theorem 4.1 has an order of ε^{-1}. Technically, (Chen et al., 2022) directly uses the closed form of least square solution and proves that $|\hat{\nu}^2(t)| = \Theta(|\nu^2(t)|),\ \forall t \in [T]$ if $n_t \geq c'' \cdot \varepsilon^{-1}$. However, for Lasso-based L1-A-MTRL method, we can choose some proper parameter λ_k which can upper bound not only the noise term but also the l_1-error between Lasso solution and true vector as $\|\hat{\nu}^1 - \nu^1\|_1 = \Theta(\|\nu^1\|_1)$ if $n_t \geq c' \cdot \varepsilon^0$ (Lemma E.3). Here $c', c'' > 0$ are model-related constants.

Moreover, we remark that we have a similar limitation as (Chen et al., 2022) that we require some prior knowledge of $\underline{\sigma}$. However, since they only hit the additional constant terms, they are unlikely to dominate either of the sampling complexities for reasonable values of d, k, T and $\varepsilon \ll 1$.

Lastly, it is worth mentioning that similar results to Theorem 3.10 also apply when our L1-A-MTRL algorithm incorporates multiple sampling stages, as presented in Algorithm 2 in Appendix. The reason is that we only need to ensure that the minimum sampling budget is larger than $\underline{N}$ which is independent of the stage, and the additional proof follows a similar approach to that of Theorem E.4 in (Chen et al., 2022).

4. Extentsion: Cost-sensitive Task Selection

In Section 3, we proved that the L_1 strategy can minimize the total number of samples from the source tasks. Implicitly, this assumes the cost of each task is equal, and the first sample costs the same as the n-th sample. In contrast, we could also consider a non-linear cost function for the t-th source task $f_t : \mathbb{N} \to \mathbb{R}$, which takes in the number of random label query n and outputs the total required cost. For example, this could encode the notion that a long-term data subscription from one single source may result in decreasing the average cost over time.

Here we show that, even in this task-cost-sensitive setting, our L_1-A-MTRL method Algorithm 1 can still be useful under many benign cost functions. Consider the following example.

Example: Saltus Cost Function. Assume N_{tot} and $\underline{N}$ are fixed. If $n_{t,1} = \underline{N}$, $f :\equiv f_t$ for all $t \in [T]$ and f is composed by fixed cost and linear variable cost:

$$f(n) = \begin{cases} C_{fix} + C_{var}(n - \underline{N}) & , n > \underline{N} \\ 0 & , n \leq \underline{N} \end{cases} \tag{20}$$

where for each source task t we have $\underline{N}$ free data for sampling. As a reference, one practical instance for this case is programmatic weak supervision, where setting up a source requires some high cost but afterward, the query cost remains low and linear (Zhang et al., 2022a). If we want to find some proper ν to minimize the total cost $\sum_{t=1}^T f_t(n_t)$, then it's equivalent to finding the L_0 minimization solution of $\hat{W}\nu = \hat{w}_{T+1}$, where $\hat{W}, \hat{w}_{T+1}$ is estimated by free data. Of course, L_0 minimization is known to be intractable, so with proper λ_f, the L_1-A-MTRL method can be a good approximation.

Now, we are ready to give a formal definition of our goal and the characterization of when our L_1-A-MTRL method can be useful. Based on the excess risk upper bound in Theorem 3.1, to get $ER \leq \epsilon^2$, we are aimed to solve the following optimization problem.

$$\begin{aligned} \min_{n_{[T],2}} \ & \sum_{t=1}^T f_t(n_{t,1} + n_{t,2}) \\ s.t.\ & \sigma^2 k(d+T) \sum_{t=1}^T \frac{\nu(t)^2}{n_{t,1} + n_{t,2}} \lessapprox \varepsilon^2 \\ & W^* \nu = w^*_{T+1} \\ & n_{t,2} \geq 0, \quad t \in [T] \end{aligned} \tag{21}$$

Then we have the following guarantees as long as f_t satisfies the properties shown there.

Theorem 4.1 (informal)**.** *Assume* f_t *is a piecewise second-order differentiable function, and on each sub-function, it satisfies* $f_t \geq 0, \nabla f_t \geq 0, \nabla^2 f_t \leq 0$ *and* $\nabla f_t(n_{t,1} +$

$n_{t,2}) = \Omega(n_{t,2}^{-2+q})$ *for some* $q \in (0, 2]$. *Denotes the optimal solution of (21) as* $(n^*_{[T],2}, \nu^*)$. *Then under a similar data generation assumption as before, we have*

$$n^*_{t,2} = h_t(|\nu^*(t)|) \tag{22}$$

where h_t *is a monotone increasing function that satisfies:* $c_{t,1}x \leq h_t(x) \leq c_{t,2}x^{2/q}$ *where* $c_{t,1}, c_{t,2} > 0$. *Moreover, we claim A-MTRL algorithm with* $n^*_{[T],[2]}$ *sampling strategy is* k*-sparse, i.e.,* $\|n^*_{[T],2}\|_0 \leq k$.

Discussion. If $\nabla f_t(n^*_{t,2}) \equiv c > 0$, (112) is equivalent to L_1 strategy mentioned in the previous sections. However, for many other cases, it might be NP-hard to optimize (21), such as the Saltus Cost Function example shown above. Therefore, our previous algorithm L1-A-MTRL can be widely applied to these task-cost-sensitive scenarios to approximate the optimal strategy.

5. Experiments

Although our theoretical analysis only holds for a linear representation, our experiments also show the effectiveness of our algorithm on neural network representations as well in the task selection case. In this section, we follow the experimental settings in (Chen et al., 2022) and empirically evaluate L1-A-MTRL on the corrupted MNIST (MNIST-C) dataset proposed in (Mu & Gilmer, 2019). We reflect the preponderance of our algorithm on the two scenarios mentioned above. The first one is cost-agnostic, which aims to minimize the total sampling number from the source tasks and can reach all the source tasks. Another scenario is task-cost-sensitive like Section 4 and we particularly concentrate on k task-selection algorithms which correspond to cost functions like saltus cost function, and the learner is only allowed to sample from only k tasks after the initial exploration stage. We call the first case **full task** scenario and the second one **k-task selection** scenario for convenience. Please refer to Appendix G.1 for further illustration of our intuition for the k-task selection scenario.

5.1. Experimental Setup

Datasets. The MNIST-C dataset is a comprehensive suite of 16 corruptions applied to the MNIST test set. Like in (Chen et al., 2022), we divide each corruption-related sub-dataset into 10 tasks according to their labels ranging from $0 \sim 9$ and thus get 160 separate new tasks denoted by "{corruption type}_{label}". For instance, *brightness_0* denotes the data corrupted by brightness noise and relabeled to $1/0$ based on whether the data corresponds to number 0 or not. And once we choose 1 task called "{type A}_{label B}" for the target task, the other 150 tasks that don't contain "type A" corruption will be chosen as source tasks.

Experimental Setups and Comparisons. Like in (Chen et al., 2022), we replace the cross-entropy loss, which is commonly used for MNIST, with the regression l_2 loss in order to align with the theoretical setting in this paper. As the model setting, for full tasks scenario, we use the linear representation as defined in our theorem. We set $d = 28*28$, $k = 50$ and there are $T = 150$ source tasks in total. And we compare L1-A-MTRL and L2-A-MTRL(Chen et al., 2022) algorithms on the above datasets with 160 different target tasks. For the k-task selection scenario, we use a 2-layer ReLU CNN followed by a fully-connected linear layer as the representation map. Since neural networks can better capture the feature, here we set a smaller representation dimension $k = 10$ to show the advantage of the sparse task selection algorithm while other parameters follow the setting in the case of the full tasks. We compare L1-A-MTRL, which has been proved to be k-sparse from Theorem 3.7, together with vanilla k-sparse baseline that randomly selects $k = 10$ source tasks for sampling data at the second stage. Please refer to Appendix G.2 for details of algorithm implementation and Appendix G.3 for details on how to determine the value of λ_k.

5.2. Results

Full tasks scenario. In summary, L1-A-MTRL achieves 0.54% higher average accuracy among all the target tasks than L2-A-MTRL and results same or better performance in 126 out of 160 tasks. Due to the imbalanced dataset, 10% is the error rate of the baseline which randomly guesses the label, and the average prediction incorrect rate for L2-A-MTRL is 7.4%.

k-Task selection scenario. Similarly, L1-A-MTRL achieves 2.2% higher average accuracy among all the target tasks than the vanilla baseline which has the average prediction error rate of 5.7%. And our algorithm results in the same or better performance in 149 out of 160 tasks. This shows the effectiveness of our method on neural network representation.

In Section G.4 of the appendix, we provide additional comparisons of the empirical sampling budgets for different algorithms. The results demonstrate that L1-A-MTRL requires fewer samples compared to L2-A-MTRL and P-MTRL while achieving comparable performance. These findings further underscore the effectiveness of our L1-A-MTRL algorithm.

6. Conclusion

We introduced a novel active sampling strategy L1-A-MTRL to sparse sample from target-related source tasks and learn a good representation that helps the target task. From a theoretical perspective, we first showed that L1-A-MTRL is strictly better than the previous L2-A-MTRL by proving a

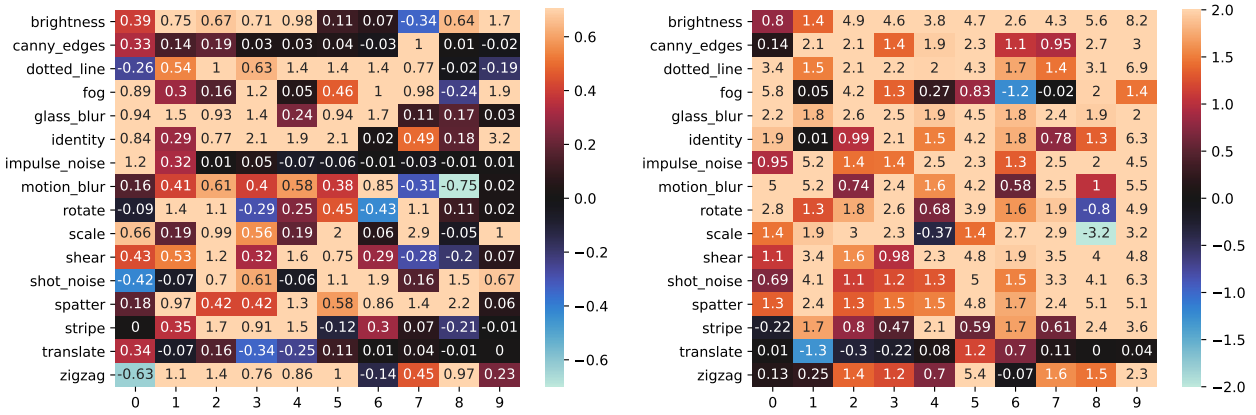

Figure 1. **Performance Comparison.** These pictures show the prediction difference (in %) between our method and baseline for all target tasks, the larger the better. The y-axis denotes the corruption type while the x-axis denotes to the binarized label, and each grid on (x, y) corresponds to the case that the target task is "$\{y\}$_$\{x\}$". **Left: full tasks scenarios.** Compare L1-A-MTRL and L2-A-MTRL using linear representation. **Right: k-task selection scenarios.** Compare two k-sparse task selection algorithms L1-A-MTRL and passive-learning baseline, which randomly selects k source tasks for the second-stage sampling, using Convnet representation.

novel sampling-strategy-dependent lower bound and then provided a tighter upper bound correspondingly. From the empirical perspective, we showed our algorithm is not only effective under the standard setting but can achieve even better results in the practical scenario where the number of source tasks is restricted.

Acknowledgements

SSD acknowledges the support of NSF IIS 2110170, NSF DMS 2134106, NSF CCF 2212261, NSF IIS 2143493, NSF CCF 2019844, NSF IIS 2229881.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Asai, A., Salehi, M., Peters, M. E., and Hajishirzi, H. Attentional mixtures of soft prompt tuning for parameter-efficient multi-task knowledge sharing. *arXiv preprint arXiv:2205.11961*, 2022.

Bai, Z., Shen, Y., Shui, N., and Guo, X. Introduction to riemann geometry, 1992.

Bertsimas, D. and Tsitsiklis, J. N. Introduction to linear optimization. 1997.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, S., Crammer, K., He, H., Roth, D., and Su, W. J. Weighted training for cross-task learning. *arXiv preprint arXiv:2105.14095*, 2021.

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

Chua, K., Lei, Q., and Lee, J. D. How fine-tuning allows for effective meta-learning. *Advances in Neural Information Processing Systems*, 34:8871–8884, 2021.

Cohen, M. B., Lee, Y. T., and Song, Z. Solving linear programs in the current matrix multiplication time. *Journal of the ACM (JACM)*, 68(1):1–39, 2021.

Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pp. 2089–2099. PMLR, 2021.

Du, S. S., Hu, W., Kakade, S. M., Lee, J., and Lei, Q. Few-shot learning via learning the representation, provably. *ArXiv*, abs/2002.09434, 2021.

Fifty, C., Amid, E., Zhao, Z., Yu, T., Anil, R., and Finn, C. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021.

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

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28, 10 2000. doi: 10.1214/aos/1015957395.

Mu, N. and Gilmer, J. Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337*, 2019.

Pajor, A. Metric entropy of the grassmann manifold. 1998.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Shachaf, G., Brutzkus, A., and Globerson, A. A theoretical analysis of fine-tuning with linear teachers. *Advances in Neural Information Processing Systems*, 34:15382–15394, 2021.

Shi, G., Azizzadenesheli, K., O' Connell, M., Chung, S.-J., and Yue, Y. Meta-adaptive nonlinear control: Theory and algorithms. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10013–10025. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/52fc2aee802efbad698503d28ebd3a1f-Paper.pdf.

Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., and Savarese, S. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pp. 9120–9132. PMLR, 2020.

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

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the royal statistical society series b-methodological*, 58:267–288, 1996.

Tripuraneni, N., Jordan, M., and Jin, C. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33: 7852–7862, 2020.

Tripuraneni, N., Jin, C., and Jordan, M. I. Provable meta-learning of linear representations. In *ICML*, 2021.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Xu, Z. and Tewari, A. Representation learning beyond linear prediction functions. *Advances in Neural Information Processing Systems*, 34:4792–4804, 2021.

Yao, X., Zheng, Y., Yang, X., and Yang, Z. Nlp from scratch without large-scale pretraining: A simple and efficient framework. In *International Conference on Machine Learning*, pp. 25438–25451. PMLR, 2022.

Zaiem, S., Parcollet, T., Essid, S., and Heba, A. Pretext tasks selection for multitask self-supervised speech representation learning. *arXiv preprint arXiv:2107.00594*, 2021.

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., and Savarese, S. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

Zhang, J., Hsieh, C.-Y., Yu, Y., Zhang, C., and Ratner, A. A survey on programmatic weak supervision. *arXiv preprint arXiv:2202.05433*, 2022a.

Zhang, Z., Wang, S., Xu, Y., Fang, Y., Yu, W., Liu, Y., Zhao, H., Zhu, C., and Zeng, M. Task compass: Scaling multi-task pre-training with task prefix. *arXiv preprint arXiv:2210.06277*, 2022b.

A. Related Work

Empirical works on P-MTRL and A-MTRL. Multi-task representation learning has been widely applied and achieved great success in the natural language domain GPT-2 (Radford et al., 2019), GPT-3(Brown et al., 2020), vision domain CLIP (Radford et al., 2019) and multi-model Flamingo (Alayrac et al.). Nevertheless, such large models are costly in both data collecting/cleaning and training. Recently, many works focus on efficiently selecting the source task. In the natural language domain, for example, (Yao et al., 2022) use a heuristic retriever method to select a subset of target-related NLP source tasks; More recently, works like (Asai et al., 2022; Zhang et al., 2022b) use prefix/prompt to capture the relation between source and target tasks. Similar topics have also been studied in the vision domain, for example, (Zamir et al., 2018) propose a transfer learning algorithm based on learning the underlying structure among visual tasks, which they called Taskonomy, and there are many following works propose different approaches on this Taxonomy dataset, including (Fifty et al., 2021; Standley et al., 2020).

Theoretical works on P-MTRL. There are also many existing works on provable P-MTRL. Tripuraneni et al. (2020; 2021); Du et al. (2021); Thekumparampil et al. (2021); Collins et al. (2021); Xu & Tewari (2021) assume there exists a ground truth shared representation across all tasks. In particular, Tripuraneni et al. (2020; 2021); Thekumparampil et al. (2021) assume a low-dimension linear representation like us while Du et al. (2021) generalize to both high-dimensional representation and 2-layer Relu network. Tripuraneni et al. (2020) also further considers any general representation with linear predictors. Both works obtain similar results. Besides, many recent works focus on fine-tuning in theoretical contexts (Shachaf et al., 2021; Chua et al., 2021; Chen et al., 2021; Kumar et al., 2022).

For the lower bound, for the first time, Tripuraneni et al. (2021) proves a minimax lower bound for the estimation error of the estimated representation layer measured by subspace angle distance. But we claim it can't directly deduce a similar lower bound of the test error on the target task, which relates to one of our main contributions. The reason is that though the estimated representation may be far away from the ground truth one, the learner can estimate a proper target predictor to achieve a sufficiently small test error as long as $B^* w_{T+1}^*$ (almost) lies in the column space of $\hat{B}$, where the notations are defined in the preliminary.

Theoretical works on A-MTRL. In order to overcome the problems in P-MTRL, some subsequent works focused on giving different priorities to the source tasks by methods like active learning (Chen et al., 2022) and weighted training (Chen et al., 2021). Representatively, Chen et al. (2022) is the first work to propose A-MTRL which calculates the proper sampling number for each source task. It iteratively estimates the relevance of each source task to the target task by estimating the relevance vector ν^*. Chen et al. (2022) utilizes the L_2 strategy defined in Def. 2.5 to decide the sampling strategy and significantly outperforms passive MTRL (P-MTRL), which uniformly samples from the source tasks, both theoretically and empirically. Nevertheless, the optimal sample strategy for A-MTRL is underexplored, and the non-sparsity of ν^2 may cause inconvenience for task-cost-sensitive scenarios. We develop our works based on the problem setting in (Chen et al., 2022) and propose a more efficient sampling strategy. As another approach, Chen et al. (2021) concentrates on learning a weighting over the tasks. The crucial difference between their work with ours is that they can attach to the whole dataset whereas we assume not but actively query new data from some large datasets (e.g., the task represented by the search terms to Wikipedia or Google). They also assume that some tasks may not only be irrelevant but even harmful and need to be down-weighted.

B. Technical Notations

We summarize the technical notations used in the appendix as follows.

Grassmann Manifold. Assume $d \geq k$, we denote by $Gr_{d,k}$ the Grassmann manifold which contains all the subspaces that are spanned by k linearly independent d-dimensional vectors. For $d \geq k$, we let $O_{d,k}$ be the set of matrices whose column contains k orthonormal vectors that are in $\mathbb{R}^d$. Then any $B \in O_{d,k}$ corresponds to an element, which is spanned by the column vectors of B, of $Gr_{d,k}$. Actually, an element in $Gr_{d,k}$ is corresponds to an equivalent class of $d \times k$ matrices that satisfies the equivalent relation $\sim$:

$$Y \sim X \Leftrightarrow Y = XA,\ \forall A \in GL(k, \mathbb{R}) \tag{23}$$

where $GL(k, \mathbb{R})$ denotes general linear group over $\mathbb{R}$ of degree k.

Subspace Distance. Finally, we use the same definition as (Tripuraneni et al., 2021) and (Pajor, 1998) to define the distance between the subspaces in the Grassmann manifold. We let $s_p(T) = (\sum_{i \geq 1} |\sigma_i(T)|^p)^{1/p}$ for any matrix T and any

$p \in [1, \infty]$. In particular, s_∞ is the operator norm of T. For $E, F \in O_{d,k}$, from Proposition 6 of (Pajor, 1998) we define $s_q(E, F) = (2\sum_{i=1}^{k} |1 - \sigma_i^2(E^T F)|^{q/2})^{1/q}$ to be the subspace distance between the spaces spanned by the column vectors of E and F, respectively. Particularly, $s_\infty(E, F) = \sqrt{1 - \sigma_k^2(E^T F)}$.

C. Proof of Theorem 3.4

Proof of Lemma 3.2. We can use the following equivalent optimization problem to prove our Lemma:

$$\begin{aligned} \min_{n_{[T]}} \quad & G(n_{[T]}) := \sum_{t=1}^{T} \frac{|\nu^*(t)|^2}{n_t} \\ s.t. \quad & c_0(n_{[T]}) := N_{tot} - \sum_{t=1}^{T} n_t = 0 \\ & c_t(n_t) := n_t - \underline{N} > 0, \quad \forall t \in [T] \end{aligned} \tag{24}$$

The corresponding Lagrangian function for (24) is

$$L(n_{[T]}) := G(n_{[T]}) - \lambda_0 c_0(n_{[T]}) - \sum_{t=1}^{T} \lambda_t c_t(n_t) \tag{25}$$

Then from the Karush-Kuhn-Tucker condition, for all $t \in [T]$ we have the necessary condition

$$\begin{aligned} \frac{\partial L}{\partial n_t} &= -\frac{|\nu^*(t)|^2}{n_t^2} + \lambda_0 - \lambda_t = 0 \\ \lambda_t &\geq 0 \\ \lambda_t c_t(n_t) &= \lambda_t (n_t - \underline{N}) = 0 \end{aligned} \tag{26}$$

So we get $\lambda_0 > \lambda_t \geq 0, \ \forall t \in [T]$ and

$$n_t = \begin{cases} \lambda_0^{-0.5} |\nu^*(t)| & , \lambda_t = 0 \Rightarrow n_t \geq \underline{N}, \\ \underline{N} & , \lambda_t > 0 \Rightarrow n_t = \underline{N}. \end{cases} \tag{27}$$

thus we finish the proof.

As a supplement, we give another proof for the special case in this Lemma where we assume $n_t > \underline{N}$ for every $t \in [T]$. Let $\beta(t) := \frac{\nu^*(t)}{\|\nu^*\|_2}$, $\alpha_t = \frac{n_t}{N_{tot}}$ and thus $\sum_{t=1}^{T} \beta^2(t) = \sum_{t=1}^{T} \alpha_t = 1$. Therefore by Cauchy inequality,

$$\begin{aligned} \|\widetilde{\nu}^*\|_2^2 &= \frac{\|\nu^*\|_2^2}{N_{tot}} \sum_{t=1}^{T} \frac{\beta^2(t)}{\alpha_t} \\ &= \frac{\|\nu^*\|_2^2}{N_{tot}} (\sum_{t=1}^{T} \frac{\beta^2(t)}{\alpha_t})(\sum_{t=1}^{T} \alpha_t) \\ &\geq \frac{\|\nu^*\|_2^2}{N_{tot}} (\sum_{t=1}^{T} |\beta(t)|)^2 = \frac{\|\nu^*\|_1^2}{N_{tot}} \end{aligned} \tag{28}$$

The equality in (28) is achieved iff $\frac{|\beta(t)|}{\sqrt{\alpha_t}} = c\sqrt{\alpha_t}$ for evert $t \in [T]$ with $c > 0$, which means that n_t is proportional to $|\nu^*(t)|$. □

Proof of Corollary 3.3. As stated in Lemma 3.2, $n_t^* = \max\{c'|\nu(t)|, \underline{N}\}$ and $c' > 0$ is some constant such that $\sum_{t=1}^{T} n_t^* = N_{tot}$, so we have $c'|\nu(t)| \leq n_t^* \leq c'|\nu(t)| + \underline{N}$. Sum up both sides of the inequality for all $t \in [T]$, then:

$$c'\|\nu\|_1 \leq N_{tot} \leq c'\|\nu\|_1 + T\underline{N} \tag{29}$$

Therefore, if we assume $N_{tot} \gg T\underline{N}$, then we get $N_{tot} = (1+o(1))c'\|\nu\|_1$. In fact, we only need to ensure that $N_{tot} > 2T\underline{N}$, which results in $c' > N_{tot}/(2|\nu|_1)$, since the coefficient in the error bound (8) is unconsidered.

Let $S_1 = \{t \in [T] | n_t \geq \underline{N}, |\nu(t)| > 0\}$ and $S_2 = \{t \in [T] | n_t < \underline{N}, |\nu(t)| > 0\}$. Then for any fixed ν, from Lemma 3.2, we have the following inequality for the optimal strategy:

$$\|\widetilde{\nu}\|_2^2 = \sum_{t\in S_1} \frac{\nu(t)^2}{n_t} + \sum_{t\in S_2} \frac{\nu(t)^2}{\underline{N}} \leq \sum_{t\in S_1\cup S_2} \frac{\nu(t)^2}{c'|\nu(t)|} = (1+o(1)) \sum_{t\in S_1\cup S_2} \frac{|\nu(t)|}{N_{tot}} \|\nu\|_1 = (1+o(1)) \frac{\|\nu\|_1^2}{N_{tot}} \tag{30}$$

Here the inequality holds if and only if S_2 is empty, which means that for all $t \in [T]$, $\nu(t)$ must satisfies $\nu(t) = 0$ or $c'|\nu(t)| \geq \underline{N}$. Combining (30) and Theorem 3.1, we get the results. □

Proof of Theorem 3.4. From (30) we know that if $c'|\nu(t)| \geq \underline{N}$ for all $t \in [T]$ such that $|\nu(t)| > 0$, then $n_t^* = N_{tot}|\nu(t)|/\|\nu\|_1$, $\forall t \in \{t' \in [T] | |\nu(t')| > 0\}$, and $\|\widetilde{\nu}\|_2^2$ attain its minimum $\|\nu\|_1^2/N_{tot}$.

We prove that such a condition can be achieved when N_{tot} is sufficiently large. Assume that $c'|\nu(t)| < \underline{N}$ always holds for some $t \in [T]$ where $\nu(t) \neq 0$. Then if we choose $N_{tot} = T\underline{N} + \underline{N}/|\nu(t)|\|\nu\|_1$, we will have $c'|\nu(t)| = c'\|\nu\|_1 \cdot |\nu(t)|/\|\nu\|_1 \geq (N_{tot} - T\underline{N})|\nu(t)|/\|\nu\|_1 \geq \underline{N}$, where we use the fact that $\underline{N}T + c'\|\nu\|_1 \geq N_{tot}$ from Eqn. 29. This is contradicted by the assumption, and thus we can always find some N_{tot} such that $c'|\nu(t)| \geq \underline{N}$ if $\nu(t) \neq 0$.

So for any given ν, the optimal sampling strategy $n_t(\nu)$(Lemma 3.2) can let $\|\widetilde{\nu}\|_2^2$ achieves its minimum $\|\nu\|_1^2/N_{tot}$. Then we vary ν among the solution candidate set of $W^*\nu = w_{T+1}^*$ and find L_1-minimization solution ν^1 can minimize $\|\nu\|_1^2/N_{tot}$. Therefore, $(\nu^1, n_{[T]}^1)$ is optimal for the original problem (9). □

D. Proof of Theorem 3.7

D.1. Preparations for minimax lower bound

First, we reclaim some concentration inequalities commonly used in the previous work (Du et al., 2021; Chen et al., 2022).

Lemma D.1. *(A variant of Lemma A.6 in (Du et al., 2021)) Let $a_1, ..., a_n$ be i.i.d. d-dimensional random vectors such that $\mathbb{E}[a_i] = 0$, $\mathbb{E}[a_i a_i^\top] = I$, and a_i is ρ^2-subgaussian. For $\delta \in (0,1), \epsilon \in (0, \frac{1}{2})$, suppose $n > \frac{1}{\epsilon^2} c_a \rho^4 (d + \ln(\frac{1}{\delta}))$ for some universal constant c_a. Then with probability at least $1-\delta$ we have*

$$(1-2\epsilon) I_d \preceq \frac{1}{n} \sum_{i=1}^n a_i a_i^\top \preceq (1+2\epsilon) I_d \tag{31}$$

□

Recall that $\Sigma_t^* = \mathbb{E}_{x_t \sim p_t}[x_t x_t^\top]$ and $\hat{\Sigma}_t := \frac{1}{n_t}(X_t)^\top X_t$ for any $t \in [T+1]$, then we have:

Lemma D.2. *(A variant of Claim A.1, A.2 in (Du et al., 2021)) Suppose for $\delta \in (0,1)$. Let $n_t > \frac{1}{\epsilon^2} c_a \rho^4 (d + \ln(\frac{2T}{\delta}))$ for all $t \in [T]$, then with probability at least $1 - \frac{\delta}{2}$ over the inputs $X_1, ..., X_T$ in the source tasks, we have*

$$(1-2\epsilon)\Sigma_t \preceq \hat{\Sigma}_t \preceq (1+2\epsilon)\Sigma_t \tag{32}$$

Here $c_a > 0$ is a universal constant. Similarly, let $n_{T+1} > \frac{1}{\epsilon^2} c_a \rho^4 (k + \ln(\frac{2}{\delta}))$. Then for any given matrix $B_1, B_2 \in \mathbb{R}^{d\times k}$ that is independent of X_{T+1} , with probability $1 - \frac{\delta}{2}$ over X_{T+1} we have

$$(1-2\epsilon) B_1^\top \Sigma_{T+1} B_2 \preceq B_1^\top \hat{\Sigma}_{T+1} B_2 \preceq (1+2\epsilon) B_1^\top \Sigma_{T+1} B_2 \tag{33}$$

□

And then, we show that $\sum_{t=1}^T |X_t(\hat{B}\hat{w}_t - B^* w_t^*)|^2 \asymp \sigma^2(kT + k(d-k))$. The upper bound has been shown in Claim A.3 in (Du et al., 2021), and the lower bound will be shown in the following theorem.

Theorem D.3. *With conditions in Theorem 3.7, with probability $1-\delta$ we have:*

$$\inf_{(\hat{B},\hat{W})} \sup_{(B^*,W^*)} \sum_{t=1}^T |X_t(\hat{B}\hat{w}_t - B^* w_t^*)|^2 \gtrsim \sigma^2(kT + k(d-k)) \tag{34}$$

The key theorems and lemmas are as follows.

Theorem D.4. *Let $G_0 := \{BW|B \in O_{d,k}\ ; W \in \mathbb{R}^{k\times T}\}$, and $G_1(\delta_1) := \{BW|B \in O_{d,k}\ ; W \in \mathbb{R}^{k\times T}\ ; \|W\|_F \leq \delta_1, t \in [T]\}$ be a local packing of G_0, where w_t is the t-th column vector of W. Then there is a lower bound for G_1's packing number:*

$$\ln M(G_1(\delta_1), \|\cdot\|_F, \Delta_1) \gtrsim k(d-k) + kT \tag{35}$$

where Δ_1 will be determined soon.

Lemma D.5. *[Adapted from (Pajor, 1998)] For any $1 \leq k \leq d$ such that $k \leq d-k$, for every $\epsilon > 0$, we have*

$$(\frac{c_1}{\epsilon})^{k(d-k)} \leq N(Gr_{d,k}, s_\infty, \epsilon) \leq (\frac{c_2}{\epsilon})^{k(d-k)} \tag{36}$$

with universal constants $c_1, c_2 > 0$. From the relation between packing number and covering number (Wainwright, 2019), we have:

$$M(Gr_{d,k}, s_\infty, \epsilon) \geq (\frac{c_1}{\epsilon})^{k(d-k)} \tag{37}$$

□

Lemma D.6. *Let $B^1, B^2 \in O_{d,k}$, $w^1, w^2 \in \mathbb{R}^k$. With SVD we get $(B^1)^\top B^2 = PDQ^T$, where $P, Q \in O_{k,k}$, $D = diag(\sigma_1, ..., \sigma_k)$. Obviously $\sigma_i \in [0,1]$, and we define $v^1 = P^\top w^1$, $v^2 = Q^\top w^2$. If subscripts denotes the index of vectors, we have:*

$$|B^1w^1 - B^2w^2|^2 = \sum_{i=1}^{k}[2|v_i^1||v_i^2|f(v_i^1, v_i^2) + (|v_i^1| - |v_i^2|)^2] \tag{38}$$

where

$$f(v_i^1, v_i^2) = \begin{cases} 1-\sigma_i, & sign(v_i^1 \cdot v_i^2) = 1 \\ 1+\sigma_i, & sign(v_i^1 \cdot v_i^2) = -1 \end{cases} \tag{39}$$

And we can get the lower bound:

$$|B^1w^1 - B^2w^2|^2 \geq 2|v_k^1||v_k^2|(1-\sigma_k) + \sum_{i=1}^{k}(|v_i^1| - |v_i^2|)^2 \geq 0 \tag{40}$$

Proof of Lemma D.6. By the calculation we get this result:

$$\begin{aligned} |B^1w^1 - B^2w^2|^2 &= (B^1w^1 - B^2w^2)^\top(B^1w^1 - B^2w^2) \\ &= |w^1|^2 + |w^2|^2 - 2(w^1)^\top(B_1)^\top B_2 w^2 \\ &= |v^1|^2 + |v^2|^2 - 2(v^1)^\top D v^2 \\ &= \sum_{i=1}^{k}((v_i^1)^2 + (v_i^2)^2 - 2v_i^1v_i^2\sigma_i) \end{aligned} \tag{41}$$

To make each term of the equation above non-negative, we use sign function:

$$\begin{aligned} |B^1w^1 - B^2w^2|^2 &= \sum_{i=1}^{k}[(v_i^1)^2 + (v_i^2)^2 - 2sign(v_i^1v_i^2) \times v_i^1v_i^2 + 2v_i^1v_i^2(sign(v_i^1v_i^2) - \sigma_i)] \\ &= \sum_{i=1}^{k}[(v_i^1)^2 + (v_i^2)^2 - 2|v_i^1||v_i^2| + 2|v_i^1||v_i^2|(1 - sign(v_i^1v_i^2)\sigma_i)] \\ &= \sum_{i=1}^{k}[(|v_i^1| - |v_i^2|)^2 + 2|v_i^1||v_i^2|f(v_i^1, v_i^2)] \end{aligned} \tag{42}$$

□

Besides, we begin to construct a separate set for G_1. Firstly we let $G_B = \{B^1, ..., B^{M_B}\}$ be a ϵ_B-separated set for metric s_∞ in $Gr_{d,k}$, where $\epsilon_B \leq \min(\frac{c_1}{2}, 1)$ as c_1 in Lemma D.5.

Then denote $(B^m)^\top B^n = P(m,n)D(m,n)Q(m,n)$, where $P(m,n), Q(m,n) \in O_{k,k}$, $D(m,n) = diag(\sigma_1(m,n), ..., \sigma_k(m,n))$, and $P(m,n) = Q(m,n) = D(m,n) = I_k$ iff $m = n$. On the other hand, for $t \in [T]$, we denote $v_{t,i}^j(P(m,n))$ to be the i-th component of $v_t^j(P(m,n)) := P(m,n)^\top w_t^j$, and similarly for $v_t^j(Q(m,n)) := Q(m,n)^\top w_t^j$.

Lemma D.7. *Suppose $G_V = \{V^j = (v_1^j, ..., v_T^j) | j \in S, v_t^j \in \mathbb{R}^k, v_t^j$ satisfy Equ. 43 and attain largest $|S|\}$:*

$$\begin{aligned} |v_{t,k}^j| &\geq \frac{\delta_V}{\sqrt{T}\epsilon_B}, && \forall j,\ \forall t \in [T] \\ \|V^j\|_F &= \sum_{t=1}^T |v_t^j|^2 \leq \frac{C_V \delta_V}{\epsilon_B}, && \forall j \\ \|V^i - V^j\|_F &= \sum_{t=1}^T |v_t^i - v_t^j|^2 \geq \frac{\delta_V}{\epsilon_B}, && \forall i, j \end{aligned} \tag{43}$$

where C_V is a universal constant and $4 < C_V < 5$. For $m, n \in [M_B]$, let $G_W(P(m,n)) := \{W^j = (w_1^j, ..., w_T^j) \mid \exists\, V^j \in G_V, s.t.\ W^j = P(m,n)V^j\}$ and similarly for $G_W(Q(m,n))$. Then let $G_{BW} = \{(B, W) | \ \exists\ m, n \in [M_B],\ W^m \in G_W(P(m,n)),\ W^n \in G_W(Q(m,n)), s.t.\ BW \in \{B^m W^m,\ B^n W^n\}\}$, and we claim that G_{BW} is a δ_V-separated subset of G_1 with Frobenius norm.

Proof of Lemma D.7. For each $t \in [T]$, we divide into 2 cases:

Case 1. For the case $m \neq n$, we will work out the lower bound of Equ. 40. Since for any $m \neq n$:

$$\begin{aligned} &s_\infty(m,n) = \sqrt{1 - \sigma_k^2((B^m)^\top B^n)} \geq \epsilon_B^2 \\ \Rightarrow &1 - \sigma_k((B^m)^\top B^n) \geq \frac{\epsilon_B^2}{1 + \sigma_k((B^m)^\top B^n)} > \frac{\epsilon_B^2}{2} \end{aligned} \tag{44}$$

combined with the first inequality of Equ. 43, we know by the definition of G_{Bw}, there exist some i, j such that:

$$\sum_{t=1}^T |B^m w_t^m - B^n w_t^n|^2 \geq 2 \sum_{t=1}^T |v_{t,k}^i||v_{t,k}^j|(1 - \sigma_k) \geq \delta_V^2 \tag{45}$$

Case 2. For the case $m = n$, note that $\sigma_i = 1$ for all $i \in [k]$. Combined Equ. 38 , Equ. 39, Equ. 43 and condition $\epsilon_B < \min(\frac{c_1}{2}, 1)$, there exist some i, j such that:

$$\sum_{t=1}^T |B^m w_t^m - B^m w_t^m|^2 = \sum_{t=1}^T \sum_{l=1}^k (v_{t,l}^i - v_{t,l}^j)^2 = \sum_{t=1}^T |v_t^i - v_t^j|^2 \geq \frac{\delta_V^2}{\epsilon_B^2} \geq \delta_V^2 \tag{46}$$

Combined them together, we see that for any $m, n \in [M_B]$, any $W^m \in G_W(P(m,n)),\ W^n \in G_W(Q(m,n))$ such that $B^m = B^n, W^m = W^n$ not hold in the meantime, we have:

$$\|B^m W^m - B^n W^n\|_F = \sum_{t=1}^T |B^m w_t^m - B^m w_t^m|^2 \geq \delta_V \tag{47}$$

□

Proof of Theorem D.4.

From the construction in Lemma D.7, we consider flattening V^j into a $k \times T$ vector $\eta^j \in \mathbb{R}^{kT}$, where $V^j \in G_V = \{V^j = (v_1^j, ..., v_T^j) | j \in S, v_t^j \in \mathbb{R}^k, v_t^j$ satisfy Equ. 43 and attain largest $|S|\}$. Then the last two conditions in (43) show that η^j is a $\frac{\delta_V}{\epsilon_B}$-separated set contained in a ball of radius $\frac{C_V \delta_V}{\epsilon_B}$ in l_2-norm. Actually, the first condition means removing the small central part along very axis of η^j in the above ball, and it's clear to see that G_V has the same order of the cardinality if we drop the first inequality of (43). So if we use $card$ to denote the cardinality of a set, we get:

$$\ln(card(G_V)) \gtrsim kT \tag{48}$$

Then from the definition of G_W and G_{BW} in Lemma D.7, we see that:

$$\begin{aligned}\ln(card(G_{BW})) &= \ln((\frac{M_B(M_B-1)}{2} \times 2 + M_B) \cdot \ln(card(G_W))) \\ &= 2\ln(M_B) + \ln(card(G_V)) \\ &\gtrsim k(d-k)\ln(c_1/\epsilon_B) + kT, \qquad (48) \\ &\gtrsim k(d-k) + kT, \qquad (\epsilon_B < \min(\frac{c_1}{2}, 1))\end{aligned} \tag{49}$$

Choose $\Delta_1 = \delta_1 \epsilon_B / C_V$ and we finish the proof. □

Proof of Theorem D.3.

Note that $\underline{\lambda} = \sigma_{\min}(\Sigma_t^{1/2})$, $\overline{\lambda} = \sigma_{\max}(\Sigma_t^{1/2})$ and $\kappa = \overline{\lambda}/\underline{\lambda}$, we can construct the local packing following Lemma D.7 by using $\widetilde{W}$ to replace W where $\widetilde{w}_t = \sqrt{n_t} w_t$. And we choose $\delta_1' = 0.9\delta_1$ where $\delta_1 = \frac{\delta_V}{\epsilon_B}$. Then we have:

$$\begin{aligned}\sqrt{\sum_{t=1}^T \|X_t(B^i w_t^i - B^j w_t^j)\|_2^2} &\le 1.1\overline{\lambda}\|B^i\widetilde{W}^i - B^j\widetilde{W}^j\|_F \\ &\le 1.1\overline{\lambda} \cdot C_V\delta_1 \cdot \frac{\delta_1'}{0.9\delta_1} \\ &< 6\overline{\lambda}\delta_1'\end{aligned} \tag{50}$$

$$\sqrt{\sum_{t=1}^T \|X_t(B^i w_t^i - B^j w_t^j)\|_2^2} \ge 0.9\underline{\lambda}\|B^i\widetilde{W}^i - B^j\widetilde{W}^j\|_F \ge \delta_1'\underline{\lambda} \tag{51}$$

Here for convenience we choose $C_V = 4.5$, and this will just influence the universal constant since C_V is $\Theta(1)$ as in Lemma D.7. Note the sum of excess risks on the source tasks in (50), (51) is actually a semi-metric between (B^i, W^i) and (B^j, W^j), and it's easy to construct the corresponding $\delta_1'\underline{\lambda}$-separated set G_{BW} from $G_{B\widetilde{W}}$ set obtained in Lemma D.7. We recall that $Y_t = X_t B^* w_t^* + Z_t$, and define $Y_t \sim \mathbb{P}_t^j$ where $\mathbb{P}_t^j = \mathcal{N}(X_t B^* w_t^*, \sigma^2 \mathbb{I}_{n_t})$. And we further let $\mathbb{P}^j := \prod_{t=1}^T \mathbb{P}_t^j$. Then by the independency among every tasks, we have the Kullaback-Leibler divergence:

$$\begin{aligned}D(\mathbb{P}^i \,\|\, \mathbb{P}^j) &= \sum_{t=1}^T D(\mathbb{P}_t^i \,\|\, \mathbb{P}_t^j) \\ &= \frac{1}{2\sigma^2}\sum_{t=1}^T \|X_t(B^i w_t^i - B^j w_t^j)\|_2^2 \\ &\le \frac{18\overline{\lambda}^2(\delta_1')^2}{\sigma^2} \qquad (50)\end{aligned} \tag{52}$$

Note that G_{BW} is a $\delta_1'\underline{\lambda}$-separated set over G_1, which is a local packing of G_0, we then let $M = M(G_0, \|\cdot\|_F, (\delta_1')^2)$ and have the following Fano's lower bound (Wainwright, 2019):

$$\begin{aligned}\inf_{(\hat{B},\hat{W})}\sup_{(B^*,W^*)}\sum_{t=1}^T \|X_t(\hat{B}\hat{w}_t - B^* w_t^*)\|_2^2 &\ge (0.9\underline{\lambda})^2 \inf_{(\hat{B},\hat{W})}\sup_{(B^*,W^*)}\sum_{t=1}^T \|\hat{B}\hat{W} - B^*W^*\|_F^2 \\ &\ge \frac{(\delta_1')^2}{4}\{1 - \frac{\frac{1}{M^2}\sum_{i,j=1}^M D(\mathbb{P}^i \,\|\, \mathbb{P}^j) + \ln 2}{\ln M}\} \\ &=: \frac{(\delta_1')^2}{4} \cdot C_{Fano}\end{aligned} \tag{53}$$

Besides, let $c_2 \geq 1$ be the universal constant in Theorem D.4. Note $d, T > k \geq 1$ and thus $\frac{c_2(k(d-k)+kT)}{3} > \frac{2}{3} > \ln 2$, we let $(\delta_1')^2 = \frac{c_2\sigma^2(k(d-k)+kT)}{108\overline{\lambda}^2}$, which enable $C_{Fano} \geq \frac{1}{2}$. Then finally we have:

$$\inf_{(\hat{B},\hat{W})} \sup_{(B^*,W^*)} \sum_{t=1}^{T} \|X_t(\hat{B}\hat{w}_t - B^* w_t^*)\|_2^2 \gtrsim \frac{\sigma^2(k(d-k)+kT)}{\kappa^2} \tag{54}$$

Then from Assumption 2.2 and our notation above, we have $\kappa^2 = \overline{\lambda}/\underline{\lambda} = \Theta(1)$, so we finish the proof.

□

D.2. Main Proof for the ER bound of P/A-MTRL

Lemma D.8. *Denote that for any $p \in \mathbb{N}^+$:*

$$\nu^p(w_{T+1}^*) = \arg\min_{\nu} \|\nu\|_p \qquad s.t.\ W^*\nu = w_{T+1}^* \tag{55}$$

and let

$$H(c_w) = \{w \in \mathbb{R}^k | \|w\|_2 = c_w\} \tag{56}$$

with constant $c_w > 0$, then for any fixed W^*, we have

$$\begin{aligned} \sup_{w_{T+1}^* \in H(c_w)} \|\nu^p(w_{T+1}^*)\|_2 &= \frac{c_w}{\sigma_{min}(W^*)} \\ \sup_{w_{T+1}^* \in H(c_w)} \|\nu^1(w_{T+1}^*)\|_1 &\leq \sqrt{k}\frac{c_w}{\sigma_{min}(W^*)} \end{aligned} \tag{57}$$

Proof of Lemma D.8.

First equality of (57) Firstly, by definition, we directly have for any w_{T+1}^*,

$$\sigma_{\min}(W^*)\|\nu^p(w_{T+1}^*)\|_2 \leq \|W^*\nu^p(w_{T+1}^*)\|_2 = \|w_{T+1}^*\|_2 \tag{58}$$

Next we are going to prove the lower bound to show the equality. Let $W^* = UDV^\top$, where $U \in O_{k\times k}, V \in O_{T\times k}, D = diag(\sigma_1(W^*), ..., \sigma_k(W^*))$ with $\sigma_1(W^*) > ... > \sigma_k(W^*)$. There always exists an w' satisfies

$$\frac{w'}{\|w'\|_2} = Ue_k \tag{59}$$

Then it is easy to see that the corresponding $\nu^p(w')$ satisfies $V^\top\nu^p(w') = \|w_{T+1}^*\|_2 \cdot (\sigma_{\min}(W^*))^{-1}e_k$. After rearranging, we have

$$\frac{\|w_{T+1}^*\|_2}{\sigma_{\min}(W^*)} = \left\|\frac{\|w_{T+1}^*\|_2}{\sigma_{\min}(W^*)}e_k\right\|_2 = \|V^\top\nu^p(w')\|_2 \leq \|\nu^p(w')\|_2 \leq \sup_{w_{T+1}^*} \|\nu^p(w_{T+1}^*)\|_2 \tag{60}$$

Combine (58) and (60) we finish the first part.

Second equality of (57). It is easy to upper bound

$$\|\nu^1(w_{T+1}^*)\|_1 \leq \sqrt{\|\nu^1(w_{T+1}^*)\|_0}\|\nu^1(w_{T+1}^*)\|_2 \leq \sqrt{\|\nu^1(w_{T+1}^*)\|_0}\frac{\|w_{T+1}^*\|_2}{\sigma_{\min}(W)} \tag{61}$$

where the last inequality again comes from (58) and the definition $W^*\nu^1(w_{T+1}^*) = w_{T+1}^*$. Now we can upper bound $\|\nu^1(w_{T+1}^*)\|_0$ by k from the following arguments.

Note that the original l_1 minimization for the undetermined linear equation $W^*\nu = w^*_{T+1}$ is equivalent to finding the solution to the following linear programming problem.

$$\begin{aligned} &\min_{\nu_\pm} \mathbf{1}^T \nu_\pm \\ &\text{s.t. } W_\pm \nu_\pm = w^*_{T+1}, \\ &\quad \nu_\pm \geq 0. \end{aligned} \tag{62}$$

where $\mathbf{1}^\top := (1, ..., 1) \in \mathbb{R}^{2T}$, $\nu_\pm^\top := (\nu^+, \nu^-)$, $\nu^+ := \max(\nu, 0)$, $\nu^- := \max(-\nu, 0)$ and $W_\pm := (W^*, -W^*) \in \mathbb{R}^{k\times 2T}$. Since $W^*\nu^* = w^*_{T+1}$ holds and there exists at least one optimal solution which is a basic feasible solution for LP (62). From Def. 2.9 and Theorem 2.3 in (Bertsimas & Tsitsiklis, 1997), we know that the cardinality for the basis of basic feasible solutions is $rank(W_\pm) = k$. so ν^1 at most k-sparse, i.e., $\|\nu^1\|_0 \leq k$.

□

We show the Lemma that reflects our motivation to get the lower bound of $\|\widetilde{\nu}^2\|_1^2$.

Lemma D.9. *Assume conditions in Theorem 3.7 hold, $N_{tot} \to \infty$, and W^* can be any matrix in $\Gamma(\sigma_k) = \{W \in \mathbb{R}^{k\times T} | \sigma_{\min}(W) \geq \sigma_k\}$, then for L2-A-MTRL and P-MTRL we have*

$$\sup_{w^*_{T+1}\in H(c_w)} \|\widetilde{\nu}^2(w^*_{T+1})\|_1^2 \gtrsim \frac{T \cdot c_w^2}{N_{tot} \cdot \sigma^2_{\min}(W^*)} \tag{63}$$

Proof of Lemma D.9. For passive learning, actually we can choose any ν^p such that $W^*\nu^p(w^*_{T+1}) = w^*_{T+1}$, then from Lemma D.8 we have:

$$\sup_{w^*_{T+1}\in H(c_w)} \|\widetilde{\nu}^p(w^*_{T+1})\|_1^2 = \frac{T}{N_{tot}} \cdot \sup_{w^*_{T+1}\in H(c_w)} \|\nu^p(w^*_{T+1})\|_1^2 = \frac{T \cdot c_w^2}{N_{tot} \cdot \sigma^2_{\min}(W^*)} \tag{64}$$

For L_2 strategy we have $n_t = \max\{c''\nu^2(t)^2, \underline{N}\}$. refer to the SVD decomposition of W^* in Lemma D.8 and the worst target vector w' defined in (60), we have

$$\nu^2(w') = VD^{-1}U^\top w' = \|w'\|_2 \cdot VD^{-1}U^\top U e_k = \|w'\|_2 \sigma^{-1}_{\min}(W^*) \cdot V_{*,k} \tag{65}$$

where $V_{*,k}$ is the k-th column vector of $V \in O_{T,k}$. Since $N_{tot} \gg T\underline{N}$ and $\|\nu^2\|_2 = \|w'\|_2\sigma^{-1}_{\min}(W^*)\|V_{*,k}\|_2^2 = \|w'\|_2\sigma^{-1}_{\min}(W^*)$, then for any $t \in S$, we have

$$n_t \approx N_{tot}\frac{|\nu^2(t)|^2}{\|\nu^2\|_2^2} = N_{tot} \cdot V_{t,k}^2 \tag{66}$$

So as $N_{tot} \to +\infty$, $t \in S \Leftrightarrow |V_{t,k}| > 0$. Note that the minimax lower bound used in Theorem 3.7 is proved by using Fano's inequality to the δ_V-separated subset as in Lemma D.7, and the corresponding separated set G_W for $W \in \mathbb{R}^{k\times T}$ is constructed from G_V. Clearly $G_{W'} := \{W \in G_W | W = UDV^\top, \exists t \in [T], \text{s.t.} V_{t,k} = 0\}$ occupy zero volume space in G_W, and thus we can use $G_W - G_{W'}$ to replace the original G_W set by excluding a corrsponding zero volume space in (43) from Lemma D.7 which has no difference to the original results. So set $\|w'\|_2 = c_w$, with probability $1 - o(1)$ we have $V_{t,k} > 0$ and thus

$$\sup_{w^*_{T+1}\in H(c_w)} \|\widetilde{\nu}^2(w^*_{T+1})\|_1^2 \overset{w^*_{T+1}=w'}{\geq} \sum_{t\in S} \frac{|\nu^2(t)|^2}{c''|\nu^2(t)|^2} + \sum_{t\notin S} \frac{|\nu^2(t)|^2}{\underline{N}} \gtrsim \frac{|S|}{c''} = \frac{T}{c''} \tag{67}$$

where $c'' = N_{tot}\sigma^2_{\min}(W^*)c_w^{-2}$.

□

We then prove a simple lemma to show that with a particular condition, we have $\|Av\| \approx \|A\|_F\|v\|$.

Lemma D.10. *Assume $v \in \mathbb{R}^b$, $A, \Delta A \in \mathbb{R}^{a\times b}$ and $\|\Delta A\|_F = c \cdot \|A\|$ for some $a, b \in \mathbb{N}^+$ and $c \in (0, 1)$. Further assume that A satisfies $\|Av\| = \|A\|_F\|v\|$, then*

$$\|(A + \Delta A)v\| \geq \frac{1-c}{1+c}\|A + \Delta A\|_F \cdot \|v\| \tag{68}$$

Proof of Lemma D.10. We proof it directly:

$$\begin{aligned}\|(A+\Delta A)v\| &\geq \|Av\| - \|\Delta A \cdot v\| = \|A\|_F\|v\| - \|\Delta A \cdot v\| \geq (\|A\|_F - \|\Delta A\|_F)\|v\| \\ &= \frac{1-c}{1+c}(\|A\|_F + \|\Delta A\|_F)\|v\| \geq \frac{1-c}{1+c}\|A+\Delta A\|_F\|v\|\end{aligned} \tag{69}$$

□

With such Lemma, we can prove an important Lemma for the lower bound of L2-A-MTRL and P-MTRL.

Lemma D.11. *Recall the definition of $H(c_w)$ and $\nu^p(w^*_{T+1})$ in (55) and (56), we have the following results for L_2-minimization solution.*

$$\inf_{(\hat{B},\widetilde{W})} \sup_{(B^*,\widetilde{W}^*,w^*_{T+1}\in H(c_w))} \|(\hat{B}\widetilde{W} - B^*\widetilde{W}^*)\tilde{\nu}^2(w^*_{T+1})\|_2^2 \gtrsim \sigma^2 \cdot k(d-k) \cdot \frac{T \cdot c_w^2}{k \cdot N_{tot} \cdot \sigma^2_{\min}(W^*)} \tag{70}$$

Proof of Lemma D.11. The key idea is that we want to find some $\widetilde{W^*}$ such that $\|(\hat{B}-B^*)\widetilde{W}^*\nu^2\| \gtrsim \|(\hat{B}-B^*)\widetilde{W}\|_F\|\nu^2\|$ when all the row vectors of $\widetilde{W}$ are almost aligned with ν^2. Without loss of generality, we assume $\nu^2(t) \neq 0,\ \forall t \in [T]$, and thus when $N_{tot} \to \infty$, we have $n_t = c'' \cdot |\nu^2(t)|^2,\ \forall t \in [T]$, where $c'' \gg 1$ is some constant satisfies $c'' = N_{tot}/\|\nu\|^2$. We prove the Lemma step by step.

First, we construct a specific $\widetilde{W^*}$, which is almost rank-1 and has rows aligned with $\widetilde{\nu}^2$, to achieve the lower bound. For any given $\nu(t)$, we define

$$\widetilde{W^*} := \frac{1}{\sqrt{c''}} u \cdot \chi^\top + \widetilde{\Delta W^*} \tag{71}$$

where $u \in \mathbb{R}^T$ is some vector to be determined later and

$$\chi(t) := \text{sgn}(\nu^2(t)) = \mathbb{I}[\nu^2(t) > 0] - \mathbb{I}[\nu^2(t) < 0], \quad \chi \in \mathbb{R}^T \tag{72}$$

$$\widetilde{\Delta W^*} := \sum_{i=2}^{k} \widetilde{\sigma}_i \widetilde{\alpha}_i \widetilde{\beta}_i^\top \tag{73}$$

Obviously, $\|\chi\| = \sqrt{T}$. Here $\widetilde{\alpha}_i, \widetilde{\beta}_i \in \mathbb{R}^T,\ \forall i \in \{2,\dots,k\}$ and we let $\{u/\|u\|_2, \widetilde{\alpha}_2, \dots, \widetilde{\alpha}_k\}$ and $\{\chi/\sqrt{T}, \widetilde{\beta}_2, \dots, \widetilde{\beta}_k\}$ to be two orthonormal bases of two k-dimensional subspace of $\mathbb{R}^T$. The reason for such a definition of $\widetilde{\Delta W^*}$ is that the Eqn. 71 will naturally be an SVD form of $\widetilde{W}^*$. And for simplicity, we let $\widetilde{\sigma}_i \equiv \widetilde{\sigma}_k,\ \forall i \in \{2,\dots,k\}$. We then let

$$\|\frac{1}{\sqrt{c''}} u \cdot \chi^\top\|_F = 2\|\widetilde{\Delta W^*}\|_F \ \Leftrightarrow\ \widetilde{\sigma}_k = \frac{\|u\|\sqrt{T}}{2\sqrt{(k-1)c''}} \tag{74}$$

Then from $\widetilde{\nu}^2(t) = \nu^2(t)/\sqrt{n_t} = \chi(t)/\sqrt{c''}$ and $\widetilde{w}^*(t) = \sqrt{n_t}w^*(t) = \sqrt{c''}|\nu^2(t)| \cdot w^*(t)$, we have

$$W^*\nu^2 = \widetilde{W^*}\widetilde{\nu}^2 = \frac{T}{c''}u + \frac{1}{\sqrt{c''}}\sum_{i=2}^{k} \widetilde{\sigma}_k \widetilde{\alpha}_i \widetilde{\beta}_i^\top \chi = \frac{T}{c''}u \tag{75}$$

Note that $W^*\nu^2 = w^*_{T+1} \in H(c_w)$, i.e., $c_w = \|W^*\nu^2\|$, we have the following conditions for $\|u\|$ and $\widetilde{\sigma}_k$.

$$\|u\|_2 = \frac{c_w \cdot c''}{T},\ \widetilde{\sigma}_k = \frac{c_w\sqrt{c''}}{2\sqrt{(k-1)T}} \tag{76}$$

We then choose $\nu^2 = \nu' := \mathbf{1} = [1,\dots,1]^\top \in \mathbb{R}^T$, thus $\chi = \nu' = \mathbf{1}, \|\nu'\| = \sqrt{T}$ and for W^* we have:

$$W^* = \frac{u}{c''}\mathbf{1}^\top + \frac{1}{\sqrt{c''}}\widetilde{\sigma}_k \sum_{i=2}^{k} \widetilde{\alpha}_i \widetilde{\beta}_i^\top = \frac{1}{\sqrt{c''}} \cdot \widetilde{W^*} \tag{77}$$

And thus

$$\sigma_{\min}(W^*) = \sigma_k = \frac{\widetilde{\sigma}_k}{\sqrt{c''}} = \frac{c_w}{2\sqrt{(k-1)T}} \tag{78}$$

which results that $T = \|\nu'\|^2 > c_w^2[4(k-1)\sigma_k^2]^{-1}$. And we get

$$\|\widetilde{\nu'}\|^2 = \sum_{t=1}^{T} \frac{|\nu'(t)|^2}{c''|\nu(t)|^2} = \frac{T\|\nu_2\|^2}{N_{tot}} \gtrsim \frac{Tc_w^2}{k\sigma_k^2 N_{tot}} \tag{79}$$

We check that ν' is a valid choice for the L_2-minimization solution. Let $W^* = UDV^\top$ be the SVD form of W^*, where $U = (u/\|u\|, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_k) \in O_{k\times k}, V = (\chi/\sqrt{T}, \widetilde{\beta}_2, \ldots, \widetilde{\beta}_k) \in O_{T\times k}, D = \text{diag}(\sigma_1, \ldots, \sigma_k), \sigma_1 \geq \ldots \geq \sigma_k \geq 0$. Note that

$$VD^{-1}U^\top W^*\nu' = VV^\top\nu' = (\frac{1}{T}\chi\chi^\top + \sum_{i=2}^{k}\widetilde{\beta}_i\widetilde{\beta}_i^\top)\nu' = \chi + 0 = \nu'. \tag{80}$$

Therefore, $\nu' = \arg\min_{W^*\nu'=W^*x}\|x\|_2$. Here we use the fact that $\widetilde{\beta}_i^\top\nu' = \widetilde{\beta}_i^\top\chi = 0,\ \forall i \in \{2, \ldots, k\}$.

Finally we have:

$$\begin{aligned}
&\inf_{(\hat{B},\widetilde{W})}\sup_{(B^*,\widetilde{W}^*,w_{T+1}^*\in H(c_w))}\|(\hat{B}\widetilde{W} - B^*\widetilde{W}^*)\tilde{\nu}^2(w_{T+1}^*)\|_2^2 \\
&\gtrsim \inf_{\hat{B}}\sup_{B^*}\|(\hat{B} - B^*)\widetilde{W}^*\tilde{\nu'}\|_2^2 \\
&\gtrsim \inf_{\hat{B}}\sup_{B^*}\|(\hat{B} - B^*)\widetilde{W}^*\|_F^2\|\tilde{\nu}'\|_2^2, \quad \text{(Eqn. 74, Lemma D.10)} \\
&\asymp \inf_{\hat{B}}\sup_{B^*}\sum_{t=1}^{T}\|X_t(\hat{B} - B^*)w_t^*\|_2^2 \cdot \frac{Tc_w^2}{k\sigma_k^2 N_{tot}}, \quad \text{(Eqn. 79)} \\
&\gtrsim \sigma^2 \cdot k(d-k) \cdot \frac{Tc_w^2}{k\sigma_k^2 N_{tot}}
\end{aligned} \tag{81}$$

For the first and last inequality, we restrict the local packing space on W and obtain the results in a manner similar to Theorem D.3. Specifically, we note that the orthonormal matrix B can be viewed as a Grassmann manifold that is diffeomorphic to a $k \times (d-k)$ dimensional linear matrix(Bai et al., 1992), and the constraint $Bw = \mathbf{0}$ introduces at most d additional limiting equations to B, which will not influence its local packing number. Therefore, it becomes straightforward to prove the last inequality using a methodology similar to the proof of Theorem D.3. And we finish the proof.

□

Finally, we turn to our main theorem in Sec. 3.2.

Proof of Theorem 3.7. From the conditions, we have $c_w = \Theta(1)$.

Upper bound of ER for L1-A-MTRL. From Eqn. 29 from the proof of Lemma 3.2, we get $\|\widetilde{\nu}^1\| \leq (1+o(1))\|\nu^1\|_1^2/N_{tot}$ when $N_{tot} \gg T\underline{N}$. Then use the second inequality of (57) in Lemma D.8, we have

$$\sup_{w_{T+1}^*\in H(c_w)}\|\widetilde{\nu}^1(w_{T+1}^*)\|_1^2 \lesssim \sup_{w_{T+1}^*\in H(c_w)}\frac{\|\nu^1(w_{T+1})\|_1^2}{N_{tot}} \leq \frac{k\cdot c_w^2}{N_{tot}\cdot\sigma_{\min}^2(W^*)} \tag{82}$$

For the upper bound, let $\widetilde{w}_t = \hat{w}_t\sqrt{n_t}$, $\widetilde{w}_t^* = \hat{w}_t^*\sqrt{n_t}$ and $\widetilde{\nu}^2(t) = \frac{\nu^*(t)}{\sqrt{n_t}}$ for all $t \in [T]$, then we have:

$$\begin{aligned}
\mathbb{E}_{x\sim\mu_{T+1}}\|x^\top(\hat{B}\hat{w}_{T+1} - B^*w_{T+1}^*)\|_2^2 &= \|(\Sigma_{T+1}^*)^{\frac{1}{2}}(\hat{B}\hat{W} - B^*W^*)\nu^1\|_2^2 \\
&\le \|(\Sigma_{T+1}^*)^{\frac{1}{2}}(\hat{B}\widetilde{W} - B^*\widetilde{W}^*)\|_F^2 \cdot \|\widetilde{\nu}^1\|^2 \\
&= \sum_{t=1}^{T} n_t\|(\Sigma_{T+1}^*)^{\frac{1}{2}}(\hat{B}\hat{w}_t - B^*w_t^*)\|^2 \cdot \|\widetilde{\nu}^1\|^2 \\
&\asymp \sum_{t=1}^{T} n_t\|(\Sigma_t^*)^{\frac{1}{2}}(\hat{B}\hat{w}_t - B^*w_t^*)\|^2 \cdot \|\widetilde{\nu}^1\|^2, \qquad \text{(Assumption 2.2)} \\
&\lesssim \sum_{t=1}^{T} \|X_t(\hat{B}\hat{w}_t - B^*w_t^*)\|^2 \cdot \|\widetilde{\nu}^1\|^2, \qquad \text{(Lemma D.2)} \\
&\le \sigma^2(kd\ln(\frac{N_{tot}}{T}) + kT + \ln(\frac{1}{\delta}))\|\widetilde{\nu}^1\|^2, \quad \text{(Claim C.1 in (Chen et al., 2022))}
\end{aligned} \tag{83}$$

Then combine (83) and (82) we prove the result for L1-A-MTRL.

Lower bound of ER for P-MTRL/L2-A-MTRL. For L2-A-MTRL, we derive the results from Lemma D.11. It can be easily verified that the same results hold for P-MTRL since we set $\nu' = [1, \dots, 1]^\top \in \mathbb{R}^T$ in Lemma D.11. □

E. Proof of Theorem 3.10

Before proofing the original Theorem, we first illustrate an assumption naturally used for the sparse linear model and Lasso Program (Wainwright, 2019):

Assumption E.1. (RE condition) Let ν^* be supported on a subset $S \in [T]$ with $|S| = s$ (From Theorem 3.7 we know $s \le k$). Then W^* satisfies *Restricted Eigenvalue* condition over S with parameters $(\kappa, 3)$ if:

$$\|W^*\Delta\|_2^2 \ge \kappa\|\Delta\|_2^2, \qquad \forall\Delta \in \mathbb{C}_3(S) \tag{84}$$

where $\mathbb{C}_\alpha(S) := \{\Delta \in \mathbb{R}^k | \|\Delta_{S^c}\|_1 \le \alpha\|\Delta_S\|_1\}$.

What should be mentioned is that in this section we just consider L1-A-MTRL, so we replace $\hat{\nu}$ and ν^* with $\hat{\nu}^1$ and ν^1, respectively.

Since $\sigma_{\max}^2(W^*) \ge \kappa \ge \sigma_{\min}^2(W^*)$, we rewrite Theorem 3.10 with RE condition as follows. Once we prove the following theorem, we can replace κ with $\sigma_{\min}^2(W^*)$ and $\sigma_{\max}^2(W^*)$ correspondingly and immediately prove the original theorem.

Theorem E.2. *Let Assumption 2.1, 2.3, 2.4, 3.8, 3.9, E.1 hold. Let Λ denote the lower bound of $\|\nu^*\|_1$, $q = \frac{\sqrt{k}R}{\underline{\sigma}}$ (so $q \ge \|\nu^*\|_1$) and $\gamma \ge \max\{2160sqC_W\Lambda^{-1}, \sqrt{2160sq\kappa\Lambda^{-1}}\}$ and $\underline{\sigma} = \sigma_{\min}(W^*) > 0$. Then in order to let $ER_{L_1} \le \varepsilon^2$ with probability $1-\delta$, the number of source samples N_{total} is at least*

$$\widetilde{\mathcal{O}}(\sigma^2(kd + kT)\|\nu^*\|_1^2\varepsilon^{-2} + T\beta) \tag{85}$$

where $\beta = \max\{\gamma^2\frac{\sigma_z^2}{\kappa^2}, \gamma^2\frac{C_W^2}{\kappa^2}\rho^4, \rho^4, \frac{\sigma_z^2}{\kappa}\} \cdot (d + \ln(\frac{4T}{\delta}))$, and target task sample complexity n_{T+1} is at least

$$\widetilde{\mathcal{O}}(\sigma^2 k\varepsilon^{-2} + \alpha) \tag{86}$$

where $\alpha = \max\{\gamma^2\frac{\sigma_z^2}{\kappa^2\Lambda^2}, \gamma^2\frac{C_W^2}{\kappa^2}\rho^4, \rho^4\} \cdot (k + \ln(\frac{4}{\delta}))$.

Lemma E.3. *(A variant of Theorem 7.13 in (Wainwright, 2019)) Assume that Assumption E.1 hold. Any solution of the Lagrangian Lasso (16) with regularization parameter lower bounded as $\lambda_k \ge 2\|\hat{W}^\top z\|_\infty$ satisfies the following bound*

$$\|\hat{\nu} - \nu^*\|_2 \le \frac{3}{\kappa}\sqrt{s}\lambda_k \tag{87}$$

$$\|\hat{\nu} - \nu^*\|_1 \le 4\sqrt{s}\|\hat{\nu} - \nu^*\|_2 \tag{88}$$

□

Remark E.4. In Theorem E.5 we want $\epsilon \leq \min(0.05, \frac{\kappa}{4\gamma C_W})$ with high probability, so from Lemma D.2, we need $n_t > \max(400, \frac{16\gamma^2 C_W^2}{\kappa^2}) c_a \rho^4 (d + \ln(\frac{2T}{\delta}))$ for all $t \in [T]$ and $n_{T+1} > \max(400, \frac{16\gamma^2 C_W^2}{\kappa^2}) c_a \rho^4 (k + \ln(\frac{2}{\delta}))$ for universal constant $c_a > 0$.

To get the bound of regularization parameter λ_k, we turn to control the bound of the noise term z since $\hat{W}$ and $\hat{w}^*_{T+1}$ are solved by original least square method.

Theorem E.5. *If* $n_t^i \geq \max\{3\gamma^2 \frac{\sigma_z^2}{\kappa^2}, 16\gamma^2 \frac{C_W^2}{\kappa^2} c_a \rho^4, 400 c_a \rho^4, \frac{12\sigma_z^2}{\kappa}\} \cdot (d + \ln(\frac{4T}{\delta}))$, $n_{M+1}^i \geq \max\{3\gamma^2 \frac{\sigma_z^2}{\kappa^2 \|\nu^*\|_1^2}, 16\gamma^2 \frac{C_W^2}{\kappa^2} c_a \rho^4, 400 c_a \rho^4\} \cdot (k + \ln(\frac{4}{\delta}))$, *and Assumption E.1 , 3.8 , 3.9 hold. Then with probability* $1 - \delta$ *we have*

$$\|\hat{\nu} - \nu^*\|_1 \leq \frac{2160}{\gamma} s \cdot \max\{C_W, \frac{\kappa}{\gamma}\} \cdot \frac{\sqrt{k}R}{\underline{\sigma}} \tag{89}$$

Remark E.6. If (89) holds and $\frac{\sqrt{k}R}{\underline{\sigma}} = \Theta(\|\nu^*\|_1)$, then active learning method with L1-minimization just multiplies an additional term $1 + \frac{2160}{\gamma} s \max\{C_W, \frac{\kappa}{\gamma}\}$, i.e.

$$ER_{active} \lesssim \sigma^2 (kd \ln(\frac{N_{tot}}{T}) + kT) \frac{\|\nu^*\|_1^2}{N_{tot}} (1 + \frac{2160}{\gamma} s \max\{C_W, \frac{\kappa}{\gamma}\})^2 + \sigma^2 \frac{(k + \ln(\frac{1}{\delta}))}{n_{T+1}} \tag{90}$$

Proof of Theorem E.5.

Substep 1: Decompose z.

As the analysis of original least square method in (Chen et al., 2022), for every $t \in [T+1]$ we have:

$$\begin{aligned}
\hat{w}_t^i &= \arg\min_w \|X_t^i \hat{B}^i w - Y_t\|_2 \\
&= ((X_t^i \hat{B}^i)^\top X_t^i \hat{B}^i)^{-1} (X_t^i \hat{B}^i)^\top Y_t \\
&= ((\hat{B}^i)^\top \hat{\Sigma}_t^i \hat{B}^i)^{-1} (\hat{B}^i)^\top \hat{\Sigma}_t^i B^* w_t^* + \frac{1}{n_t} ((\hat{B}^i)^\top \hat{\Sigma}_t^i \hat{B}^i)^{-1} (\hat{B}^i)^\top (X_t^i)^\top Z_t
\end{aligned} \tag{91}$$

Then we have

$$\begin{aligned}
z^i =& \hat{w}_{T+1}^i - \hat{W}^i \nu^* \\
=& \hat{w}_{T+1}^i - \sum_{t=1}^T \hat{w}_t^i \nu_t^* \\
=& ((\hat{B}^i)^\top \hat{\Sigma}_{T+1}^i \hat{B}^i)^{-1} (\hat{B}^i)^\top \hat{\Sigma}_{T+1}^i B^* w_{T+1}^* - \sum_{t=1}^T ((\hat{B}^i)^\top \hat{\Sigma}_t^i \hat{B}^i)^{-1} (\hat{B}^i)^\top \hat{\Sigma}_t^i B^* w_t^* \nu_t^* \\
&+ \frac{1}{n_{T+1}} ((\hat{B}^i)^\top \hat{\Sigma}_{T+1}^i \hat{B}^i)^{-1} (\hat{B}^i)^\top (X_{T+1}^i)^\top Z_{T+1} - \sum_{t=1}^T \frac{1}{n_t} ((\hat{B}^i)^\top \hat{\Sigma}_t^i \hat{B}^i)^{-1} (\hat{B}^i)^\top (X_t^i)^\top Z_t \nu_t^* \\
=& \underbrace{((\hat{B}^i)^\top \hat{\Sigma}_{T+1}^i \hat{B}^i)^{-1} (\hat{B}^i)^\top \hat{\Sigma}_{T+1}^i B^* w_{T+1}^* - ((\hat{B}^i)^\top \Sigma^* \hat{B}^i)^{-1} (\hat{B}^i)^\top \Sigma^* B^* w_{T+1}^*}_{E_1^i} \\
&\underbrace{- (\sum_{t=1}^T ((\hat{B}^i)^\top \hat{\Sigma}_t^i \hat{B}^i)^{-1} (\hat{B}^i)^\top \hat{\Sigma}_t^i B^* w_t^* \nu_t^* - \sum_{t=1}^T ((\hat{B}^i)^\top \Sigma^* \hat{B}^i)^{-1} (\hat{B}^i)^\top \Sigma^* B^* w_t^* \nu_t^*)}_{E_2^i} \\
&\underbrace{+ \frac{1}{n_{T+1}} ((\hat{B}^i)^\top \hat{\Sigma}_{T+1}^i \hat{B}^i)^{-1} (\hat{B}^i)^\top (X_{T+1}^i)^\top Z_{T+1}}_{E_3^i} \underbrace{- \sum_{t=1}^T \frac{1}{n_t} ((\hat{B}^i)^\top \hat{\Sigma}_t^i \hat{B}^i)^{-1} (\hat{B}^i)^\top (X_t^i)^\top Z_t \nu_t^*}_{E_4^i}
\end{aligned} \tag{92}$$

where the third equality of Equ. 92 use Equ. 91 and the fourth equality comes from $w^*_{T+1} = W^*\nu^*$. It's obvious that $E^i_k, k \in \{1,2,3,4\}$ all have 0 expectation, and to control the bound of z, we just need to bound these 4 term in l_2-norm for all i and use the inequality

$$\|z\|_2 = \|E^i_1 - E^i_2 + E^i_3 - E^i_4\|_2 \leq 2(\|E^i_1\|_2 + \|E^i_2\|_2 + \|E^i_3\|_2 + \|E^i_4\|_2) \tag{93}$$

Substep 2: Calculate Error Terms E^i_*.

For the first term, with Inequ. 33 and Assumption 3.9 we have

$$\begin{aligned}
\|E^i_1\|_2 &\leq \|((\hat{B}^i)^\top \hat{\Sigma}^i_{T+1}\hat{B}^i)^{-1}(\hat{B}^i)^\top \hat{\Sigma}^i_{T+1}B^* - ((\hat{B}^i)^\top \Sigma^* \hat{B}^i)^{-1}(\hat{B}^i)^\top \Sigma^* B^*\|_2 \|w^*_{T+1}\|_2 \\
&\leq \|w^*_{T+1}\|_2 \cdot \|\frac{1+2\epsilon}{1-2\epsilon}((\hat{B}^i)^\top \Sigma^* \hat{B}^i)^{-1}(\hat{B}^i)^\top \Sigma^* B^* - ((\hat{B}^i)^\top \Sigma^* \hat{B}^i)^{-1}(\hat{B}^i)^\top \Sigma^* B^*\|_2 \\
&\leq \|w^*_{T+1}\|_2 \frac{4\epsilon}{1-2\epsilon}\|((\hat{B}^i)^\top \hat{B}^i)^{-1}(\hat{B}^i)^\top B^*\|_2 \\
&\leq \frac{4\epsilon}{1-2\epsilon}\|w^*_{T+1}\|_2, \qquad (\sigma_{\max}((\hat{B}^i)^\top B^*) \leq 1) \\
&\leq \frac{4\epsilon}{1-2\epsilon}C_W\|\nu^*\|_1, \qquad (\|w^*_{T+1}\|_2 = \|\sum_{t=1}^T W^* e_t \nu^*_t\|_2 \leq \max_t \|W^* e_t\|_2 \cdot \|\nu^*\|_1)
\end{aligned} \tag{94}$$

The fourth inequality is relevant to subspace angle distance between p and q, where $\hat{B}^i$ and B^* are orthonormal matrices whose colums form orthonormal bases of p and q respectively, as section 2 in (Tripuraneni et al., 2021). The second term E^i_2 has upper bound similar to E^i_1:

$$\begin{aligned}
\|E^i_2\|_2 &\leq \sum_{t=1}^T \|((\hat{B}^i)^\top \hat{\Sigma}^i_t \hat{B}^i)^{-1}(\hat{B}^i)^\top \hat{\Sigma}^i_t B^* - ((\hat{B}^i)^\top \Sigma^* \hat{B}^i)^{-1}(\hat{B}^i)^\top \Sigma^* B^*\|_2 \|w^*_t \nu^*_t\|_2 \\
&\leq \frac{4\epsilon}{1-2\epsilon}\|((\hat{B}^i)^\top \hat{B}^i)^{-1}(\hat{B}^i)^\top B^*\|_2 \sum_{t=1}^T \|w^*_t \nu^*_t\|_2 \\
&\leq \frac{4\epsilon}{1-2\epsilon}C_W\|\nu^*\|_1
\end{aligned} \tag{95}$$

For the third term, from Lemma E.8 with probability at least $1 - \frac{\delta}{4}$ we have:

$$\begin{aligned}
\|E^i_3\|_2 &\leq \frac{1}{n_{T+1}}\|((\hat{B}^i)^\top \hat{\Sigma}^i_{T+1}\hat{B}^i)^{-1}(\hat{B}^i)^\top (X^i_{T+1})^\top Z_{T+1}\|_2 \\
&\leq \frac{1}{n_{T+1}\cdot(1-2\epsilon)}\|((\hat{B}^i)^\top \Sigma^* \hat{B}^i)^{-1}\|_2 \|(\hat{B}^i)^\top (X^i_{T+1})^\top Z_{T+1}\|_2 \\
&\leq \frac{\sqrt{1+2\epsilon}}{1-2\epsilon}\cdot\sigma_z\sqrt{\frac{2k+3\ln(\frac{4}{\delta})}{n_{T+1}}}
\end{aligned} \tag{96}$$

Analogously, from Lemma E.8 with probability at least $1 - \frac{\delta}{4}$ we have:

$$\begin{aligned}
\|E^i_4\|_2 &\leq \sum_{t=1}^T \frac{1}{n_t}\|((\hat{B}^i)^\top \hat{\Sigma}^i_t \hat{B}^i)^{-1}(\hat{B}^i)^\top (X^i_t)^\top Z_t \nu^*_t\|_2 \\
&\leq \sum_{t=1}^T \frac{1}{n_t}\|((\hat{B}^i)^\top \hat{\Sigma}^i_t \hat{B}^i)^{-1}(\hat{B}^i)^\top\|_2 \|(X^i_t)^\top Z_t\|_2 |\nu^*_t| \\
&\leq \sum_{t=1}^T \frac{\sqrt{1+2\epsilon}}{1-2\epsilon}\cdot\sigma_z\sqrt{\frac{2d+3\ln(\frac{4T}{\delta})}{n_t}}|\nu^*_t| \\
&\leq \frac{\sqrt{1+2\epsilon}}{1-2\epsilon}\cdot\sigma_z\sqrt{\frac{2d+3\ln(\frac{4T}{\delta})}{\min_t(n_t)}}\|\nu^*\|_1
\end{aligned} \tag{97}$$

Substep 3: Final Calculation.

Combining (94), (95), (96), (97) and (93), with probability at least $1-\delta$ we have

$$\begin{aligned}\|z^i\|_2 &\le \frac{16\epsilon}{1-2\epsilon}C_W\|\nu^*\|_1 + \frac{2\sqrt{1+2\epsilon}}{1-2\epsilon}\cdot\sigma_z\Big(\sqrt{\frac{2k+3\ln(\frac{4}{\delta})}{n_{T+1}}}+\sqrt{\frac{2d+3\ln(\frac{4T}{\delta})}{\min_t(n_t)}}\|\nu^*\|_1\Big)\\ &\le \frac{16}{0.9\times 4\times\gamma}\kappa\|\nu^*\|_1 + \frac{2\sqrt{1.1}}{0.9}\times\frac{\kappa\|\nu^*\|_1}{\gamma}\times 2, \qquad \text{(Conditions)}\\ &\le \frac{82}{9}\frac{\kappa\|\nu^*\|_1}{\gamma}\end{aligned} \tag{98}$$

Choose that

$$\begin{aligned}\lambda_k &:= 45\frac{\kappa\sqrt{k}R}{\gamma\underline{\sigma}}\max\{C_W,\frac{\kappa}{\gamma}\}\\ &\ge 45\frac{\kappa\|\nu^*\|_1}{\gamma}\max\{C_W,\frac{\kappa}{\gamma}\}\\ &\ge 2\times\frac{22}{9}\max\{C_W,\frac{\kappa}{\gamma}\}\times\frac{82}{9}\frac{\kappa\|\nu^*\|_1}{\gamma}\\ &\ge 2\cdot(\max_t\|\hat{w}_t^i\|_2)\cdot\|z^i\|_2, \qquad ((98),(101))\\ &\ge 2\max_t|(\hat{w}_t^i)^\top z^i| \ge 2\|\hat{W}^\top z^i\|_\infty\end{aligned} \tag{99}$$

Finally from Lemma E.3 , the solution of (16) with regularization parameter λ_k satisfies:

$$\begin{aligned}\|\hat{\nu}-\nu^*\|_1 &\le \frac{12s}{\frac{1}{4}\kappa}\lambda_k, \qquad \text{(Lemma E.3, E.9)}\\ &= \frac{2160}{\gamma}s\cdot\frac{\sqrt{k}R}{\underline{\sigma}}\cdot\max\{C_W,\frac{\kappa}{\gamma}\}, \qquad (99)\end{aligned} \tag{100}$$

□

Lemma E.7. *Assume conditions in Theorem E.5 hold, then the norms of column vectors of $\hat{W}$ have similar uppper bound to that of W^*:*

$$\|\hat{w}_t^i\|_2 \le \frac{22}{9}\max\{C_W,\frac{\kappa}{\gamma}\} \tag{101}$$

Proof of Lemma E.7. This can be done by directly calculation as (95) and (97)

$$\begin{aligned}\|\hat{w}_t^i\|_2 &= \|((\hat{B}^i)^\top\hat{\Sigma}_t^i\hat{B}^i)^{-1}(\hat{B}^i)^\top\hat{\Sigma}_t^iB^*w_t^* + \frac{1}{n_t}((\hat{B}^i)^\top\hat{\Sigma}_t^i\hat{B}^i)^{-1}(\hat{B}^i)^\top(X_t^i)^\top Z_t\|_2\\ &\le \|((\hat{B}^i)^\top\hat{\Sigma}_t^i\hat{B}^i)^{-1}(\hat{B}^i)^\top\hat{\Sigma}_t^iB^*\|_2\|w_t^*\|_2 + \frac{1}{n_t}\|((\hat{B}^i)^\top\hat{\Sigma}_t^i\hat{B}^i)^{-1}(\hat{B}^i)^\top\|_2\|(X_t^i)^\top Z_t\|_2\\ &\le \frac{1+2\epsilon}{1-2\epsilon}C_W + \frac{\sqrt{1+2\epsilon}}{1-2\epsilon}\cdot\frac{\kappa}{\gamma}\\ &\le \frac{1.1\times 2}{9}\max\{C_W,\frac{\kappa}{\gamma}\}\end{aligned} \tag{102}$$

□

Lemma E.8. *Assume Assuption 3.9 holds. For any $t\in[T]$, with probability $1-\frac{\delta}{4}$ we have*

$$\|(X_t^i)^\top Z_t\|_2 \le \sigma_z\sqrt{n_t(1+2\epsilon)(2d+3\ln(\frac{4T}{\delta}))} \tag{103}$$

As for target task, for any $B\in\mathbb{R}^{d\times k}$ that is independent of Z_{T+1} , with probability $1-\frac{\delta}{4}$ we have

$$\|B^\top(X_{T+1}^i)^\top Z_{T+1}\|_2 \le \sigma_z\sqrt{n_{T+1}(1+2\epsilon)(2k+3\ln(\frac{4}{\delta}))} \tag{104}$$

Proof of Lemma E.8. We firstly proof 104. Using SVD we have $B^\top(X_{T+1}^i)^\top = U_{BX}D_{BX}V_{BX}^\top$, where $U_{BX} \in O_{k\times k}, V_{BX} \in O_{n\times k}, D_{VX} = diag(\sigma_1(B^\top(X_{T+1}^i)^\top), ..., \sigma_k(B^\top(X_{T+1}^i)^\top))$. Let $Q := V_{BX}^\top Z_{T+1}$, we know $Q \sim \mathcal{N}(0, \sigma_z^2 I_k)$ since B, X_{T+1}^i are independent to Z_{T+1}, so does V_{BX}. Note that $\frac{1}{\sigma_z^2}\|Q\|_2^2 \sim \chi^2(k)$, and thus with probability at least $1-\frac{\delta}{4}$ we have (Laurent & Massart, 2000)

$$\frac{1}{\sigma_z^2}\|Q\|_2^2 \le k + 2\sqrt{k\ln\frac{4}{\delta}} + 2\ln\frac{4}{\delta} \tag{105}$$

Then use (105), with probability at least $1-\frac{\delta}{4}$ we have

$$\begin{aligned}
\|B^\top(X_{T+1}^i)^\top Z_{T+1}\|_2^2 &= Z_{T+1}^\top(X_{T+1}^i)BB^\top(X_{T+1}^i)^\top Z_{T+1} \\
&= Z_{T+1}^\top V_{BX}D_{BX}^2V_{BX}^\top Z_{T+1} \\
&= \sum_{j=1}^{k}\sigma_j^2(B^\top(X_{T+1}^i)^\top)Q_j^2 \\
&\le \sigma_{\max}((X_{T+1}^i)^\top X_{T+1}^i)\|Q\|_2^2 \\
&\le n_{T+1}\cdot(1+2\epsilon)\cdot\sigma_z^2(2k+3\ln(\frac{4}{\delta})), \qquad \text{(Assumption 3.9 , (105))}
\end{aligned} \tag{106}$$

As for source tasks, (104) don't hold since $\hat{B}^i$ is not independent to X_t^i and Z_t. Then in order to get (103), we just need to note that $rank(X_t^i) = d$ and others steps are similar to the proof above.

□

Lemma E.9. *If all the conditions of Theorem E.5 hold, then $\hat{W}$ satisfies RE conditions with parameter $(\frac{1}{4}\kappa, 3)$.*

Proof of Lemma E.9. Applying SVD to $\frac{1}{\sqrt{n_t}}(X_t^i)^\top = U_tD_tV_t^\top$, where $U_t \in O^{d\times d}, V_t \in O^{n\times d}, D_t = diag(\sigma_{1,t}, ..., \sigma_{d,t})$. Let $Q_t := V_t^\top Z_t\Delta_t$, we know $Q_t \sim \mathcal{N}(0, \sigma_z^2\Delta_t^2 I_d)$ since X_t^i, Δ_t are independent to Z_t, so does V_t. Furthermore, we have $\sum_{t=1}^T \frac{1}{\sqrt{n_t}}U_tD_tQ_t \sim \mathcal{N}(0, \sigma_z^2\sum_{t=1}^T\frac{1}{n_t}\Delta_t^2U_tD_t^2U_t^\top) = \mathcal{N}(0, \sigma_z^2\sum_{t=1}^T\frac{1}{n_t}\Delta_t^2\hat{\Sigma}_t^i)$ due to task independence. Notice that:

$$(1-2\epsilon)I_d \preceq \hat{\Sigma}_t^i = \frac{1}{n_t}(X_t^i)^\top X_t^i = U_tD_t^2U_t^\top \preceq (1+2\epsilon)I_d, \qquad \text{(Assumption 3.9, (32))} \tag{107}$$

We immediately have $\sigma_*(D_t) \in [\sqrt{1-2\epsilon}, \sqrt{1+2\epsilon}]$. From the density function of multivariate normal distribution, let $\hat{\Gamma} := \sum_{t=1}^T\frac{1}{n_t}\Delta_t^2\hat{\Sigma}_t^i$ and $\widetilde{\Delta}_t = \frac{\Delta_t}{\sqrt{n_t}}$, then from (107), when $\|x\|_2$ is sufficiently large we have:

$$\frac{1}{(2\pi)^{\frac{d}{2}}\|\widetilde{\Delta}\|_2\sqrt{1+2\epsilon}}\exp(-\frac{1}{2}x^\top x\frac{1}{\|\widetilde{\Delta}\|_2^2(1+2\epsilon)}) \ge \frac{1}{(2\pi)^{\frac{d}{2}}|\hat{\Gamma}|^{1/2}}\exp(-\frac{1}{2}x^\top\hat{\Gamma}^{-1}x) \tag{108}$$

Thus in order to bound the L2 norm of $\sum_{t=1}^T\frac{1}{\sqrt{n_t}}U_tD_tQ_t$ with high probability, we just need to bound the L2 norm of random vectors with distribution $\mathcal{N}(0, \sigma_z^2(1+2\epsilon)\|\widetilde{\Delta}\|_2^2)$. Let $\xi \sim \mathcal{N}(0, \sigma_z^2(1+2\epsilon)\|\widetilde{\Delta}\|_2^2)$, like (105), with probability at least $1-\frac{\delta}{4}$ we have:

$$\|\xi\|_2^2 \le \sigma_z^2(1+2\epsilon)\|\widetilde{\Delta}\|_2^2(2d+3\ln(\frac{4}{\delta})) \tag{109}$$

Then with probability at least $1-\frac{\delta}{4}$ we have the following inequality for all $\Delta \in \mathbb{R}^T$

$$\begin{aligned}
\|\hat{W}\Delta\|_2 &= \|\sum_{t=1}^{T}((\hat{B}^i)^\top\hat{\Sigma}_t^i\hat{B}^i)^{-1}(\hat{B}^i)^\top\hat{\Sigma}_t^iB^*w_t^*\Delta_t + \sum_{t=1}^{T}\frac{1}{n_t}((\hat{B}^i)^\top\hat{\Sigma}_t^i\hat{B}^i)^{-1}(\hat{B}^i)^\top(X_t^i)^\top Z_t\Delta_t\|_2 \\
&\geq |\|\sum_{t=1}^{T}((\hat{B}^i)^\top\hat{\Sigma}_t^i\hat{B}^i)^{-1}(\hat{B}^i)^\top\hat{\Sigma}_t^iB^*w_t^*\Delta_t\|_2 - \|\sum_{t=1}^{T}\frac{1}{n_t}((\hat{B}^i)^\top\hat{\Sigma}_t^i\hat{B}^i)^{-1}(\hat{B}^i)^\top(X_t^i)^\top Z_t\Delta_t\|_2| \\
&\geq |\frac{1-2\epsilon}{1+2\epsilon}\|W^*\Delta\|_2 - \frac{1}{1-2\epsilon}\|(\hat{B}^i)^\top(\sum_{t=1}^{T}\frac{1}{n_t}(X_t^i)^\top Z_t\Delta_t)\|_2| \\
&\geq |\frac{1-2\epsilon}{1+2\epsilon}\|W^*\Delta\|_2 - \frac{1}{1-2\epsilon}\|\sum_{t=1}^{T}\frac{1}{\sqrt{n_t}}U_tD_tQ_t\|_2| \\
&\geq |\frac{0.9}{1.1}\|W^*\Delta\|_2 - \frac{\sqrt{1.1}}{0.9}\sigma_z\|\Delta\|_2\sqrt{\frac{(2d+3\ln(\frac{4}{\delta}))}{\min_t(n_t)}}|, \qquad \text{(Conditions, (109))} \\
&\geq |\frac{0.9}{1.1}\sqrt{\kappa}\|\Delta\|_2 - \frac{\sqrt{1.1}}{0.9\times 4}\sqrt{\kappa}\|\Delta\|_2|, \qquad (n_t \geq 12\frac{\sigma_z^2}{\kappa}(d+\ln(\frac{4}{\delta}))) \\
&\geq 0.5\sqrt{\kappa}\|\Delta\|_2
\end{aligned} \tag{110}$$

From the definition of RE condition like Assumption E.1, we done the proof. □

Lemma E.10. *Let $q = \frac{\sqrt{k}R}{\underline{\sigma}}$ (so $q \geq \|\nu^*\|_1$). If $\gamma \geq \max\{2160sqC_W\Lambda^{-1}, \sqrt{2160sq\kappa\Lambda^{-1}}\}$, then*

$$\frac{2160}{\gamma}sq\max\{C_W, \frac{\kappa}{\gamma}\} \leq \|\nu^*\|_1 \tag{111}$$

Proof of Lemma E.10. Just note that if $\gamma \geq \max\{2160sqC_W\|\nu^*\|_1^{-1}, \sqrt{2160sq\kappa\|\nu^*\|_1^{-1}}\}$, then we can prove (111) by direct calculation. Then since $\|\nu^*\|_1 \geq \Lambda$ by definition, we get the result.

□

Proof of Theorem 3.10/E.2. Combine Theorem E.5 and Lemma E.10 and we can figure out the result like (90). For the estimation of $\beta_1 = T\beta$, we use the fact that $s \leq k$, $\|\nu\|_2 \leq R/\underline{\sigma}$ and $\|\nu\|_1 \geq \|\nu\|_2 \geq R/\overline{\sigma}$, then from the definition of γ, we can figure out that β_1 should be at least $\Theta(Tk^3C_W^6/\underline{\sigma}^6)$. □

F. Proof of Theorem 4.1

First, we rewrite the assumption and theorem formally.

Assumption F.1. (decreasing gradient) Assume f_t is a piecewise second-order differentiable function, and on each sub-function, it satisfies $f_t \geq 0, \nabla f_t \geq 0, \nabla^2 f_t \leq 0$ and $\nabla f_t(n_{t,1}+n_{t,2}) = \Omega(n_{t,2}^{-2+q})$ for some $q \in (0,2]$.

Remark F.2. Assumption F.1 covers a wide range of functions that may be used in practice, including the above example (20). The last upper bound constraint in Assumption F.1 shows that we need ∇f_t to decrease moderately, and it's used for our main theorem in this section.

And our main result for Section 4 is:

Theorem F.3. *Let $n_{t,1} \equiv n_1$ for all $t \in [T]$ and assume Assumption 2.1, 2.3, 2.4, 3.8, F.1 hold. Without loss of generality, we also assume $R = \Theta(1)$ and $C_W = \Theta(1)$ where C_W, R are defined in Assumption 3.8. Then denotes the optimal solution of (21) as $(n_{[T],2}^*, \nu^*)$, we have*

$$n_{t,2}^* = h_t(|\nu^*(t)|) \tag{112}$$

where h_t is a monotone increasing function that satisfies: $c_{t,1}x \leq h_t(x) \leq c_{t,2}x^{2/q}$ where $c_{t,1}, c_{t,2} > 0$ and q defined in Assumption F.1. Moreover, we claim A-MTRL algorithm with $n_{[T],[2]}^*$ sampling strategy is at least k-sparse task selection algorithm.

And we also rewrite the optimization problem (21) formally:

$$\begin{aligned}
\min_{n_{[T],2}} \quad & g(n_{[T],2}) := \sum_{t=1}^{T} f_t(n_{t,1}+n_{t,2}) \\
s.t. \quad & c_0(n_{[T],2},\nu) := \frac{\varepsilon^2}{C_{ER}\sigma^2 k(d+T)} - \sum_{t=1}^{T} \frac{\nu(t)^2}{n_{t,2}+n_{t,1}} \quad \geq 0, \\
& c_j(\nu) := \sum_{t=1}^{T} w^*_{j,t}\nu(t) - (w^*_{T+1})_j \quad = 0, \quad j \in [k] \\
& c_m(n_{[T],2}) := n_{m,2} \quad \geq 0, \quad m \in [T]
\end{aligned} \tag{113}$$

where $C_{ER} > 0$ is a constant.

Proof of Theorem F.3. Here we note that the main insight for such a theorem is that we want to prove the objective function is concave relative to ν. So we just prove for global second-order differentiable function and it can be easily generalized to the piecewise second-order differentiable function by showing the maintenance of concavity.

Step 1: Use KKT conditions to reduce the variable's number

Firstly we define the Lagrange function:

$$L(n_{[T],2},\nu) = g(n_{[T],2}) - \lambda_0 c_0(n_{[T],2}) - \sum_{j=1}^{k}\lambda_j c_j(\nu) - \sum_{m=1}^{T}\lambda_{m+k}c_m(n_{[T],2}) \tag{114}$$

Then from KKT conditions we have

$$\begin{aligned}
\left.\frac{\partial L}{\partial n_{t,2}}\right|_{n^*_{t,2},\nu^*(t)} &= \nabla f_t(n_{t,1}+n^*_{t,2}) - \lambda^*_0 \frac{\nu^*(t)^2}{(n^*_{t,2}+n_{t,1})^2} - \lambda^*_{t+k,2} = 0, \qquad \forall t \in [T] \\
\left.\frac{\partial L}{\partial \nu_t}\right|_{n^*_{t,2},\nu^*(t)} &= 2\lambda^*_0 \frac{\nu^*(t)}{n^*_{t,2}+n_{t,1}} - \sum_{j=1}^{k}\lambda^*_j w^*_{j,t} = 0, \qquad \forall t \in [T] \\
\lambda^*_0 &\geq 0, \qquad \lambda^*_0 c_0(n^*_{[T],2},\nu^*) = 0 \\
\lambda^*_{m+k} &\geq 0, \qquad \lambda^*_{m+k} c_m(n^*_{[T],2}) = 0, \qquad \forall m \in [T]
\end{aligned} \tag{115}$$

Note that when $n^*_{t,2} > 0$, $\lambda^*_{m+k} = 0$ and $\nabla f_t(n_{t,1}+n_{t,2}) = \Omega(n_{t,2}^{-2+q})$. then from the first equation of (115) we deduce (112) and its property immediately.

Also, with (112) we can reduce the number of variables of the original problem from $2T$ to $T+1$. To avoid confusion we denote $\alpha = \sqrt{\lambda_0}, \gamma(t) := \nu(t)$ for new optimization problem (116). It's clear that if the optimal solution of the original optimization problem (113) is $(\nu^*, n^*_{[T],2})$ and the corresponding lagrange coefficient for the first equality constraint of (113) is λ^*_0, then the optimal solution (γ^*, α^*) of the following problem (116) is equal to $(\nu^*, \sqrt{\lambda^*_0})$.

$$\begin{aligned}
\min_{\gamma,\alpha} \quad & l(\gamma,\alpha) := \sum_{t=1}^{T} f_t(n_{t,1} + h_t(\alpha|\gamma(t)|)) \\
s.t. \quad & d_0(\gamma,\alpha) := \frac{\varepsilon^2}{C_{ER}\sigma^2 k(d+T)} - \sum_{t=1}^{T}\frac{\gamma(t)^2}{h_t(\alpha|\gamma(t)|)+n_{t,1}} = 0 \\
& d_j(\gamma) := \sum_{t=1}^{T} w^*_{j,t}\gamma(t) - (w^*_{T+1})_j = 0, \quad j \in [k]
\end{aligned} \tag{116}$$

Step 2: The objective function of (116) is concave

From the KKT conditions above we know for any feasible solution (γ, α) and any $t \in [S]$, there exist a unique $x_t > 0$ such that $\alpha|\gamma(t)| = \sqrt{\nabla f_t(n_{t,1} + x)} \cdot (n_{t,1} + x)$. Then from the key Lemma F.4 we know the objective function of (116) is concave relative to $|\gamma(t)|$ for all $t \in [S]$.

Step 3: Analyze γ^* from the sub-problem of (116)

The first equality constraint of the problem (116) is non-linear relative to γ and α, which results that the feasible region of (116) having non-linear boundary. This makes it difficult for us to get the closed form of the optimal solution for (116).

Fortunately, the other equality constraints, which are equivalent to $W^*\gamma = w^*_{T+1}$, are not only linear but also have nothing to do with α. So we try to find out the optimal solution of sub-problem (117) and connect it to that of (116).

$$\begin{aligned} \min_{\xi} \quad & l(\xi, \alpha) := \sum_{t=1}^{T} f_t(n_{t,1} + h_t(\alpha|\xi(t)|)) \\ s.t. \quad & D(\xi) := W^*\xi - w^*_{T+1} = 0 \end{aligned} \tag{117}$$

In (117) α is taken as a given value and ξ plays the same role as γ as above. Define $opt(\alpha) : \mathbb{R} \to \Omega^*$, where Ω^* is the set of optimal solutions for (117) with given α.

Firstly we show that the optimal solution of (117) is k-sparse. From step 2 we know $l(\xi)$ is concave for any $|\xi(t)|, t \in [S]$, which means that the region contained by the isosurface of the objective function is concave where the axes are made up of $|\gamma(t)|$ for $t \in [S]$. Consequently, the solutions of the system of linear equations that minimize such a concave function will give out sparse results (Tibshirani, 1996).

Secondly, we say the optimal solution of the original optimization problem (116) is k-sparse. For a non-trivial case, where the algorithm achieves require performance and terminates at the first stage, we know $d_0(\gamma, 0) < 0$, and if $\alpha \to \infty$, $d_0(\gamma, \alpha) \to \frac{\varepsilon^2}{C_{ER}\sigma^2 k(d+T)} > 0$. Then from continuality of h_t we see that for any $\gamma^*(\alpha) \in opt(\alpha)$, there exist a unique α_0 such that $\gamma^*(\alpha_0)$ is a feasible solution for (116). On the other hand, every optimal solution (γ^*, α^*) of (116) should be the optimal solution of sub-problem (117),i.e. it should satisfy $\gamma^* \in opt(\alpha^*)$. Thus γ^* is k-sparse, and so as ν^*. Therefore A-MTRL with $n^*_{[T],[2]}$ strategy is k-sparse task selection algorithm.

□

Lemma F.4. *Assume $f_t, h_t, n_{t,1}$ follow the conditions and results in Theorem 4.1, $W^* \in \mathbb{R}^{k\times T}$, $w^*_{T+1} \in \mathbb{R}^k$. Then if for any feasible solution (γ, α) of (116), any $t \in [S]$, there exist a unique $x_t > 0$ such that $\alpha|\gamma(t)| = \sqrt{\nabla f_t(n_{t,1} + x)} \cdot (n_{t,1} + x)$, then the objective function of (116) relative to $|\gamma(t)|$ is concave for all $t \in [S]$.*

Proof of Lemma F.4.

Firstly we denote $n_{t,1}$ as n for convenience. Note that from the chain rule:

$$\frac{\partial l(\gamma, \alpha)}{\partial |\gamma(t)|} = \nabla f_t(n + h_t(\alpha|\gamma(t)|)) \cdot \nabla h_t(\alpha|\gamma(t)|)) \cdot \alpha \tag{118}$$

Clearly $l(\gamma, \alpha)$ is also monotone increasing relative to $|\gamma(t)|$. For the second order of $l(\gamma, \alpha)$ we have:

$$\frac{\partial^2 l(\gamma, \alpha)}{\partial |\gamma(t)|^2} = \{\nabla^2 f_t(n + h_t(\alpha|\gamma(t)|)) \cdot (\nabla h_t(\alpha|\gamma(t)|))^2 + \nabla f_t(n + h_t(\alpha|\gamma(t)|)) \cdot \nabla^2 h_t(\alpha|\gamma(t)|)\} \cdot \alpha^2 \tag{119}$$

Firstly we need to figure out the relation between the derivative of h_t and f_t. From the first equation of (115) and the definition of h_t we have:

$$h_t(\sqrt{\nabla f_t(n + x)} \cdot (n + x)) = x \tag{120}$$

Since h_t is monotone contineous function, from inverse function theory we have

$$\nabla h_t(\sqrt{\nabla f_t(n + x)} \cdot (n + x)) = \frac{2\sqrt{\nabla f_t(n + x)}}{(n + x)\nabla^2 f_t(n + x) + 2\nabla f_t(n + x)} \tag{121}$$

Let $g(x) := \sqrt{\nabla f_t(n+x)} \cdot (n+x)$, from assumption F.1 we know g is a continous monotone increasing function and $g \in (0, +\infty)$. Besides, from conditions we have that for each $t \in [S]$ there is a unique $x := x_t > 0$ such that $g(x_t) = \alpha|\gamma(t)|$, with which we can simplify the gradient:

$$\begin{aligned}\nabla^2 h_t(\alpha|\gamma(t)|) &= \nabla^2 h_t(\sqrt{\nabla f_t(n+x)} \cdot (n+x)) \\ &= d(\frac{2\sqrt{\nabla f_t(n+x)}}{(n+x)\nabla^2 f_t(n+x) + 2\nabla f_t(n+x)})/dx \cdot \nabla h_t(\sqrt{\nabla f_t(n+x)} \cdot (n+x)) \\ &= 2\frac{(\nabla^2 f_t(n+x))^2(n+x) - 4\nabla^2 f_t(n+x)\nabla f_t(n+x) - 2(n+x)\nabla^3 f_t(n+x)\nabla f_t(n+x)}{[(n+x)\nabla^2 f_t(n+x) + 2\nabla f_t(n+x)]^3}\end{aligned} \tag{122}$$

Denote $h_t^1 := \nabla h_t(\sqrt{\nabla f_t(n+x)} \cdot (n+x)), h_t^2 := \nabla^2 h_t(\sqrt{\nabla f_t(n+x)} \cdot (n+x))$. Thus we have:

$$\begin{aligned}\frac{1}{\alpha^2}\frac{\partial^2 l(\gamma,\alpha)}{\partial|\gamma(t)|^2} &= \nabla^2 f_t(n+x)(\nabla h_t(\sqrt{\nabla f_t(n+x)}(n+x)))^2 + \nabla f_t(n+x)\nabla^2 h_t(\sqrt{\nabla f_t(n+x)}(n+x)) \\ &= h_t^1 \cdot \frac{\sqrt{\nabla f_t(n+x)}(n+x)}{[(n+x)\nabla^2 f_t(n+x) + 2\nabla f_t(n+x)]^2} \cdot \{3(\nabla^2 f_t(n+x))^2 - 2\nabla^3 f_t(n+x)\nabla f_t(n+x)\} \\ &= 2\nabla f_t(n+x)(n+x) \cdot \frac{3(\nabla^2 f_t(n+x))^2 - 2\nabla^3 f_t(n+x)\nabla f_t(n+x)}{[(n+x)\nabla^2 f_t(n+x) + 2\nabla f_t(n+x)]^3} \\ &= 2\nabla f_t(n+x)(n+x) \cdot q(x), \qquad (q(x) := \frac{3(\nabla^2 f_t(n+x))^2 - 2\nabla^3 f_t(n+x)\nabla f_t(n+x)}{[(n+x)\nabla^2 f_t(n+x) + 2\nabla f_t(n+x)]^3})\end{aligned} \tag{123}$$

So if $q(x) < 0$ holds for all $x > 0$, we finish the proof. First we assume that $\nabla f_t(y) = \frac{A_t}{(B_t+y)^\delta}$ where $A_t > 0, B_t \geq 0$ and $\delta \in [0, 2-q)$. Then

$$q(x) = \frac{3\frac{\delta^2 A_t^2}{(n+x+B_t)^{2\delta+2}} - 2\frac{\delta(\delta+1)A_t^2}{(n+x+B_t)^{\delta+3+\delta}}}{\frac{2A_t}{(n+x+B_t)^\delta} - \frac{\delta A_t(n+x)}{(n+x+B_t)^{\delta+1}}} = \frac{A_t}{(n+x+B_t)^{\delta+1}} \cdot \frac{\delta(\delta-2)}{2B_t + (2-\delta)(n+x)} \tag{124}$$

Since $n + x > 0$ and $0 \leq \delta < 2$, we have $q(x) < 0, \forall x > 0$. Besides, due to the fact that ∇f_t is monotone decreasing and non-negative, together with Assumption F.1 and $n > 0$, we can find $\delta_i \in [0, 2-q), A_{t,i} > 0, B_{t,i} \geq 0$ for $i = 1, 2$ such that $\frac{A_{t,1}}{(B_{t,1}+x+n)^{\delta_1}} \leq \nabla f_t(x+n) \leq \frac{A_{t,2}}{(B_{t,2}+x+n)^{\delta_2}}$. So $q(x) < 0$ holds for any ∇f_t that satisfies Assumption F.1.

□

Remark F.5. If δ in (124) is in $(0, 2)$, then the optimization problem (113) is not computable.

G. Supplements to the Experiments Section

G.1. Explanation of k-task selection scenario

We provide an illustration of our intuition for the k-task selection scenario in Section 5. We emphasize that the specific choice of the cost function is not critical in such a scenario, since solving the exact optimization problem (Eqn. 21) can be computationally challenging. For instance, the cost functions could correspond to L_p-minimization ($0 \leq p < 1$) solutions of the relation equation $W^*\nu = w_{T+1}^*$, which is known to be NP-hard.

To address this challenge, as discussed in Theorem 4.1, we employ L1-A-MTRL as an approximation to the optimal solution of (21). This approach is justified by the fact that the time complexity for solving the approximate solution of (21) using L1-A-MTRL with relative accuracy δ is just $\text{poly}(T)\ln(T/\delta)$ from (Cohen et al., 2021), and the L_1-minimization solution is also k-sparse. Therefore, in cost-sensitive scenarios, our main focus is on addressing the question: "How well can active multi-task representation learning algorithms perform when no more than k tasks are available for further sampling?" This leads us to the setting of the k-task selection scenario.

G.2. Details of Algorithm Implementation.

In practice, $\hat{W}$ and $\hat{w}_{T+1}$ may differ at different epochs after the model converges due to the noise of data points. So to enhance the stability of $\hat{\nu}$, we calculate $\hat{\nu}$ at every epoch in the last 20 rounds and take their average as the final reference to

Algorithm 2 Multi-Stage L1-A-MTRL Method

Input: confidence δ, representation function class Φ, stage number S, scaling $L > 1$, minimum singular value $\underline{\sigma}$
Initialize $\underline{N} = \beta_1/T$ from (15) and $\hat{\nu}^1 = [1, 1, ...]$.
for $i = 1$ **to** S **do**
  Set $n_t^i = \max\{\beta_i |\hat{\nu}^i(t)| \cdot \|\hat{\nu}^i\|_1^{-1}, \underline{N}\}$.
  For each task t, draw n_t i.i.d samples from the offline dataset denoted as $\{X_t^i, Y_t^i\}_{t=1}^T$
  Estimate $\hat{\phi}^i, \hat{W}^i$ on the source tasks with Eqn.(2)
  Estimate $\hat{w}_{T+1}^i$ on the target task with Eqn.(3)
  Estimate ν^{i+1} by Lasso Program (16)
  Set $\beta_{i+1} = \beta_i \cdot L$
end for

calculate $n_{[T]}$ for both our algorithm and baselines, while the total number of epochs at each stage is no less than 2000. For full tasks scenario, note that L2-A-MTRL(Chen et al., 2022) utilize the iterative L2-A-MTRL algorithm with 4 stages to optimize the model we also run our algorithm iteratively with 4 stages for comparison, and the detailed procedure for multi-stage learning is in Algorithm 2. We mention that Chen's method requires multiple stages but we allow both single-stage (Algorithm 1) and multi-stage (Algorithm 2) versions.

Here we set $\underline{N} = 100$. We sample 500 data from the target task, while at the final stage, we sample around 30000 to 40000 data from the source tasks. For k-task selection scenario, we run the algorithm with 2 stages. Here we set $\underline{N} = 40$. We sample 200 data from the target task and around 12000 data from the source tasks.

G.3. How to choose λ_k

Determining the optimal value of λ_k requires additional knowledge of $\underline{\sigma} = \sigma_{\min}(W^*)$, which are dataset-dependent prior parameters. To address this, we explore two approaches to determine λ_k in our experiments:

- Tuning way: We roughly tune λ_k exponentially for the 2-phase L1-A-MTRL approach (Algorithm 1). And to further obtain the optimal λ_k, we can utilize grid search to find better λ_k. Once we identify a good λ_k, we can run the multi-phase L1-A-MTRL algorithm (Algorithm 2) using that λ_k and a larger N_{tot} to achieve improved results.

- Lazy way: Alternatively, we can simply choose a very small value for λ_k, such as 10^{-10}, for our algorithm.

To provide a clearer illustration of the first approach, we apply the 2-phase L1-A-MTRL on the *identity_9* dataset in full task scenarios, where $k = 50$ and $T = 150$. In the first phase, each source task is assigned $\underline{N} = 100$ data points, and in the second phase, the total budget for the source data is $N_{tot} = 33k$. The results are presented in Table 1.

Table 1. The relevance between λ_k and the second-stage test error on the target task *identity_9*.

λ_k	1.0	10^{-1}	10^{-2}	10^{-3}	10^{-4}	2×10^{-4}	10^{-5}	10^{-6}	10^{-8}	10^{-10}	10^{-16}
Error	0.0691	0.0690	0.0703	0.0694	**0.0561**	0.0570	0.0655	0.0655	0.0633	0.0631	0.0625

The optimal value for λ_k is approximately 10^{-4}. Additionally, we observe that, except for the terms 10^{-4} and 2×10^{-4}, the target error decreases as λ_k decreases. For other target tasks, although we don't find an optimal λ_k similar to that of *identity_9*, we consistently observe that smaller values of λ_k lead to better performance for L1-A-MTRL. We think this phenomenon can be attributed to our Theorem 3.7, which considers a worst-case scenario where the noise may be significant. However, in practice, smaller values of λ_k are often sufficient to control the noise. Furthermore, since smaller values of λ_k result in a smaller bias when solving the Lasso program, L1-A-MTRL with small λ_k consistently exhibits good performance.

Therefore, to save time and resources, we adopt the lazy way instead of the tuning way in the experiments presented in this paper. We set $\lambda_k = 10^{-10}$, and the empirical results demonstrate that L1-A-MTRL with such a small value of λ_k still achieves excellent performance.

G.4. Additional Experiments on Sampling Budgets

To better show the empirical difference of the sampling budget in the experiments of MNIST-C, we consider the full task scenario (mentioned in Sec. 5) and evaluate the model performance by utilizing the 5-epoch L1-A-MTRL (Algorithm 2) with fixed minimum sampling data from every source task $\underline{N} = 100$ and increasing total sampling number N_{tot}. Due to the limited time and resources, we randomly select two target tasks *shear_1* and *identity_9*, and obtained the results in Table 2.

From Table 2 we find that to achieve accuracy higher than 95% on the *shear_1* target task, P-MTRL (passive sampling) requires more than $86k$ source data, L2-A-MTRL(Chen et al., 2022) requires about $61k$ source data and L1-A-MTRL just requires about $33k$ source data. Since at the later phase, we can reuse the evenly sampled data ($T\underline{N} = 15k$ in total) from the first phase, L1-A-MTRL just requires labeling additional $18k$ source data at the later phase to achieve 95% accuracy, while L2-A-MTRL requires approximately $46k$ extra data, and P-MTRL requires no less than $71k$ extra data. Similar results apply to *identity_9* . To achieve an accuracy above 93.7% on the *identity_9* target task, P-MTRL requires more than $95k$ source data, L2-A-MTRL(Chen et al., 2022) requires about $61k$ source data, while L1-A-MTRL requires only about $33k$ source data. The above results illustrate the effectiveness of our L1-A-MTRL algorithm.

Table 2. Test error on the target task *shear_1* and *identity_9* with different N_{tot}.

shear_1	N_{tot}				
Algorithms	**15000**	**32850**	**44000**	**60700**	**86000**
P-MTRL	0.0544	0.0538	0.0536	0.0520	0.0518
L2-A-MTRL(Chen et al. (2022))	0.0544	0.0511	0.0519	0.0494	0.0488
L1-A-MTRL(**Ours**)	0.0544	**0.0496**	**0.0478**	**0.0442**	**0.0428**
identity_9	N_{tot}				
Algorithms	**15000**	**33000**	**43800**	**60900**	**95400**
P-MTRL	0.0932	0.0834	0.0778	0.0738	0.0652
L2-A-MTRL(Chen et al. (2022))	0.0932	0.0909	0.0638	0.0627	0.0621
L1-A-MTRL(**Ours**)	0.0932	**0.0631**	**0.0620**	**0.0605**	**0.0551**