Towards Sample-efficient Overparameterized Meta-learning

Yue Sun
University of Washington
yuesun@uw.edu

Adhyyan Narang
University of Washington
adhyyan@uw.edu

Halil Ibrahim Gulluk
Bogazici University
hibrahimgulluk@gmail.com

Samet Oymak
University of California, Riverside
oymak@ece.ucr.edu

Maryam Fazel
University of Washington
mfazel@uw.edu

Abstract

An overarching goal in machine learning is to build a generalizable model with few samples. To this end, overparameterization has been the subject of immense interest to explain the generalization ability of deep nets even when the size of the dataset is smaller than that of the model. While the prior literature focuses on the classical supervised setting, this paper aims to demystify overparameterization for meta-learning. Here we have a sequence of linear-regression tasks and we ask: (1) Given earlier tasks, what is the optimal linear representation of features for a new downstream task? and (2) How many samples do we need to build this representation? This work shows that surprisingly, overparameterization arises as a natural answer to these fundamental meta-learning questions. Specifically, for (1), we first show that learning the optimal representation coincides with the problem of designing a task-aware regularization to promote inductive bias. We leverage this inductive bias to explain how the downstream task actually benefits from overparameterization, in contrast to prior works on few-shot learning. For (2), we develop a theory to explain how feature covariance can implicitly help reduce the sample complexity well below the degrees of freedom and lead to small estimation error. We then integrate these findings to obtain an overall performance guarantee for our meta-learning algorithm. Numerical experiments on real and synthetic data verify our insights on overparameterized meta-learning.

Organization of the appendix

The appendix consists of the proof of our main results including the following parts:

- We included a short section and Figure 6 containing more experiments on real data. This verifies the positive correlation between the canonical task covariance and feature covariance across distinct datasets which supports the theory developed in Section 4.
- Optimal representation. The proof for optimal overparameterized representation is in Sec. B. We show that we can use an R dimensional representation of feature for few-shot learning, and it can beat typical PCA (low dimensional/underparameterized) representation with optimal weighting matrix $\mathbf{\Lambda}^*$.
  - In Sec. B.1 we first prove Observation 1. In Remark 2 we analyze the projection from d to R dimensional space, where we calculate the PCA truncation noise.
  - In Sec. B.2 and B.3 we provide the asymptotic analysis of optimal weighting. By asymptotic we refer to the regime where $n_2, d \to \infty$ and the eigenvalues of task and feature covariance

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

  - matrices converge to a fixed distribution. We show that $\hat{\beta}_{\boldsymbol{\Lambda}}$ converges to a Gaussian distribution parameterized by $\boldsymbol{\Lambda}$, and use it to express the risk.
  - We extend the asymptotic case (infinite dimensional) to the non-asymptotic (finite dimensional) regime in Sec. B.4. We define the risk function with respect to representation matrix $\boldsymbol{\Lambda}$, and in Sec. B.5 solve for the optimal representation by minimizing risk.
- Representation learning. Sec. C includes the proof for the result about representation learning in Sec. 4, including the sample complexity and error guarantee of MoM estimators.
  - We first analyze the estimation of feature covariance matrix $\boldsymbol{\Sigma}_F$ in Sec. C.1 which is the most straightforward.
  - We prove the second result of Thm. 2 in Sec. C.2.2. With the assumption that each task has $\Omega(s_T)$ corresponding samples, the sample complexity is **reduced by a factor of s_T compared to MoM**, which meets the *information theoretical lower bound* in [37].
  - We extend the Bernstein type technique for obtaining the estimation error of $\hat{\boldsymbol{M}}$ in Sec. C.2. The estimator given in [37], slightly different from ours, is also analyzed.
- End to end bound. We prove the robustness of the optimal representation in Sec. D, which leads to the overall error guarantee of the proposed meta-learning algorithm.

Contents

1 Introduction

In a multitude of machine learning (ML) tasks with limited data, it is crucial to build accurate models in a sample-efficient way. Constructing a simple yet informative representation of features is a critical component of learning a model that generalizes well to an unseen test set. The field of meta-learning dates back to [9, 5] and addresses this challenge by transferring insights across distinct but related tasks. Usually, the meta-learner first (1) learns a feature-representation from previously seen tasks and then (2) uses this representation to succeed at an unseen task. The first phase is called representation learning and the second is called few-shot learning. Such information transfer between tasks is the backbone of modern transfer and multitask learning and finds ubiquitous applications in image classification [15], machine translation [7] and reinforcement learning [18].

Recent literature in ML theory has posited that overparameterization can be beneficial to generalization in traditional single-task setups for both regression [29, 40, 4, 33, 30] and classification [32, 31] problems. Empirical literature in deep learning suggests that overparameterization is of interest for both phases of meta-learning as well. Deep networks are stellar representation learners despite containing many more parameters than the sample size. Additionally, overparameterization is observed to be beneficial in the few-shot phase for transfer-learning in Figure 1(a). A ResNet-50 network pretrained on Imagenet was utilized to obtain a representation of R features for classification on CIFAR-10. All layers except the final (softmax) layer are frozen and are treated as a fixed feature-map. We then train the final layer of the network for the downstream task which yields a linear classifier on pretrained features. The figure plots the effect of increasing R on the test error on CIFAR-10, for different choices of training size n_2. For each choice of n_2, increasing R beyond n_2 is seen to reduce the test-error. These findings are corroborated by [18] (MAML) and [39], who successfully use a transfer learning method that adapts a pre-trained model, with 112980 parameters, to downstream tasks with only 1-5 new training samples.

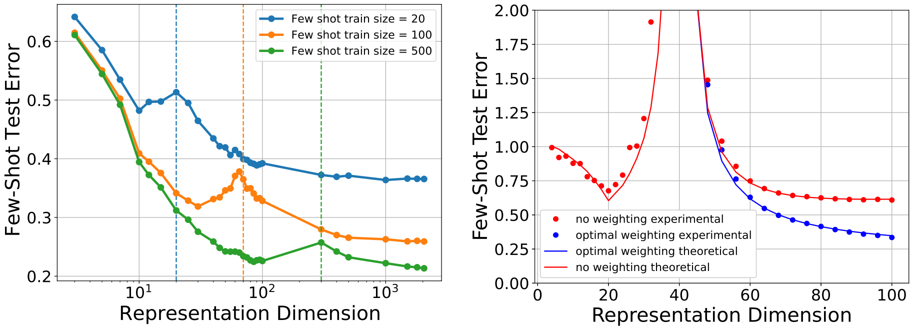

Figure 1: **Illustration of the benefit of overparameterization in the few-shot phase.** (a) Double-descent in transfer learning: dashed lines indicate the location where the number of features R exceed the number of training points; i.e., the transition from under to over-parameterization. The experimental details are contained in the supplement. (b) Illustration of the benefit of using Weighted minL2-interpolation in Definition 3 (blue). See Remark 1 for details and discussion.

In Figure 1(b), we consider a sequence of *linear* regression tasks and plot the few-shot error of our proposed projection and eigen-weighting based meta-learning algorithm for a fixed few-shot training size, but varying dimensionality of features. The resulting curve looks similar to Figure 1(a) and suggests that the observations regarding overparameterization for meta-learning in neural networks can, to a good extent, be captured by linear models, thus motivating their detailed study. This aligns with trends in recent literature: while deep nets are nonlinear, recent advances show that linearized problems such as kernel regression (e.g., via neural tangent kernel [21, 17, 24, 35, 13]) provide a good proxy to understand some of the theoretical properties of practical overparameterized deep nets.

However, existing analysis of subspace-based meta-learning algorithms for both the representation learning and few-shot phases of linear models have typically focused on the classical *underparameterized regime*. These works (see Paragraphs 2-3 of Sec. 1.2) consider the case where representation learning involves projection onto a lower-dimensional subspace. On the other hand, recent works on double descent shows that an *overparameterized* interpolator beats PCA-based method. to build upon these results to develop a theoretical understanding of overparameterized meta-learning.

1.1 Our contributions

This paper studies meta-learning when each task is a linear regression problem, similar in spirit to [37, 23]. In the representation learning phase, the learner is provided with training data from T distinct tasks, with n_1 training samples per task: using this data, it selects a matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d\times R}$ with arbitrary R to obtain a linear *representation* of features via the map $\boldsymbol{x} \to \boldsymbol{\Lambda}^\top \boldsymbol{x}$. In the few-shot learning phase, the learner faces a new task with n_2 training samples and aims to use the representation $\boldsymbol{\Lambda}^\top \boldsymbol{x}$ to aid prediction performance.

We highlight that obtaining the representation consists of two steps: first the learner projects $\boldsymbol{x}$ onto R basis directions, and then performs *eigen-weighting* of each of these directions, as shown in Figure 2(b). The overarching goal of this paper is to propose a scheme to use the knowledge gained from earlier tasks to choose $\boldsymbol{\Lambda}$ that minimizes few-shot risk. This goal enables us to engage with important questions regarding overparameterization:

Q1: What should the size R and the representation $\boldsymbol{\Lambda}$ be to minimize risk at the few-shot phase?

Q2: Can we learn the Rd dimensional representation $\boldsymbol{\Lambda}$ with $N \ll Rd$ samples?

The answers to the questions above will shed light on whether overparameterization is beneficial in few-shot learning and representation learning respectively. Towards this goal, we make several contributions to the finite-sample understanding of *linear* meta-learning, under assumptions discussed in Section 2. Our results are obtained for a general data/task model with *arbitrary task covariance* $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ *and feature covariance* $\boldsymbol{\Sigma}_F$ which allows for a rich set of observations.

Optimal representation for few-shot learning. As a stepping stone towards the goal of characterizing few-shot risk for different $\boldsymbol{\Lambda}$, in Section 3 we first consider learning with **known covariances** $\boldsymbol{\Sigma}_T$ and $\boldsymbol{\Sigma}_F$ respectively (Algorithm 1). Compared to projection-only representations in previous works (see Paragraphs 2-3 of Sec. 1.2), our scheme applies *eigen-weighting* matrix $\boldsymbol{\Lambda}^*$ to incentivize the optimizer to place higher weight on promising eigen-directions. This eigen-weighting procedure has been shown in the single-task case to be extremely crucial to avail the benefit of overparameterization [6, 30, 33]: it captures an inductive bias that promotes certain features and demotes others. We show that the importance of eigen-weighting extends to the multi-task case as well.

Canonical task covariance. Our analysis in Section 3 also reveals that, the optimal subspace and representation matrix are closed-form functions of the *canonical task covariance* $\tilde{\boldsymbol{\Sigma}}_T = \boldsymbol{\Sigma}_F^{1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F^{1/2}$, which captures the feature saliency by summarizing the feature and task distributions.

Representation learning. In practice, task and feature covariances (and hence the canonical covariance) are rarely known apriori. However, we can estimate the principal subspace of the canonical task covariance $\tilde{\boldsymbol{\Sigma}}_T$ (which has a degree of freedom (DoF) of $\Omega(Rd)$) from data. In Section 4 we first present empirical evidence that feature covariance $\boldsymbol{\Sigma}_F$ is "positively correlated" with $\tilde{\boldsymbol{\Sigma}}_T$. Then we propose an efficient algorithm based on Method-of-Moments (MoM), and show that the sample complexity of representation learning is well below $\mathcal{O}(Rd)$ due to the inductive bias. Our sample complexity bound depends on interpretable quantities such as *effective ranks* $\boldsymbol{\Sigma}_F, \tilde{\boldsymbol{\Sigma}}_T$ and improves over prior art (e.g., [23, 37]), even though the prior works were specialized to low-rank $\tilde{\boldsymbol{\Sigma}}_T$ and identity $\boldsymbol{\Sigma}_F$ (see Table 2).

$\boldsymbol{\Sigma}_F$	Feature covariance
$\boldsymbol{\Sigma}_T$	Task covariance
$\tilde{\boldsymbol{\Sigma}}_T$	Canonical task covariance
n_1	Samples per each earlier task
T	Number of earlier tasks
N	Total sample size $T \times n_1$
n_2	Samples for new task
$\boldsymbol{\Lambda}$	Eigen-weighting matrix

Table 1: Main notation

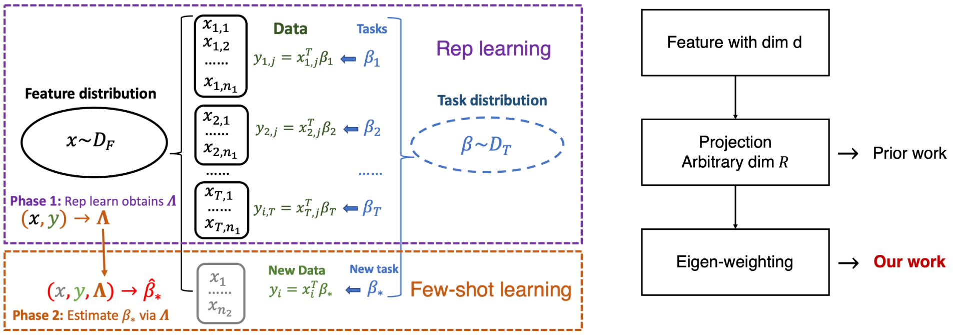
Feature distribution
$x \sim D_F$
Phase 1: Rep learn obtains Λ
$(x, y) \to \Lambda$
$(x, y, \Lambda) \to \hat{\beta}_*$
Phase 2: Estimate β_* via Λ
Data
Tasks
$x_{1,1}$ $x_{1,2}$ …… x_{1,n_1}
$y_{1,j} = x_{1,j}^T\beta_1$ ⬅ β_1
$x_{2,1}$ …… x_{2,n_1}
$y_{2,j} = x_{2,j}^T\beta_2$ ⬅ β_2
$x_{T,1}$ …… x_{T,n_1}
$y_{i,T} = x_{T,j}^T\beta_T$ ⬅ β_T
Rep learning
Task distribution
$\beta \sim D_T$
New Data
New task
x_1 …… x_{n_2}
$y_i = x_i^T\beta_*$ ⬅ β_*
Few-shot learning
Feature with dim d
Projection
Arbitrary dim R
→ Prior work
Eigen-weighting
→ Our work

Figure 2: (a) Steps of the meta-learning algorithm. (b) Our representation-learning algorithm has two steps: projection and eigen-weighting. We focus on the use of overparameterization+weighting matrix (Def. 3), and compare this with overparameterization with simple projection (no eigen-weighting), and underparameterization (for which eigen-weighting has no impact and is equivalent to projection). [37, 23, 22, 16] study underparameterized projections only. To distinguish from eigen-weighting, we will refer to simple projections as subspace-based representations.

End to end meta-learning guarantee. In Section 5, we consider the generalization of Section 3, where we have only estimates of the covariances instead of perfect knowledge. This leads to an overall meta-learning guarantee in terms of $\mathbf{\Lambda}^*$, N and n_2 and uncovers a bias-variance tradeoff: As N decreases, it becomes more preferable to use a smaller R (more bias, less variance) due to inaccurate estimate of the weak eigen-directions of $\tilde{\mathbf{\Sigma}}_T$. In other words, we find that overparameterization is only beneficial for few-shot learning if the quality of representation learning is sufficiently good. This explains why, in practice, increasing the representation dimension may not help reduce few-shot risk beyond a certain point (see Fig. 5).

1.2 Related work

Overparameterized ML and double-descent The phenomenon of double-descent was first discovered by [6]. This paper and subsequent works on this topic [4, 33, 32, 30, 11] emphasize the importance of the right prior (sometimes referred to as inductive bias or regularization) to avail the benefits of overparameterization. However, an important question that arises is: where does this prior come from? Our work shows that the prior can come from the insights learned from related previously-seen tasks. Section 3 extends the ideas in [34, 40] to depict how the optimal representation described can be learned from imperfect covariance estimates as well.

Theory for representation learning Recent papers [23, 22, 37, 16] propose the theoretical bounds of representation learning when the tasks lie in an exactly r dimensional subspace. [23, 22, 37] discuss method of moment estimators and [37, 16] discuss matrix factorized formulations. [37] shows that the number of samples that enable meaningful representation learning is $\mathcal{O}(dr^2)$. [23, 22, 37] assume the features follow a standard normal distribution. We define a canonical covariance which handles arbitrary feature and task covariances. We also show that our estimator succeeds with $\mathcal{O}(dr)$ samples when $n_1 \sim r$, and extend the bound to general covariances with effective rank defined.

Subspace-based meta learning With tasks being low rank, [23, 22, 37, 19, 16] do few-shot learning in a low dimensional space. [41, 42] study meta-learning for linear bandits. [27] gives information theoretic lower and upper bounds. [8] proposes subspace-based methods for nonlinear problems such as classification. We investigate a representation with arbitrary dimension, specifically interested in overparameterized case and show it yields a smaller error with general task/feature covariances. Related work [16] provides results on overparameterized representation learning, but [16] requires number of samples per pre-training task to obey $n_1 \gtrsim d$, whereas our results apply as soon as $n_1 \gtrsim 1$.

Mixed Linear Regression (MLR) In MLR [43, 25, 12], multiple linear regression are executed, similar to representation learning. The difference is that, the tasks are drawn from a finite set, and number of tasks can be larger than d and not necessarily low rank. [26, 10, 28] propose sample complexity bounds of representation learning for mixed linear regression. They can be combined with other structures such as binary task vectors [3] and sparse task vectors [2].

2 Problem Setup

The problem we consider consists of two phases:

1. Representation learning: Prior tasks are used to learn a suitable representation to process features.
2. Few-shot learning: A new task is learned with a few samples by using the suitable representation.

This section defines the key notations and describes the data generation procedure for the two phases. In summary, we study linear regression tasks, the features and tasks are generated randomly, i.i.d. from their associated distributions $\mathcal{D}_T$ and $\mathcal{D}_F$, and the two phases share the same feature and task distributions.The setup is summarized in Figure 2(a).

2.1 Data generation

Definition 1 (Task and feature distributions) *Throughout, $\mathcal{D}_T$ and $\mathcal{D}_F$ denote the distributions of tasks $\boldsymbol{\beta}_i$ and features $\boldsymbol{x}_{ij}$ respectively. These distributions are subGaussian, zero-mean with corresponding covariance matrices $\boldsymbol{\Sigma}_T$ and $\boldsymbol{\Sigma}_F$.*

Definition 2 (Data distribution for a single task) *Given a specific realization of task vector $\boldsymbol{\beta} \sim \mathcal{D}_T$, the corresponding label/input distribution $(y, \boldsymbol{x}) \sim \mathcal{D}_{\boldsymbol{\beta}}$ is obtained via $y = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon$ where $\boldsymbol{x} \sim \mathcal{D}_F$ and ε is zero-mean subgaussian noise with variance σ^2.*

Data for Representation Learning (Phase 1). We have T tasks, each with n_1 training examples. The task vectors $(\boldsymbol{\beta}_i)_{i=1}^T \subset \mathbb{R}^d$ are drawn i.i.d. from the distribution $\mathcal{D}_T$. The data for ith task is given by $(y_{ij}, \boldsymbol{x}_{ij})_{j=1}^{n_1} \overset{\text{i.i.d.}}{\sim} \mathcal{D}_{\boldsymbol{\beta}_i}$. In total, there are $N = T \times n_1$ examples.

Data for Few-Shot Learning (Phase 2). Sample task $\boldsymbol{\beta}_\star \sim \mathcal{D}_T$. Few-shot dataset has n_2 examples $(y_i, \boldsymbol{x}_i)_{j=1}^{n_2} \overset{\text{i.i.d.}}{\sim} \mathcal{D}_{\boldsymbol{\beta}_\star}$.

We use representation learning data to learn a representation of feature-task distribution, called eigen-weighting matrix $\boldsymbol{\Lambda}$ in Def. 3 below. The matrix $\boldsymbol{\Lambda}$ is passed to few-shot learning stage, helping learn $\boldsymbol{\beta}_\star$ with few data.

2.2 Training in Phase 2

We will define a weighted representation, called eigen-weighting matrix, and show how it is applied for few-shot learning. The matrix is learned during representation learning using the data from the T tasks. Denote $\boldsymbol{X} \in \mathbb{R}^{n_2 \times d}$ whose i^{th} row is $\boldsymbol{x}_i$, and $\boldsymbol{y} = [y_1, ..., y_m]^\top$. We are interested in studying the weighted 2-norm interpolator defined below for overparameterization regime $R \geq n_2$.

Definition 3 (Eigen-weighting matrix and Weighted ℓ_2-norm interpolator) *Let the representation dimension be R, where R is any integer between 1 and d. We define an eigen-weighting matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times R}$ and the associated weighted ℓ_2-norm interpolator*

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\Lambda}^\dagger \boldsymbol{\beta}\|_2 \quad \textit{s.t.} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} \quad \textit{and} \quad \boldsymbol{\beta} \in \text{range\_space}(\boldsymbol{\Lambda}).$$

The solution is equivalent to defining $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}^\dagger \hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}}$ and solving an unweighted minimum 2-norm regression with features $\boldsymbol{X}\boldsymbol{\Lambda}$. This corresponds to our few-shot learning problem

$$\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_2 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{\alpha}$$

from which we obtain $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}}$. When there is no confusion, we can replace $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}}$ with $\hat{\boldsymbol{\beta}}$. One can easily see that $\hat{\boldsymbol{\beta}} = \boldsymbol{\Lambda}(\boldsymbol{X}\boldsymbol{\Lambda})^\dagger \boldsymbol{y}$. We note that Definition 3 is a special case of the weighted ridge regression discussed in [40], as stated in Observation 1. An alternative equivalence between min-norm interpolation and ridge regression can be found in [33].

Observation 1 *Let $\boldsymbol{X} \in \mathbb{R}^{n_2 \times d}$ and $\boldsymbol{y} \in \mathbb{R}^{n_2}$, define*

$$\hat{\boldsymbol{\beta}}_1 = \lim_{t \to 0} \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + t\boldsymbol{\beta}^\top (\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^\dagger \boldsymbol{\beta}, \ \boldsymbol{\beta} \in \text{column space of } \boldsymbol{\Lambda}. \tag{2.1}$$

We have that $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}$.

Algorithm 1 Constructing the optimal representation

Require: Projection dimension R, noise level σ, canonical covariance $\tilde{\mathbf{\Sigma}}_T$, task covariance $\mathbf{\Sigma}_F$.

```
function ComputeOptimalRep($R, \mathbf{\Sigma}_F, \tilde{\mathbf{\Sigma}}_T, \sigma, n_2$)
    $\boldsymbol{U}_1, \mathbf{\Sigma}_F^R, \tilde{\mathbf{\Sigma}}_T^R, \sigma_R$ = ComputeReduction($R, \mathbf{\Sigma}_F, \tilde{\mathbf{\Sigma}}_T, \sigma$)
    Optimization: Get $\boldsymbol{\theta}^*$ from (OPT-REP).
    Map to eigenvalues: Set diagonal $\mathbf{\Lambda}_R^* \in \mathbb{R}^{R \times R}$ with entries $\mathbf{\Lambda}_{R,i}^* = (1/\boldsymbol{\theta}_i^* - 1)^{-2}$.
    Lifting and feature whitening: $\mathbf{\Lambda}^* \leftarrow \boldsymbol{U}_1 (\mathbf{\Sigma}_F^R)^{-1/2} \mathbf{\Lambda}_R^*$.
    return $\mathbf{\Lambda}^*$
function ComputeReduction($R, \mathbf{\Sigma}_F, \tilde{\mathbf{\Sigma}}_T, \sigma$)
    Get eigen-decomposition $\tilde{\mathbf{\Sigma}}_T = \boldsymbol{U}\mathbf{\Sigma}\boldsymbol{U}^\top$.
    Principal eigenspace $\boldsymbol{U}_1 \in \mathbb{R}^{d \times R}$ = the first $R$ columns of $\boldsymbol{U}$.
    Top eigenvalues: Set $\tilde{\mathbf{\Sigma}}_T^R = \boldsymbol{U}_1^\top \tilde{\mathbf{\Sigma}}_T \boldsymbol{U}_1, \mathbf{\Sigma}_F^R = \boldsymbol{U}_1^\top \mathbf{\Sigma}_F \boldsymbol{U}_1$
    Equivalent noise level: $\sigma_R^2 \leftarrow \sigma^2 + \mathbf{tr}(\tilde{\mathbf{\Sigma}}_T) - \mathbf{tr}(\tilde{\mathbf{\Sigma}}_T^R)$.
    return $\boldsymbol{U}_1, \mathbf{\Sigma}_F^R, \tilde{\mathbf{\Sigma}}_T^R, \sigma_R$
```

3 Canonical Covariance and Optimal Representation

In this section, we ask the simpler question: if the covariances $\mathbf{\Sigma}_T$ and $\mathbf{\Sigma}_F$ are known, what is the best choice of $\mathbf{\Lambda}$ to minimize the risk of the interpolator from Definition 3? In general, the covariances are not known; however, the insights from this section help us study the more general case in Section 5. Define the risk as the expected error of inferring the label on the few-shot dataset,

$$\text{risk}(\mathbf{\Lambda}, \mathbf{\Sigma}_T, \mathbf{\Sigma}_F) = \boldsymbol{E}_{\boldsymbol{x}, y, \boldsymbol{\beta}}(y - \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_{\mathbf{\Lambda}})^2 = \boldsymbol{E}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_{\mathbf{\Lambda}} - \boldsymbol{\beta})^\top \mathbf{\Sigma}_F (\hat{\boldsymbol{\beta}}_{\mathbf{\Lambda}} - \boldsymbol{\beta}) + \sigma^2. \tag{3.1}$$

The natural choice of optimization for choosing $\mathbf{\Lambda}$ would be to choose the weighting that minimizes the eventual risk of the learned interpolator.

$$\mathbf{\Lambda}^* = \arg \min_{\mathbf{\Lambda}' \in \mathbb{R}^{d \times R}} \text{risk}(\mathbf{\Lambda}', \mathbf{\Sigma}_T, \mathbf{\Sigma}_F) \tag{3.2}$$

Since the label y is bilinear in x and β, we introduce whitened features $\tilde{\boldsymbol{x}} = \mathbf{\Sigma}_F^{-1/2} \boldsymbol{x}$ and associated task vector $\tilde{\boldsymbol{\beta}} = \mathbf{\Sigma}_F^{1/2} \boldsymbol{\beta}$. This change of variables ensures $\boldsymbol{x}^T \boldsymbol{\beta} = \tilde{\boldsymbol{x}}^T \tilde{\boldsymbol{\beta}}$; now, the task covariance in the transformed coordinates takes the form

$$\tilde{\mathbf{\Sigma}}_T = \mathbf{\Sigma}_F^{1/2} \mathbf{\Sigma}_T \mathbf{\Sigma}_F^{1/2},$$

which we call the **canonical task covariance**; it captures the joint behavior of feature and task covariances $\mathbf{\Sigma}_F, \mathbf{\Sigma}_T$. Below, we observe that the risk in Equation (3.1) is invariant to the change of co-ordinates that we have described above i.e it does not change when $\mathbf{\Sigma}_F^{1/2} \mathbf{\Sigma}_T \mathbf{\Sigma}_F^{1/2}$ is fixed and we vary $\mathbf{\Sigma}_F$ and $\mathbf{\Sigma}_T$.

Observation 2 (Equivalence to problem with whitened features) *Let data be generated as in Phase 1. Denote* $\tilde{\mathbf{\Sigma}}_T = \mathbf{\Sigma}_F^{1/2} \mathbf{\Sigma}_T \mathbf{\Sigma}_F^{1/2}$*. Then* $risk(\mathbf{\Sigma}_F^{-1/2} \mathbf{\Lambda}, \mathbf{\Sigma}_T, \mathbf{\Sigma}_F) = risk(\mathbf{\Lambda}, \tilde{\mathbf{\Sigma}}_T, \boldsymbol{I})$.

This observation can be easily verified by substituting the change-of-coordinates into Equation (3.1) and evaluating the risk.

The risk in (3.1) quantifies the quality of representation $\mathbf{\Lambda}$; however it is not a manageable function of $\mathbf{\Lambda}$ that can be straightforwardly optimized. In this subsection, we show that it is asymptotically equivalent to a different optimization problem, which can be easily solved by analyzing KKT optimality conditions. Theorem 1 characterizes this equivalence; the COMPUTEREDUCTION subroutine of Algorithm 1 calculates key quantities that are used in specifying the reduction, and the COMPUTEOPTIMALREP subroutine of Algorithm 1 uses the solution of the simpler problem to obtain a solution for the original.

Assumption 1 (Bounded feature covariance) *There exist positive constants* $\Sigma_{\min}$, $\Sigma_{\max}$ *such that* $\mathbf{\Sigma}_F$ *is lower/upper bounded as follows:* $\mathbf{0} \prec \Sigma_{\min} \boldsymbol{I} \preceq \mathbf{\Sigma}_F \preceq \Sigma_{\max} \boldsymbol{I}$.

Assumption 2 (Joint diagonalizability) $\mathbf{\Sigma}_F$ *and* $\mathbf{\Sigma}_T$ *are diagonal matrices.*[1]

[1] This is equivalent to the more general scenario where $\mathbf{\Sigma}_F$ and $\mathbf{\Sigma}_T$ are jointly diagonalizable.

Assumption 3 (Double asymptotic regime) *We let the dimensions and the sample size grow as $d, R, n_2 \to \infty$ at fixed ratios $\bar{\kappa} := d/n_2$ and $\kappa := R/n_2$.*

Assumption 4 *The joint empirical distribution of the eigenvalues of $\mathbf{\Lambda}_R$ and $\tilde{\mathbf{\Sigma}}_T^R$ is given by the average of Dirac δ's: $\frac{1}{R}\sum_{i=1}^R \delta_{\mathbf{\Lambda}_{R,i}, \sqrt{R}\tilde{\mathbf{\Sigma}}_{T,i}^R}$. It converges to a fixed distribution as $d \to \infty$.*

With these assumptions, we can derive an analytical expression to quantify the risk of a representation $\mathbf{\Lambda}$. We will then optimize this analytic expression to obtain a formula for the optimal representation.

Theorem 1 (Asymptotic risk equivalence) *Suppose Assumptions 1, 2, 3, 4 hold. Let $\xi > 0$ be the unique number obeying $n_2 = \sum_{i=1}^R \left(1 + (\xi \mathbf{\Lambda}_i^2)^{-1}\right)^{-1}$. Define $\boldsymbol{\theta} \in \mathbb{R}^R$ with entries $\boldsymbol{\theta}_i = \frac{\xi \mathbf{\Lambda}_i^2}{1+\xi \mathbf{\Lambda}_i^2}$ and calculate $\tilde{\mathbf{\Sigma}}_T^R, \sigma_R$ using the* ComputeReduction *procedure of Algorithm 1. Then, define the analytic risk formula*

$$f(\boldsymbol{\theta}, \tilde{\mathbf{\Sigma}}_T^R, n_2) = \frac{1}{n_2 - \|\boldsymbol{\theta}\|_2^2}\left(n_2 \sum_{i=1}^R (1-\boldsymbol{\theta}_i)^2 \tilde{\mathbf{\Sigma}}_{T,i}^R + (\|\boldsymbol{\theta}\|_2^2 + 1)\sigma_R^2\right). \tag{3.3}$$

We have that

$$\lim_{n_2 \to \infty} f(\boldsymbol{\theta}, \tilde{\mathbf{\Sigma}}_T^R, n_2) = \lim_{n_2 \to \infty} risk(\mathbf{\Sigma}_F^{-1/2}\mathbf{\Lambda}, \mathbf{\Sigma}_T, \mathbf{\Sigma}_F) \tag{3.4}$$

The proof of Theorem 1 applies the convex Gaussian Min-max Theorem (CGMT) in [36] and can be found in the Appendix B.2.We show that as dimension grows, the distribution of the estimator $\hat{\beta}$ converges to a Gaussian distribution and we can calculate the expectation of risk.

Theorem 1 provides us with a closed-form risk for any linear representation. Now, one can solve for the optimal representation by computing (OPT-REP) below. In order to do this, we propose an algorithm for the optimization problem in Appendix B.5 via a study of the KKT conditions for the problem [2].

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \; f(\boldsymbol{\theta}, \mathbf{\Sigma}_T, \mathbf{\Sigma}_F), \text{ s.t. } 0 \le \boldsymbol{\theta} < 1, \sum_{i=1}^R \boldsymbol{\theta}_i = n_2 \tag{OPT-REP}$$

The optimal representation is[3] $\mathbf{\Lambda}_{R,i}^* = ((1/\boldsymbol{\theta}_i^* - 1)\xi)^{-2}$. The subroutine ComputeOptimalRep in Algorithm 1 summarizes this procedure.

Remark 1 *Thm. 1 states that risk$(\mathbf{\Sigma}_F^{-1/2}\mathbf{\Lambda}, \mathbf{\Sigma}_T, \mathbf{\Sigma}_F)$ can be arbitrarily well-approximated by $f(\boldsymbol{\theta}, \tilde{\mathbf{\Sigma}}_T^R, n_2)$ if n_2 is sufficiently large. In Fig. 1(b), we set $\mathbf{\Sigma}_F = \boldsymbol{I}_{100}$, $\mathbf{\Sigma}_T =$ diag$(\boldsymbol{I}_{20}, 0.1\boldsymbol{I}_{80})$, $n_2 = 40$. The curves in Fig1(b) are the finite dimensional approximation of f (LHS of* (3.4)*); the dots are empirical approximations of the risk (RHS of* (3.4)*). We tested two cases when $\mathbf{\Lambda}$ is the optimal eigen-weighting or projection matrix with no weighting. Our theorem is corroborated by the observation that the dots and curves are visibly very close. The approximation is already accurate for the finite dimensional problem with just $n_2 = 40$.*

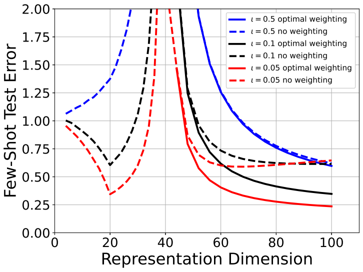

Figure 3: Theoretical risk of optimal representation. $\mathbf{\Sigma}_F = \boldsymbol{I}_{100}$, $\mathbf{\Sigma}_T =$ diag$(\boldsymbol{I}_{20}, \iota\boldsymbol{I}_{80})$, $n_2 = 40$.

The benefit of overparameterization. Theorem 1 leads to an optimal eigen-weighting strategy via asymptotic analysis. In Figure 3, we plot the effect on the risk of increasing R for different shapes of task covariance; the parameter ι controls how spiked $\mathbf{\Sigma}_T$ is, with a smaller value for ι indicating increased spikedness. For the underparameterized problem, the weighting does not have any impact on the risk. In the overparameterized regime, the eigen-weighted learner achieves lower few-shot error than its unweighted ($\mathbf{\Lambda} = \boldsymbol{I}$) counterpart, showing that eigen-weighting becomes critical.

[2] In Sec. 5 the constraint is $\underline{\theta} \le \boldsymbol{\theta} \le 1 - \frac{d-n_2}{n_2}\underline{\theta}$ for robustness concerns.

[3] In the algorithm, $\xi = 1$ and $\mathbf{\Lambda}_{R,i} = (1/\boldsymbol{\theta}_i^* - 1)^{-2}$, because $c\mathbf{\Lambda}^*$ for any constant c gives the same $\hat{\boldsymbol{\beta}}$.

The eigen-weighting procedure can introduce inductive bias during few-shot learning, and helps explain how optimal representation minimizing the few-shot risk can be overparameterized with $R \gg n_2$. We note that, an R dimensional representation can be recovered by a d dimensional representation matrix of rank R, thus the underparameterized case can never beat d dimensional case in theory. The error with optimal eigen-weighting in overparameterized regime is smaller than the respective underparameterized counterpart. The error is lower with smaller ι. It implies that, while $\tilde{\boldsymbol{\Sigma}}_T$ gets closer to low-rank, the excess error caused by choosing small dimension R (equal to the gap $\sigma_R^2 - \sigma^2$ in Algo 1) is not as significant.

Low dimensional representations zero out features and cause bias. By contrast, when $\tilde{\boldsymbol{\Sigma}}_T \in \mathbb{R}^{d\times d}$ is not low rank, every feature contributes to learning with the importance of the features reflected by the weights. This viewpoint is in similar spirit to that of [20] where the authors devise a misspecified linear regression to demonstrate the benefits of overparameterization. Our algorithm allows arbitrary representation dimension R and eigen-weighting.

4 Representation Learning

In this section, we will show how to estimate the useful distribution in representation learning phase that enables us to calculate eigen-weighting matrix $\boldsymbol{\Lambda}^*$. Note that $\boldsymbol{\Lambda}^*$ depends on the canonical covariance $\tilde{\boldsymbol{\Sigma}}_T = \boldsymbol{\Sigma}_F^{1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F^{1/2}$. Learning the R-dimensional principal subspace of $\tilde{\boldsymbol{\Sigma}}_T$ enables us[4] to calculate $\boldsymbol{\Lambda}^*$. Denote this subspace by $\tilde{\boldsymbol{S}}_T$.

Subspace estimation vs. inductive bias. The subspace-based representation $\tilde{\boldsymbol{S}}_T$ has degrees of freedom= Rd. When $\tilde{\boldsymbol{\Sigma}}_T$ is exactly rank R and features are whitened, [37] provides a sample-complexity lower bound of $\Omega(Rd)$ examples and gives an algorithm achieving $\mathcal{O}(R^2 d)$ samples. However, in practice, deep nets learn good representations despite overparameterization. In this section, recalling our **Q2**, we argue that the inductive bias of the feature distribution can implicitly accelerate learning the canonical covariance. This differentiates our results from most prior works such as [23, 22, 37] in two aspects:

1. Rather than focusing on a *low dimensional* subspace and assuming $N \gtrsim Rd$, we can estimate $\tilde{\boldsymbol{\Sigma}}_T$ or $\tilde{\boldsymbol{S}}_T$ in the overparameterized regime $N \lesssim Rd$.
2. Rather than assuming whitened features $\boldsymbol{\Sigma}_F = \boldsymbol{I}$ and achieving a sample complexity of $R^2 d$, our learning guarantee holds for arbitrary covariance matrices $\boldsymbol{\Sigma}_F, \boldsymbol{\Sigma}_T$. The sample complexity depends on *effective rank* and can be arbitrarily smaller than DoF. We showcase our bounds via a spiked covariance setting in Example 1 below.

For learning $\tilde{\boldsymbol{\Sigma}}_T$ or its subspace $\tilde{\boldsymbol{S}}_T$, we investigate the method-of-moments (MoM) estimator.

Definition 4 (MoM Estimator) *For* $1 \leq i \leq T$*, define* $\hat{\boldsymbol{b}}_{i,1} = 2n_1^{-1}\sum_{j=1}^{n_1/2} y_{ij}\boldsymbol{x}_{ij}$*,* $\hat{\boldsymbol{b}}_{i,2} = 2n_1^{-1}\sum_{j=n_1/2+1}^{n_1} y_{ij}\boldsymbol{x}_{ij}$*. Set*

$$\hat{\boldsymbol{M}} = n_1^{-1}\sum_{i=1}^{T}(\boldsymbol{b}_{i,1}\boldsymbol{b}_{i,2}^\top + \boldsymbol{b}_{i,2}\boldsymbol{b}_{i,1}^\top),$$

The expectation of $\hat{\boldsymbol{M}}$ *is equal to* $\boldsymbol{M} = \boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F$*.*

Inductive bias in representation learning: Recall that canonical covariance $\tilde{\boldsymbol{\Sigma}}_T = \boldsymbol{\Sigma}_F^{1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F^{1/2}$ is the attribute of interest. However, feature covariance $\boldsymbol{\Sigma}_F^{1/2}$ term implicitly modulates the estimation procedure because the population MoM is not $\tilde{\boldsymbol{\Sigma}}_T$ but $\boldsymbol{M} = \boldsymbol{\Sigma}_F^{1/2}\tilde{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_F^{1/2}$. For instance, when estimating the principle canonical subspace $\tilde{\boldsymbol{S}}_T$, the degree of alignment between $\boldsymbol{\Sigma}_F$ and $\tilde{\boldsymbol{\Sigma}}_T$ can make or break the estimation procedure: If $\boldsymbol{\Sigma}_F$ and $\tilde{\boldsymbol{\Sigma}}_T$ have *well-aligned* principal subspaces, $\tilde{\boldsymbol{S}}_T$ will be easier to estimate since $\boldsymbol{\Sigma}_F$ will amplify the $\tilde{\boldsymbol{S}}_T$ direction within $\boldsymbol{M}$.

We verify the inductive bias on practical image dataset, reported in Appendix A. We assessed correlation coefficient between covariances $\tilde{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F$ via the canonical-feature alignment score defined

[4] We also need to estimate $\boldsymbol{\Sigma}_F$ for whitening. Estimating $\boldsymbol{\Sigma}_F$ is rather easy and incurs smaller error compared to $\tilde{\boldsymbol{\Sigma}}_T$. The analysis is provided in the first part of Appendix B.

feature cov	$\boldsymbol{\Sigma}_F = \boldsymbol{I}, \boldsymbol{\Sigma}_T = \text{diag}(\boldsymbol{I}_{s_T}, \boldsymbol{0})$			$\boldsymbol{\Sigma}_F = \text{diag}(\boldsymbol{I}_{s_F}, \iota_F \boldsymbol{I}_{d-s_F}),$ $\boldsymbol{\Sigma}_T = \text{diag}(\boldsymbol{I}_{s_T}, \iota_T \boldsymbol{I}_{d-s_T})$		
estimator	sample N	sample n_1	error	sample N	sample n_1	error
MoM	ds_T^2	1	$(ds_T^2/N)^{1/2}$	$r_F r_T^2$	1	$(r_F r_T^2/N)^{1/2}$
MoM	ds_T	s_T	$(s_T/n_1)^{1/2}$	$r_F r_T$	r_T	$(r_T/n_1)^{1/2}$

Table 2: **Right side:** Sample complexity and error of MoM estimators. s_F (s_T) is the dimension of the principal eigenspace of the feature (task) covariance. $r_F = s_F + \iota_F(d - s_F)$, $r_T = s_T + \iota_T(d - s_T)$ are the effective ranks. **Left side:** This is the well-studied setting of identity feature covariance and low-rank task covariance. Our bound in the second row is the first result to achieve optimal sample complexity of $\mathcal{O}(ds_T)$ (cf. [37, 23]).

as the correlation coefficient

$$\rho(\boldsymbol{\Sigma}_F, \tilde{\boldsymbol{\Sigma}}_T) := \frac{\langle \boldsymbol{\Sigma}_F, \tilde{\boldsymbol{\Sigma}}_T \rangle}{\|\boldsymbol{\Sigma}_F\|_F \|\tilde{\boldsymbol{\Sigma}}_T\|_F} = \frac{\text{trace}(\boldsymbol{M})}{\|\boldsymbol{\Sigma}_F\|_F \|\tilde{\boldsymbol{\Sigma}}_T\|_F}.$$

Observe that, the MoM estimator $\boldsymbol{M}$ naturally shows up in the alignment definition because the inner product of $\tilde{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F$ is equal to trace$(\boldsymbol{M})$. This further supports our inductive bias intuition. As reference, we compared it to canonical-identity alignment defined as $\frac{\text{trace}(\tilde{\boldsymbol{\Sigma}}_T)}{\sqrt{d}\|\tilde{\boldsymbol{\Sigma}}_T\|_F}$ (replacing $\boldsymbol{\Sigma}_F$ with $\boldsymbol{I}$). The canonical-feature alignment score is higher than the canonical-identity alignment score. This significant score difference exemplifies how $\boldsymbol{\Sigma}_F$ and $\tilde{\boldsymbol{\Sigma}}_T$ can synergistically align with each other (inductive bias). This alignment helps our MoM estimator defined below, illustrated by Example 1 (spiked covariance).

In the following subsections, let $N = n_1 T$ refer to the total tasks in representation-learning phase. Let $r_F = \mathbf{tr}(\boldsymbol{\Sigma}_F)$, $r_T = \mathbf{tr}(\boldsymbol{\Sigma}_T)$, and $\tilde{r}_T = \mathbf{tr}(\tilde{\boldsymbol{\Sigma}}_T)$. Define the approximate low-rankness measure of feature covariance by[5]

$$s_F = \min\ s'_F, \text{ s.t. } s'_F \in \{1, ..., d\},\ s'_F/d \geq \lambda_{s'_F+1}(\boldsymbol{\Sigma}_F)$$

We have two results for this estimator.

1. Generally, we can estimate $\boldsymbol{M}$ with $\mathcal{O}(r_F \tilde{r}_T^2)$ samples.
2. Let $n_1 \geq s_T$, we can estimate $\boldsymbol{M}$ with $\mathcal{O}(s_F \tilde{r}_T)$ samples.

Paper [37] has sample complexity $\mathcal{O}(dr^2)$ (r is exact rank). Our sample complexity is $\mathcal{O}(r_F \tilde{r}_T^2)$. $r_F, \tilde{r}_T$ can be seen as effective ranks and our bounds are always smaller than [37]. We will discuss later in Example 1. Our second result says when $n_1 \geq s_T$, our sample complexity achieves the $\mathcal{O}(dr)$ which is proven a lower bound in [37].

Theorem 2 *Let data be generated as in Phase 1. Assume* $\|\boldsymbol{\Sigma}_F\|, \|\boldsymbol{\Sigma}_T\| = 1$ *for normalization*[6].

1. Let n_1 *be a even number. Then with probability at least* $1 - N^{-100}$,

$$\|\hat{\boldsymbol{M}} - \boldsymbol{M}\| \lesssim (\tilde{r}_T + \sigma^2)\sqrt{\frac{r_F}{N}} + \sqrt{\frac{r_T}{T}}.$$

2. Assume $T \geq s_F$. *If* $n_1 \gtrsim \tilde{r}_T + \sigma^2$, *then with probability at least*

$$\|\hat{\boldsymbol{M}} - \boldsymbol{M}\| \lesssim \left((\tilde{r}_T + \sigma^2)/n_1\right)^{1/2}.$$

Denote the top-R *principal subspaces of* $\boldsymbol{M}, \hat{\boldsymbol{M}}$ *by* $\boldsymbol{M}_{top}, \hat{\boldsymbol{M}}_{top}$ *and assume the eigen-gap condition* $\lambda_R(\boldsymbol{M}) - \lambda_{R+1}(\boldsymbol{M}) > 2\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|$. *Then a direct application of Davis-Kahan Theorem [14] bounds the subspace angle as follows*

$$\textit{angle}(\boldsymbol{M}_{top}, \hat{\boldsymbol{M}}_{top}) \lesssim \|\hat{\boldsymbol{M}} - \boldsymbol{M}\|/(\lambda_R(\boldsymbol{M}) - \lambda_{R+1}(\boldsymbol{M})).$$

[5] The $(s_F + 1)$-th eigenvalue is smaller than s_F/d. Note the top eigenvalue is 1.

[6] This is simply equivalent to scaling y_{ij}, which does not affect the normalized error $\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|/\|\boldsymbol{M}\|$. In the appendix we define $\mathcal{S} = \max\{\|\boldsymbol{\Sigma}_F\|, \|\boldsymbol{\Sigma}_T\|\}$ and prove the theorem for general $\mathcal{S}$.

Estimating eigenspace of canonical covariance. Note that if $\boldsymbol{\Sigma}_F$ and $\boldsymbol{\Sigma}_T$ are aligned, (e.g. Example 1 below with $s_F = s_T = R$), then $\boldsymbol{M}_{\text{top}} = \tilde{\boldsymbol{S}}_T$ is exactly the principal subspace of $\tilde{\boldsymbol{\Sigma}}_T$. Theorem 2 indeed gives estimation error for the principal subspace of $\tilde{\boldsymbol{\Sigma}}_T$. Note that, such alignment is and more general requirement compared to related works which require whitened features [37, 23].

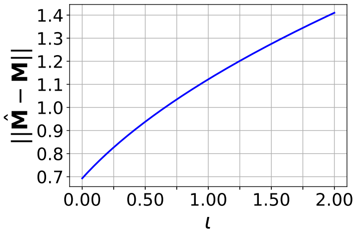

Figure 4: Error of MoM estimator

Example 1 (Spiked $\tilde{\boldsymbol{\Sigma}}_T$, Aligned principal subspaces) *Suppose the spectra of $\boldsymbol{\Sigma}_F$ and $\tilde{\boldsymbol{\Sigma}}_T$ are bimodal as follows $\boldsymbol{\Sigma}_F = diag(\boldsymbol{I}_{s_F}, \iota_F \boldsymbol{I}_{d-s_F})$, $\boldsymbol{\Sigma}_T = diag(\boldsymbol{I}_{s_T}, \iota_T \boldsymbol{I}_{d-s_T})$. Set statistical error $Err_{T,N} := \sqrt{r_T^2 r_F/N} + \sqrt{r_T/T}$. When $\iota_T, \iota_F < 1$, $s_F \geq s_T$, the recovery error of $\tilde{\boldsymbol{\Sigma}}_T$ and its principal subspace $\tilde{\boldsymbol{S}}_T$ are bounded as*

$$angle(\hat{\boldsymbol{M}}_{top}, \tilde{\boldsymbol{S}}_T) \lesssim Err_{T,N} + \iota_F^2 \iota_T \quad and \quad \|\hat{\boldsymbol{M}} - \tilde{\boldsymbol{\Sigma}}_T\| \lesssim Err_{T,N} + \iota_F \iota_T.$$

The estimation errors for $\tilde{\boldsymbol{\Sigma}}_T, \tilde{\boldsymbol{S}}_T$ are controlled in terms of the effective ranks and the spectrum tails ι_F, ι_T. Typically $s_F s_T \gtrsim n_1$ so $\sqrt{r_T^2 r_F/N}$ term dominates the statistical error in practice. In Fig. 4 we plot the error of estimating $\boldsymbol{M}$ (whose principal subspace coincides with $\tilde{\boldsymbol{\Sigma}}_T$). $\boldsymbol{\Sigma}_F = \text{diag}(\boldsymbol{I}_{30}, \iota \boldsymbol{I}_{70})$, $\boldsymbol{\Sigma}_T = \text{diag}(\boldsymbol{I}_{30}, \boldsymbol{0}_{70})$. $T = N = 100$. We can see that the error increase with ι .

5 Robustness of Optimal Representation and Overall Meta-Learning Bound

In Section 3, we described the algorithm for computing the optimal representation with *known* distributions of features and tasks. In Section 4, we proposed the MoM estimator in representation learning phase to estimate the unknown covariance matrices. In this section, we study the algorithm's behaviors when we calculate $\boldsymbol{\Lambda}$ using the *estimated* canonical covariance, rather than the full-information setting of Section 3.

Armed with the provably reliable estimators of Section 4, we can replace $\tilde{\boldsymbol{\Sigma}}_T$ and $\boldsymbol{\Sigma}_F$ in Algorithm 1 with our estimators. In this section, we inquire: how does the estimation error in covariance-estimation in representation learning stage affect the downstream few-shot learning risk? That says, we are interested in[7] $\text{risk}(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) - \text{risk}(\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F)$.

Let us replace the constraint in (OPT-REP) by $\underline{\theta} \leq \boldsymbol{\theta} \leq 1 - \frac{d-n_2}{n_2}\underline{\theta}$. This changes the "optimization" step in Algorithm 1. Theorem 3 does not require an explicit computation of the optimal representation by enforcing $\underline{\theta}$. Instead, we use the robustness of such a representation (due to its well-conditioned nature) to deduce its stability. That said, for practical computation of optimal representation, we simply use Algorithm 1. We can then evaluate $\underline{\theta}$ after-the-fact as the minimum singular value of this representation to apply Theorem 3 without assuming an explicit $\underline{\theta}$.

Let $\boldsymbol{\Lambda}_{\underline{\theta}}(R) = \text{ComputeOptimalRep}(R, \boldsymbol{\Sigma}_F, \hat{\boldsymbol{M}}, \sigma, n_2)$ denote the estimated optimal representation and $\boldsymbol{\Lambda}^*_{\underline{\theta}}(R) = \text{ComputeOptimalRep}(R, \boldsymbol{\Sigma}_F, \tilde{\boldsymbol{\Sigma}}_T, \sigma, n_2)$ denote the true optimal representation, which cannot be accessed in practice. Below we present the bound of the whole meta-learning algorithm. It shows that a bounded error in representation learning leads to a bounded increase on the downstream few-shot learning risk, thus quantifying the robustness of few-shot learning to errors in covariance estimates.

Theorem 3 *Let $\boldsymbol{\Lambda}_{\underline{\theta}}(R)$, $\boldsymbol{\Lambda}^*_{\underline{\theta}}(R)$ be as defined above, and $r_F = \mathbf{tr}(\boldsymbol{\Sigma}_F)$, $r_T = \mathbf{tr}(\boldsymbol{\Sigma}_T), \tilde{r}_T = \mathbf{tr}(\tilde{\boldsymbol{\Sigma}}_T)$. The risk of meta-learning algorithm satisfies*[8]

$$risk(\boldsymbol{\Lambda}_{\underline{\theta}}(R), \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) - risk(\boldsymbol{\Lambda}^*_{\underline{\theta}}(R), \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) \lesssim \frac{n_2^2}{d(R-n_2)(2n_2 - R\underline{\theta})\underline{\theta}} \left[(\tilde{r}_T + \sigma^2)\sqrt{\frac{r_F}{N}} + \sqrt{\frac{r_T}{T}} \right].$$

[7] Note that Sec.6 of [40] gives the exact value of $\text{risk}(\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F)$ so we have an end to end error guarantee.

[8] The bracketed expression applies first conclusion of Theorem 3. One can plug in the second as well.

Notice that as the number of previous tasks T and total representation-learning samples N observed increases, the risk of the estimated $\mathbf{\Lambda}_{\underline{\theta}}(R)$ approaches that of the optimal $\mathbf{\Lambda}^*_{\underline{\theta}}(R)$ as we expect. The result only applies to the overparameterized regime of interest $R > n_2$. The expression of risk in the underparameterized case is different, and covered by the second case of Equation(4.4) in [40]. We plot it in Fig 1(b) on the left side of the peak as a comparison.

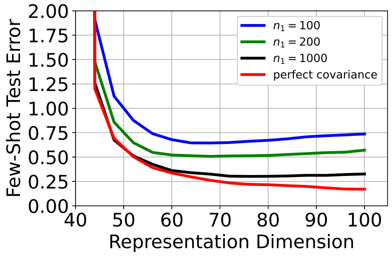

Figure 5: End to end learning guarantees. $d = 100, n_2 = 40, T = 200$, $\mathbf{\Sigma}_T = (\boldsymbol{I}_{20}, 0.05 \cdot \boldsymbol{I}_{80})$, $\mathbf{\Sigma}_F = \boldsymbol{I}_{100}$.

Risk with respect to PCA level R. In Fig. 5, we plot the error of the whole meta-learning algorithm. We simulate representation learning and get $\hat{\boldsymbol{M}}$, use it to compute $\mathbf{\Lambda}$ and plot the theoretical downstream risk (experiments match, see Fig. 1 (b)). Mainly, we compare the behavior of Theorem 3 with different R. When R grows, we search $\mathbf{\Lambda}$ in a larger space. The optimal $\mathbf{\Lambda}$ in a feasible *sub*set is always no better than searching in a larger space, thus the risk decreases with R increasing. At the same time, representation learning error increases with R since we need to fit a matrix in a larger space. In essence, this result provides a theoretical justification on a sweet-spot for the optimal representation. $d = R$ is optimal when $N = \infty$, i.e., representation learning error is 0. As N decreases, there is a tradeoff between learning error and truncating small eigenvalues. Thus choosing R adaptively with N can strike the right bias-variance tradeoff between the excess risk (variance) and the risk due to suboptimal representation.

6 Conclusion

In this paper, we study the sample efficiency of meta-learning with linear representations. We show that the optimal representation is typically overparameterized and outperforms subspace-based representations for general data distributions. We refine the sample complexity analysis for learning arbitrary distributions and show the importance of inductive bias of feature and task. Finally we provide an end-to-end bound for the meta-learning algorithm showing the tradeoff of choosing larger representation dimension v.s. robustness against representation learning error.

References

[1] Theodore W Anderson et al. Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. *Essays in probability and statistics*, pages 1–24, 1970.

[2] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

[3] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. Efficient representations for lifelong learning and autoencoding. In *Conference on Learning Theory*, pages 191–210, 2015.

[4] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

[5] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[6] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[7] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58, 2014.

[8] Quentin Bouniot, Ievgen Redko, Romaric Audigier, Angélique Loesch, Yevhenii Zotkin, and Amaury Habrard. Towards better understanding meta-learning methods through multi-task representation learning theory. *arXiv preprint arXiv:2010.01992*, 2020.

[9] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[10] Giovanni Cavallanti, Nicolo Cesa-Bianchi, and Claudio Gentile. Linear algorithms for online multitask classification. *The Journal of Machine Learning Research*, 11:2901–2934, 2010.

[11] Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. *arXiv preprint arXiv:2012.08749*, 2020.

[12] Sitan Chen, Jerry Li, and Zhao Song. Learning mixtures of linear regressions in subexponential time via fourier moments. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 587–600, 2020.

[13] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

[14] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[16] Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

[17] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

[18] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135, 2017.

[19] Halil Ibrahim Gulluk, Yue Sun, Samet Oymak, and Maryam Fazel. Sample efficient subspace-based representations for nonlinear meta-learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3685–3689. IEEE, 2021.

[20] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[21] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[22] Weihao Kong, Raghav Somani, Sham Kakade, and Sewoong Oh. Robust meta-learning for mixed linear regression with small batches. *arXiv preprint arXiv:2006.09702*, 2020.

[23] Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh. Meta-learning for mixed linear regression. *arXiv preprint arXiv:2002.08936*, 2020.

[24] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32:8572–8583, 2019.

[25] Yuanzhi Li and Yingyu Liang. Learning mixtures of linear regressions with nearly optimal complexity. In *Conference On Learning Theory*, pages 1125–1144, 2018.

[26] Karim Lounici, Massimiliano Pontil, Sara Van De Geer, Alexandre B Tsybakov, et al. Oracle inequalities and optimal inference under group sparsity. *The annals of statistics*, 39(4):2164–2204, 2011.

[27] James Lucas, Mengye Ren, Irene Kameni, Toniann Pitassi, and Richard Zemel. Theoretical bounds on estimation error for meta-learning. *arXiv preprint arXiv:2010.07140*, 2020.

[28] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884, 2016.

[29] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[30] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[31] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime, 2020.

[32] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*, 2020.

[33] Vidya Muthukumar, Kailas Vodrahalli, and Anant Sahai. Harmless interpolation of noisy data in regression. *CoRR*, abs/1903.09139, 2019.

[34] Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

[35] Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *ICML Workshop on Understanding and Improving Generalization in Deep Learning*, 2019.

[36] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. *Advances in Neural Information Processing Systems*, 28:3420–3428, 2015.

[37] Nilesh Tripuraneni, Chi Jin, and Michael I Jordan. Provable meta-learning of linear representations. *arXiv preprint arXiv:2002.11684*, 2020.

[38] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[39] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[40] Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression, 2020.

[41] Jiaqi Yang, Wei Hu, Jason D Lee, and Simon S Du. Provable benefits of representation learning in linear bandits. *arXiv preprint arXiv:2010.06531*, 2020.

[42] Jiaqi Yang, Wei Hu, Jason D Lee, and Simon S Du. Impact of representation learning in linear bandits. In *International Conference on Learning Representations*, 2021.

[43] Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Mixed linear regression with multiple components. In *Advances in neural information processing systems*, pages 2190–2198, 2016.

A Numerical verification of inductive bias for representation learning

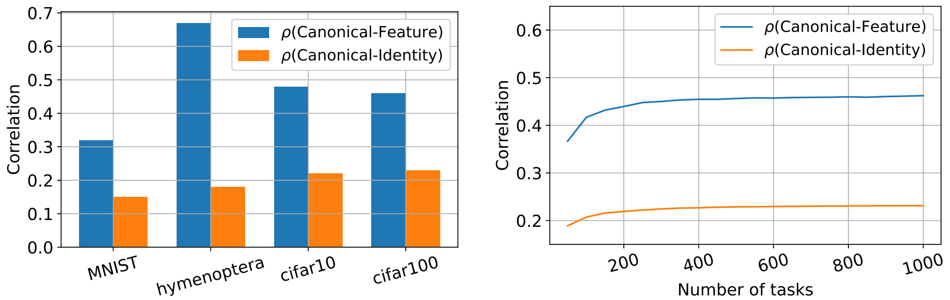

Figure 6: (a) Alignment of feature-task on image classification models. The result of MNIST uses the setting in Sec. 4. We apply the pretrained ResNet classification model on the other three datasets, compute the (last layer) feature/task covariances and get the alignments. The alignment is a measure of correlation which is denoted by ρ here. (b) We use the cifar100 dataset, take the pretrained ResNet18 network and vary the number of tasks (i.e., varying the number of output classes of the neural net, also equivalent to number of rows of the last layer matrix B defined below). The alignments increase with number of tasks.

We add a figure with experiments on a few image datasets. We take the pretrained ResNet18 neural network, and feed the images into it. For every image, we take the last (closest to output) layer output as the feature $\boldsymbol{x}$, which is of dimension $d = 512$. The weights of the last layer are the tasks, which is a $T \times d$ matrix (We call it B). $T = 1000$, each row of B is a task vector. Then $Bx \in \mathbb{R}^T$ generates the label, whose each entry corresponds to each class. We calculate the feature and task covariance, as well as the alignments defined in Sec. 4. We can clearly see the inductive bias of every dataset.

B Analysis of optimal representation

B.1 Proof of Observation 1 and equivalent noise

Observation 1 *Let $\boldsymbol{\Lambda} \in \mathbb{R}^{d\times R}$, $\boldsymbol{X} \in \mathbb{R}^{n_2\times d}$ and $\boldsymbol{y} \in \mathbb{R}^{n}_2$, and define*

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\Lambda}(\boldsymbol{X}\boldsymbol{\Lambda})^{\dagger}\boldsymbol{y}, \tag{B.1}$$

$$\hat{\boldsymbol{\beta}}_1 = \lim_{t\to 0} \operatorname{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + t\boldsymbol{\beta}^\top(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{\dagger}\boldsymbol{\beta} \tag{B.2}$$

Then $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}$.

Proof Denote the SVD $(\boldsymbol{X}\boldsymbol{\Lambda})^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{R\times R}, \boldsymbol{\Sigma} \in \mathbb{R}^{R\times n_2}, \boldsymbol{V} \in \mathbb{R}^{n_2\times n_2}$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_1 &= \lim_{t\to 0} \operatorname{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + t\boldsymbol{\beta}^\top(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{\dagger}\boldsymbol{\beta} \\
&= \lim_{t\to 0} (\boldsymbol{X}^\top\boldsymbol{X} + t(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{\dagger})^{-1}\boldsymbol{X}\boldsymbol{y} \\
&= \lim_{s\to\infty} s\boldsymbol{\Lambda}(s\boldsymbol{\Lambda}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Lambda} + I)^{-1}\boldsymbol{\Lambda}^\top\boldsymbol{X}^\top\boldsymbol{y} \\
&= \lim_{s\to\infty} s\boldsymbol{\Lambda}(s\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top + I)^{-1}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{y} \\
&= \lim_{s\to\infty} s\boldsymbol{\Lambda}(s\boldsymbol{U}\operatorname{diag}(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma} + I_{n_2}, I_{R-n_2})\boldsymbol{U}^\top)^{-1}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{y} \\
&= \lim_{s\to\infty} \boldsymbol{\Lambda}\boldsymbol{U}(\operatorname{diag}(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}, I_{R-n_2}/s))^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{y}. \\
&= \boldsymbol{\Lambda}(\boldsymbol{X}\boldsymbol{\Lambda})^{\dagger}\boldsymbol{y}
\end{aligned}$$

■

The risk of $\hat{\boldsymbol{\beta}}$ is given by

$$\text{risk}(\hat{\boldsymbol{\beta}}) = \boldsymbol{E}(y - \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}) = \boldsymbol{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{\Sigma}_F (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \sigma^2.$$

In Sec. B.2, we study the asymptotic optimal representation. Below, we characterize the properties of the problem for fixed $\boldsymbol{\beta}$ and arbitrary input covariance $\boldsymbol{\Sigma}_F$. We first go over this and then discuss how to obtain the optimal representation $\boldsymbol{\Lambda}^*$ minimizing test risk.

Remark 2 ***Projection onto R dimensional subspace.*** *For the remaining proof after this part, we will mainly analyze the relation between $\boldsymbol{\Lambda}_R$ and $\boldsymbol{\theta}$ in Thm. 1, which lie in an R dimensional subspace. Here we will build the connection from the d dimensional problem to R dimensional, mainly computing the equivalent noise below. The equivalent noise consists of original noise and the extra noise caused by PCA truncation.*

Let $\boldsymbol{x}_R$ be the projection of $\boldsymbol{x}$ onto the R-dimensional subspace spanned by columns of $\boldsymbol{U}_1$, and $\boldsymbol{x}_{R^\perp}$ is the projection of $\boldsymbol{x}$ onto the orthogonal complement. Namely, $\boldsymbol{x}_R = \boldsymbol{U}_1^\top \boldsymbol{x} \in \mathbb{R}^R$ and $\boldsymbol{x}_{R^\perp} = \boldsymbol{U}_2^\top \boldsymbol{x} \in \mathbb{R}^{(d-R)}$. Similarly we can define $\boldsymbol{\beta}_R$ and $\boldsymbol{\beta}_{R^\perp}$. Thus,

$$y = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon = \boldsymbol{x}_R^\top \boldsymbol{\beta}_R + \boldsymbol{x}_{R^\perp}^\top \boldsymbol{\beta}_{R^\perp} + \varepsilon \tag{B.3}$$

We can treat $\varepsilon_R = \boldsymbol{x}_{R^\perp}^\top \boldsymbol{\beta}_{R^\perp} + \varepsilon$ as the new noise, and try to solve for $\boldsymbol{\beta}_R$. Then define $\boldsymbol{\Sigma}_{T,R^\perp}$ as the matrix containing the same eigenvectors as $\boldsymbol{\Sigma}_T$ and the top R eigenvalues are zeroed out, our noise variance becomes $\sigma_R^2 = \sigma^2 + \boldsymbol{E}(\|\boldsymbol{x}_{R^\perp}\|^2 \|\boldsymbol{\beta}_{R^\perp}\|^2) = \sigma^2 + \mathbf{tr}(\tilde{\boldsymbol{\Sigma}}_T) - \mathbf{tr}(\tilde{\boldsymbol{\Sigma}}_T^R)$ in our algorithm. If we are still in overparameterized regime, namely $R > n_2$, then we define optimal representation on top of it.

In summary, the R-SVD truncation reduces the search space of $\boldsymbol{\Lambda}$ into R dimensional space, where the covariance of the noise in $\boldsymbol{y}$ increases from $\sigma^2 \boldsymbol{I}$ to $\sigma_R^2 \boldsymbol{I}$.

B.2 Distributional characterization of least norm solution

In this part, for simplicity of discussion, we focus on the R dimensional space while omitting the projection step, and the equivalence of a diagonal eigen-weighting matrix $\boldsymbol{\Lambda}_R \in \mathbb{R}^{R \times R}$ and $\boldsymbol{\theta} \in \mathbb{R}^R$ in Thm. 1. Here, we assume a truncated feature matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n \times R}$ where the feature is projected into an R dimensional space.

Define $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n \times R}$, $\tilde{\boldsymbol{y}} \in \mathbb{R}^n$. We study the following least norm solution of the least squares problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}'} \|\boldsymbol{\beta}'\|, \quad \text{s.t., } \tilde{\boldsymbol{X}} \boldsymbol{\beta}' = \tilde{\boldsymbol{y}} \tag{B.4}$$

Assumption 5 *Assume the rows of $\tilde{\boldsymbol{X}}$ are independently drawn from $\mathcal{N}(0, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}})$. We focus on a double asymptotic regime where $R, n \to \infty$ at fixed overparameterization ratio $\kappa := R/n > 0$.*

Assumption 6 *The covariance matrix $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}$ is diagonal and there exist constants $\Sigma_{\min}, \Sigma_{\max} \in (0, \infty)$ such that: $0 \prec \Sigma_{\min} \boldsymbol{I} \preceq \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}} \preceq \Sigma_{\max} \boldsymbol{I}$.*

Assumption 7 *The joint empirical distribution of $\{(\lambda_i(\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}), \boldsymbol{\beta}_i)\}_{i \in [R]}$ converges in Wasserstein-k distance to a probability distribution μ on $\mathbb{R}_{>0} \times \mathbb{R}$ for some $T \geq 4$. That is $\frac{1}{R} \sum_{i \in [R]} \delta_{(\lambda_i(\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}), \boldsymbol{\beta}_i)} \overset{W_k}{\Longrightarrow} \mu$.*

Definition 5 (Asymptotic distribution characterization – Overparameterized regime) *[36] Let random variables $(\Sigma, B) \sim \mu$ (where μ is defined in Assumption 7) and fix $\kappa > 1$. Define parameter ξ as the unique positive solution to the following equation*

$$\mathbb{E}_\mu \Big[\big(1 + (\xi \cdot \Sigma)^{-1}\big)^{-1} \Big] = \kappa^{-1}\,. \tag{B.5}$$

Define the positive parameter γ as follows:

$$\gamma := \Big(\sigma^2 + \mathbb{E}_\mu \Big[\frac{B^2 \Sigma}{(1 + \xi \Sigma)^2}\Big]\Big) \Big/ \Big(1 - \kappa\, \mathbb{E}_\mu \Big[\frac{1}{(1 + (\xi \Sigma)^{-1})^2}\Big]\Big). \tag{B.6}$$

With these and $H \sim \mathcal{N}(0,1)$, define the random variable

$$X_{\kappa,\sigma^2}(\Sigma, B, H) := \Big(1 - \frac{1}{1+\xi\Sigma}\Big)B + \sqrt{\kappa}\frac{\sqrt{\gamma}\,\Sigma^{-1/2}}{1+(\xi\Sigma)^{-1}}H, \tag{B.7}$$

and let Π_{κ,σ^2} be its distribution.

Theorem 4 (Asymptotic distribution characterization – Overparameterized linear Gaussian problem) *[36] Fix $\kappa > 1$ and suppose Assumptions 6 and 7 hold. Let*

$$\frac{1}{R}\sum_{i=1}^{R}\delta_{\sqrt{R}\hat{\boldsymbol{\beta}}_i,\sqrt{R}\boldsymbol{\beta}_i,\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}_{i,i}}}$$

be the joint empirical distribution of $(\sqrt{R}\hat{\boldsymbol{\beta}}, \sqrt{R}\boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}})$ and it converges to a fixed distribution as dimension grows. Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a function in $\mathrm{PL}(2)$. We have that

$$\frac{1}{R}\sum_{i=1}^{R} f(\sqrt{R}\hat{\boldsymbol{\beta}}_i, \sqrt{R}\boldsymbol{\beta}_i, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}_{i,i}}) \xrightarrow{P} \mathbb{E}\left[f(X_{\kappa,\sigma^2}, B, \Sigma)\right]. \tag{B.8}$$

In particular, the risk is given by

$$\begin{aligned} \textit{risk}(\hat{\boldsymbol{\beta}}_n) &\xrightarrow{P} \mathbb{E}[\Sigma(B - X_{\kappa,\sigma^2})] + \sigma_R^2 & \text{(B.9)}\\ &= \mathbb{E}[\frac{\Sigma}{(1+\xi\Sigma)^2}B^2 + \frac{\kappa\gamma}{(1+(\xi\Sigma)^{-1})^2}] + \sigma_R^2. & \text{(B.10)} \end{aligned}$$

B.3 Finding Optimal Representation

Now, for simplicity (and actually without losing generality) assume $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \boldsymbol{I}$. This means that empirical measure of $\boldsymbol{\Sigma}_F$ trivially converges to $\Sigma = 1$. With the representation $\boldsymbol{\Lambda}^*$ with asymptotic distribution Λ, the ML problem has the following mapping

$$\boldsymbol{\beta} \to \boldsymbol{\Lambda}_R^{-1}\boldsymbol{\beta} \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}} \to \boldsymbol{\Lambda}_R\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\boldsymbol{\Lambda}_R.$$

This means the empirical measure converges to the following mapped distributions

$$B \to \bar{B} = \Lambda^{-1}B \quad \text{and} \quad \Sigma = 1 \to \bar{\Sigma} = \Lambda^2\Sigma = \Lambda^2.$$

Our question: Craft the optimal distribution Λ to minimize the representation learning risk. Specifically, for a given (B, Λ) pair, we know from the theorem above that

$$\begin{aligned} \mathrm{risk}^{\boldsymbol{\Lambda}_R}(\hat{\boldsymbol{\beta}}_n) &\xrightarrow{P} \mathbb{E}[\frac{\bar{\Sigma}}{(1+\xi\bar{\Sigma})^2}\bar{B}^2 + \frac{\kappa\gamma}{(1+(\xi\bar{\Sigma})^{-1})^2}] + \sigma_R^2 & \text{(B.11)}\\ &= \mathbb{E}[\frac{B^2}{(1+\xi\Lambda^2)^2} + \frac{\kappa\gamma}{(1+(\xi\Lambda^2)^{-1})^2}] + \sigma_R^2. & \text{(B.12)} \end{aligned}$$

Thus, the optimal weighting strategy (asymptotically) is given by the distribution

$$\Lambda^* = \arg\min_{\Lambda} \mathbb{E}[\frac{B^2}{(1+\xi\Lambda^2)^2} + \frac{\kappa\gamma}{(1+(\xi\Lambda^2)^{-1})^2}],$$

where γ, ξ are strictly positive scalars that are also functions of Λ.

B.4 Non-asymptotic Analysis (for simpler insights)

We apply the discussion iin Sec. B.2 non-asymptotically in few-shot learning. Remember we define $\boldsymbol{X} \in \mathbb{R}^{n_2 \times R}, \boldsymbol{y} \in \mathbb{R}^{n_2}$, each row of $\boldsymbol{X}$ is independently drawn from $\mathcal{N}(0, \boldsymbol{\Sigma}_F)$. We study the following least norm solution of the least squares problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}'} \|\boldsymbol{\beta}'\|, \quad \text{s.t.}, \boldsymbol{X}\boldsymbol{\beta}' = \boldsymbol{y}. \tag{B.13}$$

Definition 6 (Non-asymptotic distribution characterization) *Set $\kappa = R/n_2 > 1$. Given $\sigma_R > 0$, covariance $\mathbf{\Sigma}_F$ and latent vector $\boldsymbol{\beta}$ and define the unique non-negative terms $\xi, \gamma, \boldsymbol{z} \in \mathbb{R}^R$ and $\boldsymbol{\phi} \in \mathbb{R}^R$ as follows:*

$$\xi > 0 \quad \textit{is the solution of} \quad \kappa^{-1} = R^{-1} \sum_{i=1}^{R} \left(1 + (\xi \mathbf{\Sigma}_{F,i})^{-1}\right)^{-1},$$

$$\gamma = \frac{\sigma_R^2 + \frac{1}{R} \sum_{i=1}^{R} \frac{\mathbf{\Sigma}_{F,i} \boldsymbol{\beta}_i^2}{(1+\xi \mathbf{\Sigma}_F)^2}}{1 - \frac{\kappa}{R} \sum_{i=1}^{R} (1 + (\xi \mathbf{\Sigma}_{F,i})^{-1})^{-2}}.$$

Let $\boldsymbol{h} \sim \mathcal{N}(0, \mathrm{I}/R)$. The non-asymptotic distributional prediction is given by the following random vector

$$\hat{\boldsymbol{\beta}}(\mathbf{\Sigma}_F) = \frac{1}{1 + (\xi \mathbf{\Sigma}_F)^{-1}} \odot \boldsymbol{\beta} + \frac{\sqrt{\kappa\gamma} \mathbf{\Sigma}_F^{-1/2}}{1 + (\xi \mathbf{\Sigma}_F)^{-1}} \odot \boldsymbol{h}.$$

Note that, the above formulas can be slightly simplified to have a cleaner look by introducing an additional variable $\boldsymbol{z} = \frac{1}{1+(\xi \mathbf{\Sigma}_F)^{-1}}$.

Also note that, the terms in the non-asymptotic distribution characterization and asymptotic distribution characterization have one to one correspondence. Non-asymptotic distribution characterization is essentially a discretized version of asymptotic DC where instead of expectations (which is integral over pdf) we have summations.

Now, we can use this distribution to predict the test risk by using Def. 6 in the risk expression.

Going back to representation question, without losing generality, assume $\mathbf{\Sigma}_F = \boldsymbol{I}$ and let us find optimal $\mathbf{\Lambda}_R$. Then

$$\hat{\boldsymbol{\beta}} = \mathbf{\Lambda}_R \left[\frac{1}{1 + (\xi \mathbf{\Lambda}_R^2)^{-1}} \odot \mathbf{\Lambda}_R^{-1} \boldsymbol{\beta} + \frac{\sqrt{\kappa\gamma} \mathbf{\Lambda}_R^{-1}}{1 + (\xi \mathbf{\Lambda}_R^2)^{-1}} \odot \boldsymbol{h} \right].$$

The risk is given by (using $\boldsymbol{h} \sim \mathcal{N}(0, \boldsymbol{I}_p)$)

$$\text{risk}^{\mathbf{\Lambda}_R}(\hat{\boldsymbol{\beta}}_n) - \sigma_R^2 = \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{\Sigma}_F (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \tag{B.14}$$

$$= \sum_{i=1}^{R} \frac{\mathbf{\Sigma}_{T,i}}{(1 + \xi (\mathbf{\Lambda}_{R,i})^2)^2} + \sum_{i=1}^{R} \frac{\kappa\gamma}{(1 + (\xi (\mathbf{\Lambda}_{R,i})^2)^{-1})^2}. \tag{B.15}$$

Here, note that ξ is function of $\mathbf{\Lambda}^*$ and γ is function of $\boldsymbol{\beta}, \mathbf{\Lambda}^*$. If we don't know $\mathbf{\Sigma}_T$, we use the estimation from representation learning $\hat{\mathbf{\Sigma}}_T$ instead.

To find the optimal representation, we will solve the following optimization problem that minimizes the risk.

$$\begin{aligned} \min_{\mathbf{\Lambda}^*} \quad & \textstyle\sum_{i=1}^{R} \frac{\boldsymbol{\beta}_i^2}{(1 + \xi (\mathbf{\Lambda}_{R,i})^2)^2} + \textstyle\sum_{i=1}^{R} \frac{\kappa\gamma}{(1 + (\xi (\mathbf{\Lambda}_{R,i})^2)^{-1})^2} \\ \text{s.t.} \quad & \kappa^{-1} = \frac{1}{R} \textstyle\sum_{i=1}^{R} (1 + (\xi (\mathbf{\Lambda}_{R,i})^2)^{-1})^{-1} \\ & \gamma = \frac{\sigma_R^2 + \sum_{i=1}^{R} \frac{\boldsymbol{\beta}_i^2}{(1+\xi (\mathbf{\Lambda}_{R,i})^2)^2}}{1 - \frac{\kappa}{R} \sum_{i=1}^{R} (1 + (\xi (\mathbf{\Lambda}_{R,i})^2)^{-1})^{-2}}. \end{aligned} \tag{B.16}$$

So we plug in the expression of γ and get

$$\kappa\gamma = \frac{\sigma_R^2 + \frac{1}{R} \sum_{i=1}^{R} \frac{\boldsymbol{\beta}_i^2}{(1+\xi (\mathbf{\Lambda}_{R,i})^2)^2}}{\kappa^{-1} - \frac{1}{R} \sum_{i=1}^{R} (1 + (\xi (\mathbf{\Lambda}_{R,i})^2)^{-1})^{-2}} = \frac{R\sigma_R^2 + \sum_{i=1}^{R} \frac{\boldsymbol{\beta}_i^2}{(1+\xi (\mathbf{\Lambda}_{R,i})^2)^2}}{\sum \frac{\xi (\mathbf{\Lambda}_{R,i})^2}{(1+\xi (\mathbf{\Lambda}_{R,i})^2)^2}}. \tag{B.17}$$

Let $\boldsymbol{\theta}_i = \frac{\xi(\boldsymbol{\Lambda}_{R,i})^2}{1+\xi(\boldsymbol{\Lambda}_{R,i})^2}$, then the objective function becomes

$$\sum_{i=1}^{R} \boldsymbol{\Sigma}_{T,i}(1-\boldsymbol{\theta}_i)^2+(\sum_{i=1}^{R}\boldsymbol{\theta}_i^2)\frac{R\sigma_R^2+\sum \boldsymbol{\Sigma}_{T,i}(1-\boldsymbol{\theta}_i)^2}{\sum_{i=1}^{R}\boldsymbol{\theta}_i(1-\boldsymbol{\theta}_i)} = \frac{n_2(\sum_{i=1}^{R}\boldsymbol{\Sigma}_{T,i}(1-\boldsymbol{\theta}_i)^2)+R\sigma_R^2(\sum_{i=1}^{R}\boldsymbol{\theta}_i^2)}{n_2-\sum_{i=1}^{R}\boldsymbol{\theta}_i^2}$$

such that $0 \le \boldsymbol{\theta}_i < 1$ and $\sum_{i=1}^{R}\boldsymbol{\theta}_i = \frac{R}{\kappa} = n_2$. This quantity is same as the objective (B.16). We divide this quantity by d to get the risk function, which is same as the definition of f in (3.3).

B.5 Solving the optimization problem.

Here, we propose the algorithm for minimizing $f(\boldsymbol{\theta})$. We explore the KKT condition for its optimality.

The objective function is

$$f(\boldsymbol{\theta}) = \sum_{i=1}^{R}\boldsymbol{\Sigma}_{T,i}(1-\boldsymbol{\theta}_i)^2 + (\sum_{i=1}^{R}\boldsymbol{\theta}_i^2)\frac{R\sigma_R^2+\sum \boldsymbol{\Sigma}_{T,i}(1-\boldsymbol{\theta}_i)^2}{\sum_{i=1}^{R}\boldsymbol{\theta}_i(1-\boldsymbol{\theta}_i)}. \tag{B.18}$$

Lemma 1 *Let $C, S, V \in \mathbb{R}$. Define*

$$\phi(\boldsymbol{\Sigma}_{T,i}; C, V, S) := \frac{Cp(R-n_2-S)^2}{2n_2(V+R\sigma_R^2+(R-n_2-S)\boldsymbol{\Sigma}_{T,i}{}^2)}$$

and we find the root of the following equations:

$$\sum_{i=1}^{R}\phi(\boldsymbol{\Sigma}_{T,i}; C, V, S) = R-n_2,$$
$$\sum_{i=1}^{R}\phi^2(\boldsymbol{\Sigma}_{T,i}; C, V, S) = S-(2n_2-R),$$
$$\sum_{i=1}^{R}\boldsymbol{\Sigma}_{T,i}\phi^2(\boldsymbol{\Sigma}_{T,i}; C, V, S) = V.$$

Let $\boldsymbol{\theta}_i = 1-\phi(\boldsymbol{\Sigma}_{T,i}; C^*, V^*, S^*)$ where C^*, V^*, S^* are the roots, then

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}'} \ f(\boldsymbol{\theta}'), \quad \textit{s.t.}, \ 0 \le \boldsymbol{\theta}' < 1, \ \sum_{i=1}^{R}\boldsymbol{\theta}'_i = n_2.$$

Proof Define $s = \sum_{i=1}^{R}\boldsymbol{\theta}_i^2$, $\phi_i = 1-\boldsymbol{\theta}_i$. Define $Q = \dfrac{1}{R}\sum_{i=1}^{R}\boldsymbol{\Sigma}_{T,i}\phi_i^2$. Then

$$\begin{aligned} f(\phi) &= \sum_{i=1}^{R}\boldsymbol{\Sigma}_{T,i}\phi_i^2 + \frac{s}{n_2-s}(R\sigma_R^2+\sum_{i=1}^{R}\boldsymbol{\Sigma}_{T,i}\phi_i^2) \\ &= R(Q+\frac{s}{n_2-s}(\sigma_R^2+Q)) \\ &= \frac{Rn_2}{R-n_2-\sum_{i=1}^{R}\phi_i^2}(Q+\sigma_R^2). \end{aligned}$$

The last line uses

$$s = \sum_{i=1}^{R}(1-\phi^2) = R-2\sum_{i=1}^{R}\phi_i+\sum_{i=1}^{R}\phi_i^2 = R-2(R-n_2)+\sum_{i=1}^{R}\phi_i^2 = 2n_2-R+\sum_{i=1}^{R}\phi_i^2.$$

Now define $\sum_{i=1}^{R}\phi_i^2 = S$, and we compute the gradient of f, we have

$$\frac{df}{R\phi_i} = \left(2n_2(\sum_{j=1}^{R}\boldsymbol{\Sigma}_{Tj}\phi_j^2+(R-n_2-s)\boldsymbol{\Sigma}_{T,i})+2Rn_2\sigma_R^2\right)\phi_i.$$

Suppose $0 < \phi_i < 1$, then we need $\frac{df}{R\phi_i}$ equal to each other for all i. Suppose $\frac{df}{R\phi_i} = C$, and denote $\sum \boldsymbol{\Sigma}_{Tj}\phi_j^2 = V$, we can solve for ϕ_i from $\frac{df}{R\phi_i} = C$ as

$$\phi_i = \frac{Cd(R - n_2 - S)^2}{2n_2(V + R\sigma_R^2 + (R - n_2 - S)\boldsymbol{\Sigma}_{T,i}{}^2)} := \phi(\boldsymbol{\Sigma}_{T,i}; C, V, S). \tag{B.19}$$

We define the function $\phi(\boldsymbol{\Sigma}_{T,i}; C, V, S)$ as above, and use the fact that

$$\sum_{i=1}^{R} \phi(\boldsymbol{\Sigma}_{T,i}; C, V, S) = R - n_2,$$
$$\sum_{i=1}^{R} \phi^2(\boldsymbol{\Sigma}_{T,i}; C, V, S) = S - (2n_2 - R),$$
$$\sum_{i=1}^{R} \boldsymbol{\Sigma}_{T,i}\phi^2(\boldsymbol{\Sigma}_{T,i}; C, V, S) = V.$$

We can solve[9] C, V, S and retrieve ϕ_i by (B.19). $\boldsymbol{\theta}_i = 1 - \phi_i$. ■

C Analysis of MoM estimators

C.1 Covariance estimator

We will first present the estimation error of the feature covariance $\boldsymbol{\Sigma}_F$, which is not covered in the main paper due to limitation of space. Note that if $\boldsymbol{\Sigma}_F$ is fully aligned with $\boldsymbol{\Sigma}_T$, e.g., $\boldsymbol{\Sigma}_F = \boldsymbol{\Sigma}_T$, then estimating $\boldsymbol{\Sigma}_F$ is enough for getting optimal representation, and we will show it has lower sample complexity and error compared to estimating canonical covariance $\tilde{\boldsymbol{\Sigma}}_T$. That is a naive case, if it does not work, this intermediate result will help in our latter proof.

We will use the following Bernstein type concentration lemma, generalized from [37, Lemma 29]:

Lemma 2 *Let $\boldsymbol{Z} \in \mathbb{R}^{n_1 \times n_2}$. Choose T_0, σ^2 such that*

1. $\boldsymbol{P}(\|\boldsymbol{Z}\| \geq C_0T_0 + t) \leq \exp(-c\sqrt{t/T_0})$.
2. $\|\boldsymbol{E}(\boldsymbol{Z}\boldsymbol{Z}^\top)\|, \|\boldsymbol{E}(\boldsymbol{Z}^\top\boldsymbol{Z})\| \leq \sigma^2$.

Then with probability at least $1 - (nT_0)^{-c}$, $c > 10$,

$$\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{Z}_i)\| \lesssim \log(nT_0)\left(\frac{T_0\log(nT_0)}{n} + \frac{\sigma}{\sqrt{n}}\right).$$

Proof Define $K = \log^2(C_K nT_0)$ for $C_K > 0$, $\boldsymbol{Z}' = \boldsymbol{Z}\boldsymbol{1}(\|\boldsymbol{Z}\| \leq KT_0)$, then

$$\|\boldsymbol{E}(\boldsymbol{Z} - \boldsymbol{Z}')\| \leq \int_{KT_0}^{\infty} \exp(-c\sqrt{t/T_0})dt \lesssim (1 + \sqrt{K})\exp(-c\sqrt{K})T_0$$
$$\lesssim (1 + \log(C_K nT_0))(nT_0)^{-C}.$$

We can choose C_K large enough so that $C > 10$. We will use [37, Lemma 29]. Set $R = \log^2(C_K nT_0)T_0 + C_0T_0$, $\Delta = (1 + \log(C_K nT_0))(nT_0)^{-C}$, $t = C_t\log(nT_0)(\frac{T_0\log(nT_0)}{n} + \frac{\sigma}{\sqrt{n}})$ for some $C_t > 0$, plugging in the last inequality of [37, Lemma 29], the LHS is smaller than $(nT_0)^{-c}$ for some c. We can also check $\boldsymbol{P}(\|\boldsymbol{Z}\| \geq R) \leq (nT_0)^{-c}$ for some c, thus we prove the lemma. ■

[9] For the root of 3-dim problem, the worst case we can grid the space and search with time complexity $\mathcal{O}(\varepsilon^{-3})$.

Feature Covariance. We can directly estimate the covariance of features by

$$\hat{\mathbf{\Sigma}}_F = \frac{1}{N}\sum_{j=1}^{n_1}\sum_{i=1}^{T} \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top, \tag{C.1}$$

The mean of this estimator is $\mathbf{\Sigma}_F$ and we can estimate the top r eigenvector of $\mathbf{\Sigma}_F$ with $\tilde{\mathcal{O}}(r)$ samples.

As we have defined in Phase 1, features $\boldsymbol{x}_{ij}$ are generated from $\mathcal{N}(0, \mathbf{\Sigma}_F)$. We aim to estimate the covariance $\mathbf{\Sigma}_F$. Although there are different kinds of algorithms, such as maximum likelihood estimator [1], to be consistent with the algorithms in the latter sections, we study the sample covariance matrix defined by (C.1).

Lemma 3 *Suppose $\boldsymbol{x}_i$, $i = 1, ..., N$ are generated independently from $\mathcal{N}(0, \mathbf{\Sigma}_F)$. We estimate* (C.1)*, then when $N \gtrsim r_F$, with probability $1 - \mathcal{O}((N\mathbf{tr}(\mathbf{\Sigma}_F))^{-C})$,*

$$\|\hat{\mathbf{\Sigma}}_F - \mathbf{\Sigma}_F\| \lesssim \sqrt{\frac{\|\mathbf{\Sigma}_F\|\mathbf{tr}(\mathbf{\Sigma}_F)}{N}}.$$

Denote the span of top s_F eigenvectors of $\mathbf{\Sigma}_F$ as $\boldsymbol{W}$ and the span of top s_F eigenvectors of $\hat{\mathbf{\Sigma}}_F$ as $\hat{\boldsymbol{W}}$. Let $\delta_\lambda = \lambda_{s_F}(\mathbf{\Sigma}_F) - \lambda_{s_F+1}(\mathbf{\Sigma}_F)$. Then if $N \gtrsim \frac{\|\mathbf{\Sigma}_F\|\mathbf{tr}(\mathbf{\Sigma}_F)}{\delta_\lambda^2}$, we have

$$\sin(\angle \boldsymbol{W}, \hat{\boldsymbol{W}}) \lesssim \sqrt{\frac{\|\mathbf{\Sigma}_F\|\mathbf{tr}(\mathbf{\Sigma}_F)}{N\delta_\lambda^2}}$$

Example 2 *When $\mathbf{\Sigma}_F = diag(\boldsymbol{I}_{s_F}, 0)$, we have $\sin(\angle \boldsymbol{W}, \hat{\boldsymbol{W}}) \lesssim \sqrt{\frac{s_F}{N}}$.*

Lemma 3 gives the quality of the estimation of the covariance of features $\boldsymbol{x}$. When the condition number of the matrix $\mathbf{\Sigma}_F$ is close to 1, we need $N \gtrsim d$ to get an estimation with error $\mathcal{O}(1)$. However, when the matrix $\mathbf{\Sigma}_F$ is close to rank r_F, the amount of samples to achieve the same error is smaller, and we can use $N \gtrsim r_F$ samples to get $\mathcal{O}(1)$ estimation error.

We will use Bernstein type concentration results to bound its error, and a similar technique will be used for $\hat{\boldsymbol{M}}$ in the next sections.

Proof First we observe that, the features $\boldsymbol{x}_{ij}$ among different tasks are generated i.i.d. from $\mathcal{N}(0, \mathbf{\Sigma}_F)$. So we can rewrite (C.1) as

$$\hat{\mathbf{\Sigma}}_F = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i^\top \tag{C.2}$$

where $\boldsymbol{x}_i \sim \mathcal{N}(0, \mathbf{\Sigma}_F)$. The error of $\hat{\mathbf{\Sigma}}_F$ depends on N regardless of T and n_1 respectively.

First, we know by concentration inequality

$$\boldsymbol{P}(\|\boldsymbol{x}\boldsymbol{x}^\top\| - \mathbf{tr}(\mathbf{\Sigma}_F) \geq t) = \boldsymbol{P}(\|\boldsymbol{x}\|^2 - \mathbf{tr}(\mathbf{\Sigma}_F) \geq t) \leq \exp(-c\min\{\frac{t^2}{\mathbf{tr}(\mathbf{\Sigma}_F^2)}, \frac{t}{\|\mathbf{\Sigma}_F\|}\}). \tag{C.3}$$

We will use the fact $\sqrt{\mathbf{tr}(\mathbf{\Sigma}_F^2)} \leq \mathbf{tr}(\mathbf{\Sigma}_F)$. Define $K = C_0\log(N\mathbf{tr}(\mathbf{\Sigma}_F))\mathbf{tr}(\mathbf{\Sigma}_F)$, $\boldsymbol{Z} = \boldsymbol{x}\boldsymbol{x}^\top$, $\boldsymbol{Z}' = \boldsymbol{Z}\cdot\mathbf{1}\{\|\boldsymbol{Z}\| \leq K\}$ where $\mathbf{1}$ means indicator function ($\mathbf{1}$(True) $= 1, \mathbf{1}$(False) $= 0$), for some positive number C_0. Then

$$\begin{aligned}
\|\boldsymbol{E}(\boldsymbol{Z} - \boldsymbol{Z}')\| &\leq \int_{t=K}^{\infty}(\exp(-c\frac{t^2}{\mathbf{tr}^2(\mathbf{\Sigma}_F)}) + \exp(-c\frac{t}{\|\mathbf{\Sigma}_F\|}))dt \\
&\leq \int_{t=K}^{\infty}(\exp(-c\frac{t}{\mathbf{tr}(\mathbf{\Sigma}_F)}) + \exp(-c\frac{t}{\|\mathbf{\Sigma}_F\|}))dt \\
&\leq 2\frac{\mathbf{tr}(\mathbf{\Sigma}_F)}{c}\exp(-c\frac{K}{\mathbf{tr}(\mathbf{\Sigma}_F)}) \\
&\leq \frac{\sqrt{K\mathbf{tr}^2(\mathbf{\Sigma}_F)}}{c}\exp(-\frac{cK}{\mathbf{tr}(\mathbf{\Sigma}_F)}) \\
&\lesssim (N\mathbf{tr}(\mathbf{\Sigma}_F))^{-C}
\end{aligned}$$

where $C \geq C_0 - 3/2$. Then we compute $(\boldsymbol{x}\boldsymbol{x}^\top)^2 = \|\boldsymbol{x}\|^2\boldsymbol{x}\boldsymbol{x}^\top$. Let $\boldsymbol{\Sigma}_F$ be diagonal (the proof is invariant from the basis. In other words, if $\boldsymbol{\Sigma}_F$ is not diagonal, then we can make the eigenvectors of $\boldsymbol{\Sigma}_F$ as basis and the proof applies). Then

$$\boldsymbol{E}(\|\boldsymbol{x}\|^2\boldsymbol{x}\boldsymbol{x}^\top)_{ij} = \begin{cases} \boldsymbol{\Sigma}_{Fii}(\mathbf{tr}(\boldsymbol{\Sigma}_F) + 2\boldsymbol{\Sigma}_{Fii}), & i = j, \\ 0, & i \neq j. \end{cases} \tag{C.4}$$

So $\|\boldsymbol{E}(\|\boldsymbol{x}\|^2\boldsymbol{x}\boldsymbol{x}^\top)\| \leq \|\boldsymbol{\Sigma}_F\|(\mathbf{tr}(\boldsymbol{\Sigma}_F) + 2\|\boldsymbol{\Sigma}_F\|) \approx \|\boldsymbol{\Sigma}_F\|\mathbf{tr}(\boldsymbol{\Sigma}_F)$. $\approx$ means $\gtrsim$ and $\lesssim$.

Using Lemma 2, with (C.3) and the inequality above, we get that with probability $1 - \mathcal{O}((N\mathbf{tr}(\boldsymbol{\Sigma}_F))^{-C})$,

$$\|\hat{\boldsymbol{\Sigma}}_F - \boldsymbol{\Sigma}_F\| \lesssim \log(N\mathbf{tr}(\boldsymbol{\Sigma}_F))\left(\frac{\log(N\mathbf{tr}(\boldsymbol{\Sigma}_F))\mathbf{tr}(\boldsymbol{\Sigma}_F)}{N} + \sqrt{\frac{\|\boldsymbol{\Sigma}_F\|\mathbf{tr}(\boldsymbol{\Sigma}_F)}{N}}\right). \tag{C.5}$$

If the number above is smaller than $\lambda_r - \lambda_{r+1}$, we have that

$$N \gtrsim \frac{\|\boldsymbol{\Sigma}_F\|\mathbf{tr}(\boldsymbol{\Sigma}_F)}{(\lambda_r - \lambda_{r+1})^2} \tag{C.6}$$

which is $\mathcal{O}(r)$ if condition number is 1.

The bound of the angle of top R eigenvector subspace is a direct application of the following lemma.

Lemma 4 *[14] Let $\boldsymbol{A}$ be a square matrix. Let $\hat{\boldsymbol{W}}$, $\boldsymbol{W}$ denote the span of top r singular vectors of $\hat{A}$ and $\boldsymbol{A}$. Suppose $\|\hat{\boldsymbol{A}} - \boldsymbol{A}\| \leq \Delta$, and $\sigma_r(\boldsymbol{A}) - \sigma_{r+1}(\boldsymbol{A}) \geq \Delta$, then*

$$\sin(\angle\boldsymbol{W}, \hat{\boldsymbol{W}}) \leq \frac{\Delta}{\sigma_r(\boldsymbol{A}) - \sigma_{r+1}(\boldsymbol{A}) - \Delta}.$$

So that the error of principle subspace recovery of feature covariance is upper bounded by $\frac{\|\hat{\boldsymbol{\Sigma}}_F - \boldsymbol{\Sigma}_F\|}{\sigma_r(\boldsymbol{\Sigma}_F) - \sigma_{r+1}(\boldsymbol{\Sigma}_F) - \|\hat{\boldsymbol{\Sigma}}_F - \boldsymbol{\Sigma}_F\|}$, where $\|\hat{\boldsymbol{\Sigma}}_F - \boldsymbol{\Sigma}_F\|$ is calculated in (C.5). ■

C.2 Method of moment

This section contains three parts. We first bound the norm of task vectors. Then we analyze the second result of Thm. 2, where n_1 is lower bounded by effective rank. Last we prove the first result of Thm. 2 which is a generalization of [37].

C.2.1 Property of task vectors

We first study the property of the tasks $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_T$. We know that, for any $\boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_T)$,

$$\boldsymbol{P}(\|\boldsymbol{\beta}\|^2 - \mathbf{tr}(\boldsymbol{\Sigma}_T) \geq t) \leq \exp(-c\min\{\frac{t^2}{\mathbf{tr}(\boldsymbol{\Sigma}_T^2)}, \frac{t}{\|\boldsymbol{\Sigma}_T\|}\}).$$

So that with probability at least $1 - \delta$, we have

$$\begin{aligned}\|\boldsymbol{\beta}_i\|^2 &\lesssim \mathbf{tr}(\boldsymbol{\Sigma}_T) + \sqrt{(\log(1/\delta) + \log(T))\mathbf{tr}(\boldsymbol{\Sigma}_T^2)} + (\log(1/\delta) + \log(T))\|\boldsymbol{\Sigma}_T\| \\ &\lesssim \mathbf{tr}(\boldsymbol{\Sigma}_T) + \log(T/\delta)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_T^2)} \lesssim \mathbf{tr}(\boldsymbol{\Sigma}_T)\log(T/\delta),\ \forall i = 1, ..., T. \end{aligned} \tag{C.7}$$

With similar technique we know that with probability at least $1 - \delta$,

$$\|\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\|^2 \lesssim \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \log(T/\delta)\sqrt{\mathbf{tr}((\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F)^2)},\ \forall i = 1, ..., T. \tag{C.8}$$

$$\|\boldsymbol{\Sigma}_F^{1/2}\boldsymbol{\beta}_i\|^2 \lesssim \mathbf{tr}(\boldsymbol{\Sigma}_F^{1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F^{1/2}) + \log(T/\delta)\sqrt{\mathbf{tr}((\boldsymbol{\Sigma}_F^{1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F^{1/2})^2)},\ \forall i = 1, ..., T. \tag{C.9}$$

We will use $\delta = T^{-c}$ for some constant c so that $\log(T/\delta) = (c+1)\log(T) \approx \log(T)$. Later, we will use the norm bounds of above quantities which happen with probability at least $1 - T^{-c}$.

C.2.2 Estimating with fewer samples when each task contains enough samples

In this part we will prove Theorem 6, which is the second case of Theorem 2. First we will give a description of standard normal features, then prove the general version.

Theorem 5 *(Standard normal feature, noiseless) Let data be generated as in Phase 1, let* $\mathcal{S} = \max\{\|\boldsymbol{\Sigma}_F\|, \|\boldsymbol{\Sigma}_T\|\}$ *in this theorem and the following section*[10]*,* $\tilde{r}_T = \mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F)$*,* $r_F = \mathbf{tr}(\boldsymbol{\Sigma}_F)$*,* $r_T = \mathbf{tr}(\boldsymbol{\Sigma}_T)$*. Suppose* $\sigma = 0$*,* $\boldsymbol{\Sigma}_F = \boldsymbol{I}$*, and suppose the rank of* $\boldsymbol{\Sigma}_T$ *is* s_T*. Define* $\hat{\boldsymbol{\beta}}_i = n_1^{-1}\sum_{j=1}^{n_1} y_{ij}\boldsymbol{x}_{ij}$*,* $\boldsymbol{B} = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_T]$*, and* $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_T]$*. Let* $n_1 > c_1 r_T \lambda_{s_T}^{-1}(\boldsymbol{\Sigma}_T)$*, with probability* $1 - \mathcal{O}(T^{-C})$*, where C is constant,*

$$\sigma_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B}) \lesssim \sqrt{\frac{T r_T}{n_1}}.$$

Denote the span of top s_T *singular column vectors of* $\hat{\boldsymbol{B}}$ *and* $\boldsymbol{\Sigma}_T$ *as* $\hat{\boldsymbol{W}}, \boldsymbol{W}$*, then*

$$\sin(\angle \hat{\boldsymbol{W}}, \boldsymbol{W}) \lesssim \sqrt{\frac{r_T}{n_1 \lambda_{s_T}(\boldsymbol{\Sigma}_T)}}.$$

For example, if $\boldsymbol{\Sigma}_T = \mathrm{diag}(\boldsymbol{I}_{s_T}, 0)$, then $\sin(\angle \hat{\boldsymbol{W}}, \boldsymbol{W}) \lesssim \sqrt{s_T/n_1}$.

Proof We first estimate $\boldsymbol{\beta}_i$ with

$$\hat{\boldsymbol{\beta}}_i = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{ij}\boldsymbol{x}_{ij}.$$

Then we fix $\boldsymbol{\beta}_i$ and compute the covariance of $y_{ij}\boldsymbol{x}_{ij}$ (its mean is $\boldsymbol{\beta}_i$).

$$\mathrm{Cov}(y_{ij}\boldsymbol{x}_{ij} - \boldsymbol{\beta}_i) = \boldsymbol{E}(\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top) - \boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top \precsim \|\boldsymbol{\beta}_i\|^2\boldsymbol{I}.$$

The first term is similar to (C.4), where the bound can is in [37, Lemma 5]. The vector $\hat{\boldsymbol{\beta}}_i$ is the average of $y_{ij}\boldsymbol{x}_{ij}$ over all j. With concentration we know that

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) \precsim \frac{\|\boldsymbol{\beta}_i\|^2}{n_1}\boldsymbol{I}. \tag{C.10}$$

Let $\boldsymbol{B} = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_T]$, and $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_T]$. Then we know the covariance of each column of $\hat{\boldsymbol{B}} - \boldsymbol{B}$ is bounded by (C.10). Thus with a constant c and probability $1 - \exp(-cT^2)$,

$$\sigma^2_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B}) \lesssim \frac{T\|\boldsymbol{\beta}_i\|^2}{n_1}. \tag{C.11}$$

We have proved in (C.7) that $\|\boldsymbol{\beta}_i\|^2 \leq \log(T)\mathbf{tr}(\boldsymbol{\Sigma}_T)$ with probability $1 - T^{-c}$. The columns of $\boldsymbol{B}$ is generated from $\mathcal{N}(0, \boldsymbol{\Sigma}_T)$, so that

$$\sigma_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B}) \lesssim \sqrt{\frac{T\log(T)\mathbf{tr}(\boldsymbol{\Sigma}_T)}{n_1}}.$$

Now we study $\boldsymbol{B}$. We know that $\boldsymbol{E}(\boldsymbol{B}\boldsymbol{B}^\top) = \boldsymbol{E}(\sum_{i=1}^T \boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top) = T\boldsymbol{\Sigma}_T$. $\boldsymbol{B}$ is a matrix with independent columns. Thus let $n_1 > c_1\mathbf{tr}(\boldsymbol{\Sigma}_T)\lambda_{s_T}^{-1}(\boldsymbol{\Sigma}_T)$, $T > \max\{c_2 d, \frac{\|\boldsymbol{\Sigma}_T\|\mathbf{tr}(\boldsymbol{\Sigma}_T)}{\lambda^2_{s_T}(\boldsymbol{\Sigma}_T)}\}$, then with Lemma 3, for Gaussian matrix with independent columns [38], with probability at least $1 - \mathcal{O}(T^{-c_3} + (T\mathbf{tr}(\boldsymbol{\Sigma}_T))^{-c_4} + \exp(-c_5T^2)) = 1 - \mathcal{O}(T^{-C})$, where c_i are constants,

$$\sigma_{s_T}(\boldsymbol{B}) \geq \sqrt{T\lambda_{s_T}(\boldsymbol{\Sigma}_T) - \mathcal{O}(\sqrt{T\|\boldsymbol{\Sigma}_T\|\mathbf{tr}(\boldsymbol{\Sigma}_T)})}.$$

Denote the span of top s_T singular vectors of $\hat{\boldsymbol{B}}$ and $\boldsymbol{\Sigma}_T$ as $\hat{\boldsymbol{W}}, \boldsymbol{W}$, with Lemma 4,

$$\sin(\angle \hat{\boldsymbol{W}}, \boldsymbol{W}) \leq \sqrt{\frac{\log(T)\mathbf{tr}(\boldsymbol{\Sigma}_T)}{n_1\lambda_{s_T}(\boldsymbol{\Sigma}_T)}}.$$

[10] in the paper we assume $\mathcal{S} = 1$ for simplicity.

■

Next, we will propose a theorem with general feature covariance and noisy data, which is a generalization of Theorem 5.

Theorem 6 *Let data be generated as in Phase 1. Suppose* $\hat{\boldsymbol{b}}_i = n_1^{-1}\sum_{j=1}^{n_1} y_{ij}\boldsymbol{x}_{ij}$, $\boldsymbol{B} = \boldsymbol{\Sigma}_F[\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_T]$, *and* $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{b}}_1, ..., \hat{\boldsymbol{b}}_T]$. *Let* $\delta_\lambda = \lambda_{s_T}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) - \lambda_{s_T+1}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F))$, *suppose* $\boldsymbol{\Sigma}_F$ *is approximately rank* s_F,

$$n_1 \gtrsim (\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|,$$
$$T \gtrsim \max\{s_F, \frac{d\lambda_{s_F+1}(\boldsymbol{\Sigma}_F)}{\|\boldsymbol{\Sigma}_F\|}\},$$

then with probability $1 - \mathcal{O}(T^{-C})$, *where* C *is constant,*

$$\sigma_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B}) \lesssim \sqrt{\frac{T(\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1}}.$$

Denote the span of top s_T *singular vectors of* $\hat{\boldsymbol{B}}$ *and* $\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F$ *as* $\hat{\boldsymbol{W}}, \boldsymbol{W}$, *if further we assume* $T \gtrsim \frac{\|\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F\|\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F)}{\delta_\lambda^2}$, *then*

$$\sin(\angle\hat{\boldsymbol{W}}, \boldsymbol{W}) \lesssim \sqrt{\frac{(\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1\delta_\lambda^2}}.$$

Example 3 *Suppose* $\boldsymbol{\Sigma}_F = diag(\boldsymbol{I}_{s_F}, \iota\boldsymbol{I}_{d-s_F})$, *and* $\boldsymbol{\Sigma}_T = diag(\boldsymbol{I}_{s_T}, 0)$, $\sigma = 0$. *Suppose* $\iota d < s_F$. *Then with* $T \gtrsim s_F$, $n_1 \gtrsim s_T$ *so that* $N \gtrsim s_F s_T$,

$$\sin(\angle\hat{\boldsymbol{W}}, \boldsymbol{W}) \lesssim \sqrt{s_T/n}.$$

Proof We let $\boldsymbol{x}_{ij} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_F)$. For the ith task, let

$$\hat{\boldsymbol{b}}_i = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{ij}\boldsymbol{x}_{ij}.$$

We fix $\boldsymbol{\beta}_i$ and compute

$$\boldsymbol{E}(y_{ij}\boldsymbol{x}_{ij}) \precsim \boldsymbol{E}(\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i) = \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i, \tag{C.12}$$

and

$$\mathrm{Cov}(y_{ij}\boldsymbol{x}_{ij} - \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i) \precsim (\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i)\boldsymbol{\Sigma}_F + \sigma^2\boldsymbol{\Sigma}_F. \tag{C.13}$$

To get the bound above, we can adopt the technique in [37, Lemma 5] such that, write $\boldsymbol{x}_{ij} = \boldsymbol{\Sigma}_F^{1/2}\boldsymbol{z}$, and reduce to $\boldsymbol{E}((\boldsymbol{z}^\top\boldsymbol{\Sigma}_F^{1/2}\boldsymbol{\beta}_i)^2\boldsymbol{\Sigma}_F^{1/2}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{\Sigma}_F^{1/2})$. The proof of [37, Lemma 5] gives the explicit bound of $\|\boldsymbol{E}((\boldsymbol{z}^\top\boldsymbol{\alpha})^2\boldsymbol{z}\boldsymbol{z}^\top)\|$ for any $\boldsymbol{\alpha}$ that equals above. The vector $\hat{\boldsymbol{b}}_i$ is the average of $y_{ij}\boldsymbol{x}_{ij}$ over all $j = 1, ..., n_1$. With concentration we know that

$$\mathrm{Cov}(\hat{\boldsymbol{b}}_i - \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i) \precsim \frac{\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i + \sigma^2}{n_1}\boldsymbol{\Sigma}_F. \tag{C.14}$$

Suppose $\boldsymbol{B} = \boldsymbol{\Sigma}_F[\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_T]$, and $\hat{\boldsymbol{B}} = [\boldsymbol{b}_1, ..., \boldsymbol{b}_T]$. $\hat{\boldsymbol{B}} - \boldsymbol{B}$ is a matrix with independent columns. Suppose $\boldsymbol{X}$ is approximately rank s_F, Let $\boldsymbol{V}_{s_F} \in \mathbb{R}^{d\times d}$ be the projection onto the top-R sigular vector space of $\boldsymbol{\Sigma}_F$ and $\boldsymbol{V}_{s_F^\perp} \in \mathbb{R}^{d\times d}$ be the projection onto the $s_F + 1$ to dth sigular vector space of $\boldsymbol{\Sigma}_F$. With T columns and $T \geq s_F$, we know that

$$\sigma_{\max}(\boldsymbol{V}_{s_F}(\hat{\boldsymbol{B}} - \boldsymbol{B})) \lesssim \frac{T(\max_i \boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1}$$
$$\sigma_{\max}(\boldsymbol{V}_{s_F^\perp}(\hat{\boldsymbol{B}} - \boldsymbol{B})) \lesssim \frac{\max\{T, d\}(\max_i \boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i + \sigma^2)\lambda_{s_T+1}(\boldsymbol{\Sigma}_F)}{n_1}$$

With similar argument as before, with probability $1 - \exp(-cT^2)$ for constant c,

$$\sigma_{\max}^2(\hat{\boldsymbol{B}} - \boldsymbol{B}) \lesssim \frac{\max\{T\|\boldsymbol{\Sigma}_F\|, d\lambda_{s_F+1}(\boldsymbol{\Sigma}_F)\}(\max_i \boldsymbol{\beta}_i^\top \boldsymbol{\Sigma}_F \boldsymbol{\beta}_i + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1}. \tag{C.15}$$

We know in (C.9) that $\|\boldsymbol{\Sigma}_F^{1/2}\boldsymbol{\beta}_i\|^2 \leq \mathcal{O}(\log(T)\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F))$ with probability $1 - T^{-c}$ for constant c. So that

$$\sigma_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B}) \lesssim \sqrt{\frac{\max\{T\|\boldsymbol{\Sigma}_F\|, d\lambda_{s_F+1}(\boldsymbol{\Sigma}_F)\}(\log(T)\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1}}. \tag{C.16}$$

Now we study $\boldsymbol{B}$. $\boldsymbol{E}(\boldsymbol{B}\boldsymbol{B}^\top) = \boldsymbol{E}(\boldsymbol{\Sigma}_F(\sum_{i=1}^T \boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top)\boldsymbol{\Sigma}_F) = T\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F$.

Thus let

$$n_1 > C_1(\log(T)\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|.$$

Now apply the concentration of Gaussian matrix with independent columns [38]. With probability $1 - \mathcal{O}(T^{-C_1} + (T\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F))^{-C_2} + \exp(-C_3T^2))$, where C_i are constants (the probability can be simplified as $1 - \mathcal{O}(T^{-C})$),

$$\sigma_{s_T}(\boldsymbol{B}) \geq \sqrt{T(\lambda_{s_T}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) - \lambda_{s_T+1}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F)) - \mathcal{O}(\sqrt{T\|\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F\|\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F)})}.$$

Denote the span of top s_T singular vectors of $\hat{\boldsymbol{B}}$ and $\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F$ as $\hat{\boldsymbol{W}}, \boldsymbol{W}$, let

$$T \gtrsim \max\{s_F, \frac{d\lambda_{s_F+1}(\boldsymbol{\Sigma}_F)}{\|\boldsymbol{\Sigma}_F\|}, \frac{\|\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F\|\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F)}{(\lambda_{s_T}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) - \lambda_{s_T+1}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F))^2}\} \tag{C.17}$$

we plug in (C.16) and Lemma 4,

$$\begin{aligned}\sin(\angle\hat{\boldsymbol{W}}, \boldsymbol{W}) &\lesssim \sqrt{(\frac{d\lambda_{s_F+1}(\boldsymbol{\Sigma}_F)}{T\|\boldsymbol{\Sigma}_F\|} + 1) \cdot \frac{(\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1(\lambda_{s_T}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) - \lambda_{s_T+1}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F))}} \\ &\approx \sqrt{\frac{(\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1(\lambda_{s_T}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) - \lambda_{s_T+1}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F))}}.\end{aligned}$$

■

C.2.3 Method of moments with arbitrary n_1

In this subsection we will analyze $\hat{\boldsymbol{B}}$ with any n_1, and propose the error of MoM estimator.

First, suppose there are at least two samples per task, we can separate the samples into two halves, and compute the following estimator.

Theorem 7 *Let data be generated as in Phase 1, and let n_1 be a even number. Define $\hat{\boldsymbol{b}}_{i,1} = 2n_1^{-1}\sum_{j=1}^{n_1/2} y_{ij}\boldsymbol{x}_{ij}$, $\hat{\boldsymbol{b}}_{i,2} = 2n_1^{-1}\sum_{j=n_1/2+1}^{n_1} y_{ij}\boldsymbol{x}_{ij}$. Define*

$$\begin{aligned}\hat{\boldsymbol{M}} &= n_1^{-1}\sum_{i=1}^T(\boldsymbol{b}_{i,1}\boldsymbol{b}_{i,2}^\top + \boldsymbol{b}_{i,2}\boldsymbol{b}_{i,1}^\top), \\ \boldsymbol{M} &= \boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F.\end{aligned}$$

Then there is a constant $c > 10$, with probability $1 - N^{-c}$,

$$\|\hat{\boldsymbol{M}} - \boldsymbol{M}\| \lesssim (\tilde{r}_T + \sigma^2)\sqrt{\frac{r_F}{N}} + \sqrt{\frac{r_T}{T}}.$$

Proof For simplicity of notation, we will define a random vector $\boldsymbol{x}$ with zero mean and covariance $\boldsymbol{\Sigma}_F$, a random vector $\boldsymbol{\beta}$ with zero mean and covariance $\boldsymbol{\Sigma}_T$, a random variable ε with zero mean and covariance σ, and they are subGaussian[11]. Let $y = \boldsymbol{x}^\top\boldsymbol{\beta} + \varepsilon$. We first estimate the mean of $\hat{\boldsymbol{M}}$.

[11] We remove the subscripts when there is no confusion.

Note that if we fix $\boldsymbol{\beta}$, $\hat{\boldsymbol{b}}_{i,1}, \hat{\boldsymbol{b}}_{i,2}$ are i.i.d., so

$$\begin{aligned}
\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1}) &= \boldsymbol{E}_{\boldsymbol{x},\varepsilon}(y\boldsymbol{x}) = \boldsymbol{E}_{\boldsymbol{x},\varepsilon}((\boldsymbol{x}^\top\boldsymbol{\beta}+\varepsilon)\boldsymbol{x}) = \boldsymbol{\Sigma}_F\boldsymbol{\beta},\\
\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{M}}) &= \frac{1}{2}(\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1})\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,2})^\top + \boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,2})\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1})^\top)\\
&= \boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1})\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1})^\top = \frac{1}{T}\boldsymbol{\Sigma}_F(\sum_{i=1}^T \boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top)\boldsymbol{\Sigma}_F.
\end{aligned}$$

We take expectation over $\boldsymbol{\beta}_i$ and get $\boldsymbol{M}$. We define the right hand side as $\bar{\boldsymbol{M}}$ for the proof below.

Next, we will bound $\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|$.

[37, Lemma 3] proposes that, with probability $1-\delta$,

$$\begin{aligned}
\|\boldsymbol{x}_{ij}\|^2 &\lesssim \log(1/\delta)\mathbf{tr}(\boldsymbol{\Sigma}_F),\\
(\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i)^2 &\lesssim \log(1/\delta)\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T),\\
\varepsilon_{ij}^2 &\lesssim \log(1/\delta)\sigma^2.
\end{aligned}$$

If we enumerate $i = 1, ..., T$ and $j = 1, ..., n_1$, there are in total $Tn_1 = N$ terms. So we set $\delta = N^{-c+1}$ for a constant $c > 1$, then with probability $1 - N^{-c}$, for all i, j we have

$$\|y_{ij}\boldsymbol{x}_{ij}\| = \|(\boldsymbol{x}_{ij}\boldsymbol{\beta}_i + \varepsilon_{ij})\boldsymbol{x}_{ij}\| \lesssim \log^{3/2}(N)\sqrt{(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\mathbf{tr}(\boldsymbol{\Sigma}_F)}.$$

Define $\boldsymbol{\delta}_{i,l} = \hat{\boldsymbol{b}}_{i,l} - \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i$ for $l = 1, 2$ (we will use $l = 1$ below, the result for $l = 2$ is the same). Note that $\boldsymbol{\delta}_i$ is zero mean. With [23, Prop. 5.1] we have with probability $1 - N^{-c}$,

$$\|\boldsymbol{\delta}_{i,1}\| \lesssim n_1^{-1/2}\log^{5/2}(N)\sqrt{(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\mathbf{tr}(\boldsymbol{\Sigma}_F)} \tag{C.18}$$

Define

$$\begin{aligned}
\boldsymbol{Z}_i &= \hat{\boldsymbol{b}}_{i,1}\hat{\boldsymbol{b}}_{i,2}^\top - \boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1}\hat{\boldsymbol{b}}_{i,2}^\top)\\
&= (\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i + \boldsymbol{\delta}_{i,1})(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i + \boldsymbol{\delta}_{i,2})^\top - \boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1}\hat{\boldsymbol{b}}_{i,2}^\top)\\
&= \boldsymbol{\delta}_{i,1}(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i)^\top + \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\delta}_{i,2}^\top + \boldsymbol{\delta}_{i,1}\boldsymbol{\delta}_{i,2}^\top - \boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\boldsymbol{\delta}_{i,1}\boldsymbol{\delta}_{i,2}^\top).
\end{aligned}$$

Then

$$\begin{aligned}
\|\boldsymbol{E}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\| \le{}& \|\boldsymbol{E}(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\delta}_{i,2}^\top + \boldsymbol{\delta}_{i,1}(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i)^\top)(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\delta}_{i,2}^\top + \boldsymbol{\delta}_{i,1}(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i)^\top)^\top\|\\
&+ \|\boldsymbol{E}\boldsymbol{\delta}_{i,1}\boldsymbol{\delta}_{i,2}^\top\boldsymbol{\delta}_{i,2}\boldsymbol{\delta}_{i,1}^\top\|.
\end{aligned} \tag{C.19}$$

Then we can use (C.18) and (C.8) to bound the first term by

$$n_1^{-1}\log^6(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma)\mathbf{tr}(\boldsymbol{\Sigma}_F)\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)\|\boldsymbol{\Sigma}_F\|^2.$$

And

$$\begin{aligned}
\boldsymbol{E}_{\boldsymbol{x},\varepsilon}\boldsymbol{\delta}_{i,1}\boldsymbol{\delta}_{i,2}^\top\boldsymbol{\delta}_{i,2}\boldsymbol{\delta}_{i,1}^\top &= (\boldsymbol{E}_{\boldsymbol{x}}\boldsymbol{\delta}_{i,2}^\top\boldsymbol{\delta}_{i,2})\|\boldsymbol{E}_{\boldsymbol{x}}\boldsymbol{\delta}_{i,1}\boldsymbol{\delta}_{i,1}^\top\|\\
&\lesssim n_1^{-2}(\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\boldsymbol{x}^\top\boldsymbol{\beta}+\varepsilon)^2\boldsymbol{x}^\top\boldsymbol{x})\|\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\boldsymbol{x}^\top\boldsymbol{\beta}+\varepsilon)^2\boldsymbol{x}\boldsymbol{x}^\top\|\\
&\lesssim n_1^{-2}(\mathbf{tr}^2(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^4)\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|.
\end{aligned}$$

The second line is due to the fact that $\boldsymbol{\delta}_{i,l}$ is the difference of $(\boldsymbol{x}^\top\boldsymbol{\beta}+\varepsilon)\boldsymbol{x}$ and its mean, and covariance is upper bounded by variance (not subtracting the mean). The n_1^{-2} factor comes from the average over n_1 terms. The reasoning of the last line is same as (C.13). Now we can go back to (C.19) and get

$$\|\boldsymbol{E}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\| \lesssim n_1^{-1}\log^6(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)^2\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|^2.$$

Next we need to bound the norm of $\boldsymbol{Z}_i$. We use (C.18) and (C.8), with probability $1 - N^{-c}$,

$$\begin{aligned}
\|\boldsymbol{Z}_i\| \le{}& n_1^{-1/2}\log^3(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F)}\|\boldsymbol{\Sigma}_F\|\\
&+ n_1^{-1}\log^5(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\mathbf{tr}(\boldsymbol{\Sigma}_F).
\end{aligned}$$

Define the upper bound for $\|\boldsymbol{E}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\|, \|\boldsymbol{Z}_i\|$ as Z_1, Z_2 (the right hand side of two above inequalities). Now we apply Bernstein type inequality (Lemma 2), with probability $1 - N^{-c}$,

$$\begin{aligned}
&\|\hat{\boldsymbol{M}} - \bar{\boldsymbol{M}}\| \\
=& \|T^{-1}\sum_{i=1}^{T}\boldsymbol{Z}_i - \boldsymbol{E}_{\boldsymbol{x}}\boldsymbol{Z}_i\| \\
\lesssim& \log(TZ_2)\left(T^{-1/2}\log(N)Z_1^{1/2} + T^{-1}Z_2\log(TZ_2)\right) \\
\lesssim& \log(TZ_2)\Big(\sqrt{\frac{\log^6(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T) + \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)^2\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|^2}{n_1 T}} \\
&+ \frac{\log^3(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T) + \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F)}\|\boldsymbol{\Sigma}_F\|}{n_1^{1/2}T} \\
&+ \frac{\log^5(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\mathbf{tr}(\boldsymbol{\Sigma}_F)}{T}\Big) \\
=& \log(TZ_2)\cdot\Big(\log^3(N)\|\boldsymbol{\Sigma}_F\|(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T) + \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\sqrt{\frac{\mathbf{tr}(\boldsymbol{\Sigma}_F)}{N}} \\
&+ \frac{\log^5(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T) + \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F)}\|\boldsymbol{\Sigma}_F\|}{N^{1/2}T^{1/2}}\Big).
\end{aligned}$$

The term

$$\|\boldsymbol{\Sigma}_F\|(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T) + \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\sqrt{\frac{\mathbf{tr}(\boldsymbol{\Sigma}_F)}{N}}$$

is the dominant term as shown in the theorem. ■

The following method of moment estimator is used in [37], where $n_1 \geq 1$. In other words, if there is one sample per task, one can use the following estimator.

Theorem 8 *Let data be generated as in Phase 1. Define* $\hat{\boldsymbol{b}}_i = n_1^{-1}\sum_{j=1}^{n_1} y_{ij}\boldsymbol{x}_{ij}$*,* $\boldsymbol{B} = \boldsymbol{\Sigma}_F[\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_T]$*, and* $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{b}}_1, ..., \hat{\boldsymbol{b}}_T]$*. Define*

$$\begin{aligned}
\hat{\mathbf{G}} &= \hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top = T^{-1}\sum_{i=1}^{T}\hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top, \\
\mathbf{G} &= \boldsymbol{E}(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top) = \boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F + n_1^{-1}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F + \mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F)\boldsymbol{\Sigma}_F + \sigma^2\boldsymbol{\Sigma}_F), \\
\bar{\boldsymbol{\Sigma}}_T &= \sum_{i=1}^{T}\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top, \\
\bar{\mathbf{G}} &= \boldsymbol{\Sigma}_F\bar{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_F + n_1^{-1}(\boldsymbol{\Sigma}_F\bar{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_F + \mathbf{tr}(\bar{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_F)\boldsymbol{\Sigma}_F + \sigma^2\boldsymbol{\Sigma}_F)
\end{aligned}$$

With probability $1 - N^c$*,*

$$\|\hat{\mathbf{G}} - \bar{\mathbf{G}}\| \lesssim \|\boldsymbol{\Sigma}_F\|(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T) + \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\sqrt{\frac{\mathbf{tr}(\boldsymbol{\Sigma}_F)}{N}}.$$

Proof First, we compute the expectation of $\hat{\mathbf{G}}$.

$$\begin{aligned}
\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\hat{\mathbf{G}} &= \boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}T^{-1}(\sum_{i=1}^{T}\hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top), \\
\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top &= \boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\left(n_1^{-1}\sum_{j=1}^{n_1}(\boldsymbol{\beta}_i^\top\boldsymbol{x}_{ij} + \varepsilon_{ij})\boldsymbol{x}_{ij}\right)\left(n_1^{-1}\sum_{j=1}^{n_1}(\boldsymbol{\beta}_i^\top\boldsymbol{x}_{ij} + \varepsilon_{ij})\boldsymbol{x}_{ij}\right)^\top \\
&= n_1^{-1}\sigma^2\boldsymbol{\Sigma}_F + \boldsymbol{E}_{\boldsymbol{x}}(n_1^{-1}\sum_{j=1}^{n_1}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i)(n_1^{-1}\sum_{j=1}^{n_1}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i)^\top. \qquad \text{(C.20)}
\end{aligned}$$

Now we will study the second term. (C.12) states that $\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}(\hat{\boldsymbol{b}}_i) = \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i$. And $\hat{\boldsymbol{b}}_i$ is an average of n_1 terms, we use the expression of the covariance of sample means to get

$$
\begin{aligned}
\mathbf{Cov}(\hat{\boldsymbol{b}}_i) &= n_1^{-1}\mathbf{Cov}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i), && \text{(C.21)}\\
\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top &= \boldsymbol{E}_{\boldsymbol{x}}(n_1^{-1}\sum_{j=1}^{n_1}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i)(n_1^{-1}\sum_{j=1}^{n_1}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i)^\top\\
&= \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F + n_1^{-1}\mathbf{Cov}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i) && \text{(C.22)}
\end{aligned}
$$

Now we study $\mathbf{Cov}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i)$.

$$
\begin{aligned}
\mathbf{Cov}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i) &= \boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i - \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i)(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i - \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i)^\top\\
&= \boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i)(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i)^\top - \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F
\end{aligned}
$$

Let $\boldsymbol{x} = \sqrt{\boldsymbol{\Sigma}_F}\boldsymbol{z}$ so that $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$. Let two indices $k, l \in [\mathrm{d}]$. When $k \neq l$,

$$
\begin{aligned}
\boldsymbol{E}_{\boldsymbol{x}}[(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i)(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i)^\top]_{kl} &= \boldsymbol{E}_{\boldsymbol{z}}(\sum_{j=1}^{d}\boldsymbol{\beta}_{i,j}\sigma_j\boldsymbol{z}_j)^2\sigma_k\boldsymbol{z}_k\sigma_l\boldsymbol{z}_l\\
&= 2\sigma_k^2\sigma_l^2\boldsymbol{\beta}_{i,k}\boldsymbol{\beta}_{i,l}
\end{aligned}
$$

And

$$
\begin{aligned}
\boldsymbol{E}_{\boldsymbol{x}}[(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i)(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i)^\top]_{kk} &= \boldsymbol{E}_{\boldsymbol{z}}(\sum_{j=1}^{d}\boldsymbol{\beta}_{i,j}\sigma_j\boldsymbol{z}_j)^2\sigma_k^2\boldsymbol{z}_k^2\\
&= \mathbf{tr}(\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i)\sigma_k^2 + 2\sigma_k^4\boldsymbol{\beta}_{i,k}^2.
\end{aligned}
$$

So that

$$
\begin{aligned}
\boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i)(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i)^\top &= 2\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F + \mathbf{tr}(\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i),\\
\mathbf{Cov}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i) &= \boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i)(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}_i)^\top - \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F\\
&= \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F + \mathbf{tr}(\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i)\boldsymbol{\Sigma}_F.
\end{aligned}
$$

We plug it back into (C.22) and (C.20) and get

$$
\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top = \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F + n_1^{-1}(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F + \mathbf{tr}(\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i)\boldsymbol{\Sigma}_F + \sigma^2\boldsymbol{\Sigma}_F).
$$

Define $\bar{\boldsymbol{\Sigma}}_T = \frac{1}{T}\sum_{j=1}^{T}\boldsymbol{\beta}_j\boldsymbol{\beta}_j^\top$. So that

$$
\begin{aligned}
\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\hat{\mathbf{G}} &= \boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}T^{-1}(\sum_{i=1}^{T}\hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top)\\
&= \boldsymbol{\Sigma}_F\bar{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_F + n_1^{-1}(\boldsymbol{\Sigma}_F\bar{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_F + \mathbf{tr}(\bar{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_F)\boldsymbol{\Sigma}_F + \sigma^2\boldsymbol{\Sigma}_F) := \bar{\mathbf{G}}.\\
\boldsymbol{E}_{\boldsymbol{\beta}}\hat{\mathbf{G}} &= \mathbf{G}.
\end{aligned}
$$

We fix all $\boldsymbol{\beta}_i$ and study $\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\hat{\mathbf{G}}$. Now we need to show how fast $\hat{\mathbf{G}}$ converges to $\bar{\mathbf{G}}$.

Define

$$
\begin{aligned}
\boldsymbol{Z}_i &= \hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top - \boldsymbol{E}_{\boldsymbol{x}}(\hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top)\\
&= (\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i + \boldsymbol{\delta}_i)(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i + \boldsymbol{\delta}_i)^\top - \boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i + \boldsymbol{\delta}_i)(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i + \boldsymbol{\delta}_i)^\top\\
&= \boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\delta}_i^\top + \boldsymbol{\delta}_i(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i)^\top + \boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top - \boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\delta}_i^\top + \boldsymbol{\delta}_i(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i)^\top + \boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top).
\end{aligned}
$$

Then

$$
\|\boldsymbol{E}\boldsymbol{Z}_i^2\| \le \|\boldsymbol{E}(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i\boldsymbol{\delta}_i^\top + \boldsymbol{\delta}_i(\boldsymbol{\Sigma}_F\boldsymbol{\beta}_i)^\top)^2\| + \|\boldsymbol{E}\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\|.
$$

Then we can use (C.18) and (C.8) to bound the first term

$$
\|\boldsymbol{E}\boldsymbol{Z}_i^2\| \lesssim n_1^{-1}\log^6(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma)\mathbf{tr}(\boldsymbol{\Sigma}_F)\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)\|\boldsymbol{\Sigma}_F\|^2 + \|\boldsymbol{E}\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\| \quad \text{(C.23)}
$$

So we need to bound $\|\boldsymbol{E}\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\|$. Note that $\boldsymbol{\delta}_i$ is the average of $\boldsymbol{x}_{ij}(\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i+\varepsilon_{ij})$ with respect to index $j=1,...,n_1$. So we just let $\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_F)$ and study $\boldsymbol{x}(\boldsymbol{x}^\top\boldsymbol{\beta}_i+\varepsilon_{ij})$. Denote it by $\boldsymbol{u}_i$.

$$\begin{aligned}\|\boldsymbol{E}_{\boldsymbol{x}}\boldsymbol{u}_i\boldsymbol{u}_i^\top\boldsymbol{u}_i\boldsymbol{u}_i^\top\| &= \|\boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{x}^\top\boldsymbol{\beta}_i+\varepsilon_{ij})^4\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{x}\boldsymbol{x}^\top\|\\ &\lesssim \|\boldsymbol{E}_{\boldsymbol{x}}((\boldsymbol{x}^\top\boldsymbol{\beta}_i)^4+\sigma^4)\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{x}\boldsymbol{x}^\top\|\\ &\lesssim (\mathbf{tr}^2(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^4)\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|.\end{aligned}$$

So that

$$\|\boldsymbol{E}\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\|\lesssim n_1^{-2}(\mathbf{tr}^2(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^4)\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|.$$

Now we can go back to (C.23) and get

$$\|\boldsymbol{E}\boldsymbol{Z}_i^2\|\lesssim n_1^{-1}\log^6(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)^2\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|^2.$$

Next we need to bound the norm of $\boldsymbol{Z}_i$. We use (C.18) and (C.8), with probability $1-N^{-c}$,

$$\|\boldsymbol{Z}_i\|\le n_1^{-1/2}\log^3(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F)}\|\boldsymbol{\Sigma}_F\|+n_1^{-1}\log^5(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\mathbf{tr}(\boldsymbol{\Sigma}_F).$$

Define the upper bound for $\|\boldsymbol{E}\boldsymbol{Z}_i^2\|,\|\boldsymbol{Z}_i\|$ as Z_1,Z_2 (the right hand side of two above inequalities). With Bernstein type inequality (Lemma 2),with probability $1-N^{-c}$,

$$\begin{aligned}&\|\hat{\mathbf{G}}-\bar{\mathbf{G}}\|\\ &=\|T^{-1}\sum_{i=1}^{T}\boldsymbol{Z}_i-\boldsymbol{E}_{\boldsymbol{x}}\boldsymbol{Z}_i\|\\ &\lesssim\log(TZ_2)\left(T^{-1/2}\log(N)Z_1^{1/2}+T^{-1}Z_2\log(TZ_2)\right)\\ &\lesssim\log(TZ_2)\Big(\sqrt{\frac{\log^6(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)^2\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|^2}{n_1T}}\\ &\quad+\frac{\log^3(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F)}\|\boldsymbol{\Sigma}_F\|}{n_1^{1/2}T}\\ &\quad+\frac{\log^5(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\mathbf{tr}(\boldsymbol{\Sigma}_F)}{T}\Big)\\ &=\log(TZ_2)\cdot\Big(\log^3(N)\|\boldsymbol{\Sigma}_F\|(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\sqrt{\frac{\mathbf{tr}(\boldsymbol{\Sigma}_F)}{N}}\\ &\quad+\frac{\log^5(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F)}\|\boldsymbol{\Sigma}_F\|}{N^{1/2}T^{1/2}}\Big).\end{aligned}$$

■

D Proof of Robustness of Optimal Representation

Theorem 3 *Suppose the data is generated as Phase 2, $\boldsymbol{\Lambda}$ and $\underline{\theta}$ are defined in Def. 1 and the estimated task is obtained as* (3)*. Let the upper bound of $\|\hat{\boldsymbol{M}}-\boldsymbol{M}\|$ be $\mathcal{E}$. The risk of meta-learning algorithm satisfies*

$$\textit{risk}(\boldsymbol{\Lambda}_{\underline{\theta}}(R),\boldsymbol{\Sigma}_T,\boldsymbol{\Sigma}_F)-\textit{risk}(\boldsymbol{\Lambda}^*_{\underline{\theta}}(R),\boldsymbol{\Sigma}_T,\boldsymbol{\Sigma}_F)\lesssim\frac{n_2^2\cdot\mathcal{E}}{d(R-n_2)(2n_2-R\underline{\theta})\underline{\theta}}.$$

Proof In the proof below, we use $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^*$ to replace $\boldsymbol{\Lambda}_{\underline{\theta}}(R),\boldsymbol{\Lambda}^*_{\underline{\theta}}(R)$ for simplicity. We first decompose the risk as

$$\begin{aligned}&\text{risk}(\boldsymbol{\Lambda},\boldsymbol{\Sigma}_T,\boldsymbol{\Sigma}_F)-\text{risk}(\boldsymbol{\Lambda}^*,\boldsymbol{\Sigma}_T,\boldsymbol{\Sigma}_F)\\ &=\underbrace{\text{risk}(\boldsymbol{\Lambda},\hat{\boldsymbol{\Sigma}}_T,\boldsymbol{\Sigma}_F)-\text{risk}(\boldsymbol{\Lambda}^*,\hat{\boldsymbol{\Sigma}}_T,\boldsymbol{\Sigma}_F)}_{\le 0}\\ &\quad+[\text{risk}(\boldsymbol{\Lambda},\boldsymbol{\Sigma}_T,\boldsymbol{\Sigma}_F)-\text{risk}(\boldsymbol{\Lambda},\hat{\boldsymbol{\Sigma}}_T,\boldsymbol{\Sigma}_F)]+[\text{risk}(\boldsymbol{\Lambda}^*,\hat{\boldsymbol{\Sigma}}_T,\boldsymbol{\Sigma}_F)-\text{risk}(\boldsymbol{\Lambda}^*,\boldsymbol{\Sigma}_T,\boldsymbol{\Sigma}_F)].\end{aligned}$$

We know risk$(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F) -$ risk$(\boldsymbol{\Lambda}^*, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F) \leq 0$ due to the optimality of $\boldsymbol{\Lambda}$ with task covariance $\hat{\boldsymbol{\Sigma}}_T$. Now we will bound risk$(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) -$ risk$(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F)$ for arbitrary $\boldsymbol{\Lambda}$, and it automatically works for risk$(\boldsymbol{\Lambda}^*, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F) -$ risk$(\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F)$. Note that in (3.3) we know that

$$\text{risk}(\boldsymbol{\Lambda}', \boldsymbol{\Sigma}'_T) = f(\boldsymbol{\theta}; \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) := \sum_{i=1}^{R} \frac{n_2(1-\boldsymbol{\theta}_i)^2}{R(n_2 - \|\boldsymbol{\theta}\|^2)} \tilde{\boldsymbol{\Sigma}}^R_{T,i} + \frac{n_2}{n_2 - \|\boldsymbol{\theta}\|^2}\sigma^2. \tag{D.1}$$

This function is linear in $\boldsymbol{\Sigma}_T$ thus we know that

$$|\text{risk}(\boldsymbol{\Lambda}^*, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F) - \text{risk}(\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F)| \leq \frac{n_2}{d(n_2 - \|\boldsymbol{\theta}\|^2)}\mathcal{E}. \tag{D.2}$$

Now we need to bound $\|\boldsymbol{\theta}\|^2$. With the constraint $\underline{\theta} \leq \boldsymbol{\theta} < 1 - \frac{R-n_2}{n_2}\underline{\theta}$ and $\sum \boldsymbol{\theta}_i = n_2$, we know that the maximum of $\|\boldsymbol{\theta}\|^2$ happens when $(R - n_2)$ among $\boldsymbol{\theta}_i$ are $\underline{\theta}$ and the others are $1 - \frac{R-n_2}{n_2}\underline{\theta}$. With this we have

$$\begin{aligned}
\|\boldsymbol{\theta}\|^2 &\leq (R-n_2)\underline{\theta}^2 + n_2(1 - \frac{R-n_2}{n_2}\underline{\theta})^2 \\
&= (R-n_2)\underline{\theta}^2 + n_2 - 2(R-n_2)\underline{\theta} + \frac{(R-n_2)^2}{n_2}\underline{\theta}^2 \\
&= n_2 - 2(R-n_2)\underline{\theta} + \frac{(R-n_2)R}{n_2}\underline{\theta}^2
\end{aligned}$$

Thus

$$n_2 - \|\boldsymbol{\theta}\|^2 \geq (R-n_2)\underline{\theta}(2n_2 - R\underline{\theta}).$$

Plugging it into (D.2) and (D.1) leads to the theorem.

■