Understanding Inverse Scaling and Emergence in Multitask Representation Learning

M. Emrullah Ildiz
University of Michigan, Ann Arbor
eildiz@umich.edu

Zhe Zhao
Google DeepMind
zhezhao@google.com

Samet Oymak
University of Michigan, Ann Arbor
oymak@umich.edu

Abstract

Large language models exhibit strong multitasking capabilities, however, their learning dynamics as a function of task characteristics, sample size, and model complexity remain mysterious. For instance, it is known that, as the model size grows, large language models exhibit emerging abilities where certain tasks can abruptly jump from poor to respectable performance. Such phenomena motivate a deeper understanding of how individual tasks evolve during multitasking. To this aim, we study a multitask representation learning setup where *tasks can have distinct distributions*, quantified by their covariance priors. Through random matrix theory, we precisely characterize the optimal linear representation for few-shot learning that minimizes the average test risk in terms of task covariances. When tasks have equal sample sizes, we prove a reduction to an equivalent problem with a single effective covariance from which the individual task risks of the original problem can be deduced. Importantly, we introduce "*task competition*" to explain how tasks with dominant covariance eigenspectrum *emerge* faster than others. We show that *task competition* can potentially explain the *inverse scaling* of certain tasks i.e. reduced test accuracy as the model grows. Overall, this work sheds light on the risk and emergence of individual tasks and uncovers new high-dimensional phenomena (including multiple-descent risk curves) that arise in multitask representation learning.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

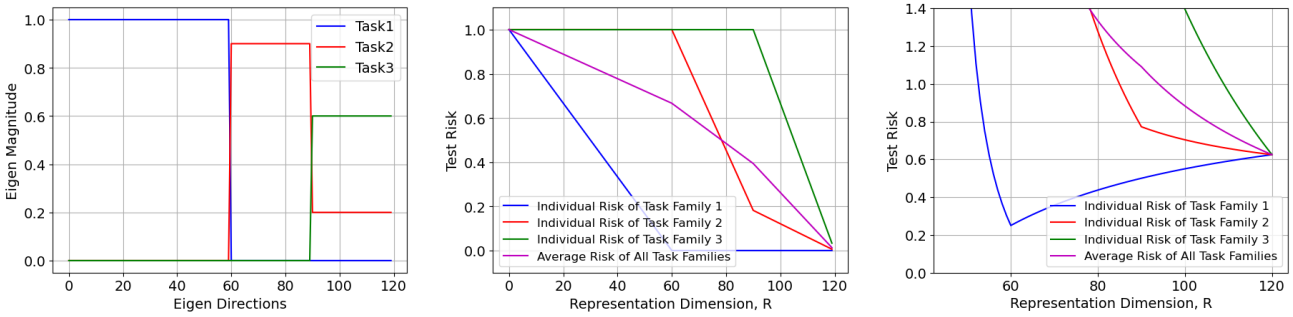
1 INTRODUCTION

Large foundation and language models have brought about a transformative impact in spanning computer vision, decision making, and natural language processing (NLP). It is now widely understood that increasing the size of language models, including aspects like computational resources and model parameters, can lead to better performance and efficiency in various NLP tasks (Wei et al., 2022; Brown et al., 2020). In numerous instances, the influence of scale on performance can be systematically anticipated through the application of scaling laws. For example, empirical evidence has shown that scaling curves for cross-entropy loss cover a substantial range that extends over seven orders of magnitude (Hoffmann et al., 2022; Kaplan et al., 2020). On the other hand, the performance of certain downstream tasks exhibits a counterintuitive pattern, where improvements do not necessarily follow increases in scale (Ganguli et al., 2022; Chowdhery et al., 2022). Specifically, it is observed in (Wei et al., 2023) and McKenzie et al. (2023) that certain tasks exhibit inverse scaling behaviors as the size of model parameters and training time increase.

The success of foundation models in several different downstream tasks (i.e. multitasking) draws enormous attention from the literature to explain the reason behind their ability. Existing multitask learning literature mostly focuses on either i.i.d. tasks or task-averaged generalization analysis. However, in reality, tasks are very diverse (e.g. sentiment analysis, mathematical reasoning, explaining humor) and each task may exhibit unique behavior such as inverse scaling or rapid emergence from low to high accuracy, as the model size grows. This motivates the fundamental question:

> **Q:** Can we characterize the individual task risks in multitask representation learning? Can this theory predict empirical phenomena such as inverse scaling and emergence?

To answer this question, we take a closer look at the scaling behaviors for the multitasking learning problem



(a) Covariance spectrum of individual tasks (intentionally non-overlapping)

(b) Optimal linear representation when tasks have infinite samples at test-time

(c) Optimal linear representation when tasks are few-shot at test-time

Figure 1: Demonstration of individual risk of each task family as well as the average risk of all task families when $\boldsymbol{\Sigma}_{\mathcal{X}} = \boldsymbol{I}_{120}$, $\boldsymbol{\Sigma}_1 = \text{diag}(\boldsymbol{I}_{60}, \boldsymbol{0}_{60})$, $\boldsymbol{\Sigma}_2 = \text{diag}(\boldsymbol{0}_{60}, 0.9\boldsymbol{I}_{30}, \boldsymbol{0}_{30})$, $\boldsymbol{\Sigma}_3 = \text{diag}(\boldsymbol{0}_{90}, 0.6\boldsymbol{I}_{30})$ as illustrated in (a), $\sigma_k^2 = 0$. For (c), the number of samples N_1, N_2 are 45.

in the context of linear regression. In the training phase, the learner is provided with training data of T tasks from K distinct task families. Each task has N training samples, $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$. Then, the learner selects a matrix $\boldsymbol{W} \in \mathbb{R}^{d \times R}$ $(R \leq d)$ to obtain a linear representation of features by the map $\boldsymbol{x}_i \to \boldsymbol{W}^\top \boldsymbol{x}_i$ similar to Hastie et al. (2020), Tripuraneni et al. (2021), Sun et al. (2021), Kong et al. (2020a). The aim of the learner is to minimize test risk with respect to the linear representation matrix $\boldsymbol{W}$ for a given R. We analyze this problem in two different settings: (1) Population risk analysis and (2) Few-shot learning. In population risk analysis, tasks have infinitely many samples during the test whereas in few-shot learning tasks have finite samples. For both of the settings, we characterize the average test risk of all tasks as well as the individual risk of each task. Specifically, we analyze the behavior of these risk curves as the model size (or the representation matrix size) R increases. For example, in Figure 1, we explore a multitask learning problem involving three different tasks. In Figure 1b and 1c, we demonstrate the average risk of all tasks and the individual risk of each task in population risk analysis setting and few-shot learning setting, respectively. As the model size R increases, the monotonicity of the risk curves is one of the main interests in this paper. In summary, our contributions are:

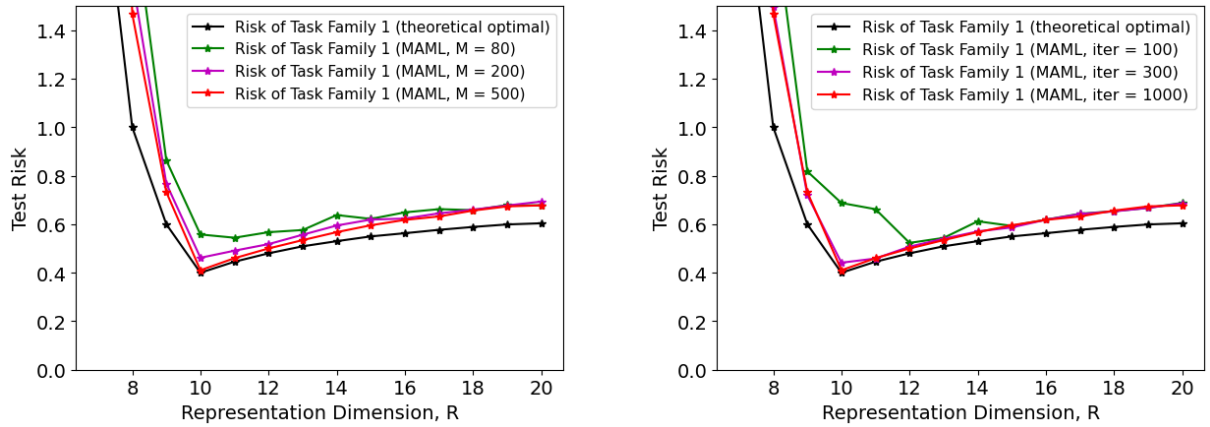
(i) **Population risk analysis.** We derive the optimal linear representation $\boldsymbol{W}$ that minimizes the multitask risk when tasks have infinite samples at test time. Through this, we characterize the task-averaged risk as well as the risks of individual tasks under optimal $\boldsymbol{W}$. We show that one can control the competition/interaction between tasks to precisely control which task emerges when and how rapidly. Concretely, one can design covariance matrices for each task to obtain arbitrary non-increasing risk curves for individual tasks. This provides a simple explanation of how the accuracy of a task can suddenly jump from poor to stellar as the model size grows.

(ii) **Few-shot learning.** We next study the optimal representation that minimizes the multitask risk when tasks have few-shot samples at test time. To do so, we reduce a multitask few-shot learning problem to an equivalent single-task problem under suitable conditions. We then provide an efficient convex optimization-based algorithm to solve single-task optimal representation. Additionally, we show that the optimal linear representation matrix is unique, which allows us to characterize the individual risk of each task.

(iii) **Understanding inverse scaling and U-shape.** Our theory can explain these phenomena at the individual task level: While task-averaged MTL risk under optimal representation has to be monotonic, the individual task risks can provably increase with model size a.k.a. inverse scaling. This is achieved by designing suitable covariance spectrums, see Fig 1c. Interestingly, the risk can also exhibit multiple ascent/descent (see our Fig 4). This explains the U-shaped behavior where inverse scaling may revert as the model size grows (Wei et al., 2023). Furthermore, we empirically justify the theoretical findings of inverse scaling behavior in multitask learning problems with MAML experiments (see Figure 2).

1.1 Related Works

Multitask Representation Learning: The idea of multitask representation learning goes back to Baxter (2000). Recently there is a resurgent interest in both statistical and optimization aspects of meta-learning theory (Denevi et al., 2018; Finn et al., 2019; Khodak et al., 2019; Zhang et al., 2023; Collins et al., 2023). A



(a) Varying task batch size M, fixed iterations

(b) Varying # of iterations, fixed batch size

Figure 2: **Linear meta learning with MAML:** MAML exhibits inverse scaling with suitable problem parameters, namely, isotropic feature covariance, noiseless labels, and orthogonal task covariances given by $\boldsymbol{\Sigma}_1 = \text{diag}(\boldsymbol{I}_{10}, \boldsymbol{0}_{10}), \boldsymbol{\Sigma}_2 = \text{diag}(\boldsymbol{0}_{10}, 0.6\boldsymbol{I}_{10})$, with few-shot sample sizes $N_1 = N_2 = 6$. The experimental details are provided in Section 4.3. There are two takeaways: (1) Inverse scaling is not unique to large language models and is reproducible under simple linear representations. (2) Our theory (black curves) provides the risk curves for optimal linear representation, formalizes inverse scaling, and matches surprisingly well with MAML training.

common of these are establishing theoretical bounds of representation learning when the tasks lie in a smaller subspace (Du et al., 2021; Tripuraneni et al., 2021; Li & Oymak, 2023b; Kong et al., 2020a; Qin et al., 2022; Kong et al., 2020b). Specifically, the recent works (Sun et al., 2021; Hastie et al., 2020; Richards et al., 2021; Wu & Xu, 2020) calculate the risk of representation learning based on the linear representation matrix, but they focus on single-task problems. Sun et al. (2021) proves that the optimal representation for the single-task problem coincides with the task of designing task-aware regularization to promote inductive bias when the regularization coefficient approaches 0. Hastie et al. (2020); Richards et al. (2021) focus on characterizing the asymptotic risk for a given task-aware regularization, but they do not minimize the test risk based on the regularization matrix. Li & Oymak (2023a) observes and formalizes related fairness challenges of multitask representations. Wu & Xu (2020) characterizes and minimizes the test risk with respect to the regularization matrix, but they minimize the test risk separately for bias and variance parts and do not provide a solution for the overall test risk. While Sun et al. (2021) considers the same problem, their analysis is restricted to single-task few-shot learning problems in the sense that there exists only one task family. Additionally, their approach to solving single-task few-shot learning problems does not take advantage of the problem's convexity and relies on the KKT conditions for the solution. In our analysis, we study the multitask few-shot learning problem and prove the existence and uniqueness of the solution, utilizing the strict convex nature of the problem. This enables us to characterize and analyze the individual risk of each task.

Double descent phenomenon: Theoretical findings of double descents were first discovered in linear regression Bartlett et al. (2020) and it was extended to ridge regression Tsigler & Bartlett (2022). These works and their subsequent works Muthukumar et al. (2019); Chang et al. (2020); Muthukumar et al. (2021); Montanari et al. (2023) study the behavior of ascend-descent risk and this double descent occurs when there is a transition from the underparametrized region to the overparameterized region. In this work, we identify a novel non-monotonic behavior of individual risk when we are inside the overparameterized region. Chen et al. (2021) provides a way to design the linear representation matrix so that it follows several different risk behaviors including double-descent and even multiple descents inside the overparameterized region. However, their selection of the linear representation matrix is not necessarily optimal, which is the main distinction from this work.

2 PROBLEM SETUP

Notation: Let $[p] = \{1, 2, \dots, p\}$. Given $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{v} \in \mathbb{R}^d$, $\|\boldsymbol{v}\|_{\boldsymbol{\Sigma}}$ represents the norm of $\boldsymbol{v}$ in the range space of $\boldsymbol{\Sigma}$, namely $\|\boldsymbol{v}\|_{\boldsymbol{\Sigma}}^2 = \boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}$. $\boldsymbol{0}_n$ and $\boldsymbol{I}_n$ represent zero and identity matrices in $\mathbb{R}^{n\times n}$. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be matrices, then $\text{diag}(\boldsymbol{A}, \boldsymbol{B})$ represents their diagonal concatenation.

In this work, we consider a multitask learning problem in the context of linear models with K task families

and T tasks where each task belongs to one of the task families. The data model for the t^{th} task is

$$\boldsymbol{\beta}_t \sim \mathcal{D}_{\mathcal{T}}, \quad (\boldsymbol{x}_i, \varepsilon_i) \sim \mathcal{D}_{\mathcal{X}} \times \mathcal{D}_{\varepsilon} \quad i \in [N] \tag{1}$$

$$\boldsymbol{y}_t = \boldsymbol{X}_t \boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t, \tag{2}$$

$$\text{where } \boldsymbol{X}_t := [\boldsymbol{x}_1^\top; \ldots; \boldsymbol{x}_N^\top], \boldsymbol{\epsilon}_t := [\varepsilon_1; \ldots; \varepsilon_N] \tag{3}$$

where $\mathcal{D}_{\mathcal{X}}$ represents the feature distribution on $\mathbb{R}^d$ such that $\mathbb{E}[\boldsymbol{x}_i] = 0, \text{Cov}[\boldsymbol{x}_i] = \boldsymbol{\Sigma}_{\mathcal{X}}$; $\mathcal{D}_{\varepsilon}$ represents the zero-mean subgaussian noise with variance σ_k^2; and $\mathcal{D}_{\mathcal{T}}$ is a mixture distribution of $(\mathcal{D}_{\mathcal{T},k})_{k=1}^K$ with prior probabilities $(\pi_k)_{k=1}^K$, where $\sum \pi_k = 1$. Note that the distribution $\mathcal{D}_{\mathcal{T},k}$ represents the distribution of k^{th} task family. If $\boldsymbol{\beta}_t \sim \mathcal{D}_{\mathcal{T},k}$, then $\mathbb{E}[\boldsymbol{\beta}_t] = 0$ and $\text{Cov}(\boldsymbol{\beta}_t) = \boldsymbol{\Sigma}_k$.

We consider a linear model that maps $d-$dimensional inputs into $R-$dimensional subspace using a representation matrix $\boldsymbol{W} \in \mathbb{R}^{d \times R}$ whose columns are orthogonal to each other and each task has its specific head $\boldsymbol{h}_t \in \mathbb{R}^R$. Specifically, given the training data consisting of $(\boldsymbol{x}_i, y_i) \in (\mathbb{R}^d, \mathbb{R})$ for a task, we regress with $\boldsymbol{W}^\top \boldsymbol{x}_i \in \mathbb{R}^R$ instead of $\boldsymbol{x}_i \in \mathbb{R}^d$ $(R \leq d)$. In the population analysis, the solution for the task-specific heads for a given representation matrix is the following:

$$\boldsymbol{h}_t^{\boldsymbol{W}} = (\boldsymbol{X}_t \boldsymbol{W})^\dagger \boldsymbol{y}_t \tag{4}$$

In the few-shot learning setting, the tasks have a few samples compared to the model size R. As a result, there are many solutions that achieve zero training error. The solution we analyze for the task-specific heads in the few-shot learning is the minimum ℓ_2 norm solution. Thus, we have the following for every $t \in [T]$:

$$\boldsymbol{h}_t^{\boldsymbol{W}} = \arg\min_{\boldsymbol{h}_t} \|\boldsymbol{h}_t\|_{\ell_2}^2 \quad \text{s.t.} \quad \boldsymbol{y}_t = \boldsymbol{X}_t \boldsymbol{W} \boldsymbol{h}_t \tag{5}$$

We define the population risk minimization problem as follows:

$$\begin{aligned} \boldsymbol{W}^* &= \arg\min_{\boldsymbol{W} \in \mathbb{R}^{d \times R}} \mathcal{R}(\boldsymbol{W}) \\ &:= \arg\min_{\boldsymbol{W} \in \mathbb{R}^{d \times R}} \frac{1}{N} \mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{X}_t \boldsymbol{W} \boldsymbol{h}_t^{\boldsymbol{W}}\|_{\ell_2}^2\right] \end{aligned}$$

where the expectation is with respect to $\boldsymbol{X}_t \sim \mathcal{D}_{\mathcal{X}}$ and $\boldsymbol{\beta}_t \sim \mathcal{D}_{\mathcal{T}}$ as shown in (1). Note that N is a normalization term, which is the number of samples in a task. In this work, we minimize the risk $\mathcal{R}(\boldsymbol{W})$ with respect to $\boldsymbol{W}$ for the population risk analysis setting (N goes to infinity) and the few-shot learning setting in Sections 3 and 4, respectively.

3 POPULATION RISK ANALYSIS

In this section, we mainly focus on finding out the optimal representation matrix $\boldsymbol{W}$ for a given R that minimizes the population risk as N goes to ∞. The task-specific heads that minimize for any $\boldsymbol{W}$ are given in (4) and utilized throughout the section.

Observation 3.1 *The optimal representation matrix for population risk is equivalent to the following:*

$$\boldsymbol{W}^* = \arg\min_{\boldsymbol{W}} \sum_{k=1}^K \pi_k \, \mathbb{E}\left[\|\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{\beta}_k - \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{W} \boldsymbol{h}_k^{\boldsymbol{W}}\|_{\ell_2}^2\right].$$

The proof is fairly straightforward and is deferred to Appendix A.

Theorem 3.2 *Assume that the eigenvalues of covariance matrices are greater than* 0. *Let* $\boldsymbol{\Sigma}$ *be the weighted covariance of all the tasks:*

$$\boldsymbol{\Sigma} = \sum_{k=1}^K \pi_k \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \tag{6}$$

Let $\boldsymbol{E} = [\boldsymbol{e}_1, \ \boldsymbol{e}_2 \ \ldots \ \boldsymbol{e}_R]$ *where* $\boldsymbol{e}_i$ *be the eigenvector corresponding to the* i^{th} *maximum eigenvalues of* $\boldsymbol{\Sigma}$. *Then, every* $\boldsymbol{W}$ *whose range space is equivalent to the range space of* $\boldsymbol{\Sigma}_{\mathcal{X}}^{-1/2} \boldsymbol{E}$ *in an optimal representation matrix.*

The proof of Theorem 3.2 is based on Observation 3.1 and provided in Appendix A. Theorem 3.2 proves that the range space of $\boldsymbol{W}$ characterizes the population risk. Then, in order to minimize the population risk, it is sufficient to find a $R-$dimensional subspace to project the $d-$dimensional inputs. Now, let $\mathcal{T}_k$ be the set of tasks that belongs to the k^{th} task family. Then, define a normalized loss function for an individual task as the following:

Lemma 3.3 *Let* $\boldsymbol{W}^*$ *be an optimal representation matrix obtained in Theorem 3.2. Let* $(\lambda_i, \boldsymbol{e}_i)$ *be the eigenvalues and eigenvectors of* $\boldsymbol{\Sigma}$ *defined in Theorem 3.2 such that* $\lambda_i \geq \lambda_{i+1}$ *and* $\|\boldsymbol{e}_i\|_{\ell_2} = 1$ *for* $i \in [d]$. *Let* $\lambda_{i,k} = \boldsymbol{e}_i^\top \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{e}_i$. *Then the population risk for the* k^{th} *task is the following:*

$$\begin{aligned} \mathcal{R}_k(\boldsymbol{W}^*) &:= \frac{1}{N} \mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{X}_t \boldsymbol{W}^* \boldsymbol{h}_t\|_{\ell_2}^2 \big| t \in \mathcal{T}_k\right] \\ &= \sigma_k^2 + \sum_{i=R+1}^d \lambda_{i,k} \end{aligned} \tag{7}$$

The proof of this lemma is provided in Appendix A. We define the emergence rate for each task family as a function of R to measure the effect of adding one dimension on the risk as follows:

Definition 3.4 *Let* $\boldsymbol{W}_R^*$ *and* $\boldsymbol{W}_{R+1}^*$ *be optimal solutions for* R *and* $R+1$ *dimensions. Then, the emergence rate* $\alpha_k(R)$ *for the* k^{th} *task family is defined as*

$$\alpha_k(R) = \mathcal{R}_k(\boldsymbol{W}_R^*) - \mathcal{R}_k(\boldsymbol{W}_{R+1}^*) \tag{8}$$

Note that $\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2}\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2}$ is a positive semidefinite matrix so $\lambda_{i,k}$ are non-negative. As a result, Lemma 3.3 reveals that the function α_k is non-negative for each task family in this setting. In addition, by the characterization of range space of $\boldsymbol{W}^*$ in Theorem 3.2, its weighted sum $\pi_k\alpha_k$ as a function of R is non-increasing. Indeed, for any given emergence rate function $(\alpha_k(R))_{k=1}^K$ (if the emergence rates satisfy the aforementioned conditions), we can devise the covariance matrices to obtain the given emergence rates as a function of R.

Theorem 3.5 *Let $(\alpha_k(R))_{k=1}^K$ be emergence rates of K task families such that $\alpha_k(R) \geq 0$ and $\sum_{k=1}^K \pi_k\alpha_k(R)$ is non-increasing as a function of R. Then, there exists a sequence of task covariances $(\Sigma_k)_{k=1}^K$, noise level σ^2, and the task family probabilities $(\pi_k)_{k=1}^K$ such that the problems follows the $(\alpha_{ki})_{i=1}^d$ emergence profile for every $k \in [K]$ under the optimal representation matrix $\boldsymbol{W}$.*

The proof of Theorem 3.5 is provided in A.

4 FEW-SHOT LEARNING

In this subsection, we consider a few-shot multitask learning problem in which we are given $K = T$ different tasks where k^{th} task belongs to the k^{th} task family. Specifically, the input/label distribution is obtained by $y_{ki} = \boldsymbol{\beta}_k^\top \boldsymbol{x}_{ki} + \varepsilon_{ki}$ for $i \in [N_k]$ where $x_{ki} \sim \mathcal{D}_{\mathcal{X}}, \boldsymbol{\beta}_k \sim \mathcal{D}_k$, and ε_{ki} is a zero-mean subgaussian random variable with variance σ_k^2. Let $\boldsymbol{\Sigma}_{\mathcal{T}}$ denote the sequences of task family covariances $(\boldsymbol{\Sigma}_k)_{k=1}^K$ and $\sigma_{\mathcal{T}}^2$ denote the sequences of noise levels $(\sigma_k^2)_{k=1}^K$. Let N_{total} be the total number of samples from all of the task families, which means $N_{total} = \sum_{k=1}^K N_k$.

We define two different regions: (1) The underparametrized region represents the case where $R < N$ and (2) the overparameterized region represents the case where $R > N$. In this part, we are interested in the asymptotic risk in the overparameterized region, thus we analyze the asymptotic risk as $N_{total} \to \infty$ while preserving the ratio between d, R, and $(N_k)_{k=1}^K$ where $N_k < R \leq d$. As we are in the over-parametrized region, the solution of task-specific heads, $\boldsymbol{h}_k^{\boldsymbol{W}}$ are given in (5). Our aim is to minimize the expected asymptotic risk by finding $\boldsymbol{W}$ assuming that the distributions $\{\mathcal{D}_k\}_{k=1}^K$ and $\mathcal{D}_{\mathcal{X}}$ are known apriori.

$$\begin{aligned}\boldsymbol{W}_{FS}^* &:= \arg\min_{\boldsymbol{W}} \mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) \qquad (9)\\ &:= \arg\min_{\boldsymbol{W}} \mathbb{E}\left[\frac{1}{N_{total}}\sum_{k=1}^K \|\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{W}\boldsymbol{h}_k^{\boldsymbol{W}}\|_{\ell_2}^2\right]\\ &= \arg\min_{\boldsymbol{W}} \sum_{k=1}^K N_k\, \mathbb{E}\left[\|\boldsymbol{W}\boldsymbol{h}_k^{\boldsymbol{W}} - \boldsymbol{\beta}_k)\|_{\boldsymbol{\Sigma}_{\mathcal{X}}}^2\right] + \sigma_k^2.\end{aligned}$$

where the expectations are taken over $\boldsymbol{\beta}_k \sim \mathcal{D}_k$, and the training samples as the task specific heads $\boldsymbol{h}_k^{\boldsymbol{W}}$ are determined based on the training samples.

In this section, we first formulate the risk and show that a few-shot multitask learning problem can be reduced to an equivalent few-shot single-task learning problem when the number of samples from different tasks is equal. Next, we provide an efficient convex optimization-based solution to the few-shot single-task learning problem. Finally, we characterize the individual risk of each task and analyze their behavior.

4.1 Transforming Multitask Problems into Single-Task Problems

In this subsection, we first reduce the dimension of $\boldsymbol{W}$ from $\mathbb{R}^{d\times R}$ to $\mathbb{R}^{R\times R}$. Then, we share a risk characterization of the few-shot multitask problem when the representation matrix is inside $\mathbb{R}^{R\times R}$. Using this characterization, we prove that there exists a few-shot single-task learning problem that is equivalent to any few-shot multitask problem.

Before we start our analysis, we restrict the set of the representation matrix $\boldsymbol{W}$ for which we minimize the risk. Note that this assumption is needed to calculate the optimal asymptotic risk and it is prevalent in the literature (Wu & Xu, 2020).

Assumption 4.1 *$\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_{\mathcal{X}}$, and $\boldsymbol{W}\boldsymbol{W}^\top$ are jointly diagonalizable matrices for $k \in [K]$.*

For any square invertible matrix $\boldsymbol{V} \in \mathbb{R}^{d\times d}$, define $\tilde{\boldsymbol{W}} = \boldsymbol{V}\boldsymbol{W}$, $\tilde{\boldsymbol{\Sigma}}_k = \boldsymbol{V}\boldsymbol{\Sigma}_k\boldsymbol{V}^\top$ for $k \in [K]$, and $\tilde{\boldsymbol{\Sigma}}_{\mathcal{X}} = (\boldsymbol{V}^\top)^{-1}\boldsymbol{\Sigma}_{\mathcal{X}}\boldsymbol{V}^{-1}$. Then, we derive the following:

$$\begin{aligned}\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) &= \mathbb{E}\left[\|\boldsymbol{W}\boldsymbol{h}_k^{\boldsymbol{W}} - \boldsymbol{\beta}_k)\|_{\boldsymbol{\Sigma}_{\mathcal{X}}}^2\right] + \sigma_k^2\\ &= \mathcal{R}_{FS}(\tilde{\boldsymbol{\Sigma}}_{\mathcal{T}}, \tilde{\boldsymbol{\Sigma}}_{\mathcal{X}}, \tilde{\boldsymbol{W}}, \sigma_{\mathcal{T}}^2) \qquad (10)\end{aligned}$$

This means that we are able to apply a linear transformation to the linear representation matrix $\boldsymbol{W}$ by applying the appropriate transformation to the covariance matrices. In addition to the linear transformation, we also reduce the dimension of $\boldsymbol{W}$ by change-of-basis and projection: Let $\boldsymbol{W}^\perp \in \mathbb{R}^{d\times R}$ be any matrix whose columns span the orthogonal complements of the subspace spanned by the columns of $\boldsymbol{W}$. Let $\boldsymbol{x}_{\boldsymbol{W}} \in \mathbb{R}^d$ and $\boldsymbol{x}_{\boldsymbol{W}^\perp} \in \mathbb{R}^d$ be the projections of $\boldsymbol{x}$ onto the column spaces of $\boldsymbol{W}$ and $\boldsymbol{W}^\perp$, respectively. We rewrite (3) for a single training sample as the following:

$$y = \boldsymbol{x}^\top\boldsymbol{\beta} + \varepsilon = \boldsymbol{x}_{\boldsymbol{W}}^\top\boldsymbol{\beta}_{\boldsymbol{W}} + \boldsymbol{x}_{\boldsymbol{W}^\perp}^\top\boldsymbol{\beta}_{\boldsymbol{W}^\perp} + \varepsilon \qquad (11)$$

In (11), we treat $x_{\boldsymbol{W}}^\top\boldsymbol{\beta}_{\boldsymbol{W}}$ as the signal and the remaining terms as noise. Note that the column rank of $\boldsymbol{W}$ is R, therefore we can obtain an equivalent

$\mathbb{R}-$dimensional vector to $\boldsymbol{x}_{\boldsymbol{W}}$ by change-of-basis. By defining a new few-shot multitask learning problem based on $\boldsymbol{x}_{\boldsymbol{W}}$ and $\boldsymbol{\beta}_{\boldsymbol{W}}$, we obtain the following proposition:

Proposition 4.2 *Suppose Assumption 4.1 holds. For every $\boldsymbol{U} \in \mathbb{R}^{d\times R}$ whose columns are the eigenvectors of the covariance matrices, there exists a unitary matrix $\boldsymbol{V} \in \mathbb{R}^{R\times R}$ such that the matrices $\bar{\boldsymbol{\Sigma}}_{\mathcal{X}} = \boldsymbol{U}^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{U}$, $\bar{\boldsymbol{\Sigma}}_k = \boldsymbol{U}^\top \boldsymbol{\Sigma}_k \boldsymbol{U}$, $\bar{\boldsymbol{W}} = \boldsymbol{U}^\top \boldsymbol{W} \boldsymbol{V}$, and the noise terms $\bar{\sigma}_k^2 = \sigma_k^2 + \text{trace}(\boldsymbol{\Sigma}_{\mathcal{X}}\boldsymbol{\Sigma}_k) - \text{trace}(\bar{\boldsymbol{\Sigma}}_{\mathcal{X}}\bar{\boldsymbol{\Sigma}}_k)$ for all $k \in [K]$ satisfy that $\bar{\boldsymbol{W}}$ is diagonal and the following equality holds:*

$$\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \mathcal{R}_{FS}(\bar{\boldsymbol{\Sigma}}_{\mathcal{T}}, \bar{\boldsymbol{\Sigma}}_{\mathcal{X}}, \bar{\boldsymbol{W}}, \bar{\sigma}_{\mathcal{T}}^2)$$

The proof of this proposition is provided in Appendix B. An important remark related to this proposition is that $\bar{\boldsymbol{\Sigma}}_{\mathcal{X}}$ and $\bar{\boldsymbol{\Sigma}}_k$ are diagonal as the columns of $\boldsymbol{U}$ are eigenvectors of the covariance matrices.

As a result Proposition 4.2 and (10), we derive the following:

$$\begin{aligned}&\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2)\\ &= \mathcal{R}_{FS}(\bar{\boldsymbol{W}}^{-1}\bar{\boldsymbol{\Sigma}}_{\mathcal{T}}(\bar{\boldsymbol{W}}^\top)^{-1}, \bar{\boldsymbol{W}}^\top\bar{\boldsymbol{\Sigma}}_{\mathcal{X}}\bar{\boldsymbol{W}}, \boldsymbol{I}_R, \bar{\sigma}_{\mathcal{T}}^2) \quad (12)\end{aligned}$$

Note that Equation (12) reduces the diagonal matrix $\bar{\boldsymbol{W}}$ to an identity matrix by modifying the feature and ground truth covariance matrices, which enables us to obtain a characterization of $\widehat{\mathcal{R}}_{FS}$. Now, we are ready to share the asymptotic risk characterization in an overparameterized region where $d, R, N_k \to \infty$ while preserving the ratios between d, R, and N_k. Fix $\kappa_0 = \frac{R}{d} < 1$, $\kappa_k = \frac{N_k}{R} < 1$ for every $k \in [K]$.

Definition 4.3 *Let $\boldsymbol{d}_{\mathcal{X}}, (\boldsymbol{d}_k)_{k=1}^K \in \mathbb{R}^R$ be the vectors such that $diag(\boldsymbol{d}_k) := \bar{\boldsymbol{W}}^{-1}\bar{\boldsymbol{\Sigma}}_k(\bar{\boldsymbol{W}}^\top)^{-1}$ and $diag(\boldsymbol{d}_{\mathcal{X}}) := \bar{\boldsymbol{W}}^\top\bar{\boldsymbol{\Sigma}}_{\mathcal{X}}\bar{\boldsymbol{W}}$*

Assumption 4.4 *Let $d_{\mathcal{X},i}$ and $d_{k,i}$ be the i^{th} element of $\boldsymbol{d}_{\mathcal{X}}$ and $\boldsymbol{d}_k$, respectively. The joint empirical distribution of $\{(d_{\mathcal{X},i}, (d_{k,i})_{k=1}^K)\}$ converges in Wasserstein-p distance to a probability distribution $(\Lambda, (M_k)_{k=1}^K)$ on $\mathbb{R}_{>0} \times \mathbb{R}$ for some $p \geq 4$. That is $\frac{1}{R}\sum_{i\in[R]} \delta_{(d_{\mathcal{X},i}, R(d_{k,i})_{k=1}^K)} \to (\Lambda, (M_k)_{k=1}^K)$. Furthermore, there exist $c_l, c_u > 0$ such that $c_l < d_{k,i}, d_{\mathcal{X},i} < c_u$ for every $k \in [K]$ and $i \in [n]$.*

Note that Assumption 4.4 (similar to Assumption 4.1) is a standard assumption in random matrix theory to obtain the asymptotic risk characterization in the overparameterized asymptotic region. The following theorem characterizes the asymptotic risk in a few-shot multitask learning problem:

Theorem 4.5 *Recall the random variables in Assumption 4.4, and fix $(\kappa_k)_{k=1}^K > 1$. Define parameter $(\xi_k)_{k=1}^K$ as the unique positive solution to the following equation:*

$$\mathbb{E}_\Lambda\left[(1 + (\xi_k\Lambda)^{-1})^{-1}\right] = \kappa_k^{-1} \qquad k \in [K] \quad (13)$$

Further define the positive parameters $(B_k)_{k=1}^K$ and $(\Omega_k)_{k=1}^K$ as follows:

$$B_k = \mathbb{E}_{(\Lambda, M_k)}\left[\frac{M_k\Lambda}{(1+\xi_k\Lambda)^2}\right], \quad (14)$$

$$\Omega_k = \mathbb{E}_\Lambda\left[\frac{\kappa_k}{(1+(\xi_k\Lambda)^{-1})^2}\right] \quad (15)$$

Then, we have the following risk characterization:

$$\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \frac{1}{N_{total}}\sum_{k=1}^K N_k \frac{B_k + \bar{\sigma}_k^2}{1-\Omega_k} \quad (16)$$

Proof of Theorem 4.5 is based on the distribution characterization for linear Gaussian problem given in Chang et al. (2020). The full proof is provided in Appendix B.

The risk characterization provided in Theorem 4.5 enables us to reduce the number of task families from K to 1 when the number of samples from each task family is equivalent. The reduction in the number of task families can be performed by collapsing all ground truth covariance matrices and noise levels into one covariance matrix and one noise level, respectively. To formalize these statements, let us define a reduced few-shot single-task problem as follows:

Definition 4.6 *Consider a few-shot multitask problem with the feature covariance $\boldsymbol{\Sigma}_{\mathcal{X}}$, ground-truth covariances $(\boldsymbol{\Sigma}_k)_{k=1}^K$, the noise levels $(\sigma_k^2)_{k=1}^K$, and the number of sample N for each task. Define $\boldsymbol{\Sigma}_{avg} = \frac{1}{K}\sum_{k=1}^K \boldsymbol{\Sigma}_k$ and $\sigma_{avg}^2 = \frac{1}{K}\sum_{k=1}^K \sigma_k^2$. Then, the reduced few-shot single-task problem of the multitask problem is defined by the feature covariance $\boldsymbol{\Sigma}_{\mathcal{X}}$, the ground-truth covariance $\boldsymbol{\Sigma}_{avg}$, the noise level σ_{avg}^2, and the number of samples N.*

Theorem 4.7 *Suppose the distributions $(\mathcal{D}_k)_{k=1}^K$ are Gaussian and Assumptions 4.1 and 4.4 hold. Then, for any $\boldsymbol{W} \in \mathbb{R}^{d\times R}$ whose column space is $R-$dimensional the population risk for the few-shot multitask problem is equivalent to that of single-task few-shot problem. That is*

$$\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \mathcal{R}_{FS}(\boldsymbol{\Sigma}_{avg}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{avg}^2)$$

The proof of Theorem 4.7 is based on the transformation of $\boldsymbol{W}$ to identity matrix in (12) and the risk characterization in Theorem 4.5, which is provided in Appendix B.

Note that the equivalence between the original few-shot multitask learning problem and the reduced of that

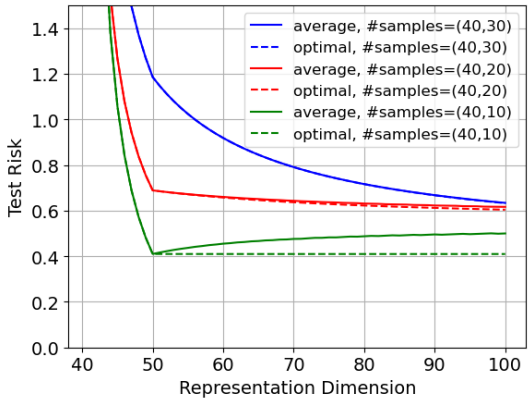


Figure 3: Comparison between average representation matrices and the optimal representation when $\boldsymbol{\Sigma}_{\mathcal{X}} = \boldsymbol{I}_{100}, \boldsymbol{\Sigma}_1 = \text{diag}(\boldsymbol{I}_{50}, \boldsymbol{0}_{50}), \boldsymbol{\Sigma}_2 = \text{diag}(\boldsymbol{0}_{50}, \boldsymbol{I}_{50}), \text{samples} = (N_1, N_2), \sigma^2 = 0$.

Algorithm 1 Solution of the Multitask Few-shot Learning Problem

Given $R, N, \boldsymbol{\Sigma}_{\mathcal{X}}, (\boldsymbol{\Sigma}_k)_{k=1}^K$, and $(\sigma_k)_{k=1}^K$
$\boldsymbol{\Sigma}_{avg} := \frac{1}{K}\sum_{k=1}^K \boldsymbol{\Sigma}_k$ and $\sigma_{avg}^2 := \frac{1}{K}\sum_{k=1}^K \sigma_k^2$
{*Multitask to Single-task Few-shot Problem*}
$\bar{\boldsymbol{U}} = \boldsymbol{U}[:, :R]$ where $\boldsymbol{U}\Sigma\boldsymbol{U}^\top = \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2}\boldsymbol{\Sigma}_{avg}\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2}$
{*Characterization of Range Space*}
$\bar{\boldsymbol{\Sigma}}_{\mathcal{X}} = \bar{\boldsymbol{U}}^\top\boldsymbol{\Sigma}_{\mathcal{X}}\bar{\boldsymbol{U}}$ and $\bar{\boldsymbol{\Sigma}}_{avg} = \bar{\boldsymbol{U}}^\top\boldsymbol{\Sigma}_{avg}\bar{\boldsymbol{U}}$
{*Reducing the dimension from d to R*}
$\bar{\sigma}_{avg}^2 = \sigma_{avg}^2 + \text{trace}(\boldsymbol{\Sigma}_{\mathcal{X}}\boldsymbol{\Sigma}_{avg}) - \text{trace}(\bar{\boldsymbol{\Sigma}}_{\mathcal{X}}\bar{\boldsymbol{\Sigma}}_{avg})$
{*Reducing the dimension from d to R*}
Find Optimal $(\zeta_i)_{i=1}^R$ by Gradient Descent in (17)
Create diagonal $\bar{\boldsymbol{W}}$ matrix: $\bar{\boldsymbol{W}}_i = (1/\zeta_i - 1)^{-1/2}$
{*Convert ζ_i to $\boldsymbol{W}_i$*}
return $\boldsymbol{W}_{FS}^* = \bar{\boldsymbol{U}}\bar{\boldsymbol{\Sigma}}_{\mathcal{X}}^{-1/2}\bar{\boldsymbol{W}}$

problem is valid only when the number of samples from each task family is equal. Figure 3 demonstrates that when the task families have different numbers of samples the optimal representation matrix for the average covariance matrices may show a non-monotonic behavior even though the optimal policy is always monotonic as R increases, which is shown in the following lemma:

Lemma 4.8 *The optimal population risk with respect to the optimal linear representation matrix $\boldsymbol{W}_{FS}^*$ is non-increasing as R increases.*

The proof of this lemma is provided in Appendix B.

4.2 Solution of Single-Task Problems

In the previous subsection, we reduce a multitask few-shot problem to an equivalent single-task few-shot learning problem. In this subsection, we provide an efficient convex optimization-based solution to a single-task few-shot learning problem for a finite dimension $R < d$. In addition, we obtain an equivalent strictly convex optimization problem to the original problem by applying a one-to-one map to the representation matrix $\boldsymbol{W}$. This enables us to obtain an efficient convex optimization-based way to solve this problem. Note that the optimization of representation matrix $\boldsymbol{W}$ for the single-task few-shot learning is studied in Sun et al. (2021) and Wu & Xu (2020). Different from these works, we prove the strong convexity of the objective function, which allows us to obtain an efficient method to solve the single-task few-shot learning problem in this section and to characterize the individual risk of each task family in the next subsection.

The range space of representation matrix $\boldsymbol{W}$ determines which R dimensional subspace of d dimensional space is utilized in the few-shot learning problem. The eigenvalues of selected dimensions appear as the terms $M_k\Lambda$ in $(B_k)_{k=1}^K$ and the remaining eigenvalues of selected features are added to noise terms $(\bar{\sigma}_k)_{k=1}^K$. The next proposition characterizes the range space of the representation matrix $\boldsymbol{W}$ with the help of Assumption 4.1.

Proposition 4.9 *Suppose Assumptions 4.1 and 4.4 hold and recall the definition of $\boldsymbol{E}$ in Theorem 3.2. The range space of $\boldsymbol{W}_{FS}^*$ is equal to the range space of $\boldsymbol{\Sigma}_{\mathcal{X}}^{-1/2}\boldsymbol{E}$.*

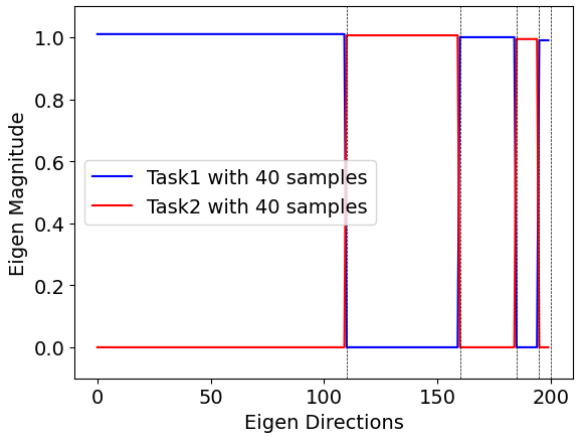
The proof of this proposition is provided in Appendix C. With the help of Assumption 4.1, Proposition 4.9 allows us to characterize the terms $(\boldsymbol{M}_k\Lambda)_{k=1}^K$ in Theorem 4.5. However, there are still free parameters of Λ in the risk expression. This means that characterizing the range space of the representation matrix $\boldsymbol{W}$ is not sufficient to minimize the average risk in the overparameterized region. In addition to the range space, it is also needed to characterize the magnitude of each column vector in $\boldsymbol{W}$. The magnitude of each column vector is determined by the distribution or the spectrum of the $(M_k\Lambda)_{k=1}^K$, which will be justified in the following theorem.

Theorem 4.10 *Suppose Assumptions 4.1 and 4.4 hold. Fix $\kappa = R/N > 1$, let $M = \frac{1}{K}\sum_{k=1}^K M_k$, and define the unique parameter $\xi \in \mathbb{R}$, the random variables $\zeta \in [0,1]$ and B as the following:*

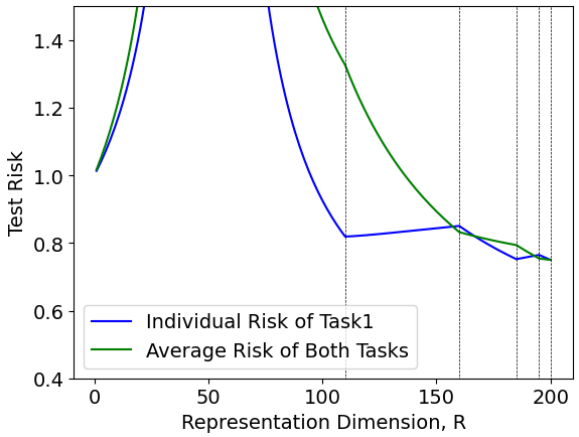
$$\mathbb{E}_\Lambda\left[(1 + (\xi\Lambda)^{-1})^{-1}\right] = \kappa^{-1} \qquad \zeta = (1 + \xi\Lambda)^{-1}$$

Further define the function $f : \mathbb{R} \to \mathbb{R}$ when the range space of the representation matrix $\boldsymbol{W}$ as stated in Proposition 4.9 as follows:

$$f(\zeta) = \frac{\mathbb{E}_{\zeta,(M\Lambda)}[M\Lambda\zeta^2] + \bar{\sigma}_{avg}^2}{1 - \kappa\,\mathbb{E}_\zeta[(1-\zeta)^2]} \tag{17}$$



(a) Covariance spectrums of individual rasks



(b) Few-shot learning risks for the optimal representation associated with the spectrums in (a)

Figure 4: Demonstration of individual risk of the first task family and the average risk of both task families when $\boldsymbol{\Sigma}_{\mathcal{X}} = \boldsymbol{I}_{200}$, $\boldsymbol{\Sigma}_k$ is illustrated in (a), $\sigma_k^2 = 0$, and $N_1 = N_2 = 50$.

Let $\mathcal{C}$ be the set of random variables ζ obeying $\Pr(\zeta \in (0,1]) = 1$ *and* $\mathbb{E}[\zeta] = 1 - \kappa^{-1}$. *Then, the following hold:*

$$\min_{\boldsymbol{W} \in \mathbb{R}^{d \times R}} \mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \min_{\zeta \in \mathcal{C}} f(\zeta) \tag{18}$$

Furthermore, $f(\zeta)$ is strongly convex on $\mathcal{C}$.

The proof of this theorem is provided in Appendix C. Utilizing the results provided in Sections 4.1 and 4.2, Algorithm 1 finds an optimal representation matrix $\boldsymbol{W}_{FS}^*$ for given finite values of representation dimension R, the number of sample N, the covariance matrices $\boldsymbol{\Sigma}_{\mathcal{X}}, (\boldsymbol{\Sigma}_k)_{k=1}^K$, and the noise levels $(\sigma_k)_{k=1}^K$. Note that the optimal representation matrix found in Algorithm 1 minimizes the risk when R, N. and d go to infinity while preserving the ratio between them.

4.3 Behaviors of Individual Task Risks

In this subsection, we prove the uniqueness of individual task family risks using the strong convexity proved in the previous subsection. In addition, we show that there are some scenarios where the individual task family risk may exhibit a non-monotonic behavior in the overparameterized region. In other words, the emergence rates of task families might be negative. Furthermore, we demonstrate some cases in which there is more than one descent-ascent region. Note that Chen et al. (2021) study a similar model and show that they are able to design their multiple descent curve by selecting an appropriate representation matrix that is not restricted to the optimal representation matrix. In this work, we prove the non-monotonic behavior of individual task family risk for the optimal representation matrix $\boldsymbol{W}$ that minimizes the average risk of all task families.

The linear representation matrix is invariant with respect to constant multiplication when the number of samples from each task family is equivalent. Namely, when $\alpha \neq 0$, we obtain the following:

$$\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \alpha\boldsymbol{W}, \sigma_{\mathcal{T}}^2)$$

This equation is valid for individual task family risk as well as the average risk of all task families and can be easily shown using (5). This implies that we are able to constrain the Frobenius norm of the representation matrix $\boldsymbol{W}$, which allows us to find a one-to-one map between ζ and $\boldsymbol{W}$. Using these relations, we prove the following proposition:

Proposition 4.11 *Suppose Assumption 4.1 holds, and recall the definition of $\boldsymbol{\Sigma}$ in Theorem 3.2. Let $(\lambda_i)_{i=1}^d$ be the eigenvalues of $\boldsymbol{\Sigma}$ where $\lambda_i \geq \lambda_{i+1}$. If $\lambda_R > \lambda_{R+1}$, then the individual risk of each task family is unique with any optimal representation matrix $\boldsymbol{W}$ that minimizes the average risk of all task families.*

The proof of this proposition is provided in Appendix D.

In Section 3, we prove that the individual risk of each task family and the average risk is always non-decreasing when N goes to infinity. Similarly, in Lemma 4.8 we prove that the average risk of all task families is non-decreasing in the overparameterized region as well. However, there are certain scenarios in which individual task family risk may exhibit non-monotonic behavior inside the overparameterized region even though their average risk is monotonic. The following lemma exemplifies such a case utilizing the definition of emergence rates in Definition 3.4 and puts a lower bound on the emergence rate of a task family:

Lemma 4.12 *Consider the scenario where* $\boldsymbol{\Sigma}_{\mathcal{X}} = \boldsymbol{I}_{p+1}$, $\boldsymbol{\Sigma}_1 = diag(\boldsymbol{I}_p, \boldsymbol{0}_1), \boldsymbol{\Sigma}_2 = diag(\boldsymbol{0}_p, \alpha \boldsymbol{I}_1)$, $\sigma_1^2 = 0$. *If* $1/2 < \alpha < 1$, *then the emergence rate of the first task family,* $\alpha_1(p+1)$, *is negative.*

The proof of this lemma is provided in Appendix D. The scenario provided in Lemma 4.12 is exemplified in Figure 1c. In addition to one descent-ascent region, we can devise covariance matrices in Figure 4 such that there are more than two ascent-descent regions even though the average risk of both task families is monotonic. This setting demonstrates the existence of U-shape behavior in the multitask representation learning problem.

Numerical Evaluations: We follow Model-Agnostic Meta-Learning (MAML) Finn et al. (2017), Collins et al. (2022) to demonstrate our findings in an experimental setting. We randomly initialize the representation matrix $\boldsymbol{W}(0)$. At each iteration t, we sample M fresh tasks from each task family with proper few-shot samples and reinitialize the task-specific heads as small random variables. We then update each head once based on its task's gradient:

$$\boldsymbol{h}_k' := \boldsymbol{h}_k - \eta_h \nabla_{h_k} \mathcal{R}_{FS}$$

Using these updated heads h_k', we update the representation matrix $\boldsymbol{W}$ following

$$\boldsymbol{W}(t+1) = \boldsymbol{W}(t) - \eta_{\boldsymbol{W}} \nabla_{\boldsymbol{W}} \mathcal{R}_{FS}$$

Following this, we conclude the iteration and start a new one. Interestingly, these MAML updates can indeed exhibit inverse scaling and match surprisingly well to the behavior of the theoretically-optimal representation. In Figure 2a, we illustrate the effect of different fresh task samples at each iteration whereas, in Figure 2b, we illustrate the effect of the number of total iterations.

5 DISCUSSION

In this work, we study the behavior of individual risk associated with each task in a linear multitask learning problem. In the population risk analysis setting in which the tasks have infinitely many samples, we find the optimal representation matrix and precisely characterize the individual risk of each task. In the few-shot learning setting in which the tasks have a few samples compared to the model size, we find the optimal representation matrix through an efficient convex optimization-based algorithm and prove the uniqueness of individual risks under mild conditions. Then, we analyze the individual risks of tasks and their emergence rate for different conditions.

An important finding of this paper is that even though the average risk of all tasks is always monotonic, the individual risk of a task can be non-monotonic as the model size grows in the few-shot learning similar to the empirical findings of inverse scaling (Wei et al., 2022, 2023). Note that this non-monotonicity is different from the double descent behavior introduced in Bartlett et al. (2020). This double descent behavior is observed on the transition from the underparameterized region to the overparameterized region. On the other hand, the non-monotonic behavior of individual risk that we point out is observed when we are completely inside the overparameterized region.

The future direction of this work includes the extension of the current analysis to the case where the number of samples from each task is not equal in the few-shot learning part. In addition, the trade-off between the cost of adding more dimensions and their effect on the individual risk of each task can be further studied utilizing the emergence rate concept.

Acknowledgment

This work was supported in part by the NSF grants CCF-2046816, CCF-2212426, CNS-1932254, UMich's MIDAS PODS program, a Google Research Scholar award, and an Adobe Data Science Research award.

References

Peter L. Bartlett, Philip M. Long, Gá bor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, apr 2020. doi: 10.1073/pnas.1907378117. URL https://doi.org/10.1073%2Fpnas.1907378117.

J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, mar 2000. doi: 10.1613/jair.731. URL https://doi.org/10.1613%2Fjair.731.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran As-

sociates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks, 2020.

Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve, 2021.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Liam Collins, Aryan Mokhtari, Sewoong Oh, and Sanjay Shakkottai. Maml and anil provably learn representations. *arXiv preprint arXiv:2202.03483*, 2022.

Liam Collins, Hamed Hassani, Mahdi Soltanolkotabi, Aryan Mokhtari, and Sanjay Shakkottai. Provable multi-task representation learning by two-layer relu neural networks, 2023.

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Incremental learning-to-learn with statistical guarantees, 2018.

Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=pW2Q2xLwIMD.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning, 2019.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, jun 2022. doi: 10.1145/3531146.3533229. URL https://doi.org/10.1145%2F3531146.3533229.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh. Meta-learning for mixed linear regression. In *International Conference on Machine Learning*, pp. 5394–5404. PMLR, 2020a.

Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh. Meta-learning for mixed linear regression, 2020b.

Yingcong Li and Samet Oymak. On the fairness of multitask representation learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Yingcong Li and Samet Oymak. Provable pathways: Learning multiple tasks over multiple paths. *AAAI Conference on Artificial Intelligence*, 2023b.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn't better, 2023.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime, 2023.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression, 2019.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter?, 2021.

Yuzhen Qin, Tommaso Menara, Samet Oymak, ShiNung Ching, and Fabio Pasqualetti. Non-stationary representation learning in sequential linear bandits. *IEEE Open Journal of Control Systems*, 1:41–56, 2022.

Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, pp. 3889–3897. PMLR, 2021.

Yue Sun, Adhyyan Narang, Ibrahim Gulluk, Samet Oymak, and Maryam Fazel. Towards sample-efficient overparameterized meta-learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28156–28168, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/ed46558a56a4a26b96a68738a0d28273-Paper.pdf.

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

A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression, 2022.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

Jason Wei, Najoung Kim, Yi Tay, and Quoc V. Le. Inverse scaling can become u-shaped, 2023.

Denny Wu and Ji Xu. On the optimal weighted \ell_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10112–10123, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/72e6d3238361fe70f22fb0ac624a7072-Paper.pdf.

Thomas T. C. K. Zhang, Leonardo F. Toso, James Anderson, and Nikolai Matni. Meta-learning operators to optimality from multi-task non-iid data, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Risk and Emergence of Individual Tasks in Multitask Representations: Supplementary Materials

A Proof of Statements in Section 3

Observation A.1 (Restated Observation 3.1) *The optimal representation matrix for population risk is equivalent to the following:*

$$\boldsymbol{W}^* = \arg\min_{\boldsymbol{W}} \min_{(\boldsymbol{h}_t)_{t=1}^T} \mathcal{R}(\boldsymbol{f}) = \arg\min_{\boldsymbol{W}} \sum_{k=1}^K \pi_k \, \mathbb{E}_{\boldsymbol{\beta}_k \sim \mathcal{D}_k} \left[\min_{\boldsymbol{h}_k} \|\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{\beta}_k - \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{W} \boldsymbol{h}_k\|_{\ell_2}^2 \right].$$

Proof.

$$\begin{aligned}
\boldsymbol{W}^* &= \arg\min_{\boldsymbol{W}} \min_{(\boldsymbol{h}_t)_{t=1}^T} \mathcal{R}(f) && (19)\\
&= \arg\min_{\boldsymbol{W}} \min_{(\boldsymbol{h}_t)_{t=1}^T} \mathbb{E}\left[\frac{1}{NT} \sum_{t=1}^T \|\boldsymbol{y}_t - \boldsymbol{X}_t \boldsymbol{W} \boldsymbol{h}_t\|_{\ell_2}^2 \right] && (20)\\
&= \arg\min_{\boldsymbol{W}} \sum_{k=1}^K N_k \, \mathbb{E}\left[\min_{\boldsymbol{h}_k} \|\boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k - \boldsymbol{X}_k \boldsymbol{W} \boldsymbol{h}_k\|_{\ell_2}^2 \right] && (21)\\
&= \arg\min_{\boldsymbol{W}} \sum_{k=1}^K \pi_k \, \mathbb{E}\left[\min_{\boldsymbol{h}_k} \boldsymbol{\beta}_k^\top \boldsymbol{X}_k^\top \boldsymbol{X}_k \boldsymbol{\beta}_k - 2\boldsymbol{\beta}_k^\top \boldsymbol{X}_k^\top \boldsymbol{X}_k \boldsymbol{W} \boldsymbol{h}_k + \boldsymbol{h}_k^\top \boldsymbol{W}^\top \boldsymbol{X}^\top \boldsymbol{X}_k \boldsymbol{W} \boldsymbol{h}_k + \boldsymbol{\epsilon}_k^2 \right] && (22)\\
&= \arg\min_{\boldsymbol{W}} \sum_{k=1}^K \pi_k \, \mathbb{E}_{\boldsymbol{\beta}_k \sim \mathcal{D}_{\mathcal{T},k}} \left[\min_{\boldsymbol{h}_k} \boldsymbol{\beta}_k^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta}_k - 2\boldsymbol{\beta}_k^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W} \boldsymbol{h}_k + \boldsymbol{h}_k^\top \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W} \boldsymbol{h}_k + \boldsymbol{\epsilon}_k^2 \right] && (23)\\
&= \arg\min_{\boldsymbol{W}} \sum_{k=1}^K \pi_k \, \mathbb{E}_{\boldsymbol{\beta}_k \sim \mathcal{D}_{\mathcal{T},k}} \left[\min_{\boldsymbol{h}_k} \|\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{\beta}_k - \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{W} \boldsymbol{h}_k\|_{\ell_2}^2 \right] && (24)
\end{aligned}$$

where the expectations in (20) and (21) are taken with respect to $\boldsymbol{\beta}_k, \boldsymbol{X}, \boldsymbol{\epsilon}$ and the expectation in (22) is taken with respect to $\boldsymbol{\beta}_k$ and $\boldsymbol{X}$. ■

Theorem A.2 (Restated Theorem 3.2) *Assume that the eigenvalues of covariance matrices are greater than 0. Let $\boldsymbol{\Sigma}$ be the weighted covariance of all the tasks:*

$$\boldsymbol{\Sigma} = \sum_{k=1}^K \pi_k \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \tag{25}$$

Let $\boldsymbol{E} = [\boldsymbol{e}_1, \ \boldsymbol{e}_2 \ \ldots \ \boldsymbol{e}_R]$ where $\boldsymbol{e}_i$ be the eigenvector corresponding to the i^{th} maximum eigenvalues of $\boldsymbol{\Sigma}$. Then, every $\boldsymbol{W}$ whose range space is equivalent to the range space of $\boldsymbol{\Sigma}_{\mathcal{X}}^{-1/2} \boldsymbol{E}$ in an optimal representation matrix.

Proof. From observation A.1, we have the following:

$$\boldsymbol{W}^* = \arg\min_{\boldsymbol{W}} \sum_{k=1}^{K} \pi_k \, \mathbb{E}_{\boldsymbol{\beta}_k \sim \mathcal{D}_{\mathcal{T},k}} \left[\min_{\boldsymbol{h}_k} \boldsymbol{\beta}_k^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{\beta}_k - 2\boldsymbol{\beta}_k^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W} \boldsymbol{h}_k + \boldsymbol{h}_k^\top \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W} \boldsymbol{h}_k \right] \tag{26}$$

The optimal task-specific heads are given in (4) as follows:

$$\boldsymbol{h}_t^{\boldsymbol{W}} = (\boldsymbol{X}_t \boldsymbol{W})^\dagger \boldsymbol{y}_t \tag{27}$$

When we plug the task-specific heads in (26), we obtain the following:

$$\begin{aligned}
\boldsymbol{W}^* &= \arg\min_{\boldsymbol{W}} \sum_{k=1}^{K} \pi_k \, \mathbb{E}_{\boldsymbol{\beta}_k \sim \mathcal{D}_{\mathcal{T},k}} \left[-\boldsymbol{\beta}_k \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{\beta}_k \right] \\
&= \arg\max_{\boldsymbol{W}} \sum_{k=1}^{K} \pi_k \mathrm{trace} \Big(\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \Big) \\
&= \arg\max_{\boldsymbol{W}} \mathrm{trace} \left(\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \left(\sum_{k=1}^{K} \pi_k \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \right) \right)
\end{aligned}$$

Note that $\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2}$ is a linear map that projects a vector onto the range space of $\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{W}$. We know that trace of a matrix is equivalent to the sum of its eigenvalues. Therefore, in order to maximize the trace, the projection matrix should select the maximum R eigen directions. As a result, if the range space of $\boldsymbol{W}$ is equivalent to the range space of $\boldsymbol{\Sigma}_{\mathcal{X}}^{-1/2} \boldsymbol{E}$, then trace is maximized. ■

Lemma A.3 (Restated Lemma 3.3) *Let $\boldsymbol{W}^*$ be an optimal representation matrix obtained in Theorem A.2. Let $(\lambda_i, \boldsymbol{e}_i)$ be the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$ defined in Theorem A.2 such that $\lambda_i \geq \lambda_{i+1}$ and $\|\boldsymbol{e}_i\|_{\ell_2} = 1$ for $i \in [d]$. Let $\lambda_{i,k} = \boldsymbol{e}_i^\top \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{e}_i$. Then the population risk for the k^{th} task is the following:*

$$\begin{aligned}
\mathcal{R}_k(\boldsymbol{W}^*) &:= \frac{1}{N} \mathbb{E} \left[\|\boldsymbol{y}_t - \boldsymbol{X}_t \boldsymbol{W}^* \boldsymbol{h}_t\|_{\ell_2}^2 \big| t \in \mathcal{T}_k \right] \\
&= \sigma_k^2 + \sum_{i=R+1}^{d} \lambda_{i,k}
\end{aligned} \tag{28}$$

Proof. As the tasks inside a task family are independent and identically distributed, we have

$$\mathcal{R}_k(\boldsymbol{W}^*) = \mathbb{E} \left[\|\boldsymbol{y}_k - \boldsymbol{X}_k \boldsymbol{W}^* \boldsymbol{h}_k^{\boldsymbol{W}}\|_{\ell_2} \right] \tag{29}$$

where $(\boldsymbol{X}_k, \boldsymbol{y}_k)$ is an input/label of k^{th} task family. Plugging in the optimal task-specific heads in (23), we obtain the following similar to the proof of Theorem A.2:

$$\mathcal{R}_k(\boldsymbol{W}^*) = \mathbb{E}_{\boldsymbol{\beta}_k \sim \mathcal{D}_{\mathcal{T},k}} \left[\boldsymbol{\beta}_k^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{\beta}_k - \boldsymbol{\beta}_k \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W}^* (\boldsymbol{W}^{*\top} \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W}^*)^{-1} \boldsymbol{W}^{*\top} \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k^2 \right] \tag{30}$$

$$= \mathrm{trace}(\boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_{\mathcal{X}}) - \mathrm{trace} \left(\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{W}^* (\boldsymbol{W}^{*\top} \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W})^{-1} \boldsymbol{W}^{*\top} \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \right) + \sigma_k^2 \tag{31}$$

$$\overset{(a)}{=} \sum_{i=1}^{d} \lambda_{i,k} - \sum_{i=1}^{R} \lambda_{i,k} + \sigma_k^2 \tag{32}$$

$$= \sum_{i=R+1}^{d} \lambda_{i,k} + \sigma_k^2 \tag{33}$$

where (a) follows from the facts that the trace is the summation of eigenvalues and the range space of $\boldsymbol{W}^*$ is $\boldsymbol{\Sigma}_{\mathcal{X}}^{-1/2} \boldsymbol{E}$. Note that $\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{W}^* (\boldsymbol{W}^{*\top} \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{W})^{-1} \boldsymbol{W}^{*\top}$ is the projection matrix onto the column space of $\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2} \boldsymbol{W}^*$. ■

Theorem A.4 (Restated Theorem 3.5) *Let $(\alpha_k(R))_{k=1}^K$ be emergence rates of K task families such that $\alpha_k(R) \geq 0$ and $\sum_{k=1}^K \pi_k \alpha_k(R)$ is non-increasing as a function of R. Then, there exists a sequence of task covariances $(\Sigma_k)_{k=1}^K$, noise level σ^2, and the task family probabilities $(\pi_k)_{k=1}^K$ such that the problems follows the $(\alpha_{ki})_{i=1}^d$ emergence profile for every $k \in [K]$ under the optimal representation matrix $\boldsymbol{W}$.*

Proof. In this proof, we provide covariance matrices such that $(\alpha_k(R))_{k=1}^K$ are achieved. Let $\boldsymbol{\Sigma}_{\mathcal{X}} = \boldsymbol{I}_d$ and the ground-truth covariance matrices, $(\boldsymbol{\Sigma}_k)_{k=1}^K$, are diagonal. The i^{th} diagonal element of $\boldsymbol{\Sigma}_k$ is $\alpha_k(i)$ for $i \in [d-1]$. For the d^{th} diagonal element of $\boldsymbol{\Sigma}_k$ is 0. The fact that $\alpha_k(R)$ is non-negative ensures that we can choose these values as the diagonal elements of a covariance matrix. The fact that $\sum_{k=1}^K \pi_k \alpha_k(R)$ is non-increasing ensures that the optimal representation matrix chooses the first R diagonal elements of covariance matrices $(\boldsymbol{\Sigma}_k)_{k=1}^K$. This concludes the proof. ■

B Proof of Statements in Subsection 4.1

Assumption B.1 (Restated Assumption 4.1) *$\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_{\mathcal{X}}$, and $\boldsymbol{W}\boldsymbol{W}^\top$ are jointly diagonalizable matrices for $k \in [K]$.*

Proposition B.2 (Restated Proposition 4.2) *Suppose Assumption B.1 holds. For every $\boldsymbol{U} \in \mathbb{R}^{d \times R}$ whose columns are the eigenvectors of the covariance matrices, there exists a unitary matrix $\boldsymbol{V} \in \mathbb{R}^{R \times R}$ such that the matrices $\bar{\boldsymbol{\Sigma}}_{\mathcal{X}} = \boldsymbol{U}^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{U}$, $\bar{\boldsymbol{\Sigma}}_k = \boldsymbol{U}^\top \boldsymbol{\Sigma}_k \boldsymbol{U}$, $\bar{\boldsymbol{W}} = \boldsymbol{U}^\top \boldsymbol{W} \boldsymbol{V}$, and the noise terms $\bar{\sigma}_k^2 = \sigma_k^2 + \operatorname{trace}(\boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{\Sigma}_k) - \operatorname{trace}(\bar{\boldsymbol{\Sigma}}_{\mathcal{X}} \bar{\boldsymbol{\Sigma}}_k)$ for all $k \in [K]$ satisfy that $\bar{\boldsymbol{W}}$ is diagonal and the following equality holds:*

$$\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \mathcal{R}_{FS}(\bar{\boldsymbol{\Sigma}}_{\mathcal{T}}, \bar{\boldsymbol{\Sigma}}_{\mathcal{X}}, \bar{\boldsymbol{W}}, \bar{\sigma}_{\mathcal{T}}^2) \tag{34}$$

Proof. Let $\boldsymbol{U}^\perp \in \mathbb{R}^{d \times (d-R)}$ be the matrix whose column space is the orthogonal complement to the column space of $\boldsymbol{U}$, whose columns are unit length, and whose columns are orthonormal to each other. Define $\boldsymbol{x}_{\boldsymbol{W}} = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}$ and $\boldsymbol{x}_{\boldsymbol{W}^\perp} = \boldsymbol{U}^\perp (\boldsymbol{U}^\perp)^\top \boldsymbol{x}$. Then, we rewrite the (3) for a single training sample as follows:

$$y = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon = \boldsymbol{x}^\top \left(\boldsymbol{U}\boldsymbol{U}^\top + \boldsymbol{U}^\perp (\boldsymbol{U}^\perp)^\top \right) \boldsymbol{\beta} + \varepsilon = \boldsymbol{x}_{\boldsymbol{W}}^\top \boldsymbol{\beta}_{\boldsymbol{W}} + \boldsymbol{x}_{\boldsymbol{W}^\perp}^\top \boldsymbol{\beta}_{\boldsymbol{W}^\perp} + \varepsilon \tag{35}$$

We treat $\boldsymbol{x}_{\boldsymbol{W}}$ and $\boldsymbol{\beta}_{\boldsymbol{W}}$ as feature and ground-truth and $\boldsymbol{x}_{\boldsymbol{W}^\perp}^\top \boldsymbol{\beta}_{\boldsymbol{W}^\perp} + \varepsilon$ as noise term. Note that $\boldsymbol{x}_{\boldsymbol{W}}$ and $\boldsymbol{\beta}_{\boldsymbol{W}}$ are inside the column space of $\boldsymbol{W}$. Then, we change the basis of $\boldsymbol{x}_{\boldsymbol{W}}$ and $\boldsymbol{\beta}_{\boldsymbol{W}}$ to the columns of $\boldsymbol{U}$. As the columns of $\boldsymbol{U}$ are the unit length and orthogonal to each other, we can change the basis by multiplication: $\bar{\boldsymbol{x}}_{\boldsymbol{W}} = \boldsymbol{U}^\top \boldsymbol{x}_{\boldsymbol{W}}$, $\bar{\boldsymbol{\beta}}_{\boldsymbol{W}} = \boldsymbol{U}^\top \boldsymbol{\beta}_{\boldsymbol{W}}$ and $\tilde{\boldsymbol{W}} = \boldsymbol{U}^\top \boldsymbol{W}$. Then, we have the following:

$$\bar{\boldsymbol{x}}_{\boldsymbol{W}} = \boldsymbol{U}^\top \boldsymbol{x}_{\boldsymbol{W}} = \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x} = \boldsymbol{U}^\top \boldsymbol{x}$$
$$\bar{\boldsymbol{\beta}}_{\boldsymbol{W}} = \boldsymbol{U}^\top \boldsymbol{\beta}_{\boldsymbol{W}} = \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{\beta} = \boldsymbol{U}^\top \boldsymbol{\beta}$$

Note that this can be considered as the projections onto the $R-$ dimensional subspace spanned by $\boldsymbol{U}$. As a result, the covariance matrices and the noise terms will be the following:

$$\bar{\boldsymbol{\Sigma}}_{\mathcal{X}} = \boldsymbol{U}^\top \boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{U} \tag{36}$$
$$\bar{\boldsymbol{\Sigma}}_k = \boldsymbol{U}^\top \boldsymbol{\Sigma}_k \boldsymbol{U} \quad \forall k \in [K] \tag{37}$$
$$\bar{\sigma}_k = \sigma_k + \operatorname{trace}(\boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{\Sigma}_k) - \operatorname{trace}(\bar{\boldsymbol{\Sigma}}_{\mathcal{X}} \bar{\boldsymbol{\Sigma}}_k) \quad \forall k \in [K] \tag{38}$$

The only remaining thing is that there exists a unitary matrix $\boldsymbol{V} \in \mathbb{R}^{R \times R}$ such that $\bar{\boldsymbol{W}} = \tilde{\boldsymbol{W}} \boldsymbol{V}$ is diagonal. Note that for any $\boldsymbol{W}$ and unitary matrix $\boldsymbol{V} \in \mathbb{R}^{R \times R}$, if we consider $\boldsymbol{W}\boldsymbol{V}$ as the representation matrix instead of $\boldsymbol{W}$, the risk does not change because the optimal header becomes $\boldsymbol{V}^\top \boldsymbol{h}_k^{\boldsymbol{W}}$ instead of $\boldsymbol{h}_k^{\boldsymbol{W}}$ using (5). Then, the population risk does not change. Let $\bar{\boldsymbol{U}} \in \mathbb{R}^{d \times d}$ be the concatenation of $\boldsymbol{U}$ and $\boldsymbol{u}^\perp$. Then, the singular decomposition of $\boldsymbol{W}$ is $\bar{\boldsymbol{U}} \bar{\boldsymbol{\Sigma}} \bar{\boldsymbol{V}}^\top$. Then, if we select $\boldsymbol{V}$ as $\bar{\boldsymbol{V}}$, we obtain that $\boldsymbol{U}^\top \boldsymbol{W} \boldsymbol{V} = \bar{\boldsymbol{W}}$, which is diagonal and gives the same risk. This completes the proof. ■

Definition B.3 *Let $\boldsymbol{d}_{\mathcal{X}}, (\boldsymbol{d}_k)_{k=1}^K \in \mathbb{R}^R$ be the vectors such that $diag(\boldsymbol{d}_k) := \bar{\boldsymbol{W}}^{-1} \bar{\boldsymbol{\Sigma}}_k (\bar{\boldsymbol{W}}^\top)^{-1}$ and $diag(\boldsymbol{d}_{\mathcal{X}}) := \bar{\boldsymbol{W}}^\top \bar{\boldsymbol{\Sigma}}_{\mathcal{X}} \bar{\boldsymbol{W}}$*

Assumption B.4 (Restated Assumption 4.4) *Let $d_{\mathcal{X},i}$ and $d_{k,i}$ be the i^{th} element of $\boldsymbol{d}_{\mathcal{X}}$ and $\boldsymbol{d}_k$, respectively. We assume that the empirical distribution of $(d_{\mathcal{X},i}, (d_{k,i})_{k=1}^K)$ jointly converges to $(\Lambda, (M_k)_{k=1}^K)$ where Λ and $(M_k)_{k=1}^K$ are non-negative random variables. Furthermore, there exists $c_l, c_u > 0$ such that $c_l < d_{k,i}, d_{\mathcal{X},i} < c_u$ for every $k \in [K]$ and $i \in [n]$.*

Definition B.5 (Asymptotic distribution characterization- overparameterized regime) *Chang et al. (2020) Recall the random variables (Λ, M_k) in Assumption B.4, and fix $\kappa_1 > 1$. Define parameter ξ as the unique positive solution to the following equation:*

$$\mathbb{E}_\Lambda\left[(1+(\xi\Lambda)^{-1})^{-1}\right] = \kappa_1^{-1} \tag{39}$$

Further define the positive parameters γ_k for every $k \in [K]$ as follows:

$$\gamma_k := \left(\sigma_k^2 + \mathbb{E}_{(\Lambda,M_k)}\left[\frac{M_k\Lambda}{(1+\xi\Lambda)^2}\right]\right) \Big/ \left(1 - \mathbb{E}_\Lambda\left[\frac{\kappa_1}{(1+(\xi\Lambda)^{-1})^2}\right]\right) \tag{40}$$

With these and $H \sim \boldsymbol{N}(0,1)$, define the random variables for every $k \in [K]$:

$$X_{\kappa_1,\sigma_k^2}(\Lambda, M_k, H) = \left(1 - \frac{1}{1+\xi\Lambda}\right)\sqrt{M_k} + \sqrt{\kappa_1}\frac{\sqrt{\gamma_k}\Lambda^{-1/2}}{1+(\xi\Lambda)^{-1}}H \tag{41}$$

Theorem B.6 *[Asymptotic distribution characterization- Linear Gaussian Problem Chang et al. (2020) Fix $\kappa_1 > 1$ and suppose Assumption B.4 holds. For every task $k \in [K]$, let*

$$\frac{1}{d}\sum_{i=1}^d \delta_{\sqrt{R}\hat{\beta}_{k,i}, \sqrt{R}\beta^*_{k,i}, \boldsymbol{\Sigma}_{\mathcal{X}\,i,i}} \tag{42}$$

be the joint empirical distribution of $(\sqrt{R}\hat{\boldsymbol{\beta}}_k, \sqrt{R}\boldsymbol{\beta}^*, \boldsymbol{\Sigma}_{\mathcal{X}})$. Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a function of PL(2) where PL(2) is defined in Chang et al. (2020). We have that for every $k \in [K]$

$$\frac{1}{R} f\left(\sqrt{R}\hat{\beta}_{k,i}, \sqrt{R}\beta^*_{k,i}, \boldsymbol{\Sigma}_{\mathcal{X}\,i,i}\right) \xrightarrow{P} \mathbb{E}\left[f\left(X_{\kappa_1,\sigma_k^2}, \sqrt{RM_k}, \Lambda\right)\right] \tag{43}$$

Theorem B.7 (Restated Theorem 4.5) *Recall the random variables in Assumption B.4, and fix $(\kappa_k)_{k=1}^K > 1$ and $\kappa_0 = 1$. Define parameter $(\xi_k)_{k=1}^K$ as the unique positive solution to the following equation:*

$$\mathbb{E}_\Lambda\left[(1+(\xi_k\Lambda)^{-1})^{-1}\right] = \kappa_k^{-1} \qquad k \in [K] \tag{44}$$

Further define the positive parameters $(B_k)_{k=1}^K$ and $(\Omega_k)_{k=1}^K$ as follows:

$$B_k = \mathbb{E}_{(\Lambda,M_k)}\left[\frac{M_k\Lambda}{(1+\xi_k\Lambda)^2}\right] \qquad \Omega_k = \mathbb{E}_\Lambda\left[\frac{\kappa_k}{(1+(\xi_k\Lambda)^{-1})^2}\right] \tag{45}$$

Then, we have the following risk characterization:

$$\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \frac{1}{N_{total}}\sum_{k=1}^K N_k \frac{B_k + \bar{\sigma}_k^2}{1-\Omega_k} \tag{46}$$

Proof. In (12), we state that

$$\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \mathcal{R}_{FS}\left(\bar{\boldsymbol{W}}^{-1}\bar{\boldsymbol{\Sigma}}_{\mathcal{T}}(\bar{\boldsymbol{W}}^\top)^{-1}, \bar{\boldsymbol{W}}^\top\bar{\boldsymbol{\Sigma}}_{\mathcal{X}}\bar{\boldsymbol{W}}, \boldsymbol{I}_R, \bar{\sigma}_{\mathcal{T}}^2\right)$$

Now, define a new few-shot multitask learning problem with the feature covariance matrix $\tilde{\boldsymbol{\Sigma}}_{\mathcal{X}} := \bar{\boldsymbol{W}}^\top\bar{\boldsymbol{\Sigma}}_{\mathcal{X}}\bar{\boldsymbol{W}}$, ground truth covariance matrices $\bar{\boldsymbol{W}}^{-1}\bar{\boldsymbol{\Sigma}}_{\mathcal{T}}(\bar{\boldsymbol{W}}^\top)^{-1}$, such that the training samples are $(\tilde{\boldsymbol{X}}_k, \tilde{\boldsymbol{y}}_k)$ and the heads are $\tilde{h}_k^{\boldsymbol{W}}$. Then, we can state the risk characterization for the diagonal $\boldsymbol{W}$ matrix as follows:

$$\begin{aligned}\mathcal{R}_{FS}\left(\bar{\boldsymbol{W}}^{-1}\bar{\boldsymbol{\Sigma}}_{\mathcal{T}}(\bar{\boldsymbol{W}}^\top)^{-1}, \bar{\boldsymbol{W}}^\top\bar{\boldsymbol{\Sigma}}_{\mathcal{X}}\bar{\boldsymbol{W}}, \boldsymbol{I}_R, \bar{\sigma}_{\mathcal{T}}^2\right) &= \mathbb{E}\left[\frac{1}{N_{total}}\sum_{k=1}^K \|\tilde{\boldsymbol{y}}_k - \tilde{\boldsymbol{X}}_k\tilde{\boldsymbol{h}}_k^{\boldsymbol{I}_R}\|_{\ell_2}^2\right] \\ &= \frac{1}{N_{total}}\sum_{k=1}^K N_k\big(\mathbb{E}\left[\|\tilde{\boldsymbol{h}}_k^{\boldsymbol{I}_R} - \tilde{\boldsymbol{\beta}}_k)\|_{\tilde{\boldsymbol{\Sigma}}_{\mathcal{X}}}^2\right] + \bar{\sigma}_k^2\big).\end{aligned}$$

Now, we are able to utilize Theorem B.6 in a way that we can consider $\boldsymbol{\beta}_k^*$ as $\bar{\boldsymbol{\beta}}_k$ and $\hat{\boldsymbol{\beta}}_k$ as $\bar{\boldsymbol{h}}_k^{\boldsymbol{I}_R}$ for each task family. Then, we obtain the following by utilizing Theorem B.6:

$$\begin{aligned}
\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{I}_R, \sigma_{\mathcal{T}}^2) &= \frac{1}{N_{total}} \sum_{k=1}^{K} N_k \big(\mathbb{E}\left[\| \tilde{\boldsymbol{h}}_k^{\boldsymbol{W}} - \tilde{\boldsymbol{\beta}}_k) \|_{\boldsymbol{\Sigma}_{\mathcal{X}}}^2 \right] + \bar{\sigma}_k^2 \big) \\
&= \frac{1}{N_{total}} \sum_{k=1}^{K} N_k \Big(\mathbb{E}\left[\| X_{\kappa_1, \sigma_k^2} - \sqrt{M_k} \|_{\boldsymbol{\Sigma}_{\mathcal{X}}}^2 \right] + \bar{\sigma}_k^2 \Big) \\
&= \frac{1}{N_{total}} \sum_{k=1}^{K} N_k \left(\mathbb{E}\left[\frac{M_k \Lambda}{(1+\xi\Lambda)^2} + \frac{\kappa_k \gamma_k}{(1+(\xi\Lambda)^{-1})^2} \right] + \bar{\sigma}_k^2 \right)
\end{aligned} \tag{47}$$

Utilizing the fact that

$$\gamma_k = \left(\bar{\sigma}_k^2 + \mathbb{E}_{(\Lambda, M_k)} \left[\frac{M_k \Lambda}{(1+\xi\Lambda)^2} \right] \right) \Big/ \left(1 - \mathbb{E}_{\Lambda} \left[\frac{\kappa_1}{(1+(\xi\Lambda)^{-1})^2} \right] \right) = \frac{B_k + \bar{\sigma}_k^2}{1 - \Omega_k} \tag{48}$$

and plugging this fact in (47) concludes the proof. ■

Definition B.8 (Restated Definition 4.6) *Consider a few-shot multitask problem with the feature covariance $\boldsymbol{\Sigma}_{\mathcal{X}}$, ground-truth covariances $(\boldsymbol{\Sigma}_k)_{k=1}^K$, the noise levels $(\sigma_k^2)_{k=1}^K$, and the number of sample N for each task. Define $\boldsymbol{\Sigma}_{avg} = \frac{1}{K}\sum_{k=1}^K \boldsymbol{\Sigma}_k$ and $\sigma_{avg}^2 = \frac{1}{K}\sum_{k=1}^K \sigma_k^2$. Then, the reduced few-shot single-task problem of the multitask problem is defined by the feature covariance $\boldsymbol{\Sigma}_{\mathcal{X}}$, the ground-truth covariance $\boldsymbol{\Sigma}_{avg}$, the noise level σ_{avg}^2, and the number of samples N.*

Theorem B.9 (Restated Theorem 4.7) *Suppose the distributions $(\mathcal{D}_k)_{k=1}^K$ are Gaussian and Assumptions B.1 and B.4 hold. Let $\boldsymbol{\Sigma}_{avg} = \sum_{k=1}^K \boldsymbol{\Sigma}_k$ and $\sigma_{avg}^2 = \sum_{k=1}^K \sigma_k^2$. Define a single task few shot learning problem for $\boldsymbol{\beta}_{avg}$ that is zero-mean Gaussian random variable with covariance $\boldsymbol{\Sigma}_{avg}$. The number of samples is N and the noise level is $\sigma_{avg}^2 := \sum_{k=1}^K \sigma_k^2$ for the single task few-shot learning problem. For any $\boldsymbol{W} \in \mathbb{R}^{d\times R}$ whose column space is $R-$dimensional, we have the following:*

$$\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \mathcal{R}_{FS}(\boldsymbol{\Sigma}_{avg}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{avg}^2) \tag{49}$$

Proof. In this proof, we apply the transformation in (12) and the result of Theorem B.7. Let $\boldsymbol{U} \in \mathbb{R}^{d\times R}$ be a matrix whose i^{th} column is the unit length and in the same direction as i^{th} column of $\boldsymbol{W}$. Let $(\Lambda, (M_k)_{k=1}^K)$ and $(\bar{\Lambda}, \bar{M})$ be the random variables defined in Assumption B.4 for multitask and single-task problems, respectively. As the feature covariance are the same for two problems, we directly say that $\Lambda = \bar{\Lambda}$ By (37) and Definition B.8, we have the following:

$$\bar{\boldsymbol{\Sigma}}_{avg} = \boldsymbol{U}^\top \left(\sum_{k=1}^K \frac{1}{K} \boldsymbol{\Sigma}_k \right) \boldsymbol{U} = \frac{1}{K} \sum_{k=1}^K \bar{\boldsymbol{\Sigma}}_k \implies \bar{M} = \frac{1}{K} \sum_{k=1}^K M_k \tag{50}$$

Using (38) and (50), we obtain the following:

$$\begin{aligned}
\bar{\sigma}_{avg} &= \sigma_{avg} + \text{trace}(\boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{\Sigma}_{avg}) - \text{trace}(\bar{\boldsymbol{\Sigma}}_{\mathcal{X}} \bar{\boldsymbol{\Sigma}}_{avg}) \\
&= \sum_{k=1}^K \frac{1}{K} \sigma_k + \text{trace}\left(\boldsymbol{\Sigma}_{\mathcal{X}} \big(\frac{1}{K} \sum_{k=1}^K \boldsymbol{\Sigma}_k \big) \right) - \text{trace}\left(\bar{\boldsymbol{\Sigma}}_{\mathcal{X}} \big(\frac{1}{K} \sum_{k=1}^K \bar{\boldsymbol{\Sigma}}_k \big) \right) \\
&= \frac{1}{K} \left(\sum_{k=1}^K \sigma_k + \text{trace}(\boldsymbol{\Sigma}_{\mathcal{X}} \boldsymbol{\Sigma}_k) - \text{trace}(\bar{\boldsymbol{\Sigma}}_{\mathcal{X}} \bar{\boldsymbol{\Sigma}}_k) \right) \\
&= \frac{1}{K} \sum_{k=1}^K \bar{\sigma}_k
\end{aligned} \tag{51}$$

Let $(\xi)_{k=1}^K, (B_k)_{k=1}^K, (\Omega_k)_{k=1}^K$, and $(\kappa_k)_{k=0}^K$ be defined for few-shot multitask learning problem. Let $\bar{\xi}, \bar{B}, \bar{\Omega}$, and $\bar{\kappa}$ be the corresponding parameters for the reduced few-shot single-task learning problem. Since the numbers of

samples from all the task families in the multitask learning problem and the number of samples from the reduced few-shot single-task learning problem are equivalent, we have $\bar{\kappa} = \kappa_k$ for all $k \in [K]$. Using (45) and (44), this implies that

$$\bar{\xi} = \xi_k \qquad \bar{\Omega} = \Omega_k \qquad \forall k \in [K]. \tag{52}$$

On the other hand, using (45), (50), and (52), we have

$$\bar{B} = \frac{1}{K} \sum_{k=1}^{K} B_k. \tag{53}$$

Using Theorem B.7 and (51), (52), (53); we obtain the following:

$$\begin{aligned} \sum_{k=1}^{K} \mathcal{R}_{FS}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_k^2) &= \sum_{k=1}^{K} \frac{B_k + \bar{\sigma}_k^2}{1 - \Omega_k} \\ &= K \frac{\bar{B} + \bar{\sigma}_{avg}}{1 - \bar{\Omega}} \\ &= K \widehat{\mathcal{R}}_{FS}(\boldsymbol{\Sigma}_{avg}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{avg}^2) \end{aligned} \tag{54}$$

Using (9), we know that

$$\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{R}_{FS}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) \tag{55}$$

which completes the proof with (54). ■

Lemma B.10 (Restated Lemma 4.8) *The optimal population risk with respect to the optimal linear representation matrix $\boldsymbol{W}_{FS}^*$ is non-increasing as R increases.*

Proof. Let $\boldsymbol{W}_{FS,R}^*$ be an optimal representation matrix when the optimal representation matrix is in $\mathbb{R}^{d \times R}$. Let $\boldsymbol{W}' \in \mathbb{R}^{d \times R+1}$ such that the first R columns are equal to $\boldsymbol{W}_{FS,R}^*$ and the last column is 0. Then, $\boldsymbol{W}'$ achieves the same risk as $\boldsymbol{W}_{FS,R}^*$. This implies that

$$\min_{\boldsymbol{W} \in \mathbb{R}^{d \times R}} \mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}', \sigma_{\mathcal{T}}^2) \geq \min_{\boldsymbol{W} \in \mathbb{R}^{d \times R+1}} \mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) \tag{56}$$

which completes the proof. ■

C Proof of Statements in Subsection 4.2

Proposition C.1 (Restated Proposition 4.9) *Suppose Assumptions B.1 and B.4 hold and recall the definition of $\boldsymbol{E}$ in Theorem A.2. The range space of $\boldsymbol{W}_{FS}^*$ is equal to the range space of $\boldsymbol{\Sigma}_{\mathcal{X}}^{-1/2}\boldsymbol{E}$.*

Proof. By Assumption B.1, we know that $\boldsymbol{\Sigma}_{\mathcal{X}}^{-1/2}\boldsymbol{E}$ and $\boldsymbol{E}$ have the same range space. We state in the theorem as $\boldsymbol{\Sigma}_{\mathcal{X}}^{1/2}$ in order to make a connection between the population risk analysis and few-shot learning settings easily. In this proof, we prove that the optimal range space is $\boldsymbol{E}$.

As the numbers of samples from each task family are equivalent, there exists Ω such that $\Omega = \Omega_k$ for $k \in [K]$. Using Theorem B.7, we have the following:

$$\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \frac{\frac{1}{K}\sum_{k=1}^{K}(B_k + \bar{\sigma}_{avg}^2)}{1 - \Omega} \tag{57}$$

Note that for any range space of $\boldsymbol{W}$, we have the same $\sum_{k=1}^{K} \mathbb{E}[M_k \Lambda] + \bar{\sigma}_{avg}^2$, because this is the total energy. The range space of $\boldsymbol{W}$ divides the total energy into two parts, where one part is reserved for the noise and the other

part is reserved for the signal. In the risk formula, the signal part is multiplied by a spectrum term $1/(1+\xi\Lambda)$ that is less than 1 (see the definition of B_k in (45)). We are going to prove the following statement: If the range space of $\boldsymbol{W}$ is the same as the range space of $\boldsymbol{E}$, then the signal part is maximized when we are allowed to use R dimensions instead of d dimensions.

Let $\boldsymbol{C}(\cdot)$ represent the range space of a matrix. Assume that there exist $\boldsymbol{W}' \in \mathbb{R}^{d\times R}$ such that $\boldsymbol{C}(\boldsymbol{W}')$ is not equal to $\boldsymbol{C}(\boldsymbol{E})$ and $\boldsymbol{W}'$ provides a smaller risk compared to any $\boldsymbol{W}$ for which $\boldsymbol{C}(\boldsymbol{W}) = \boldsymbol{C}(\boldsymbol{E})$. We are going to construct $\boldsymbol{W}$ based on $\boldsymbol{W}'$ such that $\boldsymbol{C}(\boldsymbol{W}) = \boldsymbol{C}(\boldsymbol{E})$ and $\boldsymbol{W}$ provides a risk that is smaller than or equal to the risk achieved by $\boldsymbol{W}'$. Let $\boldsymbol{W}' = \boldsymbol{U}'\boldsymbol{\Sigma}_{\boldsymbol{W}'}\boldsymbol{V}'$ be the singular value decomposition of $\boldsymbol{W}'$. By Assumption B.1, we know that the columns of $\boldsymbol{U}'$ are eigenvectors of the covariance matrices. Now, we construct $\boldsymbol{U}$. The i^{th} column of $\boldsymbol{U}$ is equal to the i^{th} column of $\boldsymbol{U}'$ if this column is a column of $\boldsymbol{E}$. If not, the i^{th} column of $\boldsymbol{U}$ is equal to an arbitrary column from $\boldsymbol{E}'$ that is not a column of $\boldsymbol{U}'$. Let $\bar{\boldsymbol{\Sigma}}_{\mathcal{X}} = \boldsymbol{U}^\top\boldsymbol{\Sigma}_{\mathcal{X}}\boldsymbol{U}$ and $\bar{\boldsymbol{\Sigma}}'_{\mathcal{X}} = \boldsymbol{U}'^\top\boldsymbol{\Sigma}_{\mathcal{X}}\boldsymbol{U}'$ similar to Proposition B.2. Similarly, we define $\bar{\boldsymbol{W}}$ and $\bar{\boldsymbol{W}}'$ as in Proposition B.2. Then, we select $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ such that $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}_{\boldsymbol{W}}$ and $\bar{\boldsymbol{W}}^\top\bar{\boldsymbol{\Sigma}}_{\mathcal{X}}\bar{\boldsymbol{W}} = \bar{\boldsymbol{W}}'^\top\bar{\boldsymbol{\Sigma}}'_{\mathcal{X}}\bar{\boldsymbol{W}}'$. This selection ensures that $\boldsymbol{d}_{\mathcal{X}} = \boldsymbol{d}'_{\mathcal{X}}$, which implies that the random variable Λ is the same for both of the $\boldsymbol{W}$ and $\boldsymbol{W}'$.

Now, we will prove that the risk induced by $\boldsymbol{W}$ is less than or equal to the risk induced by $\boldsymbol{W}'$. As we have the same Λ for both of the representation matrices, Ω of both risks are the same. Additionally, the construction of $\boldsymbol{W}$ guarantees that any realization of $\sum_{k=1}^K M_k\Lambda$ is greater than or equal to the realization of $\sum_{k=1}^K M'_k\Lambda$. (M_k and M'_k are the random variables defined in Definition B.3 for $\boldsymbol{W}$ and $\boldsymbol{W}'$, respectively.) This means that $\sum_{k=1}^K \mathbb{E}[M_k\Lambda] \geq \sum_{k=1}^K \mathbb{E}[M'_k\Lambda]$. Note that

$$B_k = \mathbb{E}_{(\Lambda,M_k)}\left[\frac{M_k\Lambda}{(1+\xi\Lambda)^2}\right]$$

In addition to that, the summation of the noise and signals are the same for every selection of the representation matrix, i.e, $\sum_{k=1}^K \mathbb{E}[M_K\Lambda] + \bar{\sigma}^2_{avg}$ is the same for every selection of the representation matrix. Since the signal part is multiplied by $1/(1+\xi\Lambda) < 1$, then the risk will be smaller when the signal part is bigger. As we prove that $\sum_{k=1}^K \mathbb{E}[M_k\Lambda] \geq \sum_{k=1}^K \mathbb{E}[M'_k\Lambda]$, the representation matrix $\boldsymbol{W}$ achieves a risk that is smaller than or equal to the risk achieved by $\boldsymbol{W}'$, which is a contradiction. This completes the proof. ■

Theorem C.2 (Restated Proposition 4.10) *Suppose Assumptions B.1 and B.4 hold. Fix $\kappa = R/N > 1$, let $M = \frac{1}{K}\sum_{k=1}^K M_k$, and define the unique parameter $\xi \in \mathbb{R}$, the random variables $\zeta \in [0,1]$ and B as the following:*

$$\mathbb{E}_\Lambda\left[(1+(\xi\Lambda)^{-1})^{-1}\right] = \kappa^{-1} \qquad \zeta = (1+\xi\Lambda)^{-1}$$

Further define the function $f : \mathbb{R} \to \mathbb{R}$ when the range space of the representation matrix $\boldsymbol{W}$ as stated in Proposition C.1 as follows:

$$f(\zeta) = \frac{\mathbb{E}_{\zeta,(M\Lambda)}[M\Lambda\zeta^2] + \bar{\sigma}^2_{avg}}{1-\kappa\,\mathbb{E}_\zeta[(1-\zeta)^2]} \tag{58}$$

Let $\mathcal{C}$ be the set of random variables ζ obeying $\Pr(\zeta \in (0,1]) = 1$ and $\mathbb{E}[\zeta] = 1-\kappa^{-1}$. Then, the following hold:

$$\min_{\boldsymbol{W}\in\mathbb{R}^{d\times R}} \mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma^2_{\mathcal{T}}) = \min_{\zeta\in\mathcal{C}} f(\zeta) \tag{59}$$

Furthermore, $f(\zeta)$ is strongly convex on $\mathcal{C}$.

Proof. First part (Risk Equivalence): We apply Theorem B.7 to prove this theorem. As the numbers of samples from each task family are equivalent, there exists Ω such that $\Omega = \Omega_k$ for $k \in [K]$. Using Theorem B.7, we have the following:

$$\mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma^2_{\mathcal{T}}) = \frac{\frac{1}{K}\sum_{k=1}^K (B_k + \bar{\sigma}^2_k)}{1-\Omega}$$

Using the definition of ζ, M, Ω, and $\bar{\sigma}^2_{avg}$; we obtain the advertised result.

Second part (Strong Convexity) : By Assumption B.4, we know that the random variables $\boldsymbol{M}_k$ and Λ have upper and lower bounds that are strictly greater than zero. Therefore, the random variables $M = \frac{1}{K}\sum_{k=1}^K M_k$

and ζ have upper and lower bounds that are strictly greater than zero. Furthermore, the multiplication of random variables $M\Lambda$ is independent of ζ even though ζ is a function of Λ. This is because $M\Lambda$ is characterized by the columns of the representation matrix. As the columns are characterized in Proposition C.1. The term ζ is characterized by the magnitude of each column. Therefore, we can treat $\mathbb{E}[M_k\Lambda]$ as a constant when we take the derivative of the function f with respect to ζ. Then, we obtain the following:

$$\begin{aligned} f(\zeta) &= \frac{\mathbb{E}_{\zeta,(M\Lambda)}[M\Lambda\zeta^2] + \bar{\sigma}^2_{avg}}{1 - \kappa\,\mathbb{E}_\zeta[(1-\zeta)^2]} \\ &= \frac{\mathbb{E}_\zeta[\zeta^2\,\mathbb{E}[M\Lambda]] + \bar{\sigma}^2_{avg}}{1 - \kappa\,\mathbb{E}_\zeta[(1-\zeta)^2]} \end{aligned}$$

As the random variable ζ is upper and lower bounded, using the Dominated Convergence Theorem, the expectation and the derivative with respect to ζ can be interchanged. Then, we obtain the following:

$$\frac{df(\zeta)}{d\zeta} = -\frac{2\kappa(\mathbb{E}_\zeta[\zeta^2\,\mathbb{E}[M\Lambda]] + \bar{\sigma}^2_{avg})\,\mathbb{E}_\zeta[1-\zeta]}{(1 - \kappa\,\mathbb{E}_\zeta[(1-\zeta)^2])^2}$$

Then, the second derivative is the following:

$$\frac{d^2f(\zeta)}{d\zeta^2} = \frac{2\kappa(\mathbb{E}_\zeta[\zeta^2\,\mathbb{E}[M\Lambda]] + \bar{\sigma}^2_{avg})}{(1 - \kappa\,\mathbb{E}_\zeta[(1-\zeta)^2])^2} + \frac{8\kappa^2(\mathbb{E}_\zeta[\zeta^2\,\mathbb{E}[M\Lambda]] + \bar{\sigma}^2_{avg})\,\mathbb{E}_\zeta[(1-\zeta)^2]}{(1 - \kappa\,\mathbb{E}_\zeta[(1-\zeta)^2])^3} \tag{60}$$

The numerators in (60) are lower bounded by a term that is greater than zero since all of the terms are lower bounded by Assumption B.4. On the other hand,

$$\begin{aligned} 1 - \kappa\,\mathbb{E}[(1-\zeta)^2] &= 1 - \kappa + \kappa(2\,\mathbb{E}[\zeta] - \mathbb{E}[\zeta^2]) \\ &\overset{(a)}{\geq} 1 - \kappa + \kappa\,\mathbb{E}[\zeta] \\ &\overset{(b)}{=} 0 \end{aligned}$$

where (a) follows from the fact that ζ lower is bounded by 0 and upper bounded by 1 and (b) follows from the fact that $\mathbb{E}_\Lambda\left[(1+(\xi\Lambda)^{-1})^{-1}\right] = \kappa^{-1}$. Therefore, (60) is lower bounded by a term that is greater than 0. This completes the proof.

■

D Proof of Statements in Subsection 4.3

Observation D.1 *The linear representation matrix is invariant with respect to constant multiplication when the number of samples from each task family is equivalent. Namely, when $\alpha \neq 0$, we have the following:*

$$\widehat{\mathcal{R}}_{FS}(\boldsymbol{\Sigma}_\mathcal{T}, \boldsymbol{\Sigma}_\mathcal{X}, \boldsymbol{W}, \sigma^2_\mathcal{T}) = \widehat{\mathcal{R}}_{FS}(\boldsymbol{\Sigma}_\mathcal{T}, \boldsymbol{\Sigma}_\mathcal{X}, \alpha\boldsymbol{W}, \sigma^2_\mathcal{T}). \tag{61}$$

Furthermore, the individual risk of each task family is also invariant to constant multiplication of $\boldsymbol{W}$.

Proof. Define a new scaled few-shot multitask learning problem in a way that $\boldsymbol{\Sigma}_{\mathcal{T},scl} = \boldsymbol{\Sigma}_\mathcal{T}$, $\boldsymbol{\Sigma}_{\mathcal{X},scl} = \boldsymbol{\Sigma}_\mathcal{X}$, $\boldsymbol{W}_{scl} = \alpha\boldsymbol{W}$, $\sigma_{\mathcal{T},scl} = \sigma^2_\mathcal{T}$. Let $(\Lambda_{scl}, (M_{k,scl})_{k=1}^K)$ be the random variables defined in Assumption B.4 for the scaled few-shot multitask learning problem. When the number of samples from each task family is equivalent, we prove in Theorem B.9 that all we need is one variable for each task family: $\kappa, \xi, \Omega, \kappa_{scl}, \xi_{scl}$, and Ω_{scl}. Using (36), (37), and (38), we derive the following:

$$\begin{aligned} \bar{\boldsymbol{\Sigma}}_\mathcal{X} &= \bar{\boldsymbol{\Sigma}}_{\mathcal{X},scl} \qquad \alpha\bar{\boldsymbol{W}} = \bar{\boldsymbol{W}}_{scl} \\ \bar{\boldsymbol{\Sigma}}_k &= \bar{\boldsymbol{\Sigma}}_{k,scl} \qquad \bar{\sigma}_k = \bar{\sigma}_{k,scl} \quad \forall k \in [K]. \end{aligned}$$

Using these equalities, we obtain that

$$\alpha^2\Lambda = \Lambda_{scl} \qquad \frac{1}{\alpha^2}M_k = M_{k,scl} \quad \forall k \in [K]. \tag{62}$$

By definition of ξ in (44), we derive $\xi = \alpha^2 \xi_{scl}$. By plugging all the corresponding variables in the definition of $(B_k)_{k=1}^K$ and Ω in (45), we obtain that

$$\Omega = \Omega_{scl} \qquad B_k = B_{k,scl} \quad \forall k \in [K] \tag{63}$$

As a result, we prove that all the terms for both plain and scaled problems in the risk characterization of the few-shot multitask learning problem are equivalent. Additionally, the individual risk of each task family is also invariant to constant multiplication, which concludes the proof. ■

Proposition D.2 (Restated Proposition 4.11) *Recall the definition of $\boldsymbol{\Sigma}$ in Theorem 3.2. Let $(\lambda_i)_{i=1}^d$ be the eigenvalues of $\boldsymbol{\Sigma}$ where $\lambda_i \geq \lambda_{i+1}$. If $\lambda_R > \lambda_{R+1}$, then the individual risk of each task family is unique with any optimal representation matrix $\boldsymbol{W}$ that minimizes the average risk of all task families.*

Proof. The fact $\lambda_R > \lambda_{R+1}$ implies that the range space of $\boldsymbol{W}_{FS}^*$ is unique. By Theorem C.2, we know that there exists a unique ζ that minimizes the average population risk. By the definition of the random variable ζ, we have $\xi\Lambda = \frac{1}{\zeta} - 1$. Note that $\xi\Lambda$ is a random variable that satisfies

$$\mathbb{E}_\Lambda \left[(1 + (\xi\Lambda)^{-1})^{-1} \right] = \kappa^{-1} \tag{64}$$

This implies that the ratios between the magnitude of each pair of column vectors in $\boldsymbol{W}_{FS}^*$ are uniquely determined, but $\alpha \boldsymbol{W}_{FS}^*$ is another optimal representation if $\boldsymbol{W}_{FS}^*$ is an optimal representation and $\alpha \neq 0$. Using Observation D.1, we know that scaling does not affect the individual risk of each task family, therefore the individual risk of each task family is unique under any optimal representation matrix $\boldsymbol{W}_{FS}^*$. ■

Lemma D.3 (Restated Lemma 4.12) *Consider the scenario where $\boldsymbol{\Sigma}_{\mathcal{X}} = \boldsymbol{I}_{p+1}$, $\boldsymbol{\Sigma}_1 = diag(\boldsymbol{I}_p, \boldsymbol{0}_1), \boldsymbol{\Sigma}_2 = diag(\boldsymbol{0}_p, \alpha \boldsymbol{I}_1)$, $\sigma_1^2 = 0$. If $1/2 < \alpha < 1$, then the emergence rate of the first task family, $\alpha_1(p)$, is negative.*

Proof. Define the following function $f : \mathbb{R} \to \mathbb{R}$ for a given p and N as follows:

$$f(z) = \frac{pz^2 + \alpha(p + 1 - N - pz)^2}{1 - \frac{1}{N}(p(1-z)^2 + (pz + N - p)^2)} \tag{65}$$

Using Theorem C.2, we obtain the following:

$$\min_{\boldsymbol{W} \in \mathbb{R}^{d \times p+1}} \mathcal{R}_{FS}(\boldsymbol{\Sigma}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{X}}, \boldsymbol{W}, \sigma_{\mathcal{T}}^2) = \min_{z \in [\frac{p-N}{p}, \frac{p-N+1}{p}]} f(z) \tag{66}$$

For any α, p, N such that $p > N$ and $1/2 < \alpha < 1$, we have $f(\frac{p-N}{p}) > f(\frac{p-N+1}{p+\alpha}) > f(\frac{p-N+1}{p+\sqrt{\alpha/2}})$ and $\frac{p-N}{p} < \frac{p-N+1}{p+\alpha} < \frac{p-N+1}{p+\sqrt{\alpha/2}}$. Using the convexity of the function obtained from Theorem C.2, we state the optimal $z^* > \frac{p-N+1}{p+\alpha}$. On the other hand, as $\alpha < 1$, while optimizing $f(z)$, we have $p - N - pz^* \leq z^*$, which shows that $z^* < \frac{p-N+1}{p+1}$. Let $g_{p+1}(z)$ represent the risk of the first task family when $R = p + 1$. Then, these two findings imply that

$$g_{p+1}(z^*) = \frac{p{z^*}^2}{1 - \frac{1}{N}(p(1-z^*)^2 + (pz^* + N - p)^2)} \tag{67}$$

$$\geq \frac{p((p - N + 1)/(p + \alpha))^2}{1 - \frac{N}{p+1}} \tag{68}$$

Note that when $R = p$, the risk of the first task family is the following:

$$g_p(z^*) = \frac{p((p - N)/p)^2}{1 - N/p} \tag{69}$$

When $1/2 < \alpha < 1$, then the term in (68) is greater than the term in (69), which shows that the emergence rate at p is negative. ■