Understanding Benign Overfitting in Nested Meta Learning

Lisha Chen  
Rensselaer Polytechnic Institute  
Troy, New York, United States  
CHENL21@RPI.EDU

Songtao Lu  
IBM Thomas J. Watson Research Center  
Yorktown Heights, New York, United States  
SONGTAO@IBM.COM

Tianyi Chen  
Rensselaer Polytechnic Institute  
Troy, New York, United States  
CHENTIANYI19@GMAIL.COM

Abstract
Meta learning has demonstrated tremendous success in few-shot learning with limited supervised data. In those settings, the meta model is usually overparameterized. While the conventional statistical learning theory suggests that overparameterized models tend to overfit, empirical evidence reveals that overparameterized meta learning methods still work well – a phenomenon often called “benign overfitting.” To understand this phenomenon, we focus on the meta learning settings with a challenging nested structure that we term the nested meta learning, and analyze its generalization performance under an overparameterized meta learning model. While our analysis uses the relatively tractable linear models, our theory contributes to understanding the delicate interplay among data heterogeneity, model adaptation and benign overfitting in nested meta learning tasks. We corroborate our theoretical claims through numerical simulations.

1 Introduction
Meta learning, also referred to as “learning to learn”, usually learns a prior model from multiple tasks so that the learned model is able to quickly adapt to unseen tasks. Meta learning has been successfully applied to few-shot learning and multi-task learning [2, 11, 19, 24, 37]. While there are many exciting meta learning methods today, in this paper, we will study a representative meta learning setting where the goal is to learn a shared initial model that can quickly adapt to task-specific models. This adaptation may take an explicit form such as the output of one gradient descent step, which is referred to as the model agnostic meta learning (MAML) method [19]. Alternatively, the adaptation step may take an implicit form such as the solution of another optimization problem, which is referred to as the implicit MAML (iMAML) method [33]. Since both MAML and iMAML will solve a nested optimization problem, we term them the nested meta learning thereafter. In many cases, overparameterized models are used as the initial models in nested meta learning for quick adaptation. However, training such initial models is often difficult in meta learning because the number of training data is much smaller than the dimension of the model parameter.

The work of L. Chen and T. Chen was partially supported by NSF MoDL-SCALE Grant 2134168 and the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons). Correspondence to: T. Chen (chentianyi19@gmail.com).
Previous works on nested meta learning mainly focus on addressing the optimization challenges or analyzing the generalization performance with sufficient data [12, 16, 17]. Different from these works, we are particularly interested in the generalization performance of the sought initial model in practical scenarios where the total number of data from all tasks is smaller than the dimension of the initial model, which we term overparameterized meta learning. In those overparameterized regimes, the generalization error of meta learning models is not fully understood. Motivated by this, we ask:

Whether overparameterized nested meta learning models would lead to overfitting?

In this paper, we take an initial step by answering this in the meta linear regression setting.

1.1 Prior art

We review prior art that we group in the following two categories.

Benign overfitting analysis. The empirical success of overparameterized deep neural networks has inspired theoretical studies of overparameterized learning. The most closest line of work is benign overfitting in linear regression [5], which provides excess risk that measures the difference between expected population risk of the empirical solution and the optimal population risk. Analysis of overparameterized linear regression model with the minimum norm solution. It concludes that certain data covariance matrices lead to benign overfitting, explaining why overparameterized models that perfectly fit the noisy training data can work well during testing. The analysis has been extended to ridge regression [40], multi-class classification [44], and adversarial learning with linear models [8]. While previous theoretical efforts on benign overfitting largely focused on linear models, most recently, the analysis of benign overfitting has been extended to two-layer neural networks [7, 20, 27]. However, existing works mainly study benign overfitting for single level empirical risk minimization problems, rather than bilevel problems such as nested MAML, which is the focus of this work.

Meta learning. Early works of meta learning build black-box recurrent models that can make predictions based on a few examples from new tasks [2, 11, 24, 37], or learn shared feature representation among multiple tasks [38, 42]. More recently, meta learning approaches aim to find the initialization of model parameters that can quickly adapt to new tasks with a few number of optimization steps such as MAML [19, 32, 34]. The empirical success of meta learning has also stimulated recent interests on building the theoretical foundation of meta learning methods.

Generalization of meta learning. The excess risk, as a metric of generalization ability of nested meta learning has been analyzed recently [3, 4, 9, 13, 17, 43]. Generalization performance has also been studied in a relevant but different setting - representation based meta learning [12, 15]. Information theoretical generalization bounds have been proposed in [10, 25], which bound the generalization error in terms of mutual information between the input training data and the output of the meta-learning algorithms. The PAC-Bayes framework has been extended to meta learning to provide a PAC-Bayes meta-population risk bound [1, 14, 18, 35]. These works mostly focus on the case where the meta learning model is underparameterized; that is, the total number of meta training data from all tasks is larger than the dimension of the model parameter. Recently, overparameterized meta learning has attracted much attention. Bernacchia [6] suggests that in overparameterized MAML, negative learning rate in the inner loop is optimal during meta training for linear models with Gaussian data. Sun et al. [39] shows that the optimal representation in representation-based meta linear regression is overparameterized and provides sample complexity bounds for the method of moment estimator.
Table 1: A comparison with closely related prior work on meta learning. “Reps.” and “Nested” refer to representation based methods and nested methods; “Per-task” refers to the per-task level overparameterization and “Meta” refers to the meta level overparameterization defined above.

| Prior work     | Type of meta learning | Overparameterization | Methods       | Focus of analysis |
|----------------|-----------------------|----------------------|---------------|------------------|
| Bai et al. [3] | ✓                     | ✓                    | iMAML         | Train-validation split |
| Chen et al. [9] | ✓                     | ✓                    | MAML, BMAML   | Test risk        |
| Bernacchia et al. [6] | ✓                 | ✓                    | MAML          | Optimal step size |
| Saunshi et al. [36] | ✓               | ✓                    | -             | Train-validation split |
| Sun et al. [39]  | ✓                     | ✓                    | -             | Optimal representation |
| Zou et al. [45]  | ✓                     | ✓                    | MAML          | Optimal step size |
| Ours            | ✓                     | ✓                    | MAML, iMAML   | Benign overfitting |

Compared to the most relevant works, our work is different in the following aspects. Compared to the works that also analyze generalization error or sample complexity in linear meta learning models such as [3, 9, 13], we focus on the overparameterized case when the total number of training data is smaller than the dimension of the model parameter. Compared to the work that focus on representation-based meta learning with a bilinear structure [39], we consider initialization-based meta learning methods with a nested structure such as MAML and iMAML. Furthermore, we provide tight analysis of the excess risk with explicit consideration of the benign overfitting condition.

A summary of key differences between our work and prior art is provided in Table 1. In Table 1, we distinguish two different overparameterization settings: i) the per-task level overparameterization where the dimension of model parameter is larger than the number of training data per task, but smaller than the total number of data across all tasks; and, ii) the meta level overparameterization where the dimension of model parameter is larger than the total number of training data from all tasks.

1.2 This work

This paper provides a unifying analysis of the generalization performance for nested meta learning problems with overparameterized meta models. To our best knowledge, this is the first work that provides the condition for benign overfitting in nested meta learning including MAML and iMAML.

Technical challenges. Before we introduce the key result of our paper, we first highlight the challenges of analyzing the generalization of nested meta learning and characterizing its benign overfitting condition, compared to the non-nested setting such as in [5, 39, 40].

T1) Due to the bilevel structure of nested meta learning, the solution to the meta training objective involves high order terms of data covariance. As a result, the dominating term in the excess risk propagated from the label noise contains higher order terms, which is harder to quantify and can potentially lead to orders of magnitude higher excess risk than the linear regression case [5, 39, 40].

T2) The existing analysis of benign overfitting in the linear regression setting [5, 40] critically relies on the monotonicity of the eigenvalues of the solution matrix on the eigenvalues of the data matrix. However, due to the structure of nested meta learning and thus the solution matrix, this monotonicity is no longer true in general.
Due to the multi-task learning nature of meta learning, the excess risk of meta learning depends on the heterogeneity across different tasks in terms of both the task data covariance and the ground truth task parameter. As a result, the data covariance matrices from different tasks have different eigenvectors. This is in contrast to the linear regression case where all the data follow the same distribution.

Contributions. In view of challenges, our contributions can be summarized as follows.

C1) We derive the upper bound of the excess risk for overparameterized nested meta learning including MAML and iMAML, with their minimum norm solution. Specifically, the excess risk upper bound adopts the following form

\[
\text{Cross-task variance} + \text{Per-task variance} + \text{Bias}
\]

where the \textit{cross-task variance} quantifies the error caused by finite task number and the variation of the ground truth task specific parameter, which is a unique term compared to single task learning. The \textit{bias} quantifies the bias resulting from the minimum norm solution; and the \textit{per-task variance} quantifies the error caused by noise in training data.

C2) We compare the benign overfitting condition for the overparameterized nested meta learning models and that for the empirical risk minimization (ERM) which learns a single shared parameter for all tasks, and show that overfitting is more likely to happen in MAML and its variants such as implicit MAML than in ERM. In addition, larger data heterogeneity across tasks will make overfitting more likely to happen.

C3) We discuss the choice of hyperparameter, e.g., the step size in MAML and the weight of the regularizer in iMAML, such that if the data leads to benign overfitting in ERM, it also leads to benign overfitting in MAML and iMAML. We show that a negative step size can preserve benign overfitting in MAML. This is complementary to the recent discovery that the optimal step size of overparameterized MAML during training is negative [6].

2 Problem Formulation and Methods

In this section, we will introduce the problem setup and the nested meta learning methods.

Problem setup. In the meta-learning setting, assume task \( m \) is drawn from a task distribution, i.e. \( m \sim \mathcal{M} \). For each task \( m \), we observe \( N \) samples with input feature \( x_m \in \mathcal{X}_m \subset \mathbb{R}^d \) and target label \( y_m \in \mathcal{Y}_m \subset \mathbb{R} \) drawn i.i.d. from a task-specific data distribution \( \mathcal{P}_m \). These samples are collected in the dataset \( \mathcal{D}_m = \{(x_{m,n}, y_{m,n})\}_{n=1}^{N} \), which is divided into the disjoint train and validation datasets, denoted as \( \mathcal{D}^{\text{tr}}_m \) and \( \mathcal{D}^{\text{va}}_m \). And \( |\mathcal{D}^{\text{tr}}_m| = N_{\text{tr}} \) and \( |\mathcal{D}^{\text{va}}_m| = N_{\text{va}} \) with \( N = N_{\text{tr}} + N_{\text{va}} \). We use the empirical loss \( \ell_m(\theta_m, \mathcal{D}_m) \) of per-task parameter \( \theta_m \in \Theta_m \) as a measure of the performance. In this paper, we consider regression problems, where \( \ell_m \) is defined as the mean squared error.

The goal for nested meta learning methods, such as MAML [19] and iMAML [33], is to learn an initial parameter \( \theta_0 \in \Theta_0 \), which, with an adaptation method \( \mathcal{A} : \Theta_0 \times (\mathcal{X}_m \times \mathcal{Y}_m)^{N_{\text{tr}}} \to \Theta_m \), can generate a per-task parameter \( \theta_m \) that performs well on the validation data for task \( m \). Given \( M \) tasks, our meta-learning objective is computed as the average of the per-task objective, given by

\[
\mathcal{L}^\mathcal{A}(\theta_0, \mathcal{D}) := \frac{1}{M} \sum_{m=1}^{M} \ell_m(\mathcal{A}(\theta_0, \mathcal{D}^{\text{tr}}_m), \mathcal{D}^{\text{va}}_m). \tag{1}
\]
Obtaining the empirical solution $\hat{\theta}_0^A$ by minimizing (1) under a meta learning method $\mathcal{A}$, in the meta testing stage, we evaluate $\hat{\theta}_0^A$ on the population risk, given by

$$\mathcal{R}^A(\hat{\theta}_0^A) := \mathbb{E}_m \left[ \mathbb{E}_{D_m} \left[ \ell_m(\mathcal{A}(\hat{\theta}_0^A, D_{tr}^m), D_{va}^m) \right] \right].$$

**Methods.** We focus on understanding the generalization performance of two representative nested meta learning methods MAML [19] and iMAML [33] in the overparameterized regime. MAML obtains the task-specific parameter $\hat{\theta}_m(\theta_0)$ by taking one step gradient descent with step size $\alpha$ of the per-task loss function $\ell_m$ from the initial parameter $\theta_0$, that is

$$\mathcal{A}(\theta_0, D_{tr}^m) = \theta_0 - \alpha \nabla_{\theta_0} \ell_m(\theta_0, D_{tr}^m).$$

(3)

On the other hand, iMAML obtains the task-specific parameter $\hat{\theta}_m$ from the initial parameter $\theta_0$ by optimizing the task-specific loss regularized by the distance between $\hat{\theta}_m$ and $\theta_0$, that is

$$\mathcal{A}(\theta_0, D_{tr}^m) = \arg \min_{\theta} \ell_m(\theta, D_{tr}^m) + \gamma/2 \| \theta - \theta_0 \|^2 \quad (4)$$

where $\gamma > 0$ is the weight of the regularizer. As summarized in Figure 1, MAML has smaller computation complexity than iMAML since iMAML requires solving an inner optimization problem during adaptation, while iMAML can achieve smaller test error on each task since it explicitly minimize the regularized per-task loss.

**Notations.** We use $\mu_i(\cdot)$ to denote the $i$-th eigenvalue of a matrix with descending order. $\mathbb{E}[\cdot]$ represents expectation and $\text{Cov}[\cdot]$ represents covariance.

### 3 Main Results: Benign Overfitting for Nested Meta Learning

In this section, we introduce the data model and some necessary assumptions for the analysis. We present the main results, highlight the key steps of the proof and conduct simulations to verify our results. Due to space limitations, we will defer the complete proofs to the supplementary document.

#### 3.1 Model and assumptions

To make a precise analysis, we will assume the following linear data model. Denoting the ground truth parameter on task $m$ as $\theta_m^* \in \mathbb{R}^d$, and the noise as $\epsilon_m$, we assume the data model for task $m$ is

$$y_m = \theta_m^* \top x_m + \epsilon_m. \quad (5)$$

Given the linear model (5), the meta training problem (1) with adaptation method (3) or (4) generally have unique solutions when $d \leq N M$. However, when the meta model $\theta_0$ and thus the per-task model $\theta_m$ are overparameterized, i.e. $d > N M$, the training problem (1) may have multiple solutions. In the subsequent analysis, we will analyze the performance of the minimum norm solution because recent advances in training overparameterized models reveal that gradient descent-based methods converge to the minimum norm solution [22, 29]. We provide a formal definition below.
Definition 1 (Minimum norm solution). Denote \( \mathbf{X}^\text{va}_m := [x_{m,1}, \ldots, x_{m,N_{\text{va}}}]^T \in \mathbb{R}^{N_{\text{va}} \times d} \), \( y_m^\text{va} := [y_{m,1}, \ldots, y_{m,N_{\text{va}}}]^T \in \mathbb{R}^{N_{\text{va}}} \). With \( \mathcal{A}(\theta, \mathcal{D}_m^{\text{tr}}) \) being either (3) or (4), the minimum norm solution to the meta training problem (1) under the linear regression loss is expressed by

\[
\min_{\theta_0} \| \theta_0 \|^2 \quad \text{s.t.} \quad \theta_0 \in \text{arg min} \sum_{m=1}^M \| \mathbf{X}^\text{va}_m \mathcal{A}(\theta, \mathcal{D}_m^{\text{tr}}) - y_m^\text{va} \|^2. \tag{6}
\]

In our analysis, we make the following basic assumptions.

Assumption 1 (Overparameterized model). The total number of meta training data is smaller than the dimension of the model parameter; i.e. \( NM < d \).

Assumption 2 (Subgaussian data). The noise \( \epsilon_m \) is subgaussian with \( \mathbb{E}[\epsilon_m] = 0 \) and \( \mathbb{E}[\epsilon_m^2] = \sigma^2 \). For the \( m \)-th task, data \( x_m = \mathbf{V}_m \Lambda_m^\frac{1}{2} z_m \), where \( z_m \) and its moments have independent, \( \sigma_x \)-subgaussian entries; \( \mathbb{E}[z_m] = 0 \), \( \mathbb{E}[z_m z_m^\top] = \mathbf{I}_d \), with \( \mathbf{I}_d \) being a \( d \times d \) identity matrix.

Assumption 3 (Data matrix). Define \( Q_m := \mathbb{E}[x_m x_m^\top] \), \( Q := \mathbb{E}_m[Q_m] = \mathbf{V} \Lambda \mathbf{V}^\top \) with orthonormal matrices \( \mathbf{V} \) and \( \mathbf{V}_m \). Denote the \( i \)-th eigenvalue of \( Q \) and \( Q_m \) in descending order as \( \lambda_i \) and \( \lambda_{m,i} \), respectively. Assume \( 0 < \lambda \leq \lambda_{m,i} \leq \lambda \) with \( \lambda \) and \( \lambda \) being the smallest and the largest eigenvalues. Also assume the cross-task data heterogeneity \( \mathbb{V}(\{Q_m\}_{m=1}^M) := \max_{i,m} |\lambda_i - \lambda_{m,i}| \) is bounded.

Assumption 4 (Ground truth task parameter). The ground truth parameter \( \theta^*_m \) is independent of \( \mathbf{X}_m \) and satisfies \( \mathbb{Cov}[(\theta^*_m)] = (R^2/d)\mathbf{I}_d \), where \( R \) is a constant, and the individual entries of \( \theta \) are i.i.d. and \( \mathcal{O}(R/\sqrt{d}) \)-subgaussian.

Assumption 1 defines the setting that the meta level is overparameterized, which has also been used in [39]. Note that Assumptions 2-4 are common in the analysis of meta learning; see e.g., [3, 9, 13, 21].

With the linear data model (5), the (minimum norm) solutions to the meta training objective (1) and the meta testing objective (2) can be computed analytically which we will summarize next.

Proposition 1 (Empirical and population level solutions). Under the data model (5), the meta testing objective of method \( \mathcal{A} \) in (2) can be equivalently written as

\[
\mathcal{R}^\mathcal{A}(\theta_0) = \mathbb{E}_m[\| \theta_0 - \theta^*_m \|^2_{W^\mathcal{A}_m}] \tag{7}
\]

where the matrix \( W^\mathcal{A}_m \) and its empirical version \( \hat{W}^\mathcal{A}_m \) are given in Table 2 with \( \hat{Q}_m := \frac{1}{N} \mathbf{X}_m \hat{X}_m^\top \). The optimal solutions to the meta-test risk and the minimum norm solutions to the empirical meta training loss are given below respectively

\[
\hat{\theta}^\mathcal{A}_m := \text{arg min}_{\theta_0} \mathcal{R}^\mathcal{A}(\theta_0) = \mathbb{E}_m[W^\mathcal{A}_m]\mathbb{I}^{-1} \mathbb{E}_m[W^\mathcal{A}_m \theta^*_m], \tag{8a}
\]

\[
\hat{\theta}_0^\mathcal{A} := \text{arg min}_{\theta_0} \mathcal{L}^\mathcal{A}(\theta_0, \mathcal{D}) = \left( \sum_{m=1}^M \hat{W}_m^\mathcal{A} \right)^\dagger \left( \sum_{m=1}^M \hat{W}_m^\mathcal{A} \theta^*_m \right) + \Delta^\mathcal{A}_M \tag{8b}
\]

where \( \dagger \) denotes the Moore–Penrose pseudo inverse; \( \Delta^\mathcal{A}_M \) is a constant that depends on \( \mathbf{X}_m, \epsilon_m \) specified in the supplementary document.
To study overfitting in the meta learning model, we quantify its generalization ability via the widely used metric - excess risk. The excess risk of method $A$ (which can be “ma” for MAML and “im” for iMAML), with an empirical solution $\theta_0^A$ and population solution $\theta_0^A$, is defined as

$$E^A(\theta_0^A) := R^A(\theta_0^A) - R^A(\theta_0^A).$$

In (9), the excess risk measures the difference between the population risk of the empirical solution, $\theta_0$ and the optimal population risk. Given total number of training samples $MN$, if $d \to \infty$, the classic learning theory implies that the excess risk $E^A(\theta_0^A)$ also grows, which leads to overfitting [23]. The larger the excess risk, the further the empirical solution $\theta_0^A$ is from the optimal population solution $\theta_0^A$, indicating more severe overfitting.

### 3.2 Main results

With the closed-form solutions given in Proposition 1, we are ready to bound the excess risk of MAML and iMAML in the overparameterized regime. For notation brevity, we first introduce some universal constants such as $c_0, c_1, c_2, \ldots$, and only present the dominating terms in the subsequent results. The precise presentation of non-dominant terms are deferred to the supplementary document.

We first decompose the excess risk into three terms in Proposition 2.

**Proposition 2.** Denote $W^A := E_m[W^A_m]$. There exists a universal constant $c_0 > 0$ that the excess risk of a meta learning method $A$ can be bounded by

$$E^A(\theta_0^A) \leq c_0 (E_{\theta_0^A} + E_{\epsilon_m^A} + E_b^A)$$

where the first term $E_{\theta_0^A}$ is a function of $\theta_0^A$; the second term $E_{\epsilon_m^A}$, as a function of $\epsilon_m^A$, is the weighted noise variance; and the third term $E_b^A$, as a function of $\theta_0^A$, $W^A_m$, $W^A_m$, is the bias of the minimum norm solution in overparameterized MAML or iMAML.

Based on this decomposition, as we will show in Section 4, the bound of the excess risk can be derived from these three terms $E_{\theta_0^A}, E_{\epsilon_m^A}, E_b^A$, respectively, which gives Theorem 1.

**Theorem 1 (Excess risk bound).** Suppose Assumptions 1-4 hold. Let $\mu_1(\cdot) \geq \mu_2(\cdot) \ldots$ denote the eigenvalues of a matrix in the descending order. Define $W^A_M := \frac{1}{M} \sum_{m=1}^M W^A_m$. For the meta linear regression problem with the minimum norm solution (6), for $0 \leq k \leq d$, define the effective ranks as

$$r_k(W^A_M) := \frac{\sum_{i>k} \mu_i(W^A_M)}{\mu_{k+1}}, \quad R_k(W^A_M) := \left(\frac{\sum_{i>k} \mu_i(W^A_M)^2}{\sum_{i>k} \mu_i^2(W^A_M)}\right)^2.$$  

With the cross-task data heterogeneity $V$ defined in Assumption 3, if there exist universal constants $c_1, c_2, c_3 > 1$ such that the effective dimension $k^* = \min\{k \geq 0 : r_k(W^A_M) \geq c_1 NM\}$, $c_2 \log(1/\delta) < NM$ and $k^* < NM/c_3$, then with probability at least $1 - \delta$, the excess risk satisfies

$$E^A(\theta_0^A) \leq c_3 \|E[\theta^*]\|^2 \sqrt{r_0(W^A_M) MN} + \sigma^2 c_2 \left(\frac{k^*}{MN} + \frac{MN}{R_k(W^A_M)}\right) \left(1 + \mathbb{V}(\{W^A_m\})\right).$$

### Table 2: Matrices $W^A_m, W^A_m$ under different method $A$.

| Method   | Weight matrices |
|----------|-----------------|
| ERM      | $W^A_m = Q_m$ |
| MAML     | $W^A_m = (I - \alpha Q_m)Q_m(I - \alpha Q_m)$ |
| iMAML    | $W^A_m = (\gamma^{-1}Q_m + I)^{-1}Q_m(\gamma^{-1}Q_m + I)^{-1}$ |
|          | $W^A_m = (\gamma^{-1}Q_m + I)^{-1}Q_m(\gamma^{-1}Q_m + I)^{-1}$ |
Theorem 1 provides the excess risk bound via the effective ranks. In (11), the effective ranks \( r_k \) and \( R_k \) of a matrix capture the distribution of the eigenvalues of this matrix, and the effective dimension \( k^* \) determines the above upper bound by considering the asymmetry of the eigenvalues of the solution matrix. The idea is to choose \( k^* \) that makes \( R_{k^*} \) large enough and keeps \( k^* \) small enough compared to \( MN \) so that the variance term of the excess risk is controlled. For example, \( r_0 \) is the trace normalized by the largest eigenvalue, which is bounded above by \( R_0 \). And both \( r_0 \) and \( R_0 \) are no larger than the rank of the matrix, and they are equal to the rank only when all non-zero eigenvalues are equal. If the eigenvalues distribute more uniformly, the effective rank will be larger, otherwise smaller.

**Remark 1.** Below we provide several remarks regarding the effective ranks.

**R1)** The definition of effective rank has been also given in [5] but only on the data matrix \( Q \). And our setting reduces to the single task ERM, or the linear regression case in [5], when \( M = 1, \theta^*_m = \theta_0, W^A_m = Q \), which implies that the cross-task variance in (10) as well as the data heterogeneity \( \mathbb{V}(\cdot) \) reduces to zero. Accordingly, Theorem 1 reduces to Theorem 4 in [5].

**R2)** Given Theorem 1, in order to control the excess risk of solution \( \hat{\theta}^A_0 \), we want \( r_0(W^A_M) \) to be small compared to the total number of training samples \( MN \), but \( k^* \) \( (W^A_M) \) and \( R_{k^*}(W^A_M) \) to be large compared to \( MN \). In addition, the cross-task heterogeneity \( \mathbb{V}(\cdot) \) should be small. Since for a matrix \( W, r_k(W) \leq R_k(W) \leq d \), this suggests the model benefits from overparameterization.

Building upon Theorem 1, we now discuss the conditions for “benign overfitting”, which refers to the situation that overparameterization does not “harm” the excess risk, or the excess risk still vanishes when \( d > MN \) and \( N, M, d \) increase.

**Definition 2 (Condition for benign overfitting in nested meta learning).** The weight matrices \( W^A_M \) for method \( A \) satisfy the benign overfitting condition in nested meta learning, if and only if

\[
\lim_{NM,d \to \infty} \frac{r_0(W^A_M)}{NM} = \lim_{NM,d \to \infty} \frac{k^*}{NM} = \lim_{NM,d \to \infty} \frac{NM}{R_{k^*}(W^A_M)} = 0. \tag{13}
\]

This guarantees the excess risk (12) goes to zero in overparameterized meta learning models with sufficient training data from all tasks. To provide an intuitive explanation, Figure 2 plots the population

\[ \begin{align*}
(a) & \text{MAML with different } \alpha. \\
(b) & \text{iMAML with different } \gamma.
\end{align*} \]

**Figure 2:** Excess risk versus number of samples \( (N) \) with MAML and iMAML under different hyperparameters \( (M = 10, d = 200). \)
risk versus the number of the training data, which demonstrates the “double descent” curve. Namely, as \( N \) increases, \( \mathcal{E}(\hat{\theta}_0) \) first decreases, then increases and then decreases again, as is discovered in overparameterized neural networks [31]. The trend in Figure 2 is similar to the trend observed in [30]. When \( d/(NM) > 1 \), the model is overparameterized, which can overfit the training data, leading to larger excess risk as \( N \) decreases. However, Figure 2 shows the excess risk does not become too large as \( N \) decreases, indicating that overfitting does not severely harm the population risk in this case.

3.3 Examples and discussion

In this section, we discuss how the benign overfitting condition (13) in nested meta learning reduces to that in single task linear regression; e.g., in [5, 40]. We also provide examples to show

\textbf{Q1)} how certain properties of meta training data affect the excess risk; and,

\textbf{Q2)} how to choose the hyperparameters that preserve benign overfitting.

\textbf{Data covariance and cross-task heterogeneity.} Theorem 1 reveals that the excess risk depends on both the eigenvalues of the data covariance matrix \( Q_m \), and the cross-task data heterogeneity, measured by \( \sqrt{\{Q_m\}_{m=1}^M} \). We give an example below to better demonstrate how these two properties of nested meta training data affect the excess risk.

\textbf{Example 1 (Data covariance).} Suppose \( Q_m = \text{diag}(I_{d_1}, \beta I_{d-d_1}) \), \( \forall m \). Set \( M = 10, d = 200, d_1 = 20, \alpha = 0.1 \) for MAML and \( \gamma = 10^3 \) for iMAML. Then the benign overfitting condition (13) is satisfied by MAML and iMAML.

We plot the excess risk under different \( \beta \) in Figure 3. From Figure 3 we can observe that given a fixed number of training data \( N \), the population risk increases with \( \beta \) for both MAML and iMAML. This observation verifies our theory since larger \( \beta \) results in a smaller \( R_k^A(\bar{W}_M) \), leading to a larger upper bound on the variance term in (12).

Example 1 demonstrates how the per-task data matrix \( Q_m \) affects the excess risk. We consider another example that demonstrates how the data across tasks affect the excess risk.

\textbf{Example 2 (Data heterogeneity).} Suppose \( Q_m = |\omega_m + 1| \text{diag}(I_{d_1}, \beta I_{d-d_1}) \) with \( \omega_m \sim N(0, \sigma^2_\omega) \) for all \( m \). Set \( M = 10, d = 200, d_1 = 20, \beta = 0.3, \alpha = 0.1 \) for MAML and \( \gamma = 0.1 \) for iMAML.
Then it satisfies the benign overfitting condition (13) for MAML and iMAML. Figure 4 plots the excess risk with different choices of \( \sigma_\omega \).

Observing from Figure 4 that the larger \( \sigma_\omega^2 \), the higher the excess risk, and the more difficult for the benign overfitting condition to be satisfied for both MAML and iMAML. Therefore, compared to ERM with a single task, the benign overfitting condition for MAML is more restrictive as it imposes constraints for both the expected data covariance \( Q_m \), and the data heterogeneity \( \mathbb{V}(\{W_m^{A}\}_{m=1}^M) \).

**Connection to multi-task ERM.** To compare benign overfitting in the nested meta learning with that in the conventional ERM, where \( \theta_m = \theta_0 \), we can set the step size \( \alpha = 0 \) in MAML, or \( \gamma \to \infty \) in iMAML, and \( N_{\text{ta}} = N \), which reduces to conventional ERM without adaptation. Compared to that of MAML and iMAML in (13), the benign overfitting condition is less restrictive for ERM since it does not impose constraints on \( \alpha \) or \( \gamma \). Intuitively, benign overfitting is more likely to happen in MAML or iMAML than in ERM. The hyperparameters \( \alpha \) and \( \gamma \) will affect the eigenvalues of \( W_{ma}^m, W_{im}^m \), respectively, thus affecting their corresponding excess risk. Here we provide a sufficient condition where the benign overfitting condition in ERM is preserved in MAML or iMAML. We summarize the results in the corollary below.

**Corollary 1 (Hyperparameters that preserve benign overfitting).** Recall \( \lambda_1 \) is the largest eigenvalue of \( Q \). For MAML, when \( \alpha \leq \frac{1}{\lambda_1^2} \), and for iMAML, when \( \gamma \geq \lambda_1 \), then the effective ranks of \( W_{ma}^{im} \) and \( W_{im}^{im} \) are bounded above and below by a positive constant times the effective rank of \( Q \), and therefore the benign overfitting condition holds for MAML and iMAML if it holds for ERM. To summarize, there are constants \( c_1, c_2, c_3, c \) such that for \( k^* = \min\{k \geq 0 : r_k(Q_M) \geq c_1 N M \} \). For \( \delta < 1, c_2 \log(1/\delta) < N M \) and \( k^* < N M / c_3 \), with probability at least \( 1 - 7e^{-2NM/c} \), it follows

\[
\mathcal{E}(\hat{\theta}_0^A) \leq c_3 \|E[\theta_m]\|^2 \lambda \sqrt{\frac{r_0(Q)}{MN}} + \sigma^2 c_2 \left( \frac{k^*}{MN} + \frac{M N}{R_{k^*}(Q_M)} \right) \left( 1 + \mathbb{V}(\{Q_m\}_{m=1}^M) \right),
\]

(14)

**Remark 2.** For MAML, let the unordered eigenvalues \( \tilde{\mu}_i(W_{ma}^{im}) = \lambda_i(1 - \alpha \lambda_i)^2 \). One challenge to control \( \tilde{\mu}_i(W_{ma}^{im}) \) is that \( \tilde{\mu}_i(W_{ma}^{im}) \) are not necessarily monotonic w.r.t. \( \lambda_i \); that is, it does not necessarily hold that \( \tilde{\mu}_1 \geq \tilde{\mu}_2 \geq \cdots \geq \tilde{\mu}_d \). For any \( \lambda_i \geq \lambda_j, \) if \( \tilde{\mu}_i(W_{ma}^{im}) \geq \tilde{\mu}_j(W_{ma}^{im}) \), then we say...
the order of the eigenvalues is preserved. For this to hold, it requires $\tilde{\mu}_i(\lambda_i)$ to be a monotonically non-decreasing function of $\lambda_i$, which yields $\alpha \leq \frac{1}{3M}$. Similar results can be obtained for iMAML by controlling the value of $\gamma$. And the bound on $\alpha$ or $\gamma$ further ensures that $\tilde{\mu}_i(W^A)$ is bounded above and below by a positive constant times the effective rank of $Q$.

4 Proof Outline

In this section, we highlight the key steps of the proof for Theorem 1. We achieve so by analyzing the three terms in Proposition 2 respectively.

The first two terms in (10) can be bounded based on the concentration inequalities on subgaussian variables, given in Lemmas 1 and 2.

**Lemma 1 (Bound on cross-task variance).** With probability at least $1 - \delta$, it follows

$$E_{\theta^*} = \left\| \left( \sum_{m=1}^{M} \hat{W}_m^A \right) \right\|^2 \leq \tilde{O}\left( \frac{1}{M} \right)$$

where $\tilde{O}(\cdot)$ hides the log polynomial dependence on $N, M, d$.

The cross-task variance term analyzed in Lemma 1 is unique in meta learning, which captures the data heterogeneity across different tasks. To elaborate Lemma 1, we plot the cross-task variance versus the task number in Figure 5(a) with the dimension of model parameter $d = 200$, training validation split parameter $s = N_{tr}/N = 0.5$, per-task data number $N = 10$. This figure demonstrates that the cross-task variance decreases with number of tasks $M$ with rate $O\left( \frac{1}{M} \right)$, which is consistent with the theoretical results in Lemma 1. Interestingly, we can observe a double descent phenomenon when $M$ increases so that the model goes from overparameterized setting to underparameterized setting. Note that the “double descent” phenomenon in Figure 5(a) is different from that in the conventional linear regression case in that increasing the number of tasks $M$ will increase data from different per-task distributions, but they still help reduce the excess risk.
Lemma 2 (Bound on bias). There is a constant $c_3$ that depends only on $\sigma_x$, such that for any $1 < \log(1/\delta) < MN_{va}$, with probability at least $1 - \delta$, we have

$$E_b \leq c_3 \|\theta_0\|^2 \|W^A\| \max \left\{ \sqrt{\frac{r_0(W^A)}{MN_{va}}}, \frac{r_0(W^A)}{MN_{va}}, \sqrt{\frac{\log(1/\delta)}{MN_{va}}} \right\}. \quad (16)$$

This term is similar to the bias term in the linear regression case, but directly depending on the solution matrix $W$ instead of the data matrix $Q$. To elaborate Lemma 2, Figure 5(b) demonstrates that the bias term decays with $M$ until it reaches zero when the model is underparameterized. These two terms in Lemma 1 and Lemma 2 do not go to infinity as $N, M, d$ increase.

Note that, the key step is the bound on $E_{\epsilon m}$, which is the dominating term in the decomposition of excess risk (10) in the overparameterized regime. We will bound it below.

Lemma 3 (Bound on per-task variance). There exist constants $c_1, c_2, c_3, c$ such that for $0 \leq k \leq 2NM/c_3$, $r_k(W^A_M) \geq c_1 NM$, and $k_0 \leq k$, with probability at least $1 - 7e^{-2NM/c}$, it follows

$$E_{\epsilon m} \leq c_2 \left( \frac{k_0}{2NM} + \frac{NM}{R_{k_0}(W^A_M)} \right) \left( 1 + \mathcal{V}(\{W^A_m\}_{m=1}^M) \right). \quad (17)$$

Note that, in the single task linear regression case, the cross-task data heterogeneity becomes none, i.e., $\mathcal{V} = 0$. This term is unique in the meta learning setting with multiple tasks. Plugging the results of Lemmas 1, 2 and 3 into (10), we will reach Theorem 1.

5 Conclusions and Limitations

This paper studies the generalization performance of the nested meta learning with an overparameterized model. For a precise analysis, we focus on linear models where the total number of data from all tasks is smaller than the dimension of the model parameter. We show that when the data heterogeneity across tasks is relatively small, the per-task data covariance matrices with certain properties lead to benign overfitting for nested meta learning with the minimum norm solution. This explains why overparameterized meta learning models can generalize well in new data and new tasks. Furthermore, our theory shows that overfitting is more likely to happen in nested meta learning than in ERM, especially when the data heterogeneity across tasks is relatively high in meta learning. Our current analysis is limited to the linear models and non-Bayesian estimate. However, it is promising to extend it for analyzing Bayesian meta learning with more general fully connected or convolutional neural networks in the neural tangent kernel regime. We will pursue those in our future work.
References

[1] Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended pac-bayes theory. In *Proc. International Conference on Machine Learning*, pages 205–214, Stockholm, Sweden, 2018.

[2] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *Proc. Advances in Neural Information Processing Systems*, pages 3981–3989, Barcelona, Spain, 2016.

[3] Yu Bai, Minshuo Chen, Pan Zhou, Tuo Zhao, Jason Lee, Sham Kakade, Huan Wang, and Caiming Xiong. How important is the train-validation split in meta-learning? In *Proc. International Conference on Machine Learning*, pages 543–553, virtual, 2021.

[4] Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *Proc. International Conference on Machine Learning*, pages 424–433, Long Beach, CA, 2019.

[5] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907378117.

[6] Alberto Bernacchia. Meta-learning with negative learning rates. In *Proc. International Conference on Learning Representations*, virtual, 2020.

[7] Yuan Cao, Zixiang Chen, Mikhail Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *arXiv preprint arXiv:2202.06526*, 2022.

[8] Jinghui Chen, Yuan Cao, and Quanquan Gu. Benign overfitting in adversarially robust linear classification. *arXiv preprint arXiv:2112.15250*, 2021.

[9] Lisha Chen and Tianyi Chen. Is bayesian model agnostic meta learning better than model agnostic meta learning, provably? In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

[10] Qi Chen, Changjian Shui, and Mario Marchand. Generalization bounds for meta-learning: An information-theoretic analysis. In *Proc. Advances in Neural Information Processing Systems*, virtual, 2021.

[11] Yutian Chen, Matthew W Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Timothy P Lillicrap, Matt Botvinick, and Nando Freitas. Learning to learn without gradient descent by gradient descent. In *Proc. International Conference on Machine Learning*, pages 748–756, Sydney, Australia, 2017.

[12] Kurtland Chua, Qi Lei, and Jason D Lee. How fine-tuning allows for effective meta-learning. In *Proc. Advances in Neural Information Processing Systems*, volume 34, 2021.

[13] Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Learning to learn around a common mean. In *Proc. Advances in Neural Information Processing Systems*, volume 31, Montreal, Canada, 2018.
[14] Nan Ding, Xi Chen, Tomer Levinboim, Sebastian Goodman, and Radu Soricut. Bridging the gap between practice and pac-bayes theory in few-shot meta-learning. In Proc. Advances in Neural Information Processing Systems, virtual, 2021.

[15] 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, 2020.

[16] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In Proc. International Conference on Artificial Intelligence and Statistics, pages 1082–1092, virtual, 2020.

[17] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks. In Proc. Advances in Neural Information Processing Systems, virtual, 2021.

[18] Alec Farid and Anirudha Majumdar. Generalization bounds for meta-learning via pac-bayes and uniform stability. In Proc. Advances in Neural Information Processing Systems, virtual, 2021.

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

[20] Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. arXiv preprint arXiv:2202.05928, 2022.

[21] Katelyn Gao and Ozan Sener. Modeling and optimization trade-off in meta-learning. In Proc. Advances in Neural Information Processing Systems, volume 33, virtual, 2020.

[22] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Proc. International Conference on Machine Learning, volume 80, pages 1832–1841, Stockholm, Sweden, 10–15 Jul 2018.

[23] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. The elements of statistical learning: data mining, inference, and prediction, volume 2. Springer, 2009.

[24] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In Proc. International Conference on Artificial Neural Networks, pages 87–94, Vienna, Austria, 2001.

[25] Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic generalization bounds for meta-learning and applications. Entropy, 23(1):126, 2021.

[26] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. Bernoulli, 23(1):110–133, 2017.

[27] Zhu Li, Zhi-Hua Zhou, and Arthur Gretton. Towards an understanding of benign overfitting in neural networks. arXiv preprint arXiv:2106.03212, 2021.
[28] L. Mirsky. A trace inequality of john von neumann. *Monatshefte für Mathematik*, 79:303–306, 1975.

[29] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

[30] Preetum Nakkiran, Prayaag Venkat, Sham M Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. In *Proc. International Conference on Learning Representations*, virtual, 2020.

[31] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

[32] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[33] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Proc. Advances in Neural Information Processing Systems*, pages 113–124, Vancouver, Canada, 2019.

[34] Jonas Rothfuss, Dennis Lee, Ignasi Clavera, Tamim Asfour, and Pieter Abbeel. Promp: Proximal meta-policy search. In *Proc. International Conference on Learning Representations*, Vancouver, Canada, 2018.

[35] Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. PACOH: Bayes-optimal meta-learning with pac-guarantees. In *Proc. International Conference on Machine Learning*, pages 9116–9126, virtual, 2021.

[36] Nikunj Saunshi, Arushi Gupta, and Wei Hu. A representation learning perspective on the importance of train-validation splitting in meta-learning. In *International Conference on Machine Learning*, pages 9333–9343, 2021.

[37] J. Schmidhuber. A neural network that embeds its own meta-levels. In *Proc. IEEE International Conference on Neural Networks*, pages 407–412 vol.1, 1993.

[38] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proc. Advances in Neural Information Processing Systems*, pages 4080–4090, Long Beach, CA, 2017.

[39] Yue Sun, Adhyyan Narang, Halil Ibrahim Gulluk, Samet Oymak, and Maryam Fazel. Towards sample-efficient overparameterized meta-learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Proc. Advances in Neural Information Processing Systems*, virtual, 2021.

[40] Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *arXiv: Statistics Theory*, 2020.
[41] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[42] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Proc. Advances in Neural Information Processing Systems*, volume 29, pages 3630–3638, Barcelona, Spain, 2016.

[43] Haoxiang Wang, Ruoyu Sun, and Bo Li. Global convergence and generalization bound of gradient-based meta-learning with deep neural nets. *arXiv preprint arXiv:2006.14606*, 2020.

[44] Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Proc. Advances in Neural Information Processing Systems*, virtual, 2021.

[45] Yingtian Zou, Fusheng Liu, and Qianxiao Li. Unraveling model-agnostic meta-learning via the adaptation learning rate. In *Proc. International Conference on Learning Representations*, 2022.
Appendix

Table of Contents

A Notations 17
B Proof of Proposition 1 17
B.1 Model agnostic meta learning method 18
B.2 Implicit model agnostic meta learning method 19
C Proof of Theorem 1 21
C.1 Proof of Proposition 2 21
C.2 Proof of Lemma 1 24
C.3 Proof of Lemma 2 24
C.4 Proof of Lemma 3 25
D Auxiliary Lemmas 33
D.1 Basic algebraic properties 33
D.2 Concentration inequalities 33
D.3 Other intermediate results 33

A Notations.

We use $[X_m]_M = \mathbb{R}^{M \times d}$ to represent row concatenation of matrices $X_m$ with indices $m$, i.e.

$$[X_m]_M = [X_1^T, X_2^T, \ldots, X_M^T]^T.$$  

For any matrix $M = \mathbb{R}^{N \times d}$ denote $M_{0:k}$ to be the matrix which comprises the first $k$ columns of $M$, and $M_{k:\infty}$ to be the matrix comprised of the rest of the columns of $M$. For any vector $v \in \mathbb{R}^d$ denote $v_{0:k}$ to be the vector comprised of the first $k$ elements of $v$, and $v_{k:\infty}$ to be the vector comprised of the rest of the elements of $v$. Finally, denote $\Lambda_{0:k} = \text{diag}(\lambda_1, \ldots, \lambda_k)$, and $\Lambda_{k:\infty} = \text{diag}(\lambda_{k+1}, \lambda_{k+2}, \ldots)$.

B Proof of Proposition 1

Proposition 3 (Empirical and population level solutions). Under the data model (5), the meta-test risk of method $A$ defined in (2) can be computed by

$$\mathcal{R}^A(\theta_0) = \mathbb{E}_m [\|\theta_0 - \theta^*_m\|_{\mathcal{W}^A_m}^2] + c.$$

The optimal solutions to the meta-test risk and the minimum norm solution are given below respectively

$$\theta^0_A := \arg\min_{\theta_0} \mathcal{R}^A(\theta_0) = \mathbb{E}_m \left[ \mathcal{W}^A_m \right]^{-1} \mathbb{E}_m \left[ \mathcal{W}^A_m \theta^*_m \right]$$  \hspace{1cm} (18a)

$$\hat{\theta}_0^A := \arg\min_{\theta_0} \mathcal{L}^A(\theta_0, D) = \left( \sum_{m=1}^M \mathcal{W}^A_m \right)^\dagger \left( \sum_{m=1}^M \mathcal{W}^A_m \theta^*_m \right) + \Delta^A_M $$  \hspace{1cm} (18b)
where \( \dagger \) denotes the Moore-Penrose pseudo inverse, the error term \( \Delta_M^A \) is a polynomial function of \( M, N, d \), which will be specified in the following sections for MAML and iMAML. And \( \hat{Q}_{m}^a := \frac{1}{N} X_m^a \top X_m^a \). The weight matrices of different methods, \( W_m^A \) and \( \hat{W}_m^A \), are given in Table 2.

### B.1 Model agnostic meta learning method

Without loss of generality, assume \( \sigma = 1 \) to simplify notation. We use meta-test risk \( \mathcal{R}_{N}^{ma} \) to represent expected test risk with finite number of adaptation data \( N \) during testing, which is slightly different compared to population risk \( \mathcal{R}^{A} = \lim_{N \to \infty} \mathcal{R}_{N}^{ma} \). The MAML meta-test risk is defined as [21]

\[
\mathcal{R}_{N}^{ma}(\theta_0) := \mathbb{E} \left[ (y_m - \theta_m^*(\theta_0, D_{m,N})^\top x_m)^2 \right]
= \mathbb{E}_m \left[ \|\theta_0 - \theta_m^*\|^2_{W_{m,N}^{ma}} \right] + 1 + \frac{\alpha^2}{N} \mathbb{E}_m [\text{Tr}(Q_{m}^2)]
\]  

where the matrix is defined as

\[
W_{m,N}^{ma} = \mathbb{E}_m \left[ (I - \alpha \hat{Q}_m)(I - \alpha \hat{Q}_m^a) \right]
= (I - \alpha \hat{Q}_m)(I - \alpha \hat{Q}_m^a) + \frac{\alpha^2}{N} \left( \mathbb{E}_{x_{m,i}} \left[ x_{m,i}^\top Q_m x_{m,i}^\top x_{m,i}^\top \right] - Q_m^3 \right).
\]  

Assume during meta testing, we have infinite adaptation data, i.e., \( N \to \infty \), then the optimal population risk of MAML is

\[
\mathcal{R}^{ma}(\theta_0) = \lim_{N \to \infty} \mathcal{R}_{N}^{ma}(\theta_0) = \mathbb{E}_m \left[ \|\theta_0 - \theta_m^*\|^2_{W_{m}^{ma}} \right] + 1.  
\]  

In MAML, define \( \theta_0^{ma} \) as the minimizer of the optimal population risk of MAML, given by

\[
\theta_0^{ma} = \arg \min_{\theta_0} \mathcal{R}_{N}^{ma}(\theta_0) = \arg \min_{\theta_0} \mathbb{E}_m \left[ \|\theta_0 - \theta_m^*\|^2_{W_{m}^{ma}} \right] = \mathbb{E}_m \left[ W_{m}^{ma} \right]^{-1} \mathbb{E}_m \left[ W_{m}^{ma} \theta_m^* \right].
\]  

Using the optimality condition of \( \mathcal{L}^{ma}(\theta_0, D) \) given in (1), we have

\[
\hat{\theta}_0^{ma} = \left( \sum_{m=1}^{M} \hat{W}_{m}^{ma} \right) \top \left( \sum_{m=1}^{M} \hat{W}_{m}^{ma} \theta_m^* + (I - \alpha \hat{Q}_m)(\frac{1}{N_{va}} X_m^a \top e_m - \frac{\alpha}{N_{tr}} \hat{Q}_m^a X_{m}^a \top e_m^a) \right)
\]  

\[
\hat{W}_{m}^{ma} = (I - \alpha \hat{Q}_m^a)(I - \alpha \hat{Q}_m^tr).
\]  

Therefore, we can arrive at (8b) by defining

\[
\Delta_{M}^{ma} := \left( \sum_{m=1}^{M} \hat{W}_{m}^{ma} \right) \top \left( \sum_{m=1}^{M} (I - \alpha \hat{Q}_m^tr) \frac{1}{N_{va}} X_m^a \top e_m^a - (I - \alpha \hat{Q}_m^tr) Q_m^a \frac{\alpha}{N_{tr}} X_{m}^a \top e_m^a \right)
\]  

It is worth noting that, from (20) and the above discussion, we have the property

\[
\mathbb{E}_{x_m}[\hat{W}_{m}^{ma}] = W_{m,N}, \quad \text{and} \quad \lim_{N \to \infty} W_{m,N}^{ma} = W_{m}^{ma}
\]  

which will be used in later sections to derive the excess risk.
B.2 Implicit model agnostic meta learning method

For the iMAML method, the task-specific parameter $\hat{\theta}_m^{im}$ is computed from the initial parameter $\theta_0$ by optimizing the regularized task-specific empirical loss, given by

$$\hat{\theta}_m^{im}(\theta_0, D_m) = \arg\min_{\theta_m} \frac{1}{N} \| y_m - X_m \theta_m \|^2 + \gamma \| \theta_m - \theta_0 \|^2$$

where $\gamma$ is the weight of the regularizer, and $D_m$ is the adaptation data during meta-testing or training data during meta-training.

The empirical loss of iMAML is defined as the average per-task loss, given by

$$\mathcal{L}_M^{im}(\theta_0, D) = \frac{1}{MN_{va}} \sum_{m=1}^{M} \left\| y_{va}^m - X_{va}^m \hat{\theta}_m^{im}(\theta_0, D_m^{tr}) \right\|^2$$

whose minimizer is

$$\hat{\theta}_0^{im} = \arg\min_{\theta_0} \frac{1}{MN_{va}} \sum_{m=1}^{M} \left\| X_{va}^m \theta_m^* + e_{va}^m - X_{va}^m \hat{\theta}_m^{im}(\theta_0, D_m^{tr}) \right\|^2.$$  

Using the optimality condition of the above problem, we obtain

$$\hat{\theta}_0^{im} = \left( \sum_{m=1}^{M} \tilde{W}_m^{im} \right)^\dagger \left( \sum_{m=1}^{M} \tilde{W}_m^{im} \theta_m^* \right) + \Delta_M^{im}$$

with

$$\Delta_M^{im} = \left( \sum_{m=1}^{M} \tilde{W}_m^{im} \right)^\dagger \left( \sum_{m=1}^{M} \gamma \Sigma_{\theta_m} \frac{1}{N_{va}} X_{va}^m e_{va}^m - \gamma^{-1} \tilde{W}_m^{im} \frac{1}{N_{tr}} X_{tr}^m e_{m}^{tr} \right)$$

where we define

$$\Sigma_{\theta_m} := \left( \frac{1}{N_{tr}} X_{tr}^m X_{tr}^m + \gamma I \right)^{-1} = (Q_m^{tr} + \gamma I)^{-1}$$

$$\tilde{W}_m^{im} := \gamma^{2} \Sigma_{\theta_m} \frac{1}{N_{va}} X_{va}^m X_{va}^m \Sigma_{\theta_m} = \gamma^{2} \Sigma_{\theta_m} Q_m^{va} \Sigma_{\theta_m}.$$  

The meta-test risk of iMAML is defined as

$$\mathcal{R}_{Na}^{im}(\theta_0) = \mathbb{E}\left[ \frac{1}{N_a} \| y_{m} - \hat{\theta}_m^{im}(\theta_0, D_m, N_a) \| x_m \|^2 \right]$$

$$= \mathbb{E}_m \left[ \| \theta_0 - \theta_m^* \| \tilde{W}_m^{im} \| \right] + 1 + \frac{1}{N_a} \mathbb{E}\left[ \gamma^{-2} \text{tr}(\tilde{W}_m^{im} Q_m^{va} \Sigma_{\theta_m}) \right]$$

where the weight matrix is defined as

$$\tilde{W}_m^{im} = \mathbb{E}_x \left[ (Q_m^{va} + \gamma I)^{-1} Q_m (Q_m^{va} + \gamma I)^{-1} \right]$$

$$= \tilde{W}_m^{im} \mathbb{E}_x \left[ \Sigma_{\theta_m} (Q_m - Q_m^{va}) \right] \mathbb{E}_x \left[ (Q_m - Q_m^{va}) \Sigma_{\theta_m} + \Sigma_{\theta_m} (Q_m - Q_m^{va}) \right]$$

$$+ \mathbb{E}_x \left[ (Q_m - Q_m^{va}) \Sigma_{\theta_m} \right]$$

$$= \tilde{W}_m^{im} \mathbb{E}_x \left[ \Sigma_{\theta_m} (Q_m - Q_m^{va}) \right] \mathbb{E}_x \left[ (Q_m - Q_m^{va}) \Sigma_{\theta_m} + \Sigma_{\theta_m} (Q_m - Q_m^{va}) \right]$$

$$+ \mathbb{E}_x \left[ (Q_m - Q_m^{va}) \Sigma_{\theta_m} \right]$$
where $W^\text{im}_m = \left(\gamma^{-1}Q_m + I\right)^{-1}Q_m\left(\gamma^{-1}Q_m + I\right)^{-1}$.

Simplify the notation of $X_{m,N_a}$, $y_{m,N_a}$, $Q_{m,N_a}$ as $X_m, y_m, Q_m$. The derivation of (31) is given by

$$R^\text{im}_{N_a}(\theta_0) = \mathbb{E}\left[\|\hat{\theta}^\text{im}_m(\theta_0, D_{m,N_a}) - \theta^*\|^2 \|Q_m\| + 1\right]$$

$$= \mathbb{E}\left[\|Q_m + \gamma I\|^{-1} \left(\frac{1}{N_a}X_m^T y_m + \gamma \theta_0\right) - \theta^*\|^2 \|Q_m\| + 1\right]$$

$$= \mathbb{E}\left[\theta_0^T W^\text{im}_{m,N_a} \theta_0 + 2\gamma \left(\frac{1}{N_a}X_m^T X_m \Sigma_{\theta_m} - \theta^*\right) Q_m \Sigma_{\theta_m} \theta_0 + \frac{1}{N_a} y_m^T X_m \Sigma_{\theta_m} Q_m \Sigma_{\theta_m} X_m y_m + 2\gamma^2 \|Q_m\|^2 \|\theta^*\| + 1\right]$$

where (a) follows from the definition of $\Sigma_{\theta_m} = \left(\hat{Q}_m + \gamma I\right)^{-1}$, and $W^\text{im}_{m,N_a} = \gamma^2 \Sigma_{\theta_m} Q_m \Sigma_{\theta_m}$.

Applying the fact that $y_m = X_m \theta^* + e_m$ and $\mathbb{E}_{e_m}[e_m] = 0$, one can further derive from (32) that

$$R^\text{im}_{N_a}(\theta_0) = \mathbb{E}\left[\theta_0^T W^\text{im}_{m,N_a} \theta_0 + 2\gamma \left(\theta^* Q_m \Sigma_{\theta_m} \theta_0 - \theta^* \right) Q_m \Sigma_{\theta_m} \theta_0 + \theta^* Q_m \Sigma_{\theta_m} Q_m \Sigma_{\theta_m} \Sigma_{\theta_m} \Sigma_{\theta_m} Q_m \theta^* + \frac{1}{N_a} e_m^T X_m \Sigma_{\theta_m} Q_m \Sigma_{\theta_m} X_m e_m\right] + 1.$$  

(33)

Based on the linearity of trace and expectation, and the cyclic property of trace, the last term inside the expectation in the above equation can be computed as

$$\mathbb{E}_{e_m}[e_m^T X_m \Sigma_{\theta_m} Q_m \Sigma_{\theta_m} X_m e_m] = \text{Tr}(X_m \Sigma_{\theta_m} Q_m \Sigma_{\theta_m} X_m \Sigma_{\theta_m} X_m)= \text{Tr}(X_m \Sigma_{\theta_m} Q_m \Sigma_{\theta_m}) = N_a \text{Tr}(\Sigma_{\theta_m} Q_m \Sigma_{\theta_m} Q_m).$$

To derive all the terms related to $\theta^*$, based on the Woodbury matrix identity, $I - \hat{Q}_m \Sigma_{\theta_m} = I - \Sigma_{\theta_m} \hat{Q}_m = \gamma \Sigma_{\theta_m}$, we have

$$(\theta^* \hat{Q}_m \Sigma_{\theta_m} \theta^* - \theta^* \hat{Q}_m \Sigma_{\theta_m} \theta^*) = \theta^* (\hat{Q}_m \Sigma_{\theta_m} - I) = -\gamma \theta^* \Sigma_{\theta_m}$$  

(34)

and then the terms related to $\theta^*$ in (33) can be computed by

$$\theta^* Q_m \Sigma_{\theta_m} Q_m \Sigma_{\theta_m} \Sigma_{\theta_m} \Sigma_{\theta_m} Q_m \theta^* + \theta^* Q_m \Sigma_{\theta_m} Q_m \theta^* + \theta^* Q_m \Sigma_{\theta_m} Q_m \theta^* + \theta^* Q_m \Sigma_{\theta_m} Q_m \theta^*$$

$$(a) = \gamma \theta^* Q_m \Sigma_{\theta_m} Q_m \Sigma_{\theta_m} Q_m \theta^* + \theta^* Q_m \Sigma_{\theta_m} Q_m \theta^* + \theta^* Q_m \Sigma_{\theta_m} Q_m \theta^* + \theta^* Q_m \Sigma_{\theta_m} Q_m \theta^*$$

(35)

where (a) follows from (34), and (b) follows from the definition of $W^\text{im}_{m,N_a}$.

Combining (33) and (35) and rearranging the equations, we obtain

$$R^\text{im}_{N_a}(\theta_0) = \mathbb{E}\left[\theta_0^T W^\text{im}_{m,N_a} \theta_0 - 2\gamma^2 \|Q_m\|^2 \|\theta^*\| + 1\right]$$

$$= \mathbb{E}\left[\|\theta_0 - \theta^*\|^2 W^\text{im}_{m,N_a} + \gamma^{-1} \theta^* \left(- W^\text{im}_{m,N_a} \hat{Q}_m + \hat{Q}_m W^\text{im}_{m,N_a}\right) \theta^* + \frac{1}{N_a \gamma^2} \text{Tr}(W^\text{im}_{m,N_a} \hat{Q}_m)\right] + 1$$

(36)
where (c) follows from rearranging the equations; (d) follows from the fact that
\[ \theta_m^+ \left( W_{m_N_m}^{im} \hat{Q}_m \right) \theta_m = \left( \theta_m^+ \left( W_{m_N_m}^{im} \hat{Q}_m \right) \theta_m^+ \right)^T = \theta_m^+ \left( \hat{Q}_m W_{m_N_m}^{im} \right) \theta_m^+. \] (37)
Since \( \lim_{N_m \to \infty} \frac{1}{N_m} \mathbb{E} \gamma^{-2} \text{Tr}(W_{m_N_m}^{im} \hat{Q}_m \hat{Q}_m) = 0 \), from the definition of the population risk in (2), the population risk of iMAML is given by
\[ R_{im}^m(\theta_0) := \lim_{N_m \to \infty} R_{im}^m(\theta_0) = \mathbb{E}_m [\|\theta_0 - \theta_m^*\|_W^{im}] + 1 \] (38a)
where \( W_m^{im} = (\gamma^{-1}Q_m + I)^{-1}Q_m(\gamma^{-1}Q_m + I)^{-1} \) (38b)
whose minimizer is given by
\[ \theta_0^{im} = \arg \min_{\theta_0} R_{im}^m(\theta_0) = \mathbb{E}_m \left[ W_m^{im} \right]^{-1} \mathbb{E}_m \left[ W_m^{im} \theta_m^* \right]. \] (39)

It is worth noting that, we have the property \( \mathbb{E}_{x_m}[\hat{W}_m^{im}] = W_m^{im}, \lim_{N_m \to \infty} W_m^{im} = W_m^{im} \), which will be used in later sections to derive the specific optimal population risk and statistical error. The above discussion provides proof for Proposition 1.

C Proof of Theorem 1

Section B gives solutions to the empirical and population risks. In this section, we provide proof to the main theorem, starting with the decomposition of the excess risk in Proposition 2. Note that our proof of the bound on the variance follows the idea of [5] by separately bounding the terms related to the first \( k \) largest eigenvalues and the rest eigenvalues of the per-task weight matrices.

C.1 Proof of Proposition 2

Next we analyze the excess risk defined in (9) based on the solutions of MAML and iMAML. First we restate the complete version of Proposition 2 in Lemma 4.

Lemma 4 (Restatement of Proposition 2). With probability at least \( 1 - \delta \), the excess risk of the MAML with the minimum norm solution is bounded by
\[
\mathcal{E}^A(\hat{\theta}_0) \leq 4 \left( \sum_{m=1}^{M} \mathcal{E}^{A}_m \right)^{\frac{1}{2}} \left( \sum_{m=1}^{M} \mathcal{E}^{im}_m (\theta_m^* - \theta_m) \right)^{\frac{1}{2}} W^A + 4 \theta_0^T B^A \theta_0 + 2 c_1 \sigma^2 \log \frac{1}{\delta} \text{Tr}(B^A) \] (40)
where the weight matrix and the constants are defined as
\[ W^A := \mathbb{E}_m[ W^A_m], \quad X^{ma} := [X^m_{va}(I - \alpha \hat{Q}^m_{tr})]_M, \quad X^{im} := [X^m_{va}(I + \gamma^{-1} \hat{Q}^m_{tr})^{-1}]_M \]
\[ B^A := \left( \bar{X}^A (\bar{X}^A \bar{X}^A)^{-1} \bar{X}^A - I \right) W^A \left( \bar{X}^A (\bar{X}^A \bar{X}^A)^{-1} \bar{X}^A - I \right), \]
\[ C^A = C_1^A + C_2^A, \quad C_1^A := (\bar{X}^A \bar{X}^A)^{-1} \bar{X}^A W^A \bar{X}^A (\bar{X}^A \bar{X}^A)^{-1}, \]
\[ C_2^{ma} := \frac{\alpha^2}{N_m} C_1^{ma} \text{diag}[X^m_{va} \hat{Q}^m_{tr} X^m_{va}]_M, \]
\[ C_2^{im} := \frac{1}{N_m} C_1^{im} \text{diag}[X^m_{va}(I + \gamma^{-1} \hat{Q}^m_{tr})^{-1} \hat{Q}^m_{tr}(I + \gamma^{-1} \hat{Q}^m_{tr})^{-1} X^m_{va}]_M. \]

Note that \( C_2^A \) can be either \( C_2^{ma} \) for MAML or \( C_2^{im} \) for iMAML.
**Proof.** The excess risk $\mathcal{E}^A$ can be derived as

$$
\mathcal{E}^A(\hat{\theta}_0) := \mathcal{R}(\hat{\theta}_0) - \mathcal{R}(\theta_0) = \mathbb{E}_m \left[ \|\hat{\theta}_0 - \theta_0^*\|_2^2 \right] - \mathbb{E}_m \left[ \left\|\theta_0 - \theta_0^*\right\|^2_{W_m} \right]
$$

$$
= \hat{\theta}_0^\top W \hat{\theta}_0 - \theta_0^\top W \theta_0 - 2(\hat{\theta}_0 - \theta_0)\top \mathbb{E}_m [W_m \theta_0^*]
$$

$$
= \hat{\theta}_0^\top W \hat{\theta}_0 - \theta_0^\top W \theta_0 - 2(\hat{\theta}_0 - \theta_0)\top W \theta_0
$$

$$
= \hat{\theta}_0^\top W \theta_0 - 2\hat{\theta}_0^\top W \theta_0 + \theta_0^\top W \theta_0 = \|\hat{\theta}_0 - \theta_0\|^2_{W_A}
$$

$$
= \left\| \left( \sum_{m=1}^M \hat{W}_m \right)^\top \left( \sum_{m=1}^M \hat{W}_m \theta_0^* \right) + \Delta_M - \theta_0 \right\|^2_{W_A}
$$

$$
\leq 2 \left\| \left( \sum_{m=1}^M \hat{W}_m \right)^\top \left( \sum_{m=1}^M \hat{W}_m \theta_0^* \right) - \theta_0 \right\|^2_{W_A} + 2 \|\Delta_M\|^2_{W_A}.
$$

(41)

In (41), $I_1$ can be bounded by

$$
I_1 = \left\| \left( \sum_{m} \hat{W}_m \right)^\top \left( \sum_{m} \hat{W}_m \theta_0^* \right) - \theta_0 \right\|^2_{W_A}
$$

$$
= \left\| \left( \sum_{m} \hat{W}_m \right)^\top \left( \sum_{m} \hat{W}_m \theta_0^* - \theta_0 \right) + \left( \sum_{m} \hat{W}_m \right)^\top \left( \sum_{m} \hat{W}_m \right) - I \right\| \theta_0 \|^2_{W_A}
$$

$$
\leq 2 \left\| \left( \sum_{m} \hat{W}_m \right)^\top \left( \sum_{m} \hat{W}_m \theta_0^* - \theta_0 \right) \right\|^2_{W_A} + 2 \left\| \left( \sum_{m} \hat{W}_m \right)^\top \left( \sum_{m} \hat{W}_m \right) - I \right\| \theta_0 \|^2_{W_A}
$$

$$
= 2 \left\| \left( \sum_{m} \hat{W}_m \right)^\top \left( \sum_{m} \hat{W}_m \theta_0^* - \theta_0 \right) \right\|^2_{W_A} + 2\theta_0^\top B \theta_0
$$

(42)

with the matrix $B$ defined as

$$
B = \left( \left( \sum_{m} \hat{W}_m \right)^\top \left( \sum_{m} \hat{W}_m \theta_0^* - \theta_0 \right) \right) W_A \left( \left( \sum_{m} \hat{W}_m \right)^\top \left( \sum_{m} \hat{W}_m \right) - I \right)
$$

$$
= (X^\top \hat{X})^\top \hat{X}^\top \hat{X} - I
$$

$$
= (X^\top (\hat{X}^\top \hat{X}))^{-1} \hat{X} - I)
$$

(43)

And (a) is from the relationship of $\hat{W}$ and $\hat{X}$, recall we use $[\cdot]_M$ to represent row concatenation of matrices or vectors.

In (41), $I_2$ can be bounded by

$$
I_2 = \left\| \left( \sum_{m=1}^M \hat{W}_m \right)^\top \left( \sum_{m=1}^M (I - \alpha \hat{Q}_m^{tr}) \frac{1}{N_2} X_m^v a^T e_m - (I - \alpha \hat{Q}_m^{tr}) \hat{Q}_m^a \alpha N_{tr} X_m^T e_m^T \right) \right\|^2_{W_A}
$$

$$
\leq (a) [e_m^v]^T M C_1^A [e_m^v]_M + [e_m^{tr}]^T M C_2^A [e_m^{tr}]_M - 2 [e_m^{tr}]^T M C_3^A [e_m^{tr}]_M
$$

$$
\leq 2 [e_m^{tr}]^T M C_1^A [e_m^{tr}]_M + 2 [e_m^{tr}]^T M C_2^A [e_m^{tr}]_M
$$

$$
= 2 \text{Tr} (C_1^A [e_m^{tr}]_M [e_m^{tr}]^T) + 2 \text{Tr} (C_2^A (e_m^{tr})_M [e_m^{tr}]^T)
$$

$$
= 2 \text{Tr} (C_1^A + C_2^A) + 2 \text{Tr} (C_1^A (e_m^{tr})_M [e_m^{tr}]^T - I) + C_2^A ([e_m^{tr}]_M [e_m^{tr}]^T - I)
$$

22
where \((b)\) follows from expanding the quadratic terms, and

\[
C_{1}^{A} = \frac{1}{N^2} \mathbf{X}(\sum_{m=1}^{M} \mathbf{W}_m^A)\mathbf{w}^A(\sum_{m=1}^{M} \mathbf{W}_m^A)^\top \mathbf{X}^\top
\]

\[
= \mathbf{X}(\mathbf{X}^\top \mathbf{X})\mathbf{w}^A(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top = (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X}\mathbf{w}^A\mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1},
\]

\[
C_{2}^{ma} = \frac{\alpha^2}{N_{tr}}\mathbf{X}_m^{\top} \mathbf{X}_m^A \tilde{\mathbf{X}}_m\mathbf{X}_m^A(\sum_{m=1}^{M} \mathbf{W}_m^A)^\top \mathbf{w}^A(\sum_{m=1}^{M} \mathbf{W}_m^A)^\top [\mathbf{X}_m^\top \mathbf{X}_m^A \tilde{\mathbf{X}}_m]^\top
\]

\[
= \frac{\alpha^2}{N_{tr}}[\mathbf{X}_m^\top \mathbf{X}_m^A \tilde{\mathbf{X}}_m]M \tilde{\mathbf{X}}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}\mathbf{w}^A \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}[\mathbf{X}_m^\top \mathbf{X}_m^A \tilde{\mathbf{X}}_m]^\top.
\]

By taking the expectation w.r.t. \(e_m\), we need to bound \(\text{Tr}(C_1), \text{Tr}(C_2)\). Based on the cyclic property of trace, \(\text{Tr}(C_{2}^{ma})\) can be further derived as

\[
\text{Tr}(C_{2}^{ma}) = \frac{\alpha^2}{N_{tr}} \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}\mathbf{w}^A \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}[\mathbf{X}_m^\top \mathbf{X}_m^A \tilde{\mathbf{X}}_m]^\top M
\]

\[
= \frac{\alpha^2}{N_{tr}} \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}\mathbf{w}^A \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}\mathbf{X}_m^\top \mathbf{X}_m^A \tilde{\mathbf{X}}_m
\]

\[
= \frac{\alpha^2}{N_{tr}} \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}\mathbf{w}^A \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}[\mathbf{X}_m^\top \mathbf{X}_m^A \tilde{\mathbf{X}}_m]^\top.
\]

For a given square matrix \(P_m\), denote

\[
\text{diag}[P_m] = \begin{bmatrix} P_1 & 0 & \ldots & 0 \\ 0 & P_2 & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & P_M \end{bmatrix}.
\]

Then \(\text{Tr}(C_{2}^{ma})\) can be further written as

\[
\text{Tr}(C_{2}^{ma}) = \frac{\alpha^2}{N_{tr}} \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}\mathbf{w}^A \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}\mathbf{X}_m^\top \mathbf{X}_m^A \tilde{\mathbf{X}}_m
\]

\[
= \frac{\alpha^2}{N_{tr}} \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}\mathbf{w}^A \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}\mathbf{X}_m^\top \mathbf{X}_m^A \tilde{\mathbf{X}}_m
\]

\[
= \frac{\alpha^2}{N_{tr}} \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}\mathbf{w}^A \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-2} \mathbf{X}[\mathbf{X}_m^\top \mathbf{X}_m^A \tilde{\mathbf{X}}_m]^\top M.
\]

Since we have

\[
E_{\epsilon}[I_2] = E_{\epsilon}[\mathbf{e}_m^\top \mathbf{C}_1 \mathbf{e}_m^\top] + E_{\epsilon}[\mathbf{e}_m^\top \mathbf{C}_2 \mathbf{e}_m^\top]
\]

\[
= \text{Tr}(C_1 \text{Cov}([e_m^\top]_M)) + \text{Tr}(C_2 \text{Cov}([e_m^\top]_M)) = \sigma^2 \text{Tr}(C_1 + C_2)
\]

by the subGaussian concentration inequality [41], it holds with probability at least \(1 - \delta\) over \(\epsilon\) that

\[
2I_2 \leq c_1 \sigma^2 \text{log} \frac{1}{\delta} \text{Tr}(C_1 + C_2).
\]

Combining the bounds for \(I_1\) and \(I_2\) in (42) and (48) completes the proof.
### C.2 Proof of Lemma 1

Define

\[
\mathbf{z}_A := \left[ (\theta_1^* - \theta_0^A)^\top, \ldots, (\theta_M^* - \theta_0^A)^\top \right]^\top \in \mathbb{R}^{dM},
\]

\[
\mathbf{U}_A := \left[ \hat{\mathbf{W}}_1^A \left( \sum_{m=1}^M \hat{\mathbf{W}}_m^A \right)^\top, \ldots, \hat{\mathbf{W}}_M^A \left( \sum_{m=1}^M \hat{\mathbf{W}}_m^A \right)^\top \right]^\top \in \mathbb{R}^{dM \times d}.
\]

Then we can derive that

\[
\left\| \left( \sum_{m=1}^M \hat{\mathbf{W}}_m^A \right)^\top \left( \sum_{m=1}^M \hat{\mathbf{W}}_m^A (\theta_m^* - \theta_0) \right) \right\|_{\mathbf{W}_A}^2 = \| \mathbf{U}_A^\top \mathbf{z}_A \|_{\mathbf{W}_A}^2.
\]

By the Hanson-Wright inequality, with probability at least \(1 - \delta\) over \(\theta_m^*\), we have

\[
\left\| \mathbf{U}_A^\top \mathbf{z}_A \right\|_{\mathbf{W}_A}^2 - \mathbb{E}_{\theta_m^*} \left[ \| \mathbf{U}_A^\top \mathbf{z}_A \|_{\mathbf{W}_A}^2 \right] = \mathcal{O} \left( \frac{R^2}{M \sqrt{d}} \right).
\]

To compute \(\mathbb{E}_{\theta_m^*} \left[ \| \mathbf{U}_A^\top \mathbf{z}_A \|_{\mathbf{W}_A}^2 \right]\), first recall \(\text{Cov} \left[ \theta_m^* \right] = \frac{R^2}{d} \mathbf{I}\), then we have

\[
\mathbb{E}_{\theta_m^*} \left[ \| \mathbf{U}_A^\top \mathbf{z}_A \|_{\mathbf{W}_A}^2 \right] = \frac{R^2}{d} \mathbf{E}_{\theta_m^*} \left[ \left( \sum_{m=1}^M \hat{\mathbf{W}}_m^A \right)^\top \mathbf{W}_A \left( \sum_{m=1}^M \hat{\mathbf{W}}_m^A \right)^\top \mathbf{W}_A \right] = \mathcal{O} \left( \frac{1}{d} \right).
\]

Combining (49) and (50) leads to the conclusion.

### C.3 Proof of Lemma 2

Recall \(\mathbf{B} := \left( \check{\mathbf{X}}^\top (\check{\mathbf{X}} \check{\mathbf{X}}^\top)^{-1} \check{\mathbf{X}} - \mathbf{I} \right) \mathbf{W} \left( \check{\mathbf{X}}^\top (\check{\mathbf{X}} \check{\mathbf{X}}^\top)^{-1} \check{\mathbf{X}} - \mathbf{I} \right)^\top\). First note that

\[
\left( \check{\mathbf{X}}^\top (\check{\mathbf{X}} \check{\mathbf{X}}^\top)^{-1} \check{\mathbf{X}} - \mathbf{I} \right)^\top \check{\mathbf{X}} = \check{\mathbf{X}}^\top - \check{\mathbf{X}}^\top = 0.
\]

Thus, for any \(\mathbf{u}\) in the column space of \(\check{\mathbf{X}}^\top\), \(\mathbf{u}\) can be represented as \(\mathbf{u} = \check{\mathbf{X}}^\top \tilde{\mathbf{u}}\), \(\tilde{\mathbf{u}} \neq 0\), then we have

\[
\left( \check{\mathbf{X}}^\top (\check{\mathbf{X}} \check{\mathbf{X}}^\top)^{-1} \check{\mathbf{X}} - \mathbf{I} \right) \mathbf{u} = 0.
\]

And for any \(\mathbf{u}\) orthogonal to the column space of \(\check{\mathbf{X}}^\top\), \(\check{\mathbf{X}} \mathbf{u} = 0\), therefore

\[
\left( \check{\mathbf{X}}^\top (\check{\mathbf{X}} \check{\mathbf{X}}^\top)^{-1} \check{\mathbf{X}} - \mathbf{I} \right) \mathbf{u} = -\mathbf{u}.
\]
Since any \( u \in \mathbb{R}^d \) can be represented as a combination of a vector in the column space of \( \tilde{X}^T \) and a vector orthogonal to the column space of \( \tilde{X}^T \), (\( \tilde{X}^T (\tilde{X} \tilde{X}^T)^{-1} \tilde{X} - I \)) has eigenvalues whose absolute values are no greater than 1, i.e.

\[
\| \tilde{X}^T (\tilde{X} \tilde{X}^T)^{-1} \tilde{X} - I \| \leq 1. \tag{54}
\]

Then let \( M = \left( \tilde{X}^T (\tilde{X} \tilde{X}^T)^{-1} \tilde{X} - I \right) \), expanding \( \theta_0^T B_0 \), we have

\[
\theta_0^T B_0 = \theta_0^T \left( \tilde{X}^T (\tilde{X} \tilde{X}^T)^{-1} \tilde{X} - I \right) W \left( \tilde{X}^T (\tilde{X} \tilde{X}^T)^{-1} \tilde{X} - I \right) \theta_0
\]

\[
\overset{(a)}{=} \theta_0^T M \left( W - \frac{1}{MN_{va}} \tilde{X}^T \tilde{X} \right) M \theta_0
\]

\[
= \theta_0^T M \left( W - \frac{1}{MN_{va}} \tilde{X}^T \tilde{X} + \frac{1}{MN_{va}} \tilde{X}^T \tilde{X} - \frac{1}{MN_{va}} \tilde{X}^T \tilde{X} \right) M \theta_0
\]

\[
\overset{(b)}{\leq} \left\| W - \frac{1}{MN_{va}} \tilde{X}^T \tilde{X} \right\| \| \theta_0 \|^2 + \frac{1}{MN_{va}} | \text{tr}(\tilde{X}^T \tilde{X} - \tilde{X}^T \tilde{X}) | \| \theta_0 \|^2 \tag{55}
\]

where \((a)\) follows from (51), and \((b)\) follows from (54). Thus, due to Theorem 9 in [26], there is an absolute constant \( c \) such that for any \( 1 < t < MN_{va} \) with probability at least \( 1 - e^{-t} \), it follows that

\[
\left\| W - \frac{1}{MN_{va}} \tilde{X}^T \tilde{X} \right\| \| \theta_0 \|^2 \leq c \| \theta_0 \|^2 \| W \| \max \left\{ \sqrt{\frac{r(W)}{MN_{va}}}, \sqrt{\frac{t}{MN_{va}}}, \sqrt{\frac{t}{MN_{va}}} \right\} \tag{56}
\]

where \( r(W) \) is defined as

\[
r(W) := \frac{(E[\|\tilde{x}\|^2])^2}{\|W\|^2} \leq \frac{E[\|\tilde{x}\|^2]}{\|W\|} = \frac{\text{tr}(W)}{\|W\|} = r_0(W). \tag{57}
\]

Based on Lemma 16, we have there exists a constant \( c \) that with probability at least \( 1 - e^{-t} \)

\[
\frac{1}{MN_{va}} | \text{tr}(\tilde{X}^T \tilde{X} - \tilde{X}^T \tilde{X}) | \leq ctr_0(W). \tag{58}
\]

Applying the union bound completes the proof.

### C.4 Proof of Lemma 3

To prove Lemma 3, we need to bound \( \text{Tr}(C) = \text{Tr}(C_1) + \text{Tr}(C_2) \). We first show in Lemma 5 that \( \text{Tr}(C_2) \) can be bounded as \( \Theta(\text{Tr}(C_1)) \). Then the key step is to bound \( \text{Tr}(C_1) \). To bound \( \text{Tr}(C_1) \), first we show in Lemma 6 that \( \text{Tr}(C_1) \) can be decomposed into terms that are related to the first \( k \) largest eigenvalues of \( W \) and the term that is only related to the rest eigenvalues of \( W \). Next we bound the term related to the \( d - k \) smallest eigenvalues of \( W \), as a function of \( \mu_n(A) \), given in Lemma 7. And then we bound the term related to the \( k \) largest eigenvalues of \( W \), given in Lemma 8. Finally, we bound the eigenvalues of \( \mu_n(A) \) in Lemma 9.

**Lemma 5 (Bound on \( \text{Tr}(C_2^m) \) in terms of \( \text{Tr}(C_1^m) \)).** Recall \( \alpha \) is the step size. Let \( c \) be a constant that depends only on \( \sigma_x \). It holds with probability at least \( 1 - \delta \) that

\[
\text{Tr}(C_2^m) \leq c \log(1/\delta) \alpha^2 \text{Tr}(C_1^m) \max_m \text{Tr}(A_m^2) / N_{tr}. \tag{59}
\]

25
Proof. Recall

\[
\text{Tr}(C_2^{\text{na}}) = \frac{\alpha^2}{N_{tr}} \text{Tr} \left( C_1^{\text{na}} \text{diag}[X_m^{\text{va}}Q_m X_m^{\text{va}}^T]_M \right)
\]

and \(X_m^{\text{va}}Q_m X_m^{\text{va}}^T \in \mathbb{R}^{N_{\text{va}} \times N_{\text{va}}},\) where \(X_m^{\text{va}} = Z_m^{\text{va}} \Lambda_m^{\frac{1}{2}} V_m^T.\) Then it satisfies that

\[
X_m^{\text{va}}Q_m X_m^{\text{va}}^T = \frac{1}{N_{tr}} Z_m^{\text{va}} \Lambda_m Z_m^{\text{va}}^T \Lambda_m Z_m^{\text{va}}^T
\]

\[
\mathbb{E}[X_m^{\text{va}}Q_m X_m^{\text{va}}^T] = \mathbb{E}[Z_m^{\text{va}} \Lambda_m Z_m^{\text{va}}^T] = \text{Tr}(\Lambda_m^2) I_{N_{\text{va}}}
\]

By Lemma 13, we have

\[
\text{Tr}(C_2) = \frac{\alpha^2}{N_{tr}} \text{Tr}(C_1 \text{diag}[X_m^{\text{va}}Q_m X_m^{\text{va}}^T]_M) \leq \frac{\alpha^2}{N_{tr}} \text{Tr}(C_1) \max_m \text{Tr}(\Lambda_m^2) \|z_{m,j}\|^2.
\]

Then by concentration inequality on \(Z,\) with probability at least \(1 - \delta\) over \(Z,\) it holds that

\[
\text{Tr}(C_2) \leq \log(1/\delta) \alpha^2 \alpha^2 \text{Tr}(C_1) \max_m \text{Tr}(\Lambda_m^2) / N_{tr}.
\]

This completes the proof.

Lemma 5 shows that \(\text{Tr}(C_2)\) can be bounded as \(\Theta(\text{Tr}(C_1)).\) Then we proceed to bound \(\text{Tr}(C_1).\) In Lemma 6, we decompose \(\text{Tr}(C_1)\) can be decomposed into terms that are related to the first \(k\) largest eigenvalues of \(W\) and the term that is only related to the rest eigenvalues of \(W.\)

**Lemma 6 (Decomposition of \(\text{Tr}(C_1)\)).** For both MAML and iMAML, \(\text{Tr}(C_1^A)\) can be decomposed into

\[
\text{Tr}(C_1^A) = \text{Tr} \left( (\bar{X}W_{0:k} X^T + \bar{X}_{0:k} W_{k:\infty} \bar{X}^T + \bar{X}_{k:\infty} W_{k:\infty} \bar{X}_{0:k}^T + \bar{X}_{k:\infty} W_{k:\infty} \bar{X}_{k:\infty}^T) A^{-2} \right)
\]

where the matrix is defined as \(A = \bar{X} \bar{X}^T.\) For notation simplicity, we omit the superscript \(A.\)

**Proof.** Recall

\[
\bar{X}_m^{\text{va}} = X_m^{\text{va}}(I - \hat{Q}_m^{\text{va}}) = Z_m^{\text{va}} \Lambda_m^{\frac{1}{2}} (I - \alpha D_m^{\text{va}} \Lambda_m^{\frac{1}{2}}) V_m^T = Z_m^{\text{va}} \hat{\Lambda}_m V_m^T.
\]

Let \(A = \bar{X} \bar{X}^T,\) and the singular value decomposition of \(W\) as \(W = V_W A_W V_W^T,\) then for any \(0 \leq k \leq d,\) \(W\) can be computed by

\[
W = V_{W,0:k} A_{W,0:k} V_{W,0:k}^T + V_{W,k:\infty} A_{W,k:\infty} V_{W,k:\infty}^T := W_{0:k} + W_{k:\infty}.
\]

Similarly, \(\bar{X}_m\) can be computed by

\[
\bar{X}_m = Z_m^{\text{va}} \hat{\Lambda}_m V_m^T = \left( Z_m^{\text{va}} A_{m,0:k} + Z_m^{\text{va}} \alpha A_m (I - \hat{D}_m^{\text{va}}) A_{m,0:k} \right) V_{m,0:k}
\]

\[
+ \left( Z_m^{\text{va}} A_{m,k:\infty} + Z_m^{\text{va}} \alpha A_m (I - \hat{D}_m^{\text{va}}) A_{m,k:\infty} \right) V_{m,k:\infty}.
\]
Plugging (63) and (64) into (44), we have \( \text{Tr}(C_1) \) can be computed by

\[
\text{Tr}(C_1) = \text{Tr}\left( (\tilde{X}X^\top)^{-1} \tilde{X}W\tilde{X}^\top (\tilde{X}X^\top)^{-1} \right) = \text{Tr}\left( \tilde{X}W\tilde{X}^\top A^{-2} \right)
\]

\[
= \text{Tr}\left( (\tilde{X}W_{0:k}\tilde{X}^\top + \tilde{X}_{0:k} W_{k:}\tilde{X}^\top + \tilde{X}_{k:}\tilde{X}^\top + \tilde{X}_{k:}\tilde{X}^\top W_{k:}\tilde{X}^\top) A^{-2} \right)
\]

from which the proof is complete.

Then we bound the term related to the \( d - k \) smallest eigenvalues of \( W \) as a function of \( \mu_n(A) \), given in Lemma 7.

**Lemma 7 (Bound on \( \text{Tr}(\tilde{X}_{k:}\tilde{X}_{k:}^\top A^{-2}) \) in \( \text{Tr}(C_1) \)).** With probability at least \( 1 - 3e^{-t} \) over \( Z \), and for \( c \geq t \), it holds that

\[
\text{Tr}(\tilde{X}_{k:}\tilde{X}_{k:}^\top A^{-2}) \leq c\mu_n^{-2}(A) \sum_{i > k} \mu_i(W) \mu_i \left( \sum_{m=1}^{M} W_m \right)
\]

where \( \mu_n \) is the smallest eigenvalue of a matrix.

**Proof.** By Von Neumann’s trace inequality in Lemma 13, \( \text{Tr}(\tilde{X}_{k:}\tilde{X}_{k:}^\top A^{-2}) \) is bounded by

\[
\text{Tr}(\tilde{X}_{k:}\tilde{X}_{k:}^\top A^{-2}) \leq \text{Tr}(\tilde{X}_{k:}^\top \tilde{X}_{k:} A^{-2}) \mu_n^{-2}(A).
\]

Denote \( \tilde{D}_m = \frac{1}{N_{va}} Z_{va}^\top Z_{va} \), \( \tilde{D}_m^\top = \frac{1}{N_{va}} Z_{va}^\top Z_{va} \), and \( \tilde{A}_m = \Lambda_m^2 (I - \alpha \Lambda_m^2 \tilde{D}_m^\top \Lambda_m^2) \). Recall

\[
\tilde{X}_{va} = X_{va} (I - \alpha \tilde{Q}_m) = Z_{va}^\top (I - \alpha \Lambda_m^2 \tilde{D}_m^\top \Lambda_m^2) V_m = Z_{va}^\top \Lambda_m V_m.
\]

Denote

\[
\tilde{X}_{va} = X_{va} (I - \alpha \tilde{Q}_m) = Z_{va}^\top \Lambda_m V_m = Z_{va}^\top \Lambda_m^2 V_m,
\]

where \( \tilde{X}_{va} = \Lambda_m^2 (I - \alpha \Lambda_m) \), \( \tilde{X} = [\tilde{X}_{va}]_{M} \), and \( \tilde{A}_m = \Lambda_m + \tilde{A}_m^A - \Lambda_m^A = \tilde{A}_m^A + \Delta_{\Lambda_m} A_m \), with \( \Delta_{\Lambda_m} = \alpha \Lambda_m (I - \tilde{D}_m^\top A_m^2) \).

Therefore, \( \text{Tr}(\tilde{W}_{k:}\tilde{X}_{k:}^\top \tilde{X}_{k:}) \) can be bounded by

\[
\frac{1}{N_{va}} \text{Tr}(\tilde{W}_{k:}\tilde{X}_{k:}^\top \tilde{X}_{k:})
\]

\[
\leq 2 \sum_{m=1}^{M} \text{Tr}\left( W_{k:} V_{m,k:} (\Lambda_{m,k:} \tilde{D}_m V_{m,k:} + \Delta_{\Lambda_m,k:} \Lambda_{m,k:} V_{m,k:}^\top) \right)
\]

\[
\equiv 2 \sum_{m=1}^{M} \text{Tr}\left( W_{k:} W_{m,k:} + W_{k:} V_{m,k:} (\Lambda_{m}(\tilde{D}_m - I) \Lambda_{m} + \Delta_{\Lambda_m} \Lambda_{m}) V_{m,k:}^\top \right)
\]

(66)

where (a) is from (65) and AM-GM inequality, and (b) is from rearranging terms.
From Lemma 13, the first term in (66), $\sum_{m=1}^{M} \text{Tr}(W_{k:\infty} W_{m,k:\infty})$ is bounded by

$$\sum_{m=1}^{M} \text{Tr}(W_{k:\infty} W_{m,k:\infty}) \leq \sum_{i>k} \mu_i(W) \mu_i\left(\sum_{m=1}^{M} W_m\right). \quad (67)$$

Recall Assumption 2 that $Z_{m,i}^{va}$ has $\sigma^2_x$-subgaussian entries and with zero mean and unit variance. By Lemma 16 and the subgaussian concentration inequality, with probability at least $1 - \delta$ over $Z_{m,i}^{va}$, the second term in (66) is bounded by

$$\sum_{m=1}^{M} \text{Tr}(WV_m A_m (D_{m}^{va} - I) A_m V_m^t) \leq \frac{\sigma_x}{\sqrt{MN}} \sum_{i>k} \mu_i(W) \mu_i\left(\sum_{m=1}^{M} W_m\right). \quad (68)$$

Similarly, based on Lemma 16, and the subgaussian concentration inequality, with probability at least $1 - e^{-t}$ over $Z_{m,i}^{va}$ and $Z_{m,i}^{tr}$, and $c \geq \max\{t, \text{tr}(A)\}$, the last term in (66) is bounded by

$$\sum_{m=1}^{M} \text{Tr}(WV_m \Delta_{m}^{va} \Delta_{m}^{tr} V_m^t) \leq \frac{\sigma_x}{\sqrt{MN}} \sum_{i>k} \mu_i(W) \mu_i\left(\sum_{m=1}^{M} W_m\right). \quad (69)$$

Combining (66)-(69), with probability at least $1 - 3e^{-t}$, for $c \geq \max\{t, \text{tr}(A)\}$, it holds that

$$\text{Tr}(\tilde{X}_{k:\infty} W_{k:\infty} \tilde{X}_{k:\infty}^t A^{-2}) \leq c \mu_n^{-2}(A) \sum_{i>k} \mu_i(W) \mu_i\left(\sum_{m=1}^{M} W_m\right).$$

This completes the proof. \qed

Next we bound the term related to the $k$ largest eigenvalues of $W$, given in Lemma 8.

**Lemma 8 (Bound on terms in $\text{Tr}(C_1)$ related to the first $k$ eigenvalues).** There exists $c$ with $0 \leq k \leq c$ such that with probability at least $1 - 2e^{MN/c}$, the following holds

$$\text{Tr}\left((\tilde{X}_0:0;k \tilde{X}^t + \tilde{X}_0:k W_{k:\infty} \tilde{X}^t + \tilde{X}_{k:\infty} W_{k:\infty} \tilde{X}_{0:k}^t) A^{-2}\right) \leq \frac{ck}{MN}.$$  

**Proof.** Recall that

$$\tilde{X}_m = \tilde{X}_{m,0:k} + \tilde{X}_{m,k:\infty}.$$  

Based on Lemma 13, disregarding the terms that has corresponding zero eigenvalues, we have

$$\text{Tr}\left(\tilde{X}_0:0;k \tilde{X}^t A^{-2}\right) \leq \text{Tr}\left(\tilde{X}_{0:k} W_{0:k} \tilde{X}^t_{0:k} A^{-2}\right),$$  

$$\text{Tr}\left(\tilde{X}_0:k W_{k:0} \tilde{X}^t A^{-2}\right) \leq \text{Tr}\left(\tilde{X}_{0:k} W_{k:0} \tilde{X}^t_{0:k} A^{-2}\right),$$  

$$\text{Tr}\left(\tilde{X}_{k:0} W_{k:0} \tilde{X}^t A^{-2}\right) \leq \text{Tr}\left(\tilde{X}_{k:2} W_{k:2} \tilde{X}^t_{0:k} A^{-2}\right).$$  

28
Therefore, we have
\[
\text{Tr} \left( \sum_{k} \left( \tilde{X}_w k \tilde{X}^\top + \tilde{X}_k \tilde{X}^\top + \tilde{X}_k \tilde{X}^\top \tilde{X}_w k \tilde{X}^\top \right) A^{-2} \right) \leq 3 \sum_{j=1}^{k} \lambda_{w,j} \tilde{x}_j^\top A^{-2} \tilde{x}_j
\]
where \( \bar{X}_j \in \mathbb{R}^{M N \times 1} \) is computed by
\[
\bar{x}_j = \bar{x}_j + \left[ Z_m^v \Delta_{\Lambda m}(V_m^T) \right] M.
\]
And based on Lemma 10, we have
\[
\tilde{x}_j^\top A^{-2} \tilde{x}_j = \tilde{x}_j^\top \sum_{j} \left( \tilde{x}_j^\top \tilde{x}_j \right)^{-2} \tilde{x}_j = \tilde{x}_j^\top \left( \tilde{x}_j^\top \tilde{x}_j + A_{-j} \right)^{-2} \tilde{x}_j = \frac{\tilde{x}_j^\top A^{-2} \tilde{x}_j}{(1 + \tilde{x}_j^\top A_{-j}^{-1} \tilde{x}_j)^2}
\]
(70)
where \( L_j \) is the span of the \( MN v_a - k \) eigenvectors with the smallest eigenvalues of \( A_{-j} \), and \( \Pi_{L_j} \) represents the projection to \( L_j \).

Since \( \bar{x}_j = \bar{x}_j + \left[ Z_m^v \Delta_{\Lambda m}(V_m^T) \right] M \) and
\[
\bar{x}_j = \left[ Z_m^v \Delta_{\Lambda m}(V_m^T) \right] M = \left[ \sum_{i=1}^{d} \lambda_{mi} v_{mji} Z_m^v \right] M.
\]
(71)
It follows that
\[
\left\| \bar{x}_j \right\|^2 \leq 2 \left\| \bar{x}_j \right\|^2 + 2 \left\| \left[ Z_m^v \Delta_{\Lambda m}(V_m^T) \right] M \right\|^2 = 2 \left\| \bar{x}_j \right\|^2 + 2 \sum_{m=1}^{M} \left\| Z_m^v \Delta_{\Lambda m}(V_m^T) \right\|^2.
\]
(72)
Based on Lemma 15, we have
\[
\left\| \bar{x}_j \right\|^2 = \sum_{m=1}^{M} \sum_{i=1}^{d} \left( \lambda_{mi} v_{mji} \right) \left\| Z_m^v \right\|^2 \leq \sum_{m=1}^{M} \sum_{i=1}^{d} \left( \lambda_{mi} v_{mji} \right) \left( N + a \sigma^2 + (t + \ln k + \sqrt{N(t + \ln k)}) \right)
\leq cN \sum_{m=1}^{M} \sum_{i=1}^{d} \left( \lambda_{mi} v_{mji} \right)^2.
\]
(73)
Each term in (72) can be further derived as
\[
\left\| Z_m^v \Delta_{\Lambda m}(V_m^T) \right\|^2 = \left( \sum_{m} \lambda_{mi} v_{mji} \right) \left( \sum_{m} \lambda_{mi} v_{mji} \right) \left( \sum_{m} \lambda_{mi} v_{mji} \right) \left( \sum_{m} \lambda_{mi} v_{mji} \right) \left( \sum_{m} \lambda_{mi} v_{mji} \right).
\]
Its expectation can be computed by
\[
\mathbb{E}[\left( \sum_{m} \lambda_{mi} v_{mji} \right)^2] = \sum_{i=1}^{d} \mathbb{E}[v_{mji}^2] \sum_{m} \lambda_{mi} \left( \sum_{m} \lambda_{mi}^2 + \lambda_{mi}^2 (\mathbb{E}[z^2] - 2) \right).
\]
Based on the subgaussian concentration inequality, it holds with probability at least $1 - e^{-t}$ that,

$$\left\| (V_m^T)_{j}^T \Delta_{\Lambda_m}^T Z_{m}^a \Delta_{\Lambda_m}^T (V_m^T)_{j}^T - E[(V_m^T)_{j}^T \Delta_{\Lambda_m}^T Z_{m}^a \Delta_{\Lambda_m}^T (V_m^T)_{j}^T] \right\| \leq c \left( \sum_{i=1}^{d} v_{mji}^2 \right) \left( \sum_{i=1}^{d} \lambda_{m,i}^2 \right) + c \sum_{m=1}^{M} v_{mj}^2 \lambda_{m,i}^2.$$ 

Let $M = \Pi_{\gamma_j}^T \Pi_{\gamma_j}$, then $\|M\| = 1, \text{Tr}(M) = \text{Tr}(M^2) = k$. With probability at least $1 - e^{-t}$, it holds that

$$\|\Pi_{\gamma_j} \bar{x}_j\|^2 = \bar{x}_j^T M \bar{x}_j \leq c(\text{Tr}(M) + 2\|M\|t + \sqrt{\text{Tr}(M^2)t + \|M\|^2 t^2})$$

$$= c(k + 2t + \sqrt{kt + t^2}) \leq c(2k + 4t).$$

Let $S = N \sum_{m=1}^{M} \text{tr}((\bar{A}_m^2)^t)$. With probability at least $1 - e^{-t}$, it follows that

$$\|\Pi_{\gamma_j} \bar{x}_j\|^2 = \|\bar{x}_j\|^2 - \|\Pi_{\gamma_j} \bar{x}_j\|^2 \geq \|\bar{x}_j\|^2 - c(2k + 4t) \geq S/c.$$

According to Lemma 12, $A_{-j} = A - \bar{x}_j \bar{x}_j^T \leq A$, which, combined with Lemma 11, leads to $\mu_{k+1}(A_{-j}) < \mu_{k+1}(A) = \mu_1(A_k)$.

Since $\mu_n(A_{-j}) \geq \mu_n(A_k)$, we have

$$\bar{x}_j (\sum_{j} \bar{x}_j \bar{x}_j^T)^{-2} \bar{x}_j \leq \mu_n^{-2}(A_{-j}) \|\bar{x}_j\|^2 \leq \frac{\mu_{k}^{-2}(A_k) S^2}{\mu_n(A_k) S^2} \leq \frac{1}{NM}$$

where (a) is because $\mu_n(A_{-j}) \geq \mu_n(A_k)$ and $\mu_{k+1}(A_{-j}) < \mu_1(A_k)$.

Finally in Lemma 9, we bound the eigenvalues of $A$ to complete the bound on the term related to the $d - k$ smallest eigenvalues of $W$.

**Lemma 9 (Bound on eigenvalues of $A$).** Denote $\mu_n(A)$ as the smallest eigenvalue of $A$, then

$$\mu_n(A) \geq \frac{1}{c} \mu_{k+1}(\bar{W}_M)^{\tau \kappa}(\bar{W}_M),$$

$$\mu_1(A_k) \leq c \mu_{k+1}(\bar{W}_M)^{\tau \kappa}(\bar{W}_M).$$

**Proof.** Recall that

$$A = \bar{X} \bar{X}^T = [\bar{X}_{m1} \bar{X}_{m2}]_{m1m2} = [Z_{m1}^a (\bar{A}_{m1} + \Delta_{\Lambda_{m1}}) V_{m1}^T V_{m2}^T (\bar{A}_{m2} + \Delta_{\Lambda_{m2}}) Z_{m2}^a]^T_{m1m2}.$$ 

For any fixed unit vector $u$, we have

$$u^T A u = u^T \bar{X} \bar{X}^T u$$

where $u^T \bar{X} = \sum_{m=1}^{M} u^T \bar{X}_{m} = \sum_{m=1}^{M} u^T Z_{m}^a (\bar{A}_{m} + \Delta_{\Lambda_{m}}) V_{m}^T$. Therefore, it follows that

$$u^T A u = \left\| \sum_{m=1}^{M} u^T Z_{m}^a (\bar{A}_{m} + \Delta_{\Lambda_{m}}) V_{m}^T \right\|_F^2 \geq \left\| \sum_{m=1}^{M} u^T Z_{m}^a \bar{A}_{m} V_{m}^T \right\|_F^2 + \left\| \sum_{m=1}^{M} u^T Z_{m}^a \Delta_{\Lambda_{m}} V_{m}^T \right\|_F^2 \geq I_1 + I_2 + I_3.$$
We first compute $E$ and the expectation of $I$.

The first term $I_1$ can be further derived as

$$\left\| \sum_{m=1}^{M} u_m^T Z_m^v A_m V_m^T \right\|_F^2 = \left\| \sum_{m=1}^{M} \sum_{i=1}^{d} u_m^T Z_m^v \bar{\lambda}_{mi} V_m^T \right\|_F^2$$

$$= \sum_{m=1}^{M} \sum_{i=1}^{d} \bar{\lambda}_{mi}^2 (u_m^T Z_m^v)^2 + \sum_{m_1 \neq m_2} (\sum_{i=1}^{d} u_m^T Z_m^v \bar{\lambda}_{mi} V_m^T) \left( \sum_{i=1}^{d} v_{m_2i} \bar{\lambda}_{m_2i} Z_{m_2i}^v u_{m_2} \right)$$

where $u_m^T Z_m^v$ is $\|u_m\|\sigma_v$-subgaussian.

First, we compute its expectation, which gives

$$E_z \left[ \left\| \sum_{m=1}^{M} u_m^T Z_m^v A_m V_m^T \right\|_F^2 \right] = \sum_{m=1}^{M} \|u_m\|^2 \sum_{i=1}^{d} \bar{\lambda}_{mi}^2. \quad (78)$$

Recall $\Delta A_m = \alpha A_m (I - \hat{D}_m^\text{tr}) A_m^\frac{1}{2}$. Then the second term $I_2$ can be derived as

$$\left\| \sum_{m=1}^{M} u_m^T Z_m^v \Delta A_m V_m^T \right\|_F^2$$

$$= \sum_{m=1}^{M} \alpha^2 u_m^T Z_m^v \Delta A_m (I - \hat{D}_m^\text{tr}) A_m (I - \hat{D}_m^\text{tr}) A_m Z_m^v u_m$$

$$+ \sum_{m_1 \neq m_2} \alpha^2 u_m^T Z_m^v \Delta A_m (I - \hat{D}_m^\text{tr}) A_m (I - \hat{D}_m^\text{tr}) A_m Z_m^v u_m$$

And the expectation of $I_2$ is

$$E \left[ \left\| \sum_{m=1}^{M} u_m^T Z_m^v \Delta A_m V_m^T \right\|_F^2 \right] = E \left[ \sum_{m=1}^{M} \alpha^2 u_m^T Z_m^v A_m (I - \hat{D}_m^\text{tr}) A_m (I - \hat{D}_m^\text{tr}) A_m Z_m^v u_m \right].$$

We first compute $E[A_m (I - \hat{D}_m^\text{tr}) A_m (I - \hat{D}_m^\text{tr}) A_m]$. By Lemma 16 we have

$$E_{Z_m^v} [A_m (I - \hat{D}_m^\text{tr}) A_m (I - \hat{D}_m^\text{tr}) A_m] = \text{diag} \left[ \frac{1}{N_{\text{tr}}} \lambda_{m_i}^2 (\lambda_{m_i} (E[z^4] - 2) + \sum_{i=1}^{d} \lambda_{mi}) \right].$$

The expectation of $I_2$ over both $Z_m^v$ and $Z_m^v$ is

$$E \left[ \sum_{m=1}^{M} \alpha^2 u_m^T Z_m^v A_m (I - \hat{D}_m^\text{tr}) A_m (I - \hat{D}_m^\text{tr}) A_m Z_m^v u_m \right]$$

$$= \sum_{m=1}^{M} \sum_{i=1}^{d} \alpha^2 \lambda_{m_i}^2 \frac{\sum_{i=1}^{d} \lambda_{mi} + \lambda_{mi} (E[z^4] - 2)}{N_{\text{tr}}} \|u_m\|^2$$

$$= \frac{c \alpha^2}{N_{\text{tr}}} \sum_{m=1}^{M} \|u_m\|^2 \left( \sum_{i=1}^{d} \lambda_{m_i}^2 \right) \leq c \sum_{m=1}^{M} \|u_m\|^2 \sum_{i=1}^{d} \lambda_{m_i}^2.$$

31
Note the expectation of $I_3$ is 0. This is because we can first take expectation w.r.t. $z^{tr}_m$, which yields that $\mathbb{E}_{z^{tr}_m}[\Delta_{\lambda_m}] = 0$, therefore $\mathbb{E}[I_3] = 0$. Combining the expectations of $I_1$, $I_2$, $I_3$ we have

$$
\mathbb{E}[u^\top Au] = \sum_{m=1}^{M} \|u_m\|^2 \left( \frac{c\sigma^2}{N} \left( \sum_{i=1}^{d} \lambda^2_{mi} \right) \right) = \sum_{i=1}^{d} \zeta_i$

where $\zeta_i = \sum_{m=1}^{M} \left( \frac{c\sigma^2}{N} \left( \sum_{i=1}^{d} \lambda^2_{mi} \right) \right) \|u_m\|^2$.

Denote the upper and lower bound of $\zeta_i$ independent of $u$ as $\zeta^u_i$ and $\zeta^l_i$, respectively, i.e., $\zeta^l_i \leq \zeta_i \leq \zeta^u_i$. Then there exists $c$ such that

$$\zeta^l_i = \frac{1}{cM} \sum_{m=1}^{M} \lambda^2_{mi}, \quad \zeta^u_i = \frac{c}{M} \sum_{m=1}^{M} \lambda^2_{mi}.$$ 

From Proposition 2.6.1 in [41], we know that for a fixed vector $u$, any $i$, given the $\sigma^2_x$-subgaussian variable $z_i$, the random variable $u^\top z_i$ is $c\|u\|^2\sigma^2_x$-subgaussian. Thus, $(u^\top z_i)^2$ is subexponential. Based on Lemma 14, for a fixed unit vector $u$, we have

$$|u^\top Au - \mathbb{E}[u^\top Au]| \leq c_4 \sigma^2_x \max \left\{ t \max_i \zeta_i, \sqrt{t \sum_{i=1}^{d} \zeta^2_i} \right\}.$$ 

Let $\mathcal{N}$ be a 1-net on $S^{N-1}$ with $|\mathcal{N}| \leq 9^N$. Based on the $\epsilon$-net argument and union bound over every $u \in \mathcal{N}$, we have with probability at least $1 - 2e^{-t}$ that

$$|u^\top Au - \mathbb{E}[u^\top Au]| \leq c_4 \sigma^2_x \max \left\{ (t + N \ln 9) \max_i \zeta^u_i, \sqrt{(t + N \ln 9) \sum_{i=1}^{d} \zeta^u_i^2} \right\}$$

$$\leq c_5 \sigma^2_x \max_i \zeta^u_i + \frac{1}{c_6 \sigma^2_x} \sum_{i=1}^{d} \zeta^u_i$$

where $c_5N \geq t + N \ln 9$.

Therefore, there is a universal constant $c$ such that with probability at least $1 - 2e^{-c'N}$,

$$\sum_{i=1}^{d} \frac{1}{Mc} \sum_{m=1}^{M} \bar{\lambda}^2_{mi} \leq \frac{1}{c} \sum_{i=1}^{d} \zeta^l_i - cN \max_i \frac{\zeta^u_i}{c} \leq \frac{c}{c'} \sum_{i=1}^{d} \zeta^l_i - \frac{1}{c} \sum_{i=1}^{d} \zeta^u_i - cN \max_i \zeta^u_i$$

$$\leq \mu_N(A) \leq \mu_1(A) \leq \frac{1}{c} \sum_{i=1}^{d} \zeta^u_i + cN \max_i \zeta^u_i \leq \sum_{i=1}^{d} \frac{c}{M} \sum_{m=1}^{M} \bar{\lambda}^2_{mi}.$$ 

Since $\frac{1}{M} \sum_{m=1}^{M} \bar{\lambda}^2_{mi} = \mu_i(\overline{W}_M)$, substituting $A$ with $A$ or $A_k$, then if $r_k(\overline{W}_M) \geq bN$, we have

$$\mu_n(A) \geq \mu_n(A_k) \geq \frac{1}{c} \mu_{k+1}(\overline{W}_M)r_k(\overline{W}_M)$$

$$\mu_1(A_k) \leq c \mu_{k+1}(\overline{W}_M)r_k(\overline{W}_M).$$

This completes the proof. \[\blacksquare\]
D Auxiliary Lemmas

D.1 Basic algebraic properties

Lemma 10. (Lemma 20 in [5]) Suppose \( k < n, A \in \mathbb{R}^{n \times n} \) is an invertible matrix, and \( Z \in \mathbb{R}^{n \times k} \) is such that \( ZZ^\top + A \) is invertible. Then
\[
Z^\top (ZZ^\top + A)^{-2}Z = (I + Z^\top A^{-1}Z)^{-1} Z^\top A^{-2}Z(I + Z^\top A^{-1}Z)^{-1}.
\] (79)

Lemma 11 (Monotonicity of eigenvalues). (Lemma 28 in [5]) If symmetric matrices \( A, B \in \mathbb{R}^{n \times n} \) satisfy \( A \preceq B \), then, for any \( i \in [n] \), we have \( \mu_i(A) \leq \mu_i(B) \).

Lemma 12. Let \( z_i \in \mathbb{R}^n, \lambda_i \geq 0, \forall i, \) if \( B \in \mathbb{R}^{n \times n} \) and \( B = A + \sum_{i=1}^d \lambda_i z_i z_i^\top \), then \( B \succeq A \).

Proof. \( B - A = zz^\top \), therefore for any \( v \neq 0, v^\top (B - A)v = \sum_{i=1}^d \lambda_i (v^\top z_i)^2 \geq 0 \), therefore \( B - A \succeq 0, B \succeq A \).

D.2 Concentration inequalities

Lemma 13 (Von Neumann’s trace inequality [28]). If \( A, B \) are \( n \times n \) matrices with singular values, \( a_1 \geq \cdots \geq a_n, \ b_1 \geq \cdots \geq b_n \) respectively, then
\[
|\text{Tr}(AB)| \leq \sum_{i=1}^n a_i b_i \leq b_1 \sum_{i=1}^n a_i.
\] (80)

D.3 Other intermediate results

Lemma 16 (Expectation of \( \Delta_{\Lambda_m} \Delta_{\tilde{\Lambda}_m}^\top \) and \( \Delta_{\tilde{\Lambda}_m}^\top \Delta_{\Lambda_m} \)).
The expectation of \( \Delta_{\Lambda_m} \Delta_{\tilde{\Lambda}_m}^\top \) and \( \Delta_{\tilde{\Lambda}_m}^\top \Delta_{\Lambda_m} \) over \( z_m^\top \) can be computed by
\[
\mathbb{E}_{z_m^\top} \left[ \Delta_{\Lambda_m} \Delta_{\tilde{\Lambda}_m}^\top \right] = \text{diag} \left[ \frac{\lambda_{mi} \sum_{i=1}^d \lambda_{mi}^2 + \lambda_{mi}^2 (\mathbb{E}[z_i^4] - 2)}{N_{\text{tr}}} \right],
\] (81a)
\[
\mathbb{E}_{z_m^\top} \left[ \Delta_{\tilde{\Lambda}_m}^\top \Delta_{\Lambda_m} \right] = \text{diag} \left[ \frac{\lambda_{mi}^2 \sum_{i=1}^d \lambda_{mi} + \lambda_{mi}(\mathbb{E}[z_i^4] - 2)}{N_{\text{tr}}} \right].
\] (81b)
**Proof.** First recall that $\Delta_{\Lambda_m} \in \mathbb{R}^{d \times d}$, the element in the $i$-th row and $j$-th column in the matrix can be represented by

$$
\Delta_{\Lambda_m,i,j} = (\frac{1}{2} (I - \hat{D}_m^T)\Lambda_m)_{i,j} = \begin{cases} 
\frac{2}{N_{tr}} \lambda_{mi} \left(1 - \frac{1}{N_{tr}} \|z_{mi}^T\|^2\right), & i = j \\
-\frac{1}{N_{tr}} \lambda_{mi} \lambda_{mj} z_{mi}^T z_{mj}^T, & i \neq j.
\end{cases}
$$

When $i = j$, taking expectation of $\Delta_{\Lambda_m,\Delta_{\Lambda_m}^T}$ leads to

$$
E_{z_m}^T [\Delta_{\Lambda_m,\Delta_{\Lambda_m}^T}] = \lambda_{mi}^2 - \frac{2}{N_{tr}} \lambda_{mi}^2 (E[z_{mi}^T z_{mi}]) + \frac{1}{N_{tr}^2} E_{z_m}^T \left[ \sum_i \lambda_{mi}^2 z_{mi}^T z_{mi}^T \right] + 1
\$$

$$
\sum_j \lambda_{mi}^2 z_{mj}^T z_{mj}^T + \frac{1}{N_{tr}} \lambda_{mi} \sum_j \lambda_{mj} z_{mi}^T z_{mj}^T z_{mj}^T z_{mi}^T
\$$

$$
= -\lambda_{mi}^2 + \frac{1}{N_{tr}^2} E_{z_m}^T \left[ \sum_{j \neq i} \lambda_{mi}^2 z_{mi}^T z_{mj}^T z_{mj}^T z_{mi}^T \right] + \frac{1}{N_{tr}} \sum_j \lambda_{mi} \lambda_{mj} z_{mi}^T z_{mj}^T z_{mj}^T z_{mi}^T
\$$

$$
= -\lambda_{mi}^2 + \frac{1}{N_{tr}^2} E_{z_m}^T \left[ \sum_{j \neq i} \lambda_{mi}^2 z_{mi}^T z_{mj}^T z_{mj}^T z_{mi}^T \right] + \sum_j \lambda_{mi} \lambda_{mj} z_{mi}^T z_{mj}^T z_{mj}^T z_{mi}^T
\$$

$$
= \frac{1}{N_{tr}} \sum_j \lambda_{mi} \lambda_{mj} z_{mi}^T z_{mj}^T z_{mj}^T z_{mi}^T
\$$

$$
= \frac{1}{N_{tr}} \sum_i \lambda_{mi}^2 z_{mi}^T z_{mi}^T z_{mi}^T z_{mi}^T
\$$

$$
= \frac{1}{N_{tr}} \sum_i \lambda_{mi}^2 \left(\text{Cov}\left[z^2\right] - 1\right)
\$$

when $i \neq j$, it follows that

$$
E_{z_m}^T \left[ \left(e_{Ii}^T - \frac{1}{N_{tr}} z_{mi}^T Z_{mi}\right) \Lambda_m^2 \left(e_{Ij} - \frac{1}{N_{tr}} Z_{mj}^T Z_{mj}\right) \right] = \frac{1}{N_{tr}^2} E_{z_m}^T \left[ \sum_i \lambda_{mi}^2 z_{mi}^T z_{mi}^T z_{mi}^T z_{mi}^T \right] = 0
\$$

Therefore, we can arrive at (81a). Similarly, we can also get (81b). This completes the proof. \qed