Is Bayesian Model-Agnostic Meta Learning Better than Model-Agnostic Meta Learning, Provably?

Lisha Chen **Tianyi Chen**
Rensselaer Polytechnic Institute

Abstract

Meta learning aims at learning a model that can quickly adapt to unseen tasks. Widely used meta learning methods include model-agnostic meta learning (MAML), implicit MAML, Bayesian MAML. Thanks to its ability of modeling uncertainty, Bayesian MAML often has advantageous empirical performance. However, the theoretical understanding of Bayesian MAML is still limited, especially on questions such as if and when Bayesian MAML has provably better performance than MAML. In this paper, we aim to provide theoretical justifications for Bayesian MAML's advantageous performance by comparing the meta test risks of MAML and Bayesian MAML. In the meta linear regression, under both the distribution agnostic and linear centroid cases, we have established that Bayesian MAML indeed has provably lower meta test risks than MAML. We verify our theoretical results through experiments, the code of which is available at https://github.com/lisha-chen/Bayesian-MAML-vs-MAML.

1 INTRODUCTION

Meta learning, also referred to as "learning to learn", usually learns a model that can quickly adapt to new tasks (Thrun and Pratt, 1998; Hospedales et al., 2020; Vilalta and Drissi, 2002; Vanschoren, 2018; Bengio et al., 1991; Schmidhuber, 1995; Hochreiter et al., 2001). The key idea of meta-learning is to learn a "prior" model from multiple existing tasks with a hope that the learned model is able to quickly adapt to unseen tasks. Meta learning has been used in various machine learning scenarios including few-shot learning (Snell et al., 2017; Obamuyide and Vlachos, 2019), continual learning (Harrison et al., 2020; Javed and White, 2019), and personalized learning (Madotto et al., 2019). In addition, meta learning has also been successfully implemented in different data limited applications including language and vision tasks (Achille et al., 2019; Li et al., 2018; Hsu et al., 2018; Liu et al., 2019; Zintgraf et al., 2019; Wang et al., 2019; Obamuyide and Vlachos, 2019). One of the popular meta-learning approaches is the model agnostic meta-learning (MAML) method (Finn et al., 2017; Vuorio et al., 2019; Yin et al., 2020; Obamuyide and Vlachos, 2019), which learns an initial model that can adapt to new tasks using one step gradient update. Despite its success, MAML still suffers from overfitting when it is trained with few data, which motivates Bayesian MAML (BaMAML) (Grant et al., 2018; Ravi and Beatson, 2019; Yoon et al., 2018). Instead of point estimation of task specific model parameters, as in MAML and its variants (Rajeswaran et al., 2019), BaMAML obtains a posterior distribution of task specific parameters as a function of the task data and the initial model parameters, as illustrated in Figure 1. For example, in 5-way 1-shot classification on TieredImageNet, BaMAML has 35.2% performance gain over MAML in terms of accuracy (Nguyen et al., 2020). In spite of BaMAML's impressive empirical performance, its theoretical understanding is still very limited, no need to mention a sound justification for its performance gain over MAML. In this context, this paper aims to answer the following question:

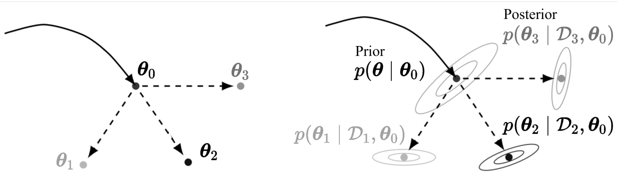

θ_0
θ_3
θ_1
θ_2
Posterior
$p(\theta_3 \mid \mathcal{D}_3, \theta_0)$
Prior
$p(\theta \mid \theta_0)$
$p(\theta_1 \mid \mathcal{D}_1, \theta_0)$
$p(\theta_2 \mid \mathcal{D}_2, \theta_0)$

Figure 1: MAML (left) and Bayesian MAML (right).

Is Bayesian Model-Agnostic Meta Learning Better than
Model-Agnostic Meta Learning, Provably?

In an attempt to provide an affirmative answer to this

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

question, our paper analyzes the meta-test risks of one-step MAML and BaMAML to make a fair comparison between them. In a high level, our theoretical results suggest that compared to one-step MAML, BaMAML 1) harnesses flexibility in the trade-off between prior and likelihood based on their quality to improve model adaptation capacity; and, 2) leverages the posterior distribution instead of a point estimation in inference, which allows model averaging to reduce variance.

1.1 Related Works

Early works of meta learning build black-box recurrent models that can make predictions based on few examples from new tasks (Schmidhuber, 1993; Hochreiter et al., 2001; Andrychowicz et al., 2016; Chen et al., 2017), or learn shared feature representation among multiple tasks (Snell et al., 2017; Vinyals et al., 2016). More recently, some methods have been developed to find the initialization of model parameters that can quickly adapt to new tasks with few optimization steps (Finn et al., 2017; Nichol et al., 2018; Rothfuss et al., 2018). The empirical success of meta learning has also stimulated recent interests on building the theoretical foundation of these methods. To put our work in context, we review prior art that is grouped into the following categories.

Theory of meta-learning. One line of theoretical works study the convergence of meta-learning algorithms under different settings. These works include analysis of the regret bound for an online meta-learning algorithm (Finn et al., 2019), the convergence and sample complexity of gradient based MAML (Fallah et al., 2020), sufficient conditions for its convergence to the exact solution for an approximate bilevel optimization method (Franceschi et al., 2018), sample complexity for a bilevel formulation for meta-learning, named implicit MAML (iMAML) (Rajeswaran et al., 2019), and the global convergence guarantee of MAML with overparameterized deep neural nets (DNNs) (Wang et al., 2020a,b). There are also works that study the convergence of general compositional (Chen et al., 2021c) or bilevel (Chen et al., 2021b; Yang et al., 2021; Liu et al., 2021) optimzation algorithms which can be applied to analyze the convergence of one-step or bilevel MAML.

Another line of works analyze the generalization error bound of meta learning methods under different settings based on their optimization trajectory. For instance, meta-learning in the linear centroid model for ridge regression (Denevi et al., 2018), MAML with sufficiently wide DNNs (Wang et al., 2020a), meta-learning in online convex optimization (Balcan et al., 2019), and MAML for strongly convex objective functions on recurring and unseen tasks (Fallah et al., 2021). Recently, information theoretical generalization error bounds of meta learning are also proposed by Jose and Simeone (2021); Rezazadeh et al. (2021); Jose et al. (2021); Chen et al. (2021a), which bounds the meta learning generalization error in terms of mutual information between the input meta-training data and the output of the meta-learning algorithms rather than gradient norm of the algorithms during optimization.

Our work is also inspired by several pioneering works that analyze the optimization, modeling and statistical errors of meta-learning methods. Gao et al. (2020) study the modeling and optimization error trade-off in MAML and compare the trade-off with that of empirical risk minimization (ERM). Collins et al. (2020) further analyze the effect of different factors on the optimal population risk, such as task hardness in task landscape. Bai et al. (2021) study how the dataset split between the training and validation affects the performance of iMAML under a noiseless realizable centroid model. But none of them tackle the meta-test risk of BaMAML. Furthermore, from the technical aspect, compared to Bai et al. (2021), our analysis does not require strong assumption on noiseless realizable model; compared to Gao and Sener (2020), our analysis provides a sharper characterization of statistical error bound in the high-dimensional asymptotic case.

Bayesian model agnostic meta-learning. From a hierarchical probabilistic modeling perspective, learning the initialization in MAML is tantamount to learning the prior distribution of model parameters shared across different tasks (Grant et al., 2018), which leads to a hierarchical Bayes formulation that we call BaMAML thereafter. Empirically, they have better performance in few-shot meta learning settings and tend to reduce over-fitting in the data-limited regimes. Several variants of BaMAML have been proposed based on different Bayesian inference methods (Grant et al., 2018; Finn et al., 2018; Yoon et al., 2018; Gordon et al., 2018; Nguyen et al., 2020). Despite the superior empirical performance of BaMAML methods compared to non-Bayesian ones, very few works study their theory. A related line of works extend the PAC-Bayes framework to meta learning (Amit and Meir, 2018; Rothfuss et al., 2021; Ding et al., 2021; Farid and Majumdar, 2021), to provide a PAC-Bayes meta-test error bound. Different from the PAC-Bayes framework that bounds the Gibbs risk, we bound the Bayes risk (Sheth and Khardon, 2017). While these works provide the meta-test error bound for BaMAML, exactly when BaMAML is provably better than non-Bayesian methods are not fully understood. Different from these works, we explicitly compare MAML and BaMAML in terms of meta-test error, consisting of the optimal population risks and statistical errors.

1.2 Our Contributions

The goal of this paper is to provide justification on the observed empirical performance gain of BaMAML over MAML. Our contributions are summarized below.

C1) Under the meta-linear regression setting, we decompose the meta-test risk into population risk and statistical error terms, which capture the bias and variance of the estimated parameter, respectively. We prove that BaMAML with proper choice of hyperparameters has smaller optimal population risk and dominating constant in statistical error than MAML, therefore smaller meta-test risk.

C2) With additional linear centroid model assumption for task data distribution, we prove that BaMAML has strictly smaller dominating constant in statistical error than MAML in the high dimensional asymptotic case.

C3) We conduct simulations on meta linear regression to verify our theory. And we also perform experiments beyond linear case, where similar conclusions can be drawn.

Our theoretical analysis justifies BaMAML for reducing the optimal population risk and statistical errors, thus the meta-test risk. And to our best knowledge, we are the first to make a comparison between MAML and BaMAML, which is complementary to existing works (Gao and Sener, 2020; Collins et al., 2020) that compare MAML against empirical risk minimization.

2 PROBLEM DEFINITION AND SOLUTIONS

In this section, we first introduce the general meta-learning setting and the formulations of two meta learning methods, MAML and BaMAML. Then we focus on meta-linear regression, where solutions to the empirical and population level risks are obtained in closed form.

2.1 Problem Setup

In our meta-learning setting, assume task τ are drawn from a task distribution, i.e. $\tau \sim \mathcal{T}$, with input features $\mathbf{x}_\tau \in \mathcal{X}_\tau \subset \mathbb{R}^d$ and target labels $y_\tau \in \mathcal{Y}_\tau \subset \mathbb{R}$. For each task τ, we observe N samples drawn i.i.d. from $\mathcal{P}_\tau$ in the dataset $\mathcal{D}_\tau = \{(\mathbf{x}_{\tau,n}, y_{\tau,n})\}_{n=1}^N$, and $\mathcal{D}_\tau$ is divided into the train and validation datasets, denoted as $\mathcal{D}_\tau^{\text{trn}}$ and $\mathcal{D}_\tau^{\text{val}}$, respectively. Here $|\mathcal{D}_\tau^{\text{trn}}| = N_1$ and $|\mathcal{D}_\tau^{\text{val}}| = N_2$ with $N = N_1 + N_2$. Given the data $\mathcal{D}_\tau$, we use the empirical loss $\ell_\tau(h_\tau, \mathcal{D}_\tau)$ of per-task hypothesis $h_\tau \in \mathcal{H}_\tau$ as a measure of the performance. And the goal for initialization based meta learning methods, such as MAML (Finn et al., 2018) and BaMAML (Yoon et al., 2018), is to learn an initial parameter $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$, which, with an adaptation method and the training data, can produce a per-task hypothesis h_τ that performs well on the validation data for task τ. Formally, for a meta-learning method, $\mathcal{A} : \boldsymbol{\Theta}_0 \times (\mathcal{X}_\tau \times \mathcal{Y}_\tau)^{N_1} \to \mathcal{H}_\tau$, represents the adaptation method or base-learner. Given T tasks with corresponding data, our meta-learning objective is to find $\boldsymbol{\theta}_0$ that minimizes the empirical loss, given by

$$\mathcal{L}^{\mathcal{A}}(\boldsymbol{\theta}_0, \mathcal{D}) := \frac{1}{T}\sum_{\tau=1}^{T} \ell_\tau(\mathcal{A}(\boldsymbol{\theta}_0, \mathcal{D}_\tau^{\text{trn}}), \mathcal{D}_\tau^{\text{val}}). \qquad (1)$$

And the corresponding meta-test risk is defined as the expectation of the per-task loss ℓ_τ over the task and data distribution, given by

$$\mathcal{R}^{\mathcal{A}}(\boldsymbol{\theta}_0) := \mathbb{E}_\tau\big[\mathbb{E}_{\mathcal{D}_\tau}\big[\ell_\tau(\mathcal{A}(\boldsymbol{\theta}_0, \mathcal{D}_\tau^{\text{trn}}), \mathcal{D}_\tau^{\text{val}})\big]\big]. \qquad (2)$$

Denote $\mathbf{X}_\tau^{\text{all}} := [\mathbf{x}_{\tau,1}, \ldots, \mathbf{x}_{\tau,N}]^\top \in \mathbb{R}^{N\times d}$, $\mathbf{y}_\tau^{\text{all}} := [y_{\tau,1}, \ldots, y_{\tau,N}]^\top \in \mathbb{R}^N$ for ease of discussion, where "all" can also be "trn" for training and "val" for validation with N_1 and N_2 data points, respectively. Throughout the discussion of this paper, we adopt a probabilistic perspective (Grant et al., 2018; Finn et al., 2018), with ℓ_τ defined as the negative log likelihood, given by

$$\begin{aligned}\ell_\tau(\mathcal{A}(\boldsymbol{\theta}_0, \mathcal{D}_\tau^{\text{trn}}), \mathcal{D}_\tau^{\text{val}}) &= -\frac{1}{N_2}\log p(\mathbf{y}_\tau^{\text{val}} | \mathbf{X}_\tau^{\text{val}}, \boldsymbol{\theta}_0, \mathcal{D}_\tau^{\text{trn}}) \\ &= -\frac{1}{N_2}\log\int p(\mathbf{y}_\tau^{\text{val}} \mid \mathbf{X}_\tau^{\text{val}}, \boldsymbol{\theta}_\tau) p_{\mathcal{A}}(\boldsymbol{\theta}_\tau \mid \boldsymbol{\theta}_0, \mathcal{D}_\tau^{\text{trn}}) d\boldsymbol{\theta}_\tau \qquad (3)\end{aligned}$$

where $p_{\mathcal{A}}(\boldsymbol{\theta}_\tau \mid \boldsymbol{\theta}_0, \mathcal{D}_\tau^{\text{trn}})$ is the posterior distribution induced by $\mathcal{A}$. And the likelihood $p(\mathbf{y}_\tau^{\text{val}} \mid \mathbf{X}_\tau^{\text{val}}, \boldsymbol{\theta}_\tau, \mathcal{D}_\tau^{\text{trn}}) = \prod_{n=1}^{N_2} p(y_{\tau,n} \mid \mathbf{x}_{\tau,n}, \boldsymbol{\theta}_\tau)$. Note that, for a point estimate method $\mathcal{A}$, such as MAML, the posterior distribution $p_{\mathcal{A}}(\boldsymbol{\theta}_\tau \mid \boldsymbol{\theta}_0, \mathcal{D}_\tau^{\text{trn}})$ reduces to a Dirac delta function $\delta(\boldsymbol{\theta}_\tau - \hat{\boldsymbol{\theta}}_\tau^{\mathcal{A}})$. And $\mathcal{A}$ specifies a mapping from the initial parameter $\boldsymbol{\theta}_0$ to the task-specific parameter $\hat{\boldsymbol{\theta}}_\tau^{\mathcal{A}}(\boldsymbol{\theta}_0, \mathcal{D}_\tau^{\text{trn}})$.

In the meta training stage, we obtain $\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}$ by minimizing (1) under each meta learning method $\mathcal{A}$. And in the meta testing stage, we evaluate the test error of $\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}$ on (2) for different methods.

Methods. We proceed to introduce the general formulations of MAML and BaMAML. Considering MAML with one step gradient update as the baseline method for meta-learning (Finn et al., 2017), the task-specific parameter $\hat{\boldsymbol{\theta}}_\tau^{\text{ma}}(\boldsymbol{\theta}_0)$ is obtained from the initial parameter $\boldsymbol{\theta}_0$ by taking one step gradient descent with step size α of the per-task loss function ℓ_τ. Combined with the empirical loss defined in (1), we have the empirical loss of MAML is given by

$$\mathcal{L}^{\text{ma}}(\boldsymbol{\theta}_0,\mathcal{D}) = \frac{1}{T}\sum_{\tau=1}^{T} \ell_\tau(\hat{\boldsymbol{\theta}}_\tau^{\text{ma}}(\boldsymbol{\theta}_0,\mathcal{D}_\tau^{\text{trn}}),\mathcal{D}_\tau^{\text{val}}) \qquad (4)$$
$$\text{s.t. } \hat{\boldsymbol{\theta}}_\tau^{\text{ma}}(\boldsymbol{\theta}_0,\mathcal{D}_\tau^{\text{trn}}) = \boldsymbol{\theta}_0 - \frac{\alpha}{2}\nabla_{\boldsymbol{\theta}_0}\ell_\tau(\boldsymbol{\theta}_0,\mathcal{D}_\tau^{\text{trn}}).$$

BaMAML obtains an approximation of the posterior distribution $p(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\text{trn}},\boldsymbol{\theta}_0)$ instead of a point estimate $\hat{\boldsymbol{\theta}}_\tau^{\mathcal{A}}(\boldsymbol{\theta}_0,\mathcal{D}_\tau^{\text{trn}})$. In general, the true posterior distribution can be difficult to compute exactly. Alternatively, the approximate distribution $\hat{p}(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\text{trn}},\boldsymbol{\theta}_0)$ can be obtained via variational inference (Nguyen et al., 2020), Markov chain Monte-Carlo sampling or Laplace approximation (Grant et al., 2018). Here we adopt the variational inference formulation, by minimizing the divergence between the approximate and the true posterior distribution. Define $\mathrm{D}_{\text{KL}}(\cdot\|\cdot)$ as the KL-divergence between two distributions, we have

the empirical loss of BaMAML is given by

$$\mathcal{L}^{\text{ba}}(\boldsymbol{\theta}_0,\mathcal{D}) = \frac{1}{T}\sum_{\tau=1}^{T} \ell_\tau(\hat{p}(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\text{trn}},\boldsymbol{\theta}_0),\mathcal{D}_\tau^{\text{val}}) \qquad (5)$$
$$\text{s.t. } \hat{p}(\boldsymbol{\theta}_\tau|\mathcal{D}_\tau^{\text{trn}},\boldsymbol{\theta}_0) = \arg\min_{q(\boldsymbol{\theta}_\tau)\in\mathcal{Q}} \mathrm{D}_{\text{KL}}\big(q(\boldsymbol{\theta}_\tau)\|p(\boldsymbol{\theta}_\tau|\mathcal{D}_\tau^{\text{trn}},\boldsymbol{\theta}_0)\big)$$

It is worth mentioning that BaMAML formulation in this paper contains iMAML, or iMAML (Rajeswaran et al., 2019) as a special case. Therefore, results obtained for BaMAML naturally implies the results for iMAML with small difference. We point out this reduction in the next remark, and provide detailed discussion in the appendix.

Remark 1 (Reduction to iMAML) *When $\mathcal{Q}$ is chosen to be the set of Dirac Delta functions and the KL-divergence in* (5) *is replaced by the cross entropy, then* (5) *reduces to*

$$\hat{p}(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\text{trn}},\boldsymbol{\theta}_0) = \delta(\boldsymbol{\theta}_\tau - \hat{\boldsymbol{\theta}}_\tau^{\text{map}}), \qquad (6)$$
$$\text{with } \hat{\boldsymbol{\theta}}_\tau^{\text{map}} = \arg\max_{\boldsymbol{\theta}_\tau} p(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\text{trn}},\boldsymbol{\theta}_0).$$

2.2 Meta Linear Regression

Data model. Under the meta linear regression setting, with the feature $\mathbf{x}_\tau \in \mathbb{R}^d$, the target $y_\tau \in \mathbb{R}$, and the ground truth parameter of task τ, $\boldsymbol{\theta}_\tau^{\text{gt}} \in \mathbb{R}^d$, we assume the data generation model for task τ is

$$y_\tau = \boldsymbol{\theta}_\tau^{\text{gt}\top}\mathbf{x}_\tau + \epsilon_\tau, \text{with } \epsilon_\tau \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\sigma_\tau^2\right), \mathbf{Q}_\tau := \mathbb{E}[\mathbf{x}_\tau\mathbf{x}_\tau^\top]. \qquad (7)$$

Given the estimate of $\boldsymbol{\theta}_\tau^{\text{gt}}$ denoted as $\hat{\boldsymbol{\theta}}_\tau^{\mathcal{A}}$, then the conditional probability $p(y_\tau \mid \mathbf{x}_\tau, \hat{\boldsymbol{\theta}}_\tau^{\mathcal{A}}) = \mathcal{N}(\hat{\boldsymbol{\theta}}_\tau^{\mathcal{A}\top}\mathbf{x}_\tau, \sigma_\tau^2)$. Thus, ignoring the constant, the negative log likelihood in (3), $-\log p(\mathbf{y}_\tau^{\text{val}} \mid \mathbf{X}_\tau^{\text{val}}, \boldsymbol{\theta}_\tau)$, becomes the squared error $\|\mathbf{y}_\tau^{\text{val}} - \mathbf{X}_\tau^{\text{val}}\boldsymbol{\theta}_\tau\|^2$. Note that σ_τ depends on task τ generally, but does not pose challenges to analysis, therefore we assume $\sigma_\tau = 1$ in this paper for simplicity.

By plugging $\hat{\boldsymbol{\theta}}_\tau^{\mathcal{A}}$ into (2), and with the squared error as the meta-linear regression loss, the empirical loss, meta-test risk along with their optimal solutions can be computed analytically with closed-form, whose derivations are deferred to the appendix. We summarize the results for different methods in Proposition 1, where the optimal solutions for MAML derived in previous work (Gao and Sener, 2020) are also included.

Proposition 1 (Empirical and population level solutions) *Under data model* (7)*, the meta-test risk of method $\mathcal{A}$ can be computed by*

$$\mathcal{R}^{\mathcal{A}}(\boldsymbol{\theta}_0) = \mathbb{E}_\tau\left[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau^{\text{gt}}\|^2_{\mathbf{W}_\tau^{\mathcal{A}}}\right] + 1. \qquad (8)$$

The optimal solutions to the meta-test risk and empirical loss are given below respectively

$$\boldsymbol{\theta}_0^{\mathcal{A}} := \arg\min_{\boldsymbol{\theta}_0} \mathcal{R}^{\mathcal{A}}(\boldsymbol{\theta}_0) = \mathbb{E}_\tau\left[\mathbf{W}_\tau^{\mathcal{A}}\right]^{-1}\mathbb{E}_\tau\left[\mathbf{W}_\tau^{\mathcal{A}}\boldsymbol{\theta}_\tau^{gt}\right] \quad (9a)$$
$$\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} := \arg\min_{\boldsymbol{\theta}_0} \mathcal{L}^{\mathcal{A}}(\boldsymbol{\theta}_0,\mathcal{D})$$
$$= \Big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_\tau^{\mathcal{A}}\Big)^{-1}\Big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_\tau^{\mathcal{A}}\boldsymbol{\theta}_\tau^{gt}\Big) + \Delta_T^{\mathcal{A}}. \qquad (9b)$$

where the error term $\Delta_T^{\mathcal{A}}$ is a polynomial function of T, N, d caused by the noise ϵ, and specified in the appendix. And $\hat{\mathbf{Q}}_{\tau,N} := \frac{1}{N}\mathbf{X}_\tau^{\text{all}\top}\mathbf{X}_\tau^{\text{all}}$, $s = N_1/N$. The weight matrices of different methods, $\mathbf{W}_\tau^{\mathcal{A}}$ and $\hat{\mathbf{W}}_\tau^{\mathcal{A}}$, are given in Table 1.

Note that, in the meta linear regression case in Proposition 1, BaMAML further assumes the prior distribution $\boldsymbol{\theta}_\tau \sim \mathcal{N}(\boldsymbol{\theta}_0, 1/\gamma_b)$ with $\gamma_b = \gamma N_1$, resulting in the weight matrices $\mathbf{W}_\tau^{\text{ba}}, \hat{\mathbf{W}}_\tau^{\text{ba}}$ in Table 1 depending on γ and s. The posterior follows a Gaussian distribution, $p(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\text{trn}},\boldsymbol{\theta}_0) = \mathcal{N}(\mu_{\boldsymbol{\theta}_\tau}, \Sigma_{\boldsymbol{\theta}_\tau})$, where the parameters $\Sigma_{\boldsymbol{\theta}_\tau}$ and $\mu_{\boldsymbol{\theta}_\tau}$ are given by

$$\Sigma_{\boldsymbol{\theta}_\tau} = (N_1\hat{\mathbf{Q}}_{\tau,N_1} + \gamma_b\mathbf{I})^{-1}, \qquad (10a)$$
$$\mu_{\boldsymbol{\theta}_\tau} = \Sigma_{\boldsymbol{\theta}_\tau}(\mathbf{X}_\tau^{\text{trn}\top}\mathbf{y}_\tau^{\text{trn}} + \gamma_b\boldsymbol{\theta}_0). \qquad (10b)$$

If $p(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\text{trn}},\boldsymbol{\theta}_0) \in \mathcal{Q}$, then $\hat{p}(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\text{trn}},\boldsymbol{\theta}_0) = p(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\text{trn}},\boldsymbol{\theta}_0)$, which holds for the meta linear regression case analyzed in this paper, with $\mathcal{Q}$ specified as the set of Gaussian distributions.

Next, we will use the closed-form solutions of different methods in Proposition 1 to compute their generalization errors in Section 3.

Table 1: Weight matrices for the closed form solutions of method $\mathcal{A}$.

Method	Weight matrices
MAML (Gao and Sener, 2020)	$\mathbf{W}_\tau^{\mathrm{ma}} = (\mathbf{I} - \alpha\mathbf{Q}_\tau)\mathbf{Q}_\tau(\mathbf{I} - \alpha\mathbf{Q}_\tau)$ $\hat{\mathbf{W}}_\tau^{\mathrm{ma}} = (\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N_1})\hat{\mathbf{Q}}_{\tau,N_2}(\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N_1})$
BaMAML	$\mathbf{W}_\tau^{\mathrm{ba}} = ((s\gamma)^{-1}\mathbf{Q}_\tau + \mathbf{I})^{-1}\mathbf{Q}_\tau(\gamma^{-1}\mathbf{Q}_\tau + \mathbf{I})^{-1}$ $\hat{\mathbf{W}}_\tau^{\mathrm{ba}} = ((s\gamma)^{-1}\hat{\mathbf{Q}}_{\tau,N} + \mathbf{I})^{-1}\hat{\mathbf{Q}}_{\tau,N_2}(\gamma^{-1}\hat{\mathbf{Q}}_{\tau,N_1} + \mathbf{I})^{-1}$

3 META-TEST RISK ANALYSIS

In this section, we will compare the meta-test risk of MAML and BaMAML. By the definition of the meta-test risk $\mathcal{R}^{\mathcal{A}}$ in (2), it can be decomposed into the optimal population risk and statistical errors, as summarized in Proposition 2.

Proposition 2 (Meta-test risk decomposition) *In meta-linear regression, the meta-test risk for method $\mathcal{A}$ can be decomposed into optimal population risks and statistical errors, given by*

$$\mathcal{R}^{\mathcal{A}}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}) = \underbrace{\mathcal{R}^{\mathcal{A}}(\boldsymbol{\theta}_0^{\mathcal{A}})}_{\substack{\textit{optimal}\\ \textit{population risk}}} + \underbrace{\|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}}\|^2_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]}}_{\textit{statistical error } \mathcal{E}^2_{\mathcal{A}}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}})} . \quad (11)$$

Invoking the definition of $\boldsymbol{\theta}_0^{\mathcal{A}}$ in (9a) as the optimal solution for $\min_{\boldsymbol{\theta}_0} \mathcal{R}^{\mathcal{A}}(\boldsymbol{\theta}_0)$, the optimal population risk $\mathcal{R}^{\mathcal{A}}(\boldsymbol{\theta}_0^{\mathcal{A}})$ is defined as the minimum meta-test risk, which captures the error resulting from limited model adaptation capacity. On the other hand, the statistical error captures error resulting from using finite samples instead of population statistics. We will next show that both errors are smaller under BaMAML than those under MAML in Sections 3.1 and 3.2.

3.1 Optimal Population Risk Analysis

We first analyze and compare the optimal population risk of different methods. Before proceeding to the theoretical results, we make the following basic assumptions.

Assumption 1 (Bounded eigenvalues) *For any τ, $0 < \underline{\lambda} \leq \lambda(\mathbf{Q}_\tau) \leq \bar{\lambda}$, where $\lambda(\mathbf{Q}_\tau)$ represents the eigenvalues of $\mathbf{Q}_\tau$.*

Assumption 2 (Sub-gaussian task parameter and bounded features) *The ground truth parameter $\boldsymbol{\theta}_\tau^{\mathrm{gt}}$ is independent of $\mathbf{X}_\tau$ and satisfies that the individual entries $\{\boldsymbol{\theta}_{\tau,i}^{\mathrm{gt}} - \boldsymbol{\theta}_{0,i}^{\mathcal{A}}\}_{i\in[d],\tau\in[T]}$ are independent and $\mathcal{O}(R/\sqrt{d})$-sub-gaussian. In addition, $\|\mathbb{E}[\boldsymbol{\theta}_\tau^{\mathrm{gt}} - \boldsymbol{\theta}_0^{\mathcal{A}}]\| \leq M$. The inputs $\|\mathbf{x}_{\tau,i}\| \leq K$. R, K are constants.*

Note that, these assumptions can be easily satisfied in data generation model (7) by controlling the hyperparameters. And they are also standard in analyzing the optimal population risks for meta-linear regression (Gao and Sener, 2020; Collins et al., 2020).

Next we will show in Theorem 2 that one can always find a range of the regularizer weight γ such that BaMAML has smaller optimal population risk than MAML.

Theorem 2 (Optimal population risks) *In the meta-linear regression with data model (7), recall that $\boldsymbol{\theta}_0^{\mathrm{ma}}$ and $\boldsymbol{\theta}_0^{\mathrm{ba}}$ are the minimizers of $\mathcal{R}^{\mathrm{ma}}(\boldsymbol{\theta};\alpha)$ and $\mathcal{R}^{\mathrm{ba}}(\boldsymbol{\theta};\gamma)$, respectively. Define $r^{\mathrm{ma}} := \min_\alpha \mathcal{R}(\boldsymbol{\theta}_0^{\mathrm{ma}};\alpha) - 1 > 0$, $C_{\boldsymbol{\theta}} := \max\{\big((M + \|\boldsymbol{\theta}_0^{\mathrm{im}}\|)^2 + R^2\big)^{\frac{1}{2}}, \big((M+\|\boldsymbol{\theta}_0^{\mathrm{ma}}\|)^2+R^2\big)^{\frac{1}{2}}\}$. Under Assumptions 1-2, when γ satisfies*

$$0 < \gamma < \big((r^{\mathrm{ma}})^{-\frac{1}{2}} C_{\boldsymbol{\theta}} \bar{\lambda}^{\frac{1}{2}} - 1\big)^{-1} \underline{\lambda} \quad (12)$$

BaMAML has smaller optimal population risk, i.e.

$$\mathcal{R}^{\mathrm{ba}}(\boldsymbol{\theta}_0^{\mathrm{ba}};\gamma) < \min_\alpha \ \mathcal{R}^{\mathrm{ma}}(\boldsymbol{\theta}_0^{\mathrm{ma}};\alpha). \quad (13)$$

Theorem 2 states that regardless of the choice of α, we can always find $\gamma > 0$ such that the BaMAML method has smaller meta-test risk than the MAML method.

Note that, the choice of γ represents trade-off between adaptation speed and optimal population risk, because $\boldsymbol{\theta}_\tau^{\mathrm{ba}}(\boldsymbol{\theta}_0)$ is a weighted average of the prior $\boldsymbol{\theta}_0$ and the ground truth paramter $\boldsymbol{\theta}_\tau^{\mathrm{gt}}$. The larger γ, the higher weight for the prior $\boldsymbol{\theta}_0$, then the closer the initial parameter $\boldsymbol{\theta}_0$ is to the optimal $\boldsymbol{\theta}_\tau^{\mathrm{ba}}(\boldsymbol{\theta}_0)$, and the faster the adaptation speed. On the other hand, the larger γ, the larger the optimal population risk is. This inspires us to select model hyperparameter based on our practical needs for the specific problem. Combined with the optimal population risk of ERM (or modeling error in the paper) established in Gao and Sener (2020), our Theorem 2 also implies that BaMAML has lower optimal population risk than ERM.

3.2 Statistical Error Analysis

We next study and compare the statistical errors of different methods defined in (11). We first bound the statistical errors of MAML and BaMAML methods.

Theorem 3 (Statistical error of MAML) *Suppose Assumptions 1-2 hold.* $T = \Omega(d)$, $M = o(R/\sqrt{T})$. *Denote* $\|\cdot\|_{\mathrm{op}}$ *as the operator norm. Define function*

$$C_0^{\mathcal{A}} := [\inf_{\tau} \lambda_{\min}(\mathbf{W}_{\tau}^{\mathcal{A}})]^{-1} [\sup_{\tau} \lambda_{\max}(\mathbf{W}_{\tau}^{\mathcal{A}})]^2 \tag{14}$$

and define ϱ *as a higher order term given by*

$$\begin{aligned} \varrho =& \frac{1}{T}\big(1 + \frac{d}{N}\big)\big(\widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\big) \\ &+ \Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big)M^2 + \frac{1}{T}\widetilde{\mathcal{O}}(\frac{d}{N}) \end{aligned} \tag{15}$$

where $\widetilde{\mathcal{O}}(\cdot)$ *hides* $\log(TNd)$ *factor. With probability at least* $1 - Td^{-10}$, *we have*

$$\mathcal{E}_{\mathrm{ma}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{ma}}) \leq \frac{R^2}{T} 2C_0^{\mathrm{ma}} + \frac{d}{TN} 2C_1^{\mathrm{ma}} + \varrho \tag{16}$$

where C_0^{ma} *is given by* (14), *and*

$$\begin{aligned} C_0^{\mathrm{ma}} =& (1 - \alpha\underline{\lambda})^4 (1 - \alpha\bar{\lambda})^{-2} \underline{\lambda}^{-1} \bar{\lambda}^2 \\ C_1^{\mathrm{ma}} =& s^{-1} + (1 - s)^{-1}(1 - \alpha\bar{\lambda})^{-2} \alpha^2 \bar{\lambda}^3 \underline{\lambda}^{-1}. \end{aligned} \tag{17a}$$

Note that $M = o(R/\sqrt{T})$ can be achieved when different tasks have similar $\mathbf{Q}_\tau$, for example, when input feature normalization is performed. Analogous to Theorem 3, we bound the BaMAML statistical error next.

Theorem 4 (Statistical error of BaMAML) *Suppose Assumptions 1-2 hold. With probability at least* $1 - Td^{-10}$, *we have*

$$\mathcal{E}_{\mathrm{ba}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{ba}}) \leq \frac{R^2}{T} 2C_0^{\mathrm{ba}} + \frac{d}{TN} 2C_1^{\mathrm{ba}} + \varrho \tag{18}$$

where ϱ *is given by* (15), C_0^{ba} *is given by* (14), *and*

$$\begin{aligned} C_0^{\mathrm{ba}} =& (1 + \gamma^{-1}\underline{\lambda})^{-4} (1 + (\gamma s)^{-1}\bar{\lambda})^2 \underline{\lambda}^{-1} \bar{\lambda}^2 \\ C_1^{\mathrm{ba}} =& 1. \end{aligned} \tag{19a}$$

Theorems 3 and 4 show that the statistical errors of MAML and BaMAML have similar decreasing rates, that is, $\mathcal{O}(T^{-1})$ and $\mathcal{O}(N^{-1})$. The difference lies in their coefficients. For the dominating constants C_0^{ma} in (17a) and C_0^{ba} in (19a), given any α, choose

$$\gamma < \min\{\bar{\lambda}, \frac{1}{2}\bar{\lambda}^{-1}\underline{\lambda}^2 s(1 - \alpha\bar{\lambda})^2(1 - \alpha\underline{\lambda})^{-1}\} \tag{20}$$

then $C_0^{\mathrm{ma}} > C_0^{\mathrm{ba}}$. In terms of the dependence on N, given any α, since $C_1^{\mathrm{ma}} > s^{-1} > 1$, thus $C_1^{\mathrm{ma}} > C_1^{\mathrm{ba}}$, i.e. MAML has larger coefficients than BaMAML. Therefore the statistical error of BaMAML is lower when N is small, which is typical in few-shot learning. Nevertheless, Theorems 3 and 4 only give the worst-case upper bounds of the statistical errors of two methods, which can be inaccurate in some cases. To precisely characterize the statistical errors of BaMAML and MAML, we will provide sharper analysis next based on an additional assumption.

3.3 Sharp Statistical Error Analysis

To precisely quantify the dominating constants in the statistical error, we further make assumptions on the task and data distributions.

Assumption 3 (Linear centroid model) *1) The inputs are standard Gaussian:* $\mathbf{x}_{\tau,i} \overset{\mathrm{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. *Then* $\mathbf{Q}_\tau = \mathbf{I}_d$, *therefore* $\mathbf{W}_\tau^{\mathcal{A}} = w_{\mathcal{A}}\mathbf{I}_d$. *This implies that for different methods, the optimal initial parameters are the same, that is,* $\boldsymbol{\theta}_0^* = \mathbb{E}_\tau[\boldsymbol{\theta}_\tau^{\mathrm{gt}}]$. *2) The ground truth parameter* $\boldsymbol{\theta}_\tau^{\mathrm{gt}}$ *is independent of* $\mathbf{X}_\tau$ *and satisfies*

$$\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}}\left[\left(\boldsymbol{\theta}_\tau^{\mathrm{gt}} - \boldsymbol{\theta}_0^*\right)\left(\boldsymbol{\theta}_\tau^{\mathrm{gt}} - \boldsymbol{\theta}_0^*\right)^\top\right] = \frac{R^2}{d}\mathbf{I}_d \tag{21}$$

where R *is a constant, and the individual entries* $\{\boldsymbol{\theta}_{\tau,i}^{\mathrm{gt}} - \boldsymbol{\theta}_{0,i}^*\}_{i\in[d],\tau\in[T]}$ *are i.i.d. mean-zero and* $\mathcal{O}(R/\sqrt{d})$*-sub-gaussian.*

Note that Assumption 3 has also been used in Bai et al. (2021); Denevi et al. (2018), whereas we do not make the noiseless realizable assumption compared to Bai et al. (2021), thus less restrictive. Based on this assumption, we can obtain the dominating constant exactly, as stated in Theorems 5 and 6.

Theorem 5 (Statistical error of MAML) *Suppose Assumptions 1,3 hold,* $T = \Omega(d), d/N = \eta > 0$, *and* $\alpha > 0$. *Define*

$$w_{\mathcal{A}} := \frac{1}{d}\mathrm{tr}(\mathbb{E}[\mathbf{W}_\tau^{\mathcal{A}}]),\;\; \tilde{C}_0^{\mathcal{A}} := \frac{1}{d}\langle \mathbb{E}^{-2}[\hat{\mathbf{W}}_\tau^{\mathcal{A}}], \mathbb{E}[(\hat{\mathbf{W}}_\tau^{\mathcal{A}})^2]\rangle$$

With probability at least $1 - Td^{-10}$, *the statistical error in* (11) *under MAML satisfies*

$$\mathcal{E}_{\mathrm{ma}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{ma}}) = \frac{R^2}{T} w_{\mathrm{ma}} \tilde{C}_0^{\mathrm{ma}} + \frac{d}{TN} w_{\mathrm{ma}} \tilde{C}_1^{\mathrm{ma}} + \varrho \tag{22}$$

where ϱ *is given by* (15). *The dominating constant* $\tilde{C}_0^{\mathrm{ma}}$ *satisfies*

$$\inf_{\substack{\alpha > 0 \\ s \in (0,1)}} \lim_{\substack{d, N \to \infty \\ d/N \to \eta}} \tilde{C}_0^{\mathrm{ma}} = 1 + \eta. \tag{23}$$

Similarly, we can obtain the concentration of the statistical error of BaMAML.

Theorem 6 (Statistical error of BaMAML) *Suppose Assumptions 1,3 hold,* $T = \Omega(d), d/N = \eta > 0$, *and* $\gamma > 0$. *Then with probability at least* $1 - Td^{-10}$, *the statistical error in* (11) *under BaMAML satisfies*

$$\mathcal{E}_{\mathrm{ba}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{ba}}) = \frac{R^2}{T} w_{\mathrm{ba}} \tilde{C}_0^{\mathrm{ba}} + \frac{d}{TN} w_{\mathrm{ba}} \tilde{C}_1^{\mathrm{ba}} + \varrho \tag{24}$$

where ϱ *is given by* (15). *In addition, the dominating constant* $\tilde{C}_0^{\mathrm{ba}}$ *satisfies*

$$\inf_{\substack{\gamma > 0 \\ s \in (0,1)}} \lim_{\substack{d, N \to \infty \\ d/N \to \eta}} \tilde{C}_0^{\mathrm{ba}} \begin{cases} = 1, & \eta \in (0,1], \\ \leq \eta, & \eta \in (1,\infty). \end{cases} \tag{25}$$

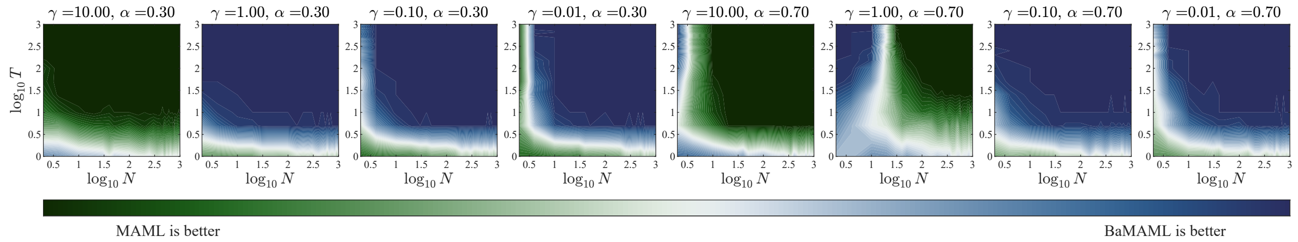


Figure 2: Contour plots of the probability that the BaMAML estimate has lower expected loss than the MAML estimate The axes are the log number of tasks ($\log_{10} T$) and the log number of data points ($\log_{10} N$) used for meta-test optimization, and the values of α, γ are given in subfigure titles.

Theorems 5 and 6 state that when T, d are large and $T = \tilde{\Omega}(d)$, the statistical errors of MAML and BaMAML are dominated by R^2/T times $\tilde{C}_0^{\mathrm{ma}}$ and $\tilde{C}_0^{\mathrm{ba}}$, respectively. Therefore we can compare the statistical errors of MAML and BaMAML based on the optimal hyperparameters α, γ and split ratio s below.

Corollary 1 (Dominating constants in statistical errors) *The dominating constants in the statistical errors of MAML and BaMAML satisfy*

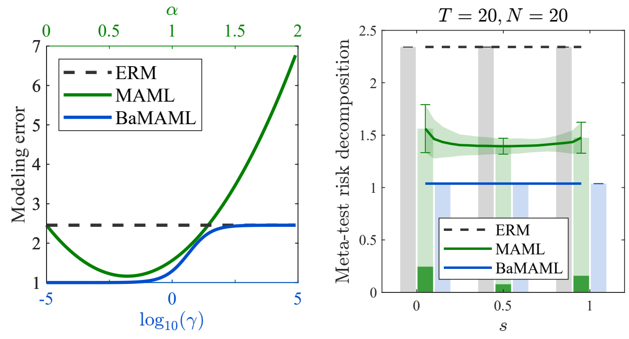
$$\inf_{\substack{\alpha > 0\\ s \in (0,1)}} \lim_{\substack{d,N \to \infty\\ d/N \to \eta}} \tilde{C}_0^{\mathrm{ma}} > \inf_{\substack{\gamma > 0\\ s \in (0,1)}} \lim_{\substack{d,N \to \infty\\ d/N \to \eta}} \tilde{C}_0^{\mathrm{ba}}. \tag{26}$$

Corollary 1 justifies the provable benefit of BaMAML in terms of strictly smaller statistical error, contributing to smaller meta-test risk. This is achieved in the high dimension limit regime as $d, N \to \infty$ and $d/N \to \eta$, which is common in the overparameterized case.

4 EXPERIMENTS

In this section, we present empirical experiments on synthetic and real datasets to verify our theorems. For synthetic datasets, we perform linear and sinusoidal regression. For real datasets, we use miniImageNet. By default, the experiments are repeated 5 times with the average, best and worst performance displayed. In our experiments, we also use ERM as a baseline for comparison. In the meta learning setting, ERM minimizes the average loss over all data, its meta-test risk and optimal solutions can be obtained by taking $\alpha = 0, N_1 = 0, N_2 = N$ in that of MAML (Gao and Sener, 2020).

All our experiments are conducted on a workstation with an Intel i9-9960x CPU with 128GB memory and four NVIDIA RTX 2080Ti GPUs each with 11GB memory. Our experiments for linear synthetic data are conducted on MATLAB R2021a with CPU only. And our experiments for sinewave regression and real classification are conducted on Python 3.7, PyTorch 1.9.1 with one GPU. More results can be found in the supplementary material.



(a) $\mathcal{R}(\boldsymbol{\theta}_0^{\mathcal{A}})$ v.s. α, γ

(b) $\mathcal{R}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}})$ v.s. s

Figure 3: Optimal population risks $\mathcal{R}(\boldsymbol{\theta}_0^{\mathcal{A}})$ v.s. α, γ, and meta-test risks $\mathcal{R}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}})$ v.s. train validation split ratio s for ERM, MAML and BaMAML. In Figure 3b, the lighter color bars represent the meta-test risks and the darker color bars represent the statistical errors.

4.1 Linear Regression

Experiment settings. For linear regression, we generate synthetic data according to the following task parameter $\mathbf{V} \sim U(\mathbb{SO}(d)), \boldsymbol{\theta}_\tau \sim U([0,2]^d)$, $\lambda_\tau \sim U([0.1,2]^d), \mathbf{Q}_\tau = \mathbf{V}\mathrm{diag}(\lambda_\tau)\mathbf{V}^\top, \mathbf{x}_\tau \sim \mathcal{N}(0, \mathbf{Q}_\tau)$, and data model in (7), where $\mathbb{SO}(d)$ is the special orthogonal group in dimension d.

Results. We present experiments for $d = 1$. To compare the meta-test performance of MAML and BaMAML, we present contour plots of probability that BaMAML has lower loss than MAML in Figure 2, where darker blue represents higher probability that BaMAML is better than MAML, and darker green vice versa. The results indicate that with sufficient adaptation tasks or data, and proper choice of γ, BaMAML performs better than MAML in terms of test error.

In Figure 3a, we report the meta-test risks (2) for MAML and BaMAML under different hyperparameters α, γ. Figure 3a shows that with proper choice of hyperparameter γ, BaMAML can achieve lower meta-test risk than MAML, verifying Theorem 1. Also, when $\gamma \to 0$, the meta-test risk of BaMAML approaches 1, and when $\gamma \to \infty$, the meta-test risk of BaMAML

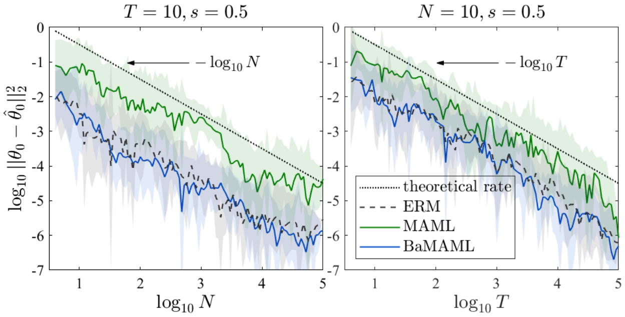


Figure 4: Statistical error v.s. N and T for ERM, MAML and BaMAML. The dotted line serves as a reference of the theoretical decay rate.

approaches that of ERM. This further demonstrates the trade-off between fast adaptation and optimal population risk (meta-test risk) that depends on γ. On the other hand, MAML is relatively more sensitive to the step size. Too large step size α can lead to very large meta-test risk of MAML, going beyond that of ERM. Besides, we can see from the empirical optimal solution that, contrary to non-Bayesian methods as analyzed in Bai et al. (2021), BaMAML is robust against different training and validation data split. This is demonstrated in Figure 3b, where the meta-test risk of BaMAML remains unchanged when s varies while that of MAML is more sensitive to the change in s due to its statistical error, verifying Theorems 3-6. Figure 3b also demonstrates the decomposition of meta-test risk into optimal population risk and statistical error, where the meta-test risks are mainly dominated by the optimal population risks in this case.

Next we fix $\alpha = 0.7, \gamma = 10^{-1}$, which are approximately optimal for each method. We vary number of data samples (N) and number of tasks (T), and compare the statistical error of ERM, MAML, and BaMAML in Figure 4. The statistical errors of all methods decrease as the number of data samples increases. When the number of data samples is small, MAML has the largest statistical error, followed by BaMAML and ERM. Similar trends exist with increasing number of tasks. Figure 4 have also shown the dotted black line as a reference, indicating the theoretical decay rate of the statistical errors. Since MAML and BaMAML has the same slope as the reference line, it verifies the theoretical decay rate in Theorems 3-6.

4.2 Sinusoidal Regression

Experiment settings. For sinusoidal regression, following Yoon et al. (2018), the N-shot dataset for each task is obtained from $x \sim U([-5.0, 5.0])$ and then by computing its corresponding y from the sinusoidal function $y = A\sin(wx+b)+\epsilon$, with task-dependent parameters amplitude A, frequency w, and phase b, and observation noise ϵ. For each task, the parameters are sampled from $A \sim U([0.1, 5.0]), b \sim U([0.0, 2\pi]), w \sim U([0.5, 2.0]), \epsilon \sim \mathcal{N}\left(0, (0.01A)^2\right)$. For all experiments under this setting, we used a neural network with 3 layers, each of which consists of 40 hidden units.

Results. Figure 5 shows the testing error v.s. the meta-train iterations for the compared methods, where we can see that BaMAML converges to a point with lowest meta-test error. For ERM and MAML, when T is small (e.g. $T = 100$), the meta test error decreases as the number of meta iterations increases in the beginning, but increases later, showing a tendency to overfit. Besides, the meta-test error decreases with increasing number of tasks or number of per-task data, with a similar trend as the linear regression even in the nonlinear sinewave regression. As the number of tasks or number of per-task data increases, the performance gap between MAML and BaMAML reduces, demonstrating that BaMAML has more significant performance gain in limited data settings.

5 CONCLUSIONS

In this paper, we study what makes BaMAML provably better than MAML under the meta linear regression setting. The meta-test risk can be decomposed into the optimal population risk and statistical error. Our analysis shows that, with proper choice of hyperparameters, BaMAML has smaller optimal population risk than MAML, demonstrating better adaptation ability to new data. And for statistical errors, MAML and BaMAML have the same dependence rate on the number of tasks and the number of data per task for training, while BaMAML has lower upper bound of the corresponding coefficients, thus lower upper bound of statistical error. And in the high dimensional asymptotic regime, BaMAML has strictly smaller statistical error than MAML. The experiments on synthetic and real datasets corroborate our theoretical findings. Building upon the current work, our future work includes analyzing the performance in nonlinear meta learning algorithms such as BaMAML with overparameterized neural networks.

Acknowledgments

This work was supported by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons), and the National Science Foundation under the project NSF 2134168.

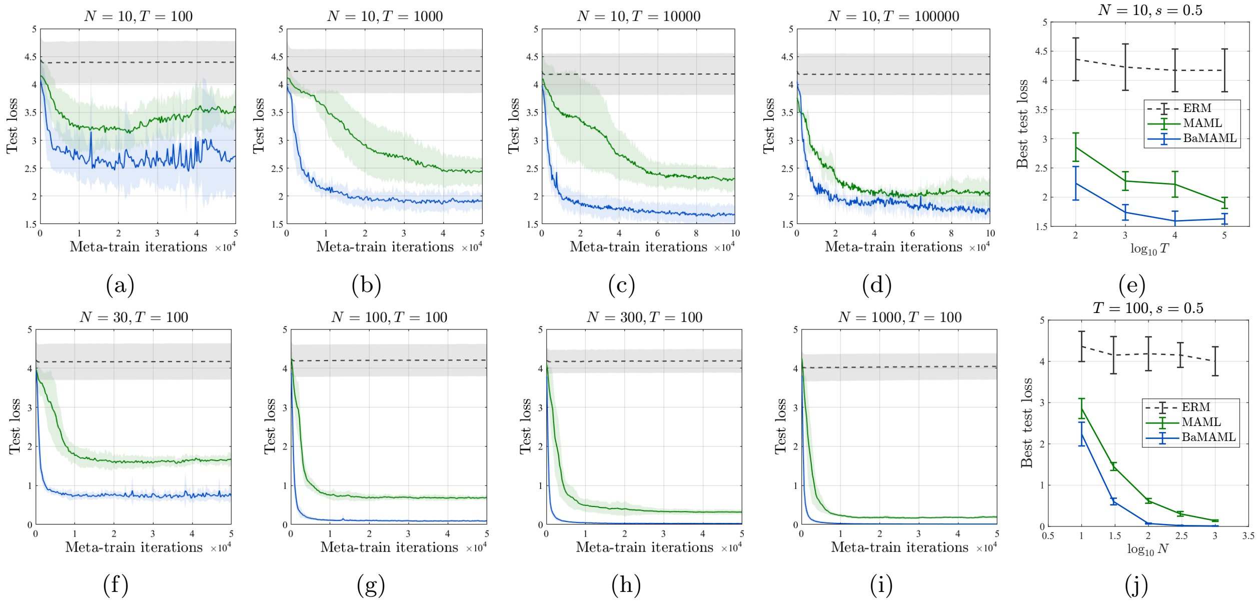


Figure 5: Testing mean squared error v.s. meta train iterations for ERM, MAML, and BaMAML in sinusoidal regression experiments with varied number of data (N) per task and the number of tasks (T) used for meta-training.

References

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proc. International Conference on Computer Vision*, pages 6430–6439, Seoul, Korea, 2019.

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.

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.

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.

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.

Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier. Learning a synaptic learning rule. In *International Joint Conference on Neural Networks*, volume 2, page 969, Seattle, WA, 1991.

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, 2021a.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Proc. Advances in Neural Information Processing Systems*, virtual, 2021b.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021c.

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.

Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. How does the task landscape affect maml performance? *arXiv preprint arXiv:2010.14672*, October 2020.

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.

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.

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.

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.

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.

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.

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, 2018.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *Proc. International Conference on Machine Learning*, pages 1920–1930, Long Beach, CA, 2019.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proc. International Conference on Machine Learning*, pages 1568–1577, Stockholm, Sweden, 2018.

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.

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard Turner. Meta-learning probabilistic inference for prediction. In *Proc. International Conference on Learning Representations*, Vancouver, Canada, 2018.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *Proc. International Conference on Learning Representations*, Vancouver, Canada, 2018.

James Harrison, Apoorva Sharma, Chelsea Finn, and Marco Pavone. Continuous meta-learning without tasks. In *Proc. Advances in Neural Information Processing Systems*, volume 33, virtual, 2020.

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.

Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. In *Proc. International Conference on Learning Representations*, Vancouver, Canada, 2018.

Khurram Javed and Martha White. Meta-learning representations for continual learning. In *Proc. Advances in Neural Information Processing Systems*, volume 32, pages 1820–1830, Vancouver, Canada, 2019.

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

Sharu Theresa Jose, Osvaldo Simeone, and Giuseppe Durisi. Transfer meta-learning: Information-theoretic bounds and information meta-risk minimization. *IEEE Transactions on Information Theory*, 2021.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proc. Association for the Advancement of Artificial Intelligence*, New Orleans, LA, 2018.

Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Learning to propagate for graph meta-learning. *Proc. Advances in Neural Information Processing Systems*, 32:1039–1050, 2019.

Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. Personalizing dialogue agents via meta-learning. In *Proc. Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy, 2019.

Cuong Nguyen, Thanh-Toan Do, and Gustavo Carneiro. Uncertainty in model-agnostic meta-learning using variational inference. In *Proc. Winter Conference on Applications of Computer Vision*, pages 3090–3100, Snowmass Village, CO, 2020.

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

Abiola Obamuyide and Andreas Vlachos. Model-agnostic meta-learning for relation classification with limited supervision. In *Proc. Annual Meeting of the Association for Computational Linguistics*, pages 5873–5879, Florence, Italy, 2019.

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.

Sachin Ravi and Alex Beatson. Amortized bayesian meta-learning. In *Proc. International Conference on Learning Representations*, New Orleans, LA, 2019.

Arezou Rezazadeh, Sharu Theresa Jose, Giuseppe Durisi, and Osvaldo Simeone. Conditional mutual information-based generalization bound for meta learning. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1176–1181. IEEE, 2021.

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.

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.

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

Jürgen Schmidhuber. On learning how to learn learning strategies. Technical report, Technical University of Munich, 1995.

Rishit Sheth and Roni Khardon. Excess risk bounds for the bayes risk using variational inference in latent gaussian models. In *Proc. Advances in Neural Information Processing Systems*, Long Beach, CA, 2017.

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.

Sebastian Thrun and Lorien Pratt. *Learning to Learn: Introduction and Overview*. Kluwer Academic Publishers, USA, 1998. ISBN 0792380479.

Joaquin Vanschoren. Meta-learning: A survey. *arXiv e-prints*, pages arXiv–1810, 2018.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95, 2002.

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.

Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim. Multimodal model-agnostic meta-learning via task-aware modulation. In *Proc. Advances in Neural Information Processing Systems*, pages 1–12, Vancouver, Canada, 2019.

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*, 2020a.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. On the global optimality of model-agnostic meta-learning. In *Proc. International Conference on Machine Learning*, pages 9837–9846, Virtual, 2020b.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *Proc. International Conference on Computer Vision*, Seoul, Korea, October 2019.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. Meta-learning without memorization. In *Proc. International Conference on Learning Representations*, virtual, 2020.

Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *Proc. Advances in Neural Information Processing Systems*, volume 31, Montreal, Canada, 2018.

Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *Proc. International Conference on Machine Learning*, pages 7693–7702, Long Beach, CA, 2019.

Supplementary Material: Is Bayesian Model-Agnostic Meta Learning Better than Model-Agnostic Meta Learning, Provably?

In this appendix, we first present the problem setting, then some basic supporting lemmas, and the missing derivations of some claims, as well as the proofs of all the lemmas and theorems in the paper, which is followed by details on our experiments along with additional experimental results.

A Formulations and closed-form solutions

In this section, we will introduce the definition and computation of meta-test (population) risks and empirical losses for the four methods that we will discuss, including ERM, MAML, iMAML, BaMAML. This prepares for the analysis of optimal population risk and statistical error of the four methods in later sections.

A.1 Empirical risk minimization formulation

In the meta learning setting, ERM minimizes the average loss over all data, its empirical loss, meta-test risk and their optimal solutions can be obtained by taking $\alpha = 0, N_1 = 0, N_2 = N$ in that of MAML (Gao and Sener, 2020), i.e. $\hat{\boldsymbol{\theta}}_\tau^{\text{er}}(\boldsymbol{\theta}_0, \mathcal{D}_\tau^{\text{trn}}) = \boldsymbol{\theta}_0$, and based on the definition in (1), the empirical loss of ERM is given by

$$\mathcal{L}^{\text{er}}(\boldsymbol{\theta}_0, \mathcal{D}) = \frac{1}{TN}\sum_{\tau=1}^{T} \|\mathbf{y}_{\tau,N}^{\text{all}} - \mathbf{X}_{\tau,N}^{\text{all}}\boldsymbol{\theta}_0\|^2. \tag{27}$$

For brevity, denote $\mathbf{e}_{\tau,N}^{\text{all}} = [\epsilon_{\tau,1}, \ldots, \epsilon_{\tau,N}]^\top \in \mathbb{R}^N$. And define $\hat{\boldsymbol{\theta}}_0^{\text{er}}$ as the minimizer of the ERM empirical loss, given by

$$\hat{\boldsymbol{\theta}}_0^{\text{er}} = \arg\min_{\boldsymbol{\theta}_0} \mathcal{L}^{\text{er}}(\boldsymbol{\theta}_0, \mathcal{D}) = \arg\min_{\boldsymbol{\theta}_0} \frac{1}{TN}\sum_{\tau=1}^{T} \|\mathbf{X}_{\tau,N}^{\text{all}}\boldsymbol{\theta}_\tau^{\text{gt}} + \mathbf{e}_{\tau,N}^{\text{all}} - \mathbf{X}_{\tau,N}^{\text{all}}\boldsymbol{\theta}_0\|^2. \tag{28}$$

Using the optimality condition, we have

$$\hat{\boldsymbol{\theta}}_0^{\text{er}} = \Big(\sum_{\tau=1}^{T} \hat{\mathbf{W}}_{\tau,N}^{\text{er}}\Big)^{-1}\Big(\sum_{\tau=1}^{T} \hat{\mathbf{W}}_{\tau}^{\text{er}}\boldsymbol{\theta}_\tau^{\text{gt}}\Big) + \Delta_T^{\text{er}} \tag{29a}$$

$$\Delta_T^{\text{er}} = \Big(\sum_{\tau=1}^{T} \hat{\mathbf{W}}_{\tau,N}^{\text{er}}\Big)^{-1}\Big(\sum_{\tau=1}^{T} \frac{1}{N}\mathbf{X}_{\tau,N}^{\text{all}\top}\mathbf{e}_{\tau,N}^{\text{all}}\Big) \tag{29b}$$

$$\hat{\mathbf{W}}_{\tau}^{\text{er}} = \frac{1}{N}\mathbf{X}_{\tau,N}^{\text{all}\top}\mathbf{X}_{\tau,N}^{\text{all}}. \tag{29c}$$

Based on the definition in (2), denote the number of adaptation data during meta testing as N_a, then the total meta-test risk of ERM can be specified by

$$\mathcal{R}_{N_a}^{\text{er}}(\boldsymbol{\theta}_0) \coloneqq \mathbb{E}_\tau\Big[\mathbb{E}_{\mathcal{D}_{\tau,N_a}}\big[\mathbb{E}_{p(\mathbf{x}_\tau, y_\tau|\tau)}[(y_\tau - \hat{\boldsymbol{\theta}}_\tau^{\text{er}}(\boldsymbol{\theta}_0, \mathcal{D}_{\tau,N_a})^\top\mathbf{x}_\tau)^2]\big]\Big] \tag{30}$$

where $\hat{\boldsymbol{\theta}}_\tau^{\text{er}}(\boldsymbol{\theta}_0, \mathcal{D}_\tau) = \boldsymbol{\theta}_0$, plugging which into (30), we have

$$\mathcal{R}_{N_a}^{\text{er}}(\boldsymbol{\theta}_0) = \mathbb{E}_\tau[(y_\tau - \boldsymbol{\theta}_0^\top\mathbf{x}_\tau)^2] = \mathbb{E}_\tau[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau^{\text{gt}}\|^2_{\mathbf{W}_{\tau,N_a}^{\text{er}}}] + 1 \tag{31a}$$

$$\mathbf{W}_{\tau,N_a}^{\text{er}} = \mathbb{E}[\mathbf{x}_\tau\mathbf{x}_\tau^\top \mid \tau] = \mathbf{Q}_\tau. \tag{31b}$$

By the general definition of optimal population risk in (2), the ERM optimal population risk is given by

$$\mathcal{R}^{\rm er}(\boldsymbol{\theta}_0) \coloneqq \lim_{N_a\to\infty} \mathcal{R}^{\rm er}_{N_a}(\boldsymbol{\theta}_0) = \mathbb{E}_\tau[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau^{\rm gt}\|^2_{\mathbf{W}_\tau^{\rm er}}] + 1 \tag{32a}$$

$$\mathbf{W}_\tau^{\rm er} = \mathbf{Q}_\tau. \tag{32b}$$

Define $\boldsymbol{\theta}_0^{\rm er}$ as the minimizer of the ERM optimal population risk, given by

$$\boldsymbol{\theta}_0^{\rm er} = \arg\min_{\boldsymbol{\theta}_0} \mathcal{R}^{\rm er}(\boldsymbol{\theta}_0) = \arg\min_{\boldsymbol{\theta}_0} \mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau^{\rm gt}\|^2_{\mathbf{W}_\tau^{\rm er}}\big] = \mathbb{E}_\tau\big[\mathbf{W}_\tau^{\rm er}\big]^{-1}\mathbb{E}_\tau\big[\mathbf{W}_\tau^{\rm er}\boldsymbol{\theta}_\tau^{\rm gt}\big]. \tag{33}$$

From (29c)(31b)(32b), we have the property $\mathbb{E}[\hat{\mathbf{W}}^{\rm er}_{\tau,N}] = \mathbf{W}^{\rm er}_{\tau,N} = \mathbf{W}^{\rm er}_\tau$, which will be used in later sections to derive the specific optimal population risks and statistical errors.

A.2 Model agnostic meta learning method

For the one-step model agnostic meta learning (MAML) method, the task-specific parameter $\hat{\boldsymbol{\theta}}_\tau^{\rm ma}$ is computed from the initial parameter $\boldsymbol{\theta}_0$ by taking one-step gradient descent of the empirical loss function as shown below

$$\hat{\boldsymbol{\theta}}_\tau^{\rm ma}(\boldsymbol{\theta}_0, \mathcal{D}_\tau) = \boldsymbol{\theta}_0 - \frac{\alpha}{2}\nabla_{\boldsymbol{\theta}_0}\ell_\tau(\boldsymbol{\theta}_0, \mathcal{D}_\tau) = (\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N})\boldsymbol{\theta}_0 + \frac{\alpha}{N}\mathbf{X}_{\tau,N}^\top\mathbf{y}_{\tau,N} \tag{34}$$

[1] where $\alpha > 0$ is twice the stepsize, and N is the number of adaptation data. During meta-training, $N = N_1$, is the number of the training data. From the definition in (1) or (2), and combined with $\hat{\boldsymbol{\theta}}_\tau^{\rm ma}$ in (34), the empirical loss of MAML is given by

$$\mathcal{L}^{\rm ma}(\boldsymbol{\theta}_0, \mathcal{D}) = \frac{1}{TN_2}\sum_{\tau=1}^{T}\|\mathbf{y}^{\rm val}_{\tau,N_2} - \mathbf{X}^{\rm val}_{\tau,N_2}\hat{\boldsymbol{\theta}}_\tau^{\rm ma}(\boldsymbol{\theta}_0, \mathcal{D}_\tau^{\rm trn})\|^2. \tag{35}$$

The minimizer of the MAML empirical loss is defined as

$$\hat{\boldsymbol{\theta}}_0^{\rm ma} = \arg\min_{\boldsymbol{\theta}_0}\mathcal{L}^{\rm ma}(\boldsymbol{\theta}_0, \mathcal{D}) = \arg\min_{\boldsymbol{\theta}_0}\frac{1}{TN_2}\sum_{\tau=1}^{T}\|\mathbf{X}^{\rm val}_{\tau,N_2}\boldsymbol{\theta}_\tau^{\rm gt} + \mathbf{e}^{\rm val}_{\tau,N_2} - \mathbf{X}^{\rm val}_{\tau,N_2}\hat{\boldsymbol{\theta}}_\tau^{\rm ma}(\boldsymbol{\theta}_0, \mathcal{D}_\tau^{\rm trn})\|^2. \tag{36}$$

Using the optimality condition, we have

$$\hat{\boldsymbol{\theta}}_0^{\rm ma} = \Big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_\tau^{\rm ma}\Big)^{-1}\Big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_\tau^{\rm ma}\boldsymbol{\theta}_\tau^{\rm gt}\Big) + \Delta_T^{\rm ma} \tag{37a}$$

$$\Delta_T^{\rm ma} = \Big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_\tau^{\rm ma}\Big)^{-1}\Big(\sum_{\tau=1}^{T}\big(\mathbf{I} - \alpha\hat{\mathbf{Q}}^{\rm trn}_{\tau,N_1}\big)\big(\frac{1}{N_2}\mathbf{X}^{\rm val\top}_{\tau,N_2}\mathbf{e}^{\rm val}_{\tau,N_2} - \frac{\alpha}{N_1}\hat{\mathbf{Q}}^{\rm val}_{\tau,N_2}\mathbf{X}^{\rm trn\top}_{\tau,N_1}\mathbf{e}^{\rm trn}_{\tau,N_1}\big)\Big) \tag{37b}$$

$$\hat{\mathbf{W}}_\tau^{\rm ma} = (\mathbf{I} - \alpha\hat{\mathbf{Q}}^{\rm trn}_{\tau,N_1})\hat{\mathbf{Q}}^{\rm val}_{\tau,N_2}(\mathbf{I} - \alpha\hat{\mathbf{Q}}^{\rm trn}_{\tau,N_1}). \tag{37c}$$

Based on (2), the MAML meta-test risk is defined as (Gao and Sener, 2020)

$$\mathcal{R}^{\rm ma}_{N_a}(\boldsymbol{\theta}_0) = \mathbb{E}[(y_\tau - \hat{\boldsymbol{\theta}}_\tau^{\rm ma}(\boldsymbol{\theta}_0, \mathcal{D}_{\tau,N_a})^\top\mathbf{x}_\tau)^2] = \mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau^{\rm gt}\|^2_{\mathbf{W}^{\rm ma}_{\tau,N_a}}\big] + 1 + \frac{\alpha^2}{N_a}\mathbb{E}_\tau[\operatorname{tr}(\mathbf{Q}_\tau^2)] \tag{38a}$$

$$\begin{aligned}\mathbf{W}^{\rm ma}_{\tau,N_a} &= \mathbb{E}_{\hat{\mathbf{Q}}_{\tau,N}}\big[(\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N})\mathbf{Q}_\tau(\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N})\big]\\ &= \big(\mathbf{I} - \alpha\mathbf{Q}_\tau\big)\mathbf{Q}_\tau\big(\mathbf{I} - \alpha\mathbf{Q}_\tau\big) + \frac{\alpha^2}{N_a}\Big(\mathbb{E}_{\mathbf{x}_{\tau,i}}\big[\mathbf{x}_{\tau,i}\mathbf{x}_{\tau,i}^\top\mathbf{Q}_\tau\mathbf{x}_{\tau,i}\mathbf{x}_{\tau,i}^\top\big] - \mathbf{Q}_\tau^3\Big).\end{aligned} \tag{38b}$$

Note that, $\lim_{N_a\to\infty}\mathbf{W}^{\rm ma}_{\tau,N_a} \to \mathbf{W}^{\rm ma}_\tau$, and $\lim_{N_a\to\infty}\frac{\alpha^2}{N_a}\operatorname{tr}(\mathbf{Q}_\tau^2) = 0$. Therefore, from the definition of optimal population risk in (2), we have the MAML optimal population risk is

$$\mathcal{R}^{\rm ma}(\boldsymbol{\theta}_0) = \lim_{N_a\to\infty}\mathcal{R}^{\rm ma}_{N_a}(\boldsymbol{\theta}_0) = \mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau^{\rm gt}\|^2_{\mathbf{W}^{\rm ma}_\tau}\big] + 1. \tag{39}$$

[1] Note that, here we define the learning rate as $\alpha/2$ to cancel the scale factor 2 from the derivative for notation simplicity.

In MAML, define $\boldsymbol{\theta}_0^{\mathrm{ma}}$ as the minimizer of the MAML optimal population risk, given by

$$\boldsymbol{\theta}_0^{\mathrm{ma}} = \arg\min_{\boldsymbol{\theta}_0} \mathcal{R}^{\mathrm{ma}}(\boldsymbol{\theta}_0) = \arg\min_{\boldsymbol{\theta}_0} \mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau^{\mathrm{gt}}\|^2_{\mathbf{W}_\tau^{\mathrm{ma}}}\big] = \mathbb{E}_\tau\big[\mathbf{W}_\tau^{\mathrm{ma}}\big]^{-1}\mathbb{E}_\tau\big[\mathbf{W}_\tau^{\mathrm{ma}}\boldsymbol{\theta}_\tau^{\mathrm{gt}}\big]. \tag{40}$$

It is worth noting that, from Lemma 1, we have the property $\mathbb{E}_{\mathbf{x}_\tau}[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}] = \mathbf{W}_{\tau,N}^{\mathrm{ma}}$, $\lim_{N\to\infty}\mathbf{W}_{\tau,N}^{\mathrm{ma}} = \mathbf{W}_\tau^{\mathrm{ma}}$, which will be used in later sections to derive the specific optimal population risk and statistical error.

A.3 Implicit model agnostic meta learning method

For the iMAML method, the task-specific parameter $\hat{\boldsymbol{\theta}}_\tau^{\mathrm{im}}$ is computed from the initial parameter $\boldsymbol{\theta}_0$ by optimizing the regularized task-specific empirical loss, given by

$$\hat{\boldsymbol{\theta}}_\tau^{\mathrm{im}}(\boldsymbol{\theta}_0) = \arg\min_{\boldsymbol{\theta}_\tau} \frac{1}{N}\|\mathbf{y}_{\tau,N} - \mathbf{X}_{\tau,N}\boldsymbol{\theta}_\tau\|^2 + \gamma\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|^2 \tag{41}$$

where γ is the weight of the regularizer, and $\mathcal{D}_{\tau,N_a}$ is the adaptation data during meta-testing or training data during meta-training. The estimated task-specific parameter can be computed by

$$\hat{\boldsymbol{\theta}}_{\tau,N}^{\mathrm{im}}(\boldsymbol{\theta}_0, \mathcal{D}_\tau) = (\hat{\mathbf{Q}}_{\tau,N} + \gamma\mathbf{I})^{-1}(\frac{1}{N}\mathbf{X}_{\tau,N}^\top\mathbf{y}_{\tau,N} + \gamma\boldsymbol{\theta}_0). \tag{42}$$

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

$$\mathcal{L}_{T,N}^{\mathrm{im}}(\boldsymbol{\theta}_0, \mathcal{D}) = \frac{1}{TN_2}\sum_{\tau=1}^T \|\mathbf{y}_{\tau,N_2}^{\mathrm{val}} - \mathbf{X}_{\tau,N_2}^{\mathrm{val}}\hat{\boldsymbol{\theta}}_\tau^{\mathrm{im}}(\boldsymbol{\theta}_0, \mathcal{D}_\tau^{\mathrm{trn}})\|^2 \tag{43}$$

whose minimizer is

$$\hat{\boldsymbol{\theta}}_0^{\mathrm{im}} = \arg\min_{\boldsymbol{\theta}_0} \mathcal{L}_{T,N}^{\mathrm{im}}(\boldsymbol{\theta}_0, \mathcal{D}) = \arg\min_{\boldsymbol{\theta}_0} \frac{1}{TN_2}\sum_{\tau=1}^T \|\mathbf{X}_{\tau,N_2}^{\mathrm{val}}\boldsymbol{\theta}_\tau^{\mathrm{gt}} + \mathbf{e}_{\tau,N_2}^{\mathrm{val}} - \mathbf{X}_{\tau,N_2}^{\mathrm{val}}\hat{\boldsymbol{\theta}}_\tau^{\mathrm{im}}(\boldsymbol{\theta}_0, \mathcal{D}_\tau^{\mathrm{trn}})\|^2. \tag{44}$$

To solve for $\boldsymbol{\theta}_0^{\mathrm{im}}$ in the above equation, using the optimality condition, we obtain

$$\hat{\boldsymbol{\theta}}_0^{\mathrm{im}} = \Big(\sum_{\tau=1}^T \hat{\mathbf{W}}_\tau^{\mathrm{im}}\Big)^{-1}\Big(\sum_{\tau=1}^T \hat{\mathbf{W}}_\tau^{\mathrm{im}}\boldsymbol{\theta}_\tau^{\mathrm{gt}}\Big) + \Delta_T^{\mathrm{im}} \tag{45a}$$

$$\Delta_T^{\mathrm{im}} = \Big(\sum_{\tau=1}^T \hat{\mathbf{W}}_\tau^{\mathrm{im}}\Big)^{-1}\Big(\sum_{\tau=1}^T \gamma\Sigma_{\boldsymbol{\theta}_\tau}\frac{1}{N_2}\mathbf{X}_{\tau,N_2}^{\mathrm{val}\top}\mathbf{e}_{\tau,N_2}^{\mathrm{val}} - \gamma^{-1}\hat{\mathbf{W}}_\tau^{\mathrm{im}}\frac{1}{N_1}\mathbf{X}_{\tau,N_1}^{\mathrm{trn}\top}\mathbf{e}_{\tau,N_1}^{\mathrm{trn}}\Big) \tag{45b}$$

$$\Sigma_{\boldsymbol{\theta}_\tau,N_1} = (\frac{1}{N_1}\mathbf{X}_{\tau,N_1}^{\mathrm{trn}\top}\mathbf{X}_{\tau,N_1}^{\mathrm{trn}} + \gamma\mathbf{I})^{-1} = (\hat{\mathbf{Q}}_{\tau,N_1} + \gamma\mathbf{I})^{-1} \tag{45c}$$

$$\hat{\mathbf{W}}_\tau^{\mathrm{im}} = \gamma^2\Sigma_{\boldsymbol{\theta}_\tau,N_1}\frac{1}{N_2}\mathbf{X}_{\tau,N_2}^{\mathrm{val}\top}\mathbf{X}_{\tau,N_2}^{\mathrm{val}}\Sigma_{\boldsymbol{\theta}_\tau,N_1} = \gamma^2\Sigma_{\boldsymbol{\theta}_\tau,N_1}\hat{\mathbf{Q}}_{\tau,N_2}\Sigma_{\boldsymbol{\theta}_\tau,N_1}. \tag{45d}$$

The iMAML meta-test risk is defined as

$$\begin{aligned}\mathcal{R}_{N_a}^{\mathrm{im}}(\boldsymbol{\theta}_0) &= \mathbb{E}\big[\big(y_\tau - \hat{\boldsymbol{\theta}}_\tau^{\mathrm{im}}(\boldsymbol{\theta}_0, \mathcal{D}_{\tau,N_a})^\top\mathbf{x}_\tau\big)^2\big]\\ &= \mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau^{\mathrm{gt}}\|^2_{\mathbf{W}_{\tau,N_a}^{\mathrm{im}}}\big] + 1 + \frac{1}{N_a}\mathbb{E}[\gamma^{-2}\operatorname{tr}(\mathbf{W}_{\tau,N_a}^{\mathrm{im}}\hat{\mathbf{Q}}_{\tau,N_a})]\end{aligned} \tag{46a}$$

$$\begin{aligned}&\mathbf{W}_{\tau,N_a}^{\mathrm{im}} = \mathbb{E}_{\mathbf{x}_\tau}\big[(\hat{\mathbf{Q}}_{\tau,N_a} + \gamma\mathbf{I})^{-1}\mathbf{Q}_\tau(\hat{\mathbf{Q}}_{\tau,N_a} + \gamma\mathbf{I})^{-1}\big] = \mathbf{W}_\tau^{\mathrm{im}} +\\ &\mathbb{E}_{\mathbf{x}_\tau}\big[\Sigma_{\theta_\tau}\big(\mathbf{Q}_\tau - \hat{\mathbf{Q}}_{\tau,N_a}\big)\mathbf{W}_\tau^{\mathrm{im}}\big(\mathbf{Q}_\tau - \hat{\mathbf{Q}}_{\tau,N_a}\big)\Sigma_{\theta_\tau} + \Sigma_{\theta_\tau}\big(\mathbf{Q}_\tau - \hat{\mathbf{Q}}_{\tau,N_a}\big)\mathbf{W}_\tau^{\mathrm{im}} + \mathbf{W}_\tau^{\mathrm{im}}\big(\mathbf{Q}_\tau - \hat{\mathbf{Q}}_{\tau,N_a}\big)\Sigma_{\theta_\tau}\big]\end{aligned} \tag{46b}$$

where $\mathbf{W}_\tau^{\mathrm{im}} = (\gamma^{-1}\mathbf{Q}_\tau + \mathbf{I})^{-1}\mathbf{Q}_\tau(\gamma^{-1}\mathbf{Q}_\tau + \mathbf{I})^{-1}$.

Let $\Sigma_{\theta_\tau} = (\hat{\mathbf{Q}}_{\tau,N} + \gamma\mathbf{I})^{-1}$, and $\mathbf{W}^{\rm im}_{\tau,N_a} = \gamma^2\Sigma_{\boldsymbol{\theta}_\tau}\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}$. And we simplify the notation of $\mathbf{X}_{\tau,N_a}, \mathbf{y}_{\tau,N_a}, \hat{\mathbf{Q}}_{\tau,N_a}$ as $\mathbf{X}_\tau, \mathbf{y}_\tau, \hat{\mathbf{Q}}_\tau$. The derivation of (46) is given below

$$\begin{aligned}
\mathcal{R}^{\rm im}_{N_a}(\boldsymbol{\theta}_0) =&\mathbb{E}\big[\|\hat{\boldsymbol{\theta}}^{\rm im}_\tau(\boldsymbol{\theta}_0, \mathcal{D}_{\tau,N_a}) - \boldsymbol{\theta}^{\rm gt}_\tau\|^2_{\mathbf{Q}_\tau}\big] + 1 = \mathbb{E}\big[\|(\hat{\mathbf{Q}}_\tau + \gamma\mathbf{I})^{-1}(\frac{1}{N_a}\mathbf{X}_\tau^\top\mathbf{y}_\tau + \gamma\boldsymbol{\theta}_0) - \boldsymbol{\theta}^{\rm gt}_\tau\|^2_{\hat{\mathbf{Q}}_\tau}\big] + 1\\
\overset{(a)}{=}&\mathbb{E}\Big[\boldsymbol{\theta}_0^\top\mathbf{W}^{\rm im}_{\tau,N_a}\boldsymbol{\theta}_0 + 2\gamma(\frac{1}{N_a}\mathbf{y}_\tau^\top\mathbf{X}_\tau\Sigma_{\boldsymbol{\theta}_\tau} - \boldsymbol{\theta}^{\rm gt\top}_\tau)\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\boldsymbol{\theta}_0 + \frac{1}{N_a}\mathbf{y}_\tau^\top\mathbf{X}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\frac{1}{N_a}\mathbf{X}_\tau^\top\mathbf{y}_\tau\\
&- 2\boldsymbol{\theta}^{\rm gt\top}_\tau\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\frac{1}{N_a}\mathbf{X}_\tau^\top\mathbf{y}_\tau + \boldsymbol{\theta}^{\rm gt\top}_\tau\mathbf{Q}_\tau\boldsymbol{\theta}^{\rm gt}_\tau\Big] + 1
\end{aligned}\tag{47}$$

where (a) follows from the definition of $\Sigma_{\boldsymbol{\theta}_\tau}$, $\mathbf{W}^{\rm im}_{\tau,N_a}$. Applying the fact that $\mathbf{y}_\tau = \mathbf{X}_\tau\boldsymbol{\theta}^{\rm gt}_\tau + \mathbf{e}_\tau$ and $\mathbb{E}_{\mathbf{e}_\tau}[\mathbf{e}_\tau] = \mathbf{0}$, one can further derive

$$\begin{aligned}
\mathcal{R}^{\rm im}_{N_a}(\boldsymbol{\theta}_0)=&\mathbb{E}\Big[\boldsymbol{\theta}_0^\top\mathbf{W}^{\rm im}_{\tau,N_a}\boldsymbol{\theta}_0 + 2\gamma(\boldsymbol{\theta}^{\rm gt\top}_\tau\hat{\mathbf{Q}}_\tau\Sigma_{\boldsymbol{\theta}_\tau} - \boldsymbol{\theta}^{\rm gt\top}_\tau)\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\boldsymbol{\theta}_0\\
&+ \boldsymbol{\theta}^{\rm gt\top}_\tau\hat{\mathbf{Q}}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\hat{\mathbf{Q}}_\tau\boldsymbol{\theta}^{\rm gt}_\tau - 2\boldsymbol{\theta}^{\rm gt\top}_\tau\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\hat{\mathbf{Q}}_\tau\boldsymbol{\theta}^{\rm gt}_\tau + \boldsymbol{\theta}^{\rm gt\top}_\tau\mathbf{Q}_\tau\boldsymbol{\theta}^{\rm gt}_\tau\\
&+ \frac{1}{N_a^2}\mathbf{e}_\tau^\top\mathbf{X}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\mathbf{X}_\tau^\top\mathbf{e}_\tau\Big] + 1.
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}_{\mathbf{e}_\tau}[\mathbf{e}_\tau^\top\mathbf{X}_\tau\Sigma_{\theta_\tau}\mathbf{Q}_\tau\Sigma_{\theta_\tau}\mathbf{X}_\tau^\top\mathbf{e}_\tau] = \mathrm{tr}(\mathbf{X}_\tau\Sigma_{\theta_\tau}\mathbf{Q}_\tau\Sigma_{\theta_\tau}\mathbf{X}_\tau^\top\mathbb{E}_{\mathbf{e}_\tau}[\mathbf{e}_\tau\mathbf{e}_\tau^\top])\\
=&\mathrm{tr}(\mathbf{X}_\tau\Sigma_{\theta_\tau}\mathbf{Q}_\tau\Sigma_{\theta_\tau}\mathbf{X}_\tau^\top) = N_a\mathrm{tr}(\Sigma_{\theta_\tau}\mathbf{Q}_\tau\Sigma_{\theta_\tau}\hat{\mathbf{Q}}_\tau) = N_a\mathrm{tr}(\mathbf{W}^{\rm im}_{\tau,N_a}\hat{\mathbf{Q}}_\tau);
\end{aligned}$$

also, based on the Woodbury matrix identity, $\mathbf{I} - \hat{\mathbf{Q}}_\tau\Sigma_{\theta_\tau} = \mathbf{I} - \Sigma_{\theta_\tau}\hat{\mathbf{Q}}_\tau = \gamma\Sigma_{\theta_\tau}$, therefore $(\boldsymbol{\theta}^{\rm gt\top}_\tau\hat{\mathbf{Q}}_\tau\Sigma_{\boldsymbol{\theta}_\tau} - \boldsymbol{\theta}^{\rm gt\top}_\tau) = \boldsymbol{\theta}^{\rm gt\top}_\tau(\hat{\mathbf{Q}}_\tau\Sigma_{\boldsymbol{\theta}_\tau} - \mathbf{I}) = -\gamma\boldsymbol{\theta}^{\rm gt\top}_\tau\Sigma_{\theta_\tau}$, and

$$\begin{aligned}
&\boldsymbol{\theta}^{\rm gt\top}_\tau\hat{\mathbf{Q}}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\hat{\mathbf{Q}}_\tau\boldsymbol{\theta}^{\rm gt}_\tau - 2\boldsymbol{\theta}^{\rm gt\top}_\tau\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\hat{\mathbf{Q}}_\tau\boldsymbol{\theta}^{\rm gt}_\tau + \boldsymbol{\theta}^{\rm gt\top}_\tau\mathbf{Q}_\tau\boldsymbol{\theta}^{\rm gt}_\tau\\
=&\boldsymbol{\theta}^{\rm gt\top}_\tau\big((\hat{\mathbf{Q}}_\tau\Sigma_{\boldsymbol{\theta}_\tau} - \mathbf{I})\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\hat{\mathbf{Q}}_\tau + \mathbf{Q}_\tau(\mathbf{I} - \Sigma_{\boldsymbol{\theta}_\tau}\hat{\mathbf{Q}}_\tau)\big)\boldsymbol{\theta}^{\rm gt}_\tau\\
=&\boldsymbol{\theta}^{\rm gt\top}_\tau\big(-\gamma\Sigma_{\boldsymbol{\theta}_\tau}\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}\hat{\mathbf{Q}}_\tau + \mathbf{Q}_\tau\gamma\Sigma_{\boldsymbol{\theta}_\tau}\big)\boldsymbol{\theta}^{\rm gt}_\tau = \gamma^{-1}\boldsymbol{\theta}^{\rm gt\top}_\tau\big(-\mathbf{W}^{\rm im}_{\tau,N_a}\hat{\mathbf{Q}}_\tau + (\hat{\mathbf{Q}}_\tau + \gamma\mathbf{I})\mathbf{W}^{\rm im}_{\tau,N_a}\big)\boldsymbol{\theta}^{\rm gt}_\tau.
\end{aligned}$$

Combining these equalities and rearranging the equations we obtain

$$\begin{aligned}
\mathcal{R}^{\rm im}_{N_a}(\boldsymbol{\theta}_0)=&\mathbb{E}\Big[\boldsymbol{\theta}_0^\top\mathbf{W}^{\rm im}_{\tau,N_a}\boldsymbol{\theta}_0 - 2\boldsymbol{\theta}^{\rm gt\top}_\tau\mathbf{W}^{\rm im}_{\tau,N_a}\boldsymbol{\theta}_0\\
&+ \gamma^{-1}\boldsymbol{\theta}^{\rm gt\top}_\tau\big(-\mathbf{W}^{\rm im}_{\tau,N_a}\hat{\mathbf{Q}}_\tau + (\hat{\mathbf{Q}}_\tau + \gamma\mathbf{I})\mathbf{W}^{\rm im}_{\tau,N_a}\big)\boldsymbol{\theta}^{\rm gt}_\tau + \frac{1}{N_a\gamma^2}\mathrm{tr}(\mathbf{W}^{\rm im}_{\tau,N_a}\hat{\mathbf{Q}}_\tau)\Big] + 1\\
\overset{(b)}{=}&\mathbb{E}\Big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^{\rm gt}_\tau\|^2_{\mathbf{W}^{\rm im}_{\tau,N_a}} + \gamma^{-1}\boldsymbol{\theta}^{\rm gt\top}_\tau\big(-\mathbf{W}^{\rm im}_{\tau,N_a}\hat{\mathbf{Q}}_\tau + \hat{\mathbf{Q}}_\tau\mathbf{W}^{\rm im}_{\tau,N_a}\big)\boldsymbol{\theta}^{\rm gt}_\tau + \frac{1}{N_a\gamma^2}\mathrm{tr}(\mathbf{W}^{\rm im}_{\tau,N_a}\hat{\mathbf{Q}}_\tau)\Big] + 1\\
\overset{(c)}{=}&\mathbb{E}\Big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^{\rm gt}_\tau\|^2_{\mathbf{W}^{\rm im}_{\tau,N_a}} + \frac{1}{N_a\gamma^2}\mathrm{tr}(\mathbf{W}^{\rm im}_{\tau,N_a}\hat{\mathbf{Q}}_\tau)\Big] + 1
\end{aligned}\tag{48}$$

where (b) follows from rearranging the equations; (c) follows from the fact that $\boldsymbol{\theta}^{gt\top}_\tau\big(\mathbf{W}^{\rm im}_{\tau,N_a}\hat{\mathbf{Q}}_\tau\big)\boldsymbol{\theta}^{\rm gt}_\tau = \big(\boldsymbol{\theta}^{\rm gt\top}_\tau(\mathbf{W}^{\rm im}_{\tau,N_a}\hat{\mathbf{Q}}_\tau)\boldsymbol{\theta}^{\rm gt}_\tau\big)^\top = \boldsymbol{\theta}^{\rm gt\top}_\tau\big(\hat{\mathbf{Q}}_\tau\mathbf{W}^{\rm im}_{\tau,N_a}\big)\boldsymbol{\theta}^{\rm gt}_\tau$.

Since $\lim_{N_a\to\infty}\frac{1}{N_a}\mathbb{E}[\gamma^{-2}\,\mathrm{tr}(\mathbf{W}^{\rm im}_{\tau,N_a}\hat{\mathbf{Q}}_{\tau,N_a})] = 0$, from the definition of optimal population risk in (2), the optimal population risk of iMAML is given by

$$\mathcal{R}^{\rm im}(\boldsymbol{\theta}_0) \coloneqq \lim_{N_a\to\infty}\mathcal{R}^{\rm im}_{N_a}(\boldsymbol{\theta}_0) = \mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^{\rm gt}_\tau\|^2_{\mathbf{W}^{\rm im}_\tau}\big] + 1 \tag{49a}$$

$$\mathbf{W}^{\rm im}_\tau = (\gamma^{-1}\mathbf{Q}_\tau + \mathbf{I})^{-1}\mathbf{Q}_\tau(\gamma^{-1}\mathbf{Q}_\tau + \mathbf{I})^{-1} \tag{49b}$$

whose minimizer is given by

$$\boldsymbol{\theta}^{\rm im}_0 = \arg\min_{\boldsymbol{\theta}_0}\mathcal{R}^{\rm im}(\boldsymbol{\theta}_0) = \mathbb{E}_\tau\big[\mathbf{W}^{\rm im}_\tau\big]^{-1}\mathbb{E}_\tau\big[\mathbf{W}^{\rm im}_\tau\boldsymbol{\theta}^{\rm gt}_\tau\big]. \tag{50}$$

It is worth noting that, from Lemma 1, we have the property $\mathbb{E}_{\mathbf{x}_\tau}[\hat{\mathbf{W}}^{\rm im}_{\tau,N}] = \mathbf{W}^{\rm im}_{\tau,N}$, $\lim_{N\to\infty}\mathbf{W}^{\rm im}_{\tau,N} = \mathbf{W}^{\rm im}_\tau$, which will be used in later sections to derive the specific optimal population risk and statistical error.

A.4 Bayes model agnostic meta learning method

For the Bayes model agnostic meta learning (BaMAML) method, instead of obtaining a point estimation of the task-specific parameter, during adaptation, it obtains the posterior distribution or its approximation, given by

$$\hat{p}(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau, \boldsymbol{\theta}_0) = \underset{q(\boldsymbol{\theta}_\tau)\in\mathcal{Q}}{\arg\min}\, \mathbb{E}_{p(\mathbf{x}_\tau, y_\tau|\tau)}\left[\mathrm{KL}(q(\boldsymbol{\theta}_\tau)\|p(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau, \boldsymbol{\theta}_0))\right] \tag{51}$$

where $p(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau, \boldsymbol{\theta}_0)$ can be computed from Bayes rule via

$$p(\boldsymbol{\theta}_\tau \mid \mathcal{D}_{\tau,N}, \boldsymbol{\theta}_0) \propto p(\mathcal{D}_{\tau,N} \mid \boldsymbol{\theta}_\tau) p(\boldsymbol{\theta}_\tau \mid \boldsymbol{\theta}_0). \tag{52}$$

Assuming $y_\tau \mid \mathbf{x}_\tau, \boldsymbol{\theta}_\tau \sim \mathcal{N}(\boldsymbol{\theta}_\tau^\top \mathbf{x}_\tau, 1)$, the likelihood $p(\mathcal{D}_\tau \mid \boldsymbol{\theta}_\tau)$ can be expressed by

$$p(\mathcal{D}_{\tau,N} \mid \boldsymbol{\theta}_\tau) = \prod_{n=1}^{N} p(y_{\tau,n} \mid \mathbf{x}_{\tau,n}, \boldsymbol{\theta}_\tau) \propto \exp\{-\frac{1}{2}\|\mathbf{y}_{\tau,N} - \mathbf{X}_{\tau,N}\boldsymbol{\theta}_\tau\|^2\}. \tag{53}$$

Assuming the prior $\boldsymbol{\theta}_\tau \mid \boldsymbol{\theta}_0 \sim \mathcal{N}(\boldsymbol{\theta}_0, \frac{1}{\gamma_b}\mathbf{I}_d)$, with γ_b being the prespecified weight of the prior or regularizer. The prior can be expressed by

$$p(\boldsymbol{\theta}_\tau \mid \boldsymbol{\theta}_0) \propto \exp\{-\frac{\gamma_b}{2}\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|^2\}. \tag{54}$$

Combining (52)-(54), the posterior distribution of the per-task parameter satisfies

$$\begin{aligned} p(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau, \boldsymbol{\theta}_0) &\propto \exp\{-\frac{1}{2}\|\mathbf{y}_{\tau,N} - \mathbf{X}_{\tau,N}\boldsymbol{\theta}_\tau\|^2 - \frac{\gamma_b}{2}\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|^2\} & (55)\\ &\propto \exp\{-\frac{1}{2}(\boldsymbol{\theta}_\tau - \mu_{\boldsymbol{\theta}_\tau,N})^\top \Sigma_{\boldsymbol{\theta}_\tau,N}^{-1}(\boldsymbol{\theta}_\tau - \mu_{\boldsymbol{\theta}_\tau,N})\} = \mathcal{N}(\mu_{\boldsymbol{\theta}_\tau,N}, \Sigma_{\boldsymbol{\theta}_\tau,N}) & (56)\\ &\text{with } \Sigma_{\boldsymbol{\theta}_\tau,N} = (\mathbf{X}_{\tau,N}^\top \mathbf{X}_{\tau,N} + \gamma_b \mathbf{I})^{-1},\ \ \mu_{\boldsymbol{\theta}_\tau,N} = \Sigma_{\boldsymbol{\theta}_\tau,N}(\mathbf{X}_{\tau,N}^\top \mathbf{y}_{\tau,N} + \gamma_b \boldsymbol{\theta}_0). & (57)\end{aligned}$$

If $p(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau, \boldsymbol{\theta}_0) \in \mathcal{Q}$, then $\hat{p}(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau, \boldsymbol{\theta}_0) = p(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau, \boldsymbol{\theta}_0)$, which holds in our analysis since $\mathcal{Q}$ is defined to be the set of Gaussian distributions. The empirical loss of BaMAML is

$$\begin{aligned} \mathcal{L}_{T,N}^{\mathrm{ba}}(\boldsymbol{\theta}_0) &:= \frac{1}{TN_2}\sum_{\tau=1}^{T}\Big[-\int \log p(\mathcal{D}_\tau^{\mathrm{val}} \mid \boldsymbol{\theta}_\tau)\hat{p}^{\mathrm{ba}}(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\mathrm{trn}}, \boldsymbol{\theta}_0)d\boldsymbol{\theta}_\tau\Big] \\ &\text{s.t.}\quad \hat{p}^{\mathrm{ba}}(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\mathrm{trn}}, \boldsymbol{\theta}_0) = \underset{q(\boldsymbol{\theta}_\tau)\in\mathcal{Q}}{\arg\min}\,\mathrm{KL}(q(\boldsymbol{\theta}_\tau)\|p(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\mathrm{trn}}, \boldsymbol{\theta}_0)) \end{aligned} \tag{58}$$

where $\hat{p}^{\mathrm{ba}}(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\mathrm{trn}}, \boldsymbol{\theta}_0) = p^{\mathrm{ba}}(\boldsymbol{\theta}_\tau \mid \mathcal{D}_\tau^{\mathrm{trn}}, \boldsymbol{\theta}_0) = \mathcal{N}(\mu_{\boldsymbol{\theta}_\tau,N_1}, \Sigma_{\boldsymbol{\theta}_\tau,N_1})$ is the solution of the inner problem. Therefore

$$\begin{aligned} \mathcal{L}_{T,N}^{\mathrm{ba}}(\boldsymbol{\theta}_0) &= \frac{1}{TN_2}\sum_{\tau=1}^{T} -\log p(\mathcal{D}_\tau^{\mathrm{val}} \mid \mathcal{D}_\tau^{\mathrm{trn}}, \boldsymbol{\theta}_0) \\ &= -\frac{1}{TN_2}\sum_{\tau=1}^{T}[\log p(\mathcal{D}_\tau^{\mathrm{val}}, \mathcal{D}_\tau^{\mathrm{trn}} \mid \boldsymbol{\theta}_0) - \log p(\mathcal{D}_\tau^{\mathrm{trn}} \mid \boldsymbol{\theta}_0)] \end{aligned}$$

where

$$\begin{aligned} \log p(\mathcal{D}_\tau^{\mathrm{val}}, \mathcal{D}_\tau^{\mathrm{trn}} \mid \boldsymbol{\theta}_0) - \log p(\mathcal{D}_\tau^{\mathrm{trn}} \mid \boldsymbol{\theta}_0) &= \log p(\mathbf{y}_{\tau,N}^{\mathrm{all}} \mid \mathbf{X}_{\tau,N}^{\mathrm{all}}, \boldsymbol{\theta}_0) - \log p(\mathbf{y}_{\tau,N_1}^{\mathrm{trn}} \mid \mathbf{X}_{\tau,N_1}^{\mathrm{trn}}, \boldsymbol{\theta}_0) \\ = -\frac{1}{2}\|\mathbf{y}_{\tau,N}^{\mathrm{all}} - \mathbf{X}_{\tau,N}^{\mathrm{all}}\boldsymbol{\theta}_0\|_{\Sigma_{y,N}^{-1}}^2 &+ \frac{1}{2}\|\mathbf{y}_{\tau,N_1}^{\mathrm{trn}} - \mathbf{X}_{\tau,N_1}^{\mathrm{trn}}\boldsymbol{\theta}_0\|_{\Sigma_{y,N_1}^{-1}}^2 \end{aligned} \tag{59}$$

with $\Sigma_{y,N}^{-1} = (\mathbf{I}_N + \gamma_b^{-1}\mathbf{X}_{\tau,N}\mathbf{X}_{\tau,N}^\top)^{-1}$. The last equation is because $p(\mathbf{y}_{\tau,N}^{\text{all}} \mid \mathbf{X}_{\tau,N}^{\text{all}}, \boldsymbol{\theta}_0)$ can be computed by

$$
\begin{aligned}
&p(\mathbf{y}_{\tau,N}^{\text{all}} \mid \mathbf{X}_{\tau,N}^{\text{all}}, \boldsymbol{\theta}_0) = \int p(\mathbf{y}_{\tau,N}^{\text{all}} \mid \mathbf{X}_{\tau,N}^{\text{all}}, \boldsymbol{\theta}_\tau) p(\boldsymbol{\theta}_\tau \mid \boldsymbol{\theta}_0) d\boldsymbol{\theta}_\tau \\
\propto &\int \exp\{-\frac{1}{2}\|\mathbf{y}_{\tau,N}^{\text{all}} - \mathbf{X}_{\tau,N}^{\text{all}}\boldsymbol{\theta}_\tau\|^2 - \frac{\gamma_b}{2}\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|^2\} d\boldsymbol{\theta}_\tau \\
= &\int \exp\{-\frac{1}{2}(\boldsymbol{\theta}_\tau - \mu_{\boldsymbol{\theta}_\tau})^\top \Sigma_{\boldsymbol{\theta}_\tau}^{-1}(\boldsymbol{\theta}_\tau - \mu_{\boldsymbol{\theta}_\tau}) + \frac{1}{2}\mu_{\boldsymbol{\theta}_\tau}^\top \Sigma_{\boldsymbol{\theta}_\tau}^{-1}\mu_{\boldsymbol{\theta}_\tau} - \frac{\gamma_b}{2}\boldsymbol{\theta}_0^\top\boldsymbol{\theta}_0 - \frac{1}{2}\mathbf{y}_{\tau,N}^{\text{all}\top}\mathbf{y}_{\tau,N}^{\text{all}}\} d\boldsymbol{\theta}_\tau \\
\propto &\exp\{-\frac{1}{2}(-\mu_{\boldsymbol{\theta}_\tau}^\top \Sigma_{\boldsymbol{\theta}_\tau}^{-1}\mu_{\boldsymbol{\theta}_\tau} + \gamma_b\boldsymbol{\theta}_0^\top\boldsymbol{\theta}_0 + \mathbf{y}_{\tau,N}^{\text{all}\top}\mathbf{y}_{\tau,N}^{\text{all}})\} = \exp\{-\frac{1}{2}\|\mathbf{y}_{\tau,N}^{\text{all}} - \mathbf{X}_{\tau,N}^{\text{all}}\boldsymbol{\theta}_0\|_{\Sigma_{y,N}^{-1}}^2\}
\end{aligned}
\tag{60}
$$

where the last equation follows from the Binomial inverse theorem. Similarly, $p(\mathbf{y}_{\tau,N}^{\text{trn}} \mid \mathbf{X}_{\tau,N}^{\text{trn}}, \boldsymbol{\theta}_0) \propto \exp\{-\frac{1}{2}\|\mathbf{y}_{\tau,N_1}^{\text{trn}} - \mathbf{X}_{\tau,N_1}^{\text{trn}}\boldsymbol{\theta}_0\|_{\Sigma_{y,N_1}^{-1}}^2\}$.

To solve for $\hat{\boldsymbol{\theta}}_0^{\text{ba}}$, using the optimality condition, we obtain

$$
\hat{\boldsymbol{\theta}}_0^{\text{ba}} = \Big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\text{ba}}\Big)^{-1}\Big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\text{ba}}\boldsymbol{\theta}_\tau^{\text{gt}}\Big) + \Delta_T^{\text{ba}} \tag{61a}
$$

$$
\Delta_T^{\text{ba}} = \Big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\text{ba}}\Big)^{-1}\frac{1}{N_2}\Big(\sum_{\tau=1}^{T}\mathbf{X}_{\tau,N}^{\text{all}\top}\Sigma_{y,N}^{-1}\mathbf{e}_{\tau,N}^{\text{all}} - \mathbf{X}_{\tau,N_1}^{\text{trn}\top}\Sigma_{y,N_1}^{-1}\mathbf{e}_{\tau,N_1}^{\text{trn}}\Big) \tag{61b}
$$

$$
\hat{\mathbf{W}}_{\tau,N}^{\text{ba}} = \Big((\frac{\gamma_b}{N})^{-1}\hat{\mathbf{Q}}_{\tau,N} + \mathbf{I}\Big)^{-1}\hat{\mathbf{Q}}_{\tau,N_2}\Big((\frac{\gamma_b}{N_1})^{-1}\hat{\mathbf{Q}}_{\tau,N_1} + \mathbf{I}\Big)^{-1} \tag{61c}
$$

$$
= \Big((\gamma s)^{-1}\hat{\mathbf{Q}}_{\tau,N} + \mathbf{I}\Big)^{-1}\hat{\mathbf{Q}}_{\tau,N_2}\Big(\gamma^{-1}\hat{\mathbf{Q}}_{\tau,N_1} + \mathbf{I}\Big)^{-1} \tag{61d}
$$

where the last equation is because we choose $\gamma_b = N_1\gamma$ for a fair comparison with iMAML.

Based on (2), the BaMAML meta-test risk is defined as

$$
\begin{aligned}
&\mathcal{R}_{N_a}^{\text{ba}}(\boldsymbol{\theta}_0) = \frac{1}{N}\mathbb{E}\big[-\log p(\mathbf{y}_{\tau,N} \mid \mathbf{X}_{\tau,N}, \mathcal{D}_{\tau,N_a}, \boldsymbol{\theta}_0)\big] \\
=&\mathbb{E}_\tau\Big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau^{\text{gt}}\|_{\mathbf{W}_{\tau,N_a}^{\text{ba}}}^2\Big] + 1 + \frac{1}{N_a}\mathbb{E}\big[\text{tr}\big((\mathbf{I}_d + (\gamma s)^{-1}\hat{\mathbf{Q}}_{\tau,N+N_a})^{-1} - (\mathbf{I}_d + \gamma^{-1}\hat{\mathbf{Q}}_{\tau,N_a})^{-1}\big)\big]
\end{aligned}
\tag{62a}
$$

$$
\begin{aligned}
\mathbf{W}_{\tau,N_a}^{\text{ba}} &= \mathbb{E}_{\mathbf{x}_\tau}\big[\frac{\gamma_b}{N}[(\mathbf{I}_d + \gamma_b^{-1}N_a\hat{\mathbf{Q}}_{\tau,N_a})^{-1} - (\mathbf{I}_d + \gamma_b^{-1}(N + N_a)\hat{\mathbf{Q}}_{\tau,N+N_a})^{-1}]\big] \\
&= \mathbb{E}_{\mathbf{x}_\tau}\big[\big((\gamma s)^{-1}\hat{\mathbf{Q}}_{\tau,N+N_a} + \mathbf{I}\big)^{-1}\hat{\mathbf{Q}}_{\tau,N}\big(\gamma^{-1}\hat{\mathbf{Q}}_{\tau,N_a} + \mathbf{I}\big)^{-1}\big]
\end{aligned}
\tag{62b}
$$

Taking limits of $\mathcal{R}_{N_a}^{\text{ba}}$ w.r.t. N_a further leads to

$$
\mathcal{R}^{\text{ba}}(\boldsymbol{\theta}_0) \coloneqq \lim_{N_a\to\infty}\mathcal{R}_{N_a}^{\text{ba}}(\boldsymbol{\theta}_0) = \lim_{N_a\to\infty}\mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau^{\text{gt}}\|_{\mathbf{W}_{\tau,N_a}^{\text{ba}}}^2\big] + 1 = \mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau^{\text{gt}}\|_{\mathbf{W}_\tau^{\text{ba}}}^2\big] + 1 \tag{63}
$$

$\boldsymbol{\theta}_0^{\text{ba}}$ is defined to be the minimizer of $\mathcal{R}^{\text{ba}}(\boldsymbol{\theta}_0)$, given by

$$
\boldsymbol{\theta}_0^{\text{ba}} = \arg\min_{\boldsymbol{\theta}_0}\mathcal{R}^{\text{ba}}(\boldsymbol{\theta}_0) = \arg\min_{\boldsymbol{\theta}_0}\mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau^{\text{gt}}\|_{\mathbf{W}_\tau^{\text{ba}}}^2\big] = \mathbb{E}_\tau\big[\mathbf{W}_\tau^{\text{ba}}\big]^{-1}\mathbb{E}_\tau\big[\mathbf{W}_\tau^{\text{ba}}\boldsymbol{\theta}_\tau^{\text{gt}}\big], \tag{64a}
$$

$$
\text{with } \mathbf{W}_\tau^{\text{ba}} = \big((\gamma s)^{-1}\mathbf{Q}_\tau + \mathbf{I}\big)^{-1}\mathbf{Q}_\tau\big(\gamma^{-1}\mathbf{Q}_\tau + \mathbf{I}\big)^{-1}. \tag{64b}
$$

It is worth noting that, from Lemma 1, we have the property $\mathbb{E}_{\mathbf{x}_\tau}[\hat{\mathbf{W}}_{\tau,N}^{\text{ba}}] = \mathbf{W}_{\tau,N}^{\text{ba}}$, $\lim_{N\to\infty}\mathbf{W}_{\tau,N}^{\text{ba}} = \mathbf{W}_\tau^{\text{ba}}$, 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. Next we analyze the optimal population risk and the statistical error based on the solutions.

B Optimal population risk and statistical error analysis

Meta-test risk decomposition. Recall the meta-test risk function for the method $\mathcal{A}$ of $\boldsymbol{\theta}_0$ in (2). For method $\mathcal{A}$, plugging $\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}$ into (2) and taking the limit over N_a, the number of data during adaptation, we have

$$\begin{aligned}
\mathcal{R}^{\mathcal{A}}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}) &= \mathbb{E}_\tau\left[\|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_\tau^{\mathrm{gt}}\|^2_{\mathbf{W}_\tau^{\mathcal{A}}}\right] + 1 \overset{(a)}{=} \mathbb{E}_\tau\left[\|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}} + \boldsymbol{\theta}_0^{\mathcal{A}} - \boldsymbol{\theta}_\tau^{\mathrm{gt}}\|^2_{\mathbf{W}_\tau^{\mathcal{A}}}\right] + 1 \\
&= \mathbb{E}_\tau\left[\|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}}\|^2_{\mathbf{W}_\tau^{\mathcal{A}}} + \|\boldsymbol{\theta}_0^{\mathcal{A}} - \boldsymbol{\theta}_\tau^{\mathrm{gt}}\|^2_{\mathbf{W}_\tau^{\mathcal{A}}} + 2(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}})^\top \mathbf{W}_\tau^{\mathcal{A}}(\boldsymbol{\theta}_0^{\mathcal{A}} - \boldsymbol{\theta}_\tau^{\mathrm{gt}})\right] + 1 \\
&\overset{(b)}{=} \mathbb{E}_\tau\left[\|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}}\|^2_{\mathbf{W}_\tau^{\mathcal{A}}} + \|\boldsymbol{\theta}_0^{\mathcal{A}} - \boldsymbol{\theta}_\tau^{\mathrm{gt}}\|^2_{\mathbf{W}_\tau^{\mathcal{A}}} + 2(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}})^\top \mathbf{W}_\tau^{\mathcal{A}}(\mathbb{E}_\tau\left[\mathbf{W}_\tau^{\mathcal{A}}\right]^{-1} \mathbb{E}_\tau\left[\mathbf{W}_\tau^{\mathcal{A}}\boldsymbol{\theta}_\tau^{\mathrm{gt}}\right] - \boldsymbol{\theta}_\tau^{\mathrm{gt}})\right] + 1 \\
&\overset{(c)}{=} \mathbb{E}_\tau\left[\|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}}\|^2_{\mathbf{W}_\tau^{\mathcal{A}}} + \|\boldsymbol{\theta}_0^{\mathcal{A}} - \boldsymbol{\theta}_\tau^{\mathrm{gt}}\|^2_{\mathbf{W}_\tau^{\mathcal{A}}}\right] + 1
\end{aligned} \tag{65}$$

where (a) follows from $\mathbf{W}_\tau^{\mathcal{A}} = \lim_{N_a\to\infty} \mathbf{W}_{\tau,N_a}^{\mathcal{A}}$, (b) is by plugging $\boldsymbol{\theta}_0^{\mathcal{A}} = \mathbb{E}_\tau\left[\mathbf{W}_\tau^{\mathcal{A}}\right]^{-1}\mathbb{E}_\tau\left[\mathbf{W}_\tau^{\mathcal{A}}\boldsymbol{\theta}_\tau^{\mathrm{gt}}\right]$, (c) is because $\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}(\mathbb{E}_\tau\left[\mathbf{W}_\tau^{\mathcal{A}}\right]^{-1}\mathbb{E}_\tau\left[\mathbf{W}_\tau^{\mathcal{A}}\boldsymbol{\theta}_\tau^{\mathrm{gt}}\right] - \boldsymbol{\theta}_\tau^{\mathrm{gt}})] = \mathbb{E}_\tau\left[\mathbf{W}_\tau^{\mathcal{A}}\boldsymbol{\theta}_\tau^{\mathrm{gt}}\right] - \mathbb{E}_\tau\left[\mathbf{W}_\tau^{\mathcal{A}}\boldsymbol{\theta}_\tau^{\mathrm{gt}}\right] = 0$. From (65) and the definition of $\mathcal{R}_{N_a}^{\mathcal{A}}(\cdot)$ in (2), we can decompose the meta-test risk to the optimal population risk and statistical error as follows

$$\lim_{N_a\to\infty} \mathcal{R}_{N_a}^{\mathcal{A}}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}) = \underbrace{\lim_{N_a\to\infty} \mathcal{R}_{N_a}^{\mathcal{A}}(\boldsymbol{\theta}_0^{\mathcal{A}})}_{\text{optimal population risk}} + \underbrace{\|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}}\|^2_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]}}_{\text{statistical error } \mathcal{E}^2_{\mathcal{A}}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}})}. \tag{66}$$

This completes the proof of Proposition 3 in the main paper. Note that, the statistical error $\mathcal{E}^2_{\mathcal{A}}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}})$ is resulted from finite random data samples during meta-training to obtain the estimation of the parameter $\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}$, but not from N_a in (66), which is the number of adaptation data during meta-testing.

B.1 Optimal population risk

The optimal population risk under different methods is given by $\mathcal{R}^{\mathcal{A}}(\boldsymbol{\theta}_0^{\mathcal{A}}) = \min_{\boldsymbol{\theta}_0} \mathcal{R}^{\mathcal{A}}(\boldsymbol{\theta}_0)$. Based on the results in Section A, we compute the optimal population risk of each method.

For **ERM**, the optimal population risk is computed by

$$\mathcal{R}^{\mathrm{er}}(\boldsymbol{\theta}_0^{\mathrm{er}}) = \mathbb{E}_\tau\left[\|\boldsymbol{\theta}_0^{\mathrm{er}} - \boldsymbol{\theta}_\tau^{\mathrm{gt}}\|^2_{\mathbf{W}_\tau^{\mathrm{er}}}\right] + 1. \tag{67}$$

For **MAML**, the optimal population risk is computed by

$$\mathcal{R}^{\mathrm{ma}}(\boldsymbol{\theta}_0^{\mathrm{ma}}, \alpha) = \mathbb{E}_\tau\left[\|\boldsymbol{\theta}_0^{\mathrm{ma}} - \boldsymbol{\theta}_\tau^{\mathrm{gt}}\|^2_{\mathbf{W}_\tau^{\mathrm{ma}}(\alpha)}\right] + 1. \tag{68}$$

Note that when $\alpha = 0$, $\mathcal{R}^{\mathrm{ma}}(\boldsymbol{\theta}_0^{\mathrm{ma}}, \alpha) = \mathcal{R}^{\mathrm{er}}(\boldsymbol{\theta}_0^{\mathrm{er}})$.

Comparison of ERM and MAML optimal population risk. To compare the optimal population risk of ERM and MAML, as shown in (Gao and Sener, 2020), when $\|\mathbf{Q}_\tau\| \leq \bar{\lambda}, 0 < \alpha \leq 1/\bar{\lambda}$, $\mathcal{R}^{\mathrm{ma}}(\boldsymbol{\theta}_0^{\mathrm{ma}}, \alpha)$ is monotonically decreasing. Therefore $\mathcal{R}^{\mathrm{ma}}(\boldsymbol{\theta}_0^{\mathrm{ma}}, \alpha) < \mathcal{R}^{\mathrm{er}}(\boldsymbol{\theta}_0^{\mathrm{er}})$ when $0 < \alpha \leq 1/\bar{\lambda}$.

For **iMAML**, the optimal population risk is computed by

$$\mathcal{R}^{\mathrm{im}}(\boldsymbol{\theta}_0^{\mathrm{im}}, \gamma) = \mathbb{E}_\tau\left[\|\boldsymbol{\theta}_0^{\mathrm{im}} - \boldsymbol{\theta}_\tau^{\mathrm{gt}}\|^2_{\mathbf{W}_\tau^{\mathrm{im}}(\gamma)}\right] + 1. \tag{69}$$

Comparison of MAML and iMAML optimal population risk. We can see that as $\gamma \to \infty, \mathbf{W}_\tau^{\mathrm{im}}(\gamma) \to \mathbf{W}_\tau^{\mathrm{er}} = \mathbf{Q}_\tau, \mathcal{R}^{\mathrm{im}}(\boldsymbol{\theta}_0, \gamma) \to \mathcal{R}^{\mathrm{er}}(\boldsymbol{\theta}_0)$; as $\gamma \to 0, \mathbf{W}_\tau^{\mathrm{im}}(\gamma) \to 0, \mathcal{R}^{\mathrm{im}}(\boldsymbol{\theta}_0) \to 1$. To explicitly compare the optimal population risk of MAML and iMAML, we will show next that when γ takes certain values, the corresponding risks satisfy $\mathcal{R}^{\mathrm{im}}(\boldsymbol{\theta}_0^{\mathrm{im}}, \gamma) < \mathcal{R}^{\mathrm{ma}}(\boldsymbol{\theta}_0^{\mathrm{ma}}, \alpha)$.

Corollary 2 *Based on Assumption 1, for any $\tau \sim p(\mathcal{T})$, and $0 < \alpha < 1/\bar{\lambda}$, $\|\mathbf{W}_\tau^{\mathrm{ma}}\| > 0$. And generally in a typical multi-task learning setting, the tasks are not all identical, therefore there exist $\tau \sim p(\mathcal{T}), \tau \in T, \boldsymbol{\theta}_0^{\mathcal{A}} - \boldsymbol{\theta}_\tau^{\mathrm{gt}} \neq \mathbf{0}$. Therefore there exists $\tau \in T$ such that $(\boldsymbol{\theta}_0^{\mathrm{ma}} - \boldsymbol{\theta}_\tau^{\mathrm{gt}})^\top \mathbf{W}_\tau^{\mathrm{ma}}(\boldsymbol{\theta}_0^{\mathrm{ma}} - \boldsymbol{\theta}_\tau^{\mathrm{gt}}) > 0$.*

First we show that the minimum value of the MAML population risk is larger than 1, i.e., $\mathcal{R}^{\rm ma}(\boldsymbol{\theta}_0^{\rm ma},\alpha) > 1$. According to Corollary 2, it is apparent that

$$\mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0^{\rm ma}-\boldsymbol{\theta}_\tau^{\rm gt}\|^2_{\mathbf{W}_\tau^{\rm ma}}\big] = \mathbb{E}_\tau\big[(\boldsymbol{\theta}_0^{\rm ma}-\boldsymbol{\theta}_\tau^{\rm gt})^\top\mathbf{W}_\tau^{\rm ma}(\boldsymbol{\theta}_0^{\rm ma}-\boldsymbol{\theta}_\tau^{\rm gt})\big] > 0. \tag{70}$$

Therefore, we have

$$\mathcal{R}^{\rm ma}(\boldsymbol{\theta}_0^{\rm ma},\alpha) = \mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0^{\rm ma}-\boldsymbol{\theta}_\tau^{\rm gt}\|^2_{\mathbf{W}_\tau^{\rm ma}}\big] + 1 > 1. \tag{71}$$

Note that $\mathcal{R}^{\rm ma}(\boldsymbol{\theta}_0^{\rm ma},\alpha)$ also depends on α, Let $r^{\rm ma} = \min_\alpha \mathcal{R}^{\rm ma}(\boldsymbol{\theta}_0^{\rm ma},\alpha) - 1$. From (71) we know that $r^{\rm ma} > 0$. We will then show one can always find certain γ such that

$$\mathcal{R}^{\rm im}(\boldsymbol{\theta}_0^{\rm im},\gamma) < \min_\alpha \mathcal{R}^{\rm ma}(\boldsymbol{\theta}_0^{\rm ma},\alpha) = r^{\rm ma} + 1 \tag{72}$$

or equivalently $\mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0^{\rm im}-\boldsymbol{\theta}_\tau^{\rm gt}\|^2_{\mathbf{W}_\tau^{\rm im}}\big] < r^{\rm ma}$.

From Assumption 1, bounded eigenvalues of per-task data matrix, we can derive

$$\|(\gamma^{-1}\mathbf{Q}_\tau+\mathbf{I})^{-1}\| = 1/\lambda_{\min}(\gamma^{-1}\mathbf{Q}_\tau+\mathbf{I}) = 1/(\gamma^{-1}\lambda_{\min}(\mathbf{Q}_\tau)+1) \le \frac{1}{\gamma^{-1}\underline{\lambda}+1} \tag{73}$$

from which we can bound the optimal population risk of iMAML by

$$\mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0^{\rm im}-\boldsymbol{\theta}_\tau^{\rm gt}\|^2_{\mathbf{W}_\tau^{\rm im}}\big] \le \mathbb{E}_\tau[\|\boldsymbol{\theta}_0^{\rm im}-\boldsymbol{\theta}_\tau^{\rm gt}\|^2]\mathrm{sup}_\tau\|\mathbf{W}_\tau^{\rm im}\|$$

where we have discussed the bound for $\mathrm{sup}_\tau\|\mathbf{W}_\tau^{\rm im}\|$ based on Assumption 1. Thus it suffices to bound $\mathbb{E}_\tau[\|\boldsymbol{\theta}_0^{\rm im}-\boldsymbol{\theta}_\tau^{\rm gt}\|^2]$, as follows

$$\begin{aligned}
&\mathbb{E}_\tau[\|\boldsymbol{\theta}_0^{\rm im}-\boldsymbol{\theta}_\tau^{\rm gt}\|^2] = \|\boldsymbol{\theta}_0^{\rm im}\|^2 - 2\boldsymbol{\theta}_0^{\rm im\top}\mathbb{E}_\tau[\boldsymbol{\theta}_\tau^{\rm gt}] + \mathrm{tr}(\mathrm{Cov}_\tau[\boldsymbol{\theta}_\tau^{\rm gt}]) + \|\mathbb{E}_\tau[\boldsymbol{\theta}_\tau^{\rm gt}]\|^2\\
\le&\|\boldsymbol{\theta}_0^{\rm im}\|^2 + 2\|\boldsymbol{\theta}_0^{\rm im}\|M + \mathrm{tr}(\mathrm{Cov}_\tau[\boldsymbol{\theta}_\tau^{\rm gt}]) + M^2 \le \|\boldsymbol{\theta}_0^{\rm im}\|^2 + 2\|\boldsymbol{\theta}_0^{\rm im}\|M + \mathrm{tr}(\mathrm{Cov}_\tau[\boldsymbol{\theta}_\tau^{\rm gt}]) + M^2\\
\le&(M+\|\boldsymbol{\theta}_0^{\rm im}\|)^2 + \frac{R^2}{d}\cdot d = (M+\|\boldsymbol{\theta}_0^{\rm im}\|)^2 + R^2
\end{aligned}$$

where the last inequality follows from Assumption 2 that the task parameter distribution is sub-gaussian. Similarly $\mathbb{E}_\tau[\|\boldsymbol{\theta}_0^{\rm im}-\boldsymbol{\theta}_\tau^{\rm gt}\|^2] \le (M+\|\boldsymbol{\theta}_0^{\rm ma}\|)^2 + R^2$. Therefore

$$\begin{aligned}
&\mathbb{E}_\tau\Big[\|\boldsymbol{\theta}_0^{\rm im}-\boldsymbol{\theta}_\tau^{\rm gt}\|^2_{\mathbf{W}_\tau^{\rm im}}\Big] \le \mathbb{E}_\tau[\|\boldsymbol{\theta}_0^{\rm im}-\boldsymbol{\theta}_\tau^{\rm gt}\|^2]\mathrm{sup}_\tau\|\mathbf{W}_\tau^{\rm im}\|\\
\le&\big((M+\|\boldsymbol{\theta}_0^{\rm im}\|)^2+R^2\big)\mathbb{E}_\tau\Big[\|(\gamma^{-1}\mathbf{Q}_\tau+\mathbf{I})^{-1}\mathbf{Q}_\tau(\gamma^{-1}\mathbf{Q}_\tau+\mathbf{I})^{-1}\|\Big]\\
\le&\big((M+\|\boldsymbol{\theta}_0^{\rm im}\|)^2+R^2\big)\mathbb{E}_\tau\Big[\|(\gamma^{-1}\mathbf{Q}_\tau+\mathbf{I})^{-1}\|\|\mathbf{Q}_\tau\|\|(\gamma^{-1}\mathbf{Q}_\tau+\mathbf{I})^{-1}\|\Big]\\
\overset{(a)}{\le}&\frac{\big((M+\|\boldsymbol{\theta}_0^{\rm im}\|)^2+R^2\big)\bar{\lambda}}{(\gamma^{-1}\underline{\lambda}+1)^2}
\end{aligned} \tag{74}$$

where (a) holds because $\|\mathbf{Q}_\tau\| \le \bar{\lambda}$, $\|(\gamma^{-1}\mathbf{Q}_\tau+\mathbf{I})^{-1}\| \le (\gamma^{-1}\underline{\lambda}+1)^{-1}$ from Assumption 1.

Let $C_{\boldsymbol{\theta}} = \max\{\big((M+\|\boldsymbol{\theta}_0^{\rm im}\|)^2+R^2\big)^{\frac{1}{2}}, \big((M+\|\boldsymbol{\theta}_0^{\rm ma}\|)^2+R^2\big)^{\frac{1}{2}}\}$. In order to ensure $\big((M+\|\boldsymbol{\theta}_0^{\rm im}\|)^2+R^2\big)\bar{\lambda}\frac{1}{(\gamma^{-1}\underline{\lambda}+1)^2} \le C_{\boldsymbol{\theta}}^2\bar{\lambda}\frac{1}{(\gamma^{-1}\underline{\lambda}+1)^2} < r^{\rm ma}$ it suffices to ensure

$$\gamma^{-1}\underline{\lambda}+1 > (r^{\rm ma})^{-\frac{1}{2}}C_{\boldsymbol{\theta}}\bar{\lambda}^{\frac{1}{2}}. \tag{75}$$

Since $0 < r^{\rm ma} = \mathbb{E}_\tau\big[\|\boldsymbol{\theta}_0^{\rm ma}-\boldsymbol{\theta}_\tau^{\rm gt}\|^2_{\mathbf{W}_\tau^{\rm ma}}\big] \le C_{\boldsymbol{\theta}}^2\mathbb{E}_\tau[\|\mathbf{W}_\tau^{\rm ma}\|] < C_{\boldsymbol{\theta}}^2\bar{\lambda}$. It follows that

$$(r^{\rm ma})^{-\frac{1}{2}}C_{\boldsymbol{\theta}}\bar{\lambda}^{\frac{1}{2}} - 1 > 0. \tag{76}$$

Then from (75) and (76) one can derive

$$0 < \gamma < \big((r^{\rm ma})^{-\frac{1}{2}}C_{\boldsymbol{\theta}}\bar{\lambda}^{\frac{1}{2}} - 1\big)^{-1}\underline{\lambda}. \tag{77}$$

In other words, by choosing (77), we have $\mathcal{R}^{\rm im}(\boldsymbol{\theta}_0^{\rm im},\gamma) < \mathcal{R}^{\rm ma}(\boldsymbol{\theta}_0^{\rm ma},\alpha) < \mathcal{R}^{\rm er}(\boldsymbol{\theta}_0^{\rm er}), \forall 0 < \alpha \le 1/\bar{\lambda}$. We summarize this conclusion in Theorem 7 below.

Theorem 7 (iMAML has lower optimal population risk than MAML) *Under Assumptions 1-2, for meta-test task τ and arbitrary $\boldsymbol{\theta}$, the population risks for MAML and iMAML, as functions of $\boldsymbol{\theta}$, are*

$$\mathcal{R}^{\mathrm{ma}}(\boldsymbol{\theta},\alpha)\equiv\mathbb{E}_{\tau}\left[\|\boldsymbol{\theta}-\boldsymbol{\theta}_{\tau}\|^2_{\mathbf{W}_{\tau}^{\mathrm{ma}}(\alpha)}\right]+1 \quad \text{and} \quad \mathcal{R}^{\mathrm{im}}(\boldsymbol{\theta},\gamma)\equiv\mathbb{E}_{\tau}\left[\|\boldsymbol{\theta}-\boldsymbol{\theta}_{\tau}\|^2_{\mathbf{W}_{\tau}^{\mathrm{im}}(\gamma)}\right]+1$$

where $\mathbf{W}_{\tau}^{\mathrm{ma}}(\alpha)=(\mathbf{I}-\alpha\mathbf{Q}_{\tau})\mathbf{Q}_{\tau}(\mathbf{I}-\alpha\mathbf{Q}_{\tau})$ and $\mathbf{W}_{\tau}^{\mathrm{im}}(\gamma)=(\gamma^{-1}\mathbf{Q}_{\tau}+\mathbf{I})^{-1}\mathbf{Q}_{\tau}(\gamma^{-1}\mathbf{Q}_{\tau}+\mathbf{I})^{-1}$. And the two functions are minimized by $\boldsymbol{\theta}_0^{\mathrm{ma}}$ and $\boldsymbol{\theta}_0^{\mathrm{im}}$ respectively. Let $r^{\mathrm{ma}}=\min_{\alpha}\mathcal{R}^{\mathrm{ma}}(\boldsymbol{\theta}_0^{\mathrm{ma}},\alpha)-c>0$, and when $0<\gamma<\left((r^{\mathrm{ma}})^{-\frac{1}{2}}C_{\boldsymbol{\theta}}\bar{\lambda}^{\frac{1}{2}}-1\right)^{-1}\underline{\lambda}$, then $\mathcal{R}^{\mathrm{im}}(\boldsymbol{\theta}_0^{\mathrm{im}},\gamma)<\mathcal{R}^{\mathrm{ma}}(\boldsymbol{\theta}_0^{\mathrm{ma}},\alpha)$.

For **BaMAML**, the optimal population risk is

$$\mathcal{R}^{\mathrm{ba}}(\boldsymbol{\theta}_0^{\mathrm{ba}},\gamma)=\mathbb{E}_{\tau}\left[\|\boldsymbol{\theta}_0^{\mathrm{ba}}-\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}\|^2_{\mathbf{W}_{\tau}^{\mathrm{ba}}(\gamma)}\right]+1 \tag{78}$$

Similar to the proof for Theorem 7, BaMAML also has lower optimal population risk than MAML, as stated in the following theorem.

Theorem 8 (BaMAML has lower optimal population risk than MAML) *Under Assumptions 1-2, for meta-test task τ and arbitrary $\boldsymbol{\theta}$, the optimal population risk for BaMAML, as functions of $\boldsymbol{\theta}$, is*

$$\mathcal{R}^{\mathrm{ba}}(\boldsymbol{\theta},\gamma)=\mathbb{E}_{\tau}\left[\|\boldsymbol{\theta}-\boldsymbol{\theta}_{\tau}\|^2_{\mathbf{W}_{\tau}^{\mathrm{ba}}(\gamma)}\right]+1.$$

And when $0<\gamma<\left((r^{\mathrm{ma}})^{-\frac{1}{2}}C_{\boldsymbol{\theta}}\bar{\lambda}^{\frac{1}{2}}-1\right)^{-1}\underline{\lambda}$, then $\mathcal{R}^{\mathrm{ba}}(\boldsymbol{\theta}_0^{\mathrm{ba}},\gamma)<\mathcal{R}^{\mathrm{ma}}(\boldsymbol{\theta}_0^{\mathrm{ma}},\alpha)$.

B.2 Statistical error

To analyze the statistical error of different meta learning methods, we begin with a looser bound under data agnostic case with Assumptions 1 and 2 only. And to give a sharper analysis of the statistical error in order to make a fair comparison among different methods, we further make Assumption 3 on the task and data distributions. In the following sections, we will first present the supporting lemmas and then the main results for different methods.

B.2.1 Supporting Lemmas

In this section, we present some supporting lemmas for the proof of the main results for statistical errors of different methods.

Lemma 1 *Suppose Assumptions 1-2 hold. Define $\mathbf{W}_{\tau,N}^{\mathcal{A}}:=\mathbb{E}_{\mathbf{x}_{\tau}}[\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}]$, then*

$$\|\mathbf{W}_{\tau,N}^{\mathcal{A}}\|\leq\|\mathbf{W}_{\tau}^{\mathcal{A}}\|+L^{\mathcal{A}}\Big(\widetilde{\mathcal{O}}(\frac{d}{N})+\widetilde{\mathcal{O}}(\sqrt{\frac{d}{N}})\Big).$$

Proof: For ERM, $\mathbf{W}_{\tau,N}^{\mathrm{er}}:=\mathbb{E}_{\mathbf{x}_{\tau}}[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{er}}]=\mathbb{E}[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{er}}]=\mathbb{E}[\hat{\mathbf{Q}}_{\tau,N}]=\mathbf{Q}_{\tau}=\mathbf{W}_{\tau}^{\mathrm{er}}$, $L^{\mathrm{er}}=0$.

For MAML, from (38b), we have

$$\begin{aligned}\mathbf{W}_{\tau,N}^{\mathrm{ma}}&=\mathbb{E}_{\hat{\mathbf{Q}}_{\tau,N}}\left[(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N})\mathbf{Q}_{\tau}(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N})\right]\\&=\mathbf{W}_{\tau}^{\mathrm{ma}}+\frac{\alpha^2}{N}\Big(\mathbb{E}_{\mathbf{x}_{\tau,i}}\left[\mathbf{x}_{\tau,i}\mathbf{x}_{\tau,i}^{\top}\mathbf{Q}_{\tau}\mathbf{x}_{\tau,i}\mathbf{x}_{\tau,i}^{\top}\right]-\mathbf{Q}_{\tau}^3\Big)\end{aligned}$$

therefore

$$\begin{aligned}\|\mathbf{W}_{\tau,N}^{\mathrm{ma}}\|&\leq\|\mathbf{W}_{\tau}^{\mathrm{ma}}\|+\frac{\alpha^2}{N}\Big\|\mathbb{E}_{\mathbf{x}_{\tau,i}}\left[\mathbf{x}_{\tau,i}\mathbf{x}_{\tau,i}^{\top}\mathbf{Q}_{\tau}\mathbf{x}_{\tau,i}\mathbf{x}_{\tau,i}^{\top}\right]-\mathbf{Q}_{\tau}^3\Big\|\\&=\|\mathbf{W}_{\tau}^{\mathrm{ma}}\|+L^{\mathrm{ma}}\Big(\widetilde{\mathcal{O}}(\frac{d}{N})+\widetilde{\mathcal{O}}(\sqrt{\frac{d}{N}})\Big).\end{aligned}$$

For iMAML, recall $\mathbf{W}^{\mathrm{im}}_{\tau,N} = \gamma^2\Sigma_{\boldsymbol{\theta}_\tau}\mathbf{Q}_\tau\Sigma_{\boldsymbol{\theta}_\tau}$, let $I_0 = \gamma\Sigma_{\theta_\tau} - (\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)^{-1}$, further derived as

$$\begin{aligned}I_0 =&\gamma\Sigma_{\theta_\tau} - (\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)^{-1} = (\mathbf{I}+\gamma^{-1}\hat{\mathbf{Q}}_\tau)^{-1} - (\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)^{-1}\\ =&\gamma^{-1}(\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)^{-1}(\mathbf{Q}_\tau-\hat{\mathbf{Q}}_\tau)(\mathbf{I}+\gamma^{-1}\hat{\mathbf{Q}}_\tau)^{-1}\end{aligned}$$

Then we have

$$\begin{aligned}\mathbf{W}^{\mathrm{im}}_{\tau,N} =&\mathbb{E}_{\mathbf{x}_\tau}[\hat{\mathbf{W}}^{\mathrm{im}}_{\tau,N}] = \mathbb{E}_{\mathbf{x}_\tau}[\gamma\Sigma_{\theta_\tau}\mathbf{Q}_\tau\gamma\Sigma_{\theta_\tau}]\\ =&\mathbb{E}_{\mathbf{x}_\tau}\Big[\big(\gamma\Sigma_{\theta_\tau} + (\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)^{-1} - (\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)^{-1}\big)\mathbf{Q}_\tau\big(\gamma\Sigma_{\theta_\tau} + (\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)^{-1} - (\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)^{-1}\big)\Big]\\ =&\mathbb{E}_{\mathbf{x}_\tau}\Big[(\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)^{-1}\mathbf{Q}_\tau(\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)^{-1}\Big] + \mathbb{E}_{\mathbf{x}_\tau}\Big[I_0\mathbf{Q}_\tau I_0\Big] + \mathbb{E}_{\mathbf{x}_\tau}\Big[I_0\mathbf{Q}_\tau(\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)^{-1} + (\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)^{-1}\mathbf{Q}_\tau I_0\Big]\\ =&\mathbf{W}^{\mathrm{im}}_\tau + \mathbb{E}_{\mathbf{x}_\tau}\Big[I_0\mathbf{Q}_\tau I_0\Big] + \mathbb{E}_{\mathbf{x}_\tau}\Big[I_0(\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)\mathbf{W}^{\mathrm{im}}_{\tau,N} + \mathbf{W}^{\mathrm{im}}_{\tau,N}(\mathbf{I}+\gamma^{-1}\mathbf{Q}_\tau)I_0\Big]\\ =&\mathbf{W}^{\mathrm{im}}_\tau + \mathbb{E}_{\mathbf{x}_\tau}\Big[\Sigma_{\theta_\tau}(\mathbf{Q}_\tau-\hat{\mathbf{Q}}_\tau)\mathbf{W}^{\mathrm{im}}_\tau(\mathbf{Q}_\tau-\hat{\mathbf{Q}}_\tau)\Sigma_{\theta_\tau} + \Sigma_{\theta_\tau}(\hat{\mathbf{Q}}_\tau-\mathbf{Q}_\tau)\mathbf{W}^{\mathrm{im}}_\tau + \mathbf{W}^{\mathrm{im}}_\tau(\mathbf{Q}_\tau-\hat{\mathbf{Q}}_\tau)\Sigma_{\theta_\tau}\Big]\end{aligned}$$

where because

$$\|\mathbb{E}_{\mathbf{x}_\tau}[\Sigma_{\boldsymbol{\theta}_\tau}(\mathbf{Q}_\tau-\hat{\mathbf{Q}}_\tau)\mathbf{W}^{\mathrm{im}}_\tau]\| \le \mathbb{E}_{\mathbf{x}_\tau}[\|\Sigma_{\boldsymbol{\theta}_\tau}\|\|\mathbf{Q}_\tau-\hat{\mathbf{Q}}_\tau\|]\|\mathbf{W}^{\mathrm{im}}_\tau\| \le \mathbb{E}_{\mathbf{x}_\tau}[\|\mathbf{Q}_\tau-\hat{\mathbf{Q}}_\tau\|]\|\mathbf{W}^{\mathrm{im}}_\tau\|$$

and $\|\mathbb{E}_{\mathbf{x}_\tau}[\Sigma_{\boldsymbol{\theta}_\tau}(\mathbf{Q}_\tau-\hat{\mathbf{Q}}_\tau)\mathbf{W}^{\mathrm{im}}_\tau(\mathbf{Q}_\tau-\hat{\mathbf{Q}}_\tau)\Sigma_{\boldsymbol{\theta}_\tau}]\| \le \mathbb{E}_{\mathbf{x}_\tau}[\|\mathbf{Q}_\tau-\hat{\mathbf{Q}}_\tau\|^2]\|\mathbf{W}^{\mathrm{im}}_\tau\|$. Based on sub-gaussian concentration inequality, it holds with probability at least $1-\delta$ that

$$\left\|\mathbf{Q}_\tau-\hat{\mathbf{Q}}_{\tau,N}\right\| \le \bar{\lambda}CK^2\Big(\sqrt{\frac{d+\log\frac{2}{\delta}}{N}} + \frac{d+\log\frac{2}{\delta}}{N}\Big).$$

Therefore choose $\delta = N^{-1}$ and since $\mathbf{x}_\tau$ is bounded, we have

$$\|\mathbf{W}^{\mathrm{im}}_{\tau,N}\| \le \|\mathbf{W}^{\mathrm{im}}_\tau\| + \Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{N}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big)L^{\mathrm{im}}.$$

Similarly, for BaMAML,

$$\begin{aligned}\mathbf{W}^{\mathrm{ba}}_{\tau,N} =& \mathbb{E}_{\mathbf{x}_\tau}\big[\frac{\gamma_b}{N_1}\big((\mathbf{I}_d+\gamma_b^{-1}N_1\hat{\mathbf{Q}}_{\tau,N_1})^{-1} - (\mathbf{I}_d+\gamma_b^{-1}(N+N_1)\hat{\mathbf{Q}}_{\tau,N+N_1})^{-1}\big)\big]\\ =&\mathbb{E}_{\mathbf{x}_\tau}\big[\frac{\gamma_b}{N_1}\big((\mathbf{I}_d+\gamma_b^{-1}N_1\mathbf{Q}_\tau)^{-1} - (\mathbf{I}_d+\gamma_b^{-1}(N+N_1)\mathbf{Q}_\tau)^{-1}\\ &+ (\mathbf{I}_d+\gamma_b^{-1}N_1\hat{\mathbf{Q}}_{\tau,N_1})^{-1} - (\mathbf{I}_d+\gamma_b^{-1}N\hat{\mathbf{Q}}_{\tau,N})^{-1} - (\mathbf{I}_d+\gamma_b^{-1}N_1\mathbf{Q}_\tau)^{-1} + (\mathbf{I}_d+\gamma_b^{-1}N\mathbf{Q}_\tau)^{-1}\big)\big]\\ =&\mathbf{W}^{\mathrm{ba}}_\tau + \frac{\gamma_b}{N_1}\mathbb{E}_{\mathbf{x}_\tau}\big[(\mathbf{I}_d+\gamma_b^{-1}N_1\hat{\mathbf{Q}}_{\tau,N_1})^{-1} - (\mathbf{I}_d+\gamma_b^{-1}N\hat{\mathbf{Q}}_{\tau,N})^{-1} - (\mathbf{I}_d+\gamma_b^{-1}N_1\mathbf{Q}_\tau)^{-1} + (\mathbf{I}_d+\gamma_b^{-1}N\mathbf{Q}_\tau)^{-1}\big]\\ =&\mathbf{W}^{\mathrm{ba}}_\tau + \frac{\gamma_b}{N_1}\mathbb{E}_{\mathbf{x}_\tau}\big[\gamma(\mathbf{I}_d+\gamma_b^{-1}N_1\mathbf{Q}_\tau)^{-1}(\mathbf{Q}_\tau-\hat{\mathbf{Q}}_{\tau,N_1})(\mathbf{I}_d+\gamma_b^{-1}N_1\hat{\mathbf{Q}}_{\tau,N_1})^{-1}\\ &\qquad - \gamma_b^{-1}N(\mathbf{I}_d+\gamma_b^{-1}N\mathbf{Q}_\tau)^{-1}(\mathbf{Q}_\tau-\hat{\mathbf{Q}}_{\tau,N})(\mathbf{I}_d+\gamma_b^{-1}N\hat{\mathbf{Q}}_{\tau,N})^{-1}\big]\\ =&\mathbf{W}^{\mathrm{ba}}_\tau + \gamma\mathbb{E}_{\mathbf{x}_\tau}\big[\gamma(\mathbf{I}_d+\gamma^{-1}\mathbf{Q}_\tau)^{-1}(\mathbf{Q}_\tau-\hat{\mathbf{Q}}_{\tau,N_1})(\mathbf{I}_d+\gamma^{-1}\hat{\mathbf{Q}}_{\tau,N_1})^{-1}\\ &\qquad - (\gamma s)^{-1}(\mathbf{I}_d+(\gamma s)^{-1}\mathbf{Q}_\tau)^{-1}(\mathbf{Q}_\tau-\hat{\mathbf{Q}}_{\tau,N})(\mathbf{I}_d+(\gamma s)^{-1}\hat{\mathbf{Q}}_{\tau,N})^{-1}\big]\end{aligned}$$

therefore

$$\|\mathbf{W}^{\mathrm{ba}}_{\tau,N}\| \le \|\mathbf{W}^{\mathrm{ba}}_\tau\| + \Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{N}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big)L^{\mathrm{ba}}.$$

□

Lemma 2 (Concentration of $\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N}$) *Denote d as the dimension of $\boldsymbol{\theta}_\tau$, T as the number of tasks. Suppose Assumption 1 holds, and $\mathbf{x}_{\tau,i}$ is sub-gaussian with parameter k, then with probability at least $1-Td^{-10}$, for $\tau = 1,\ldots,T$, we have the following bounds, given by*

$$\mathbf{0} \preceq \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \preceq \widetilde{\mathcal{O}}(c^{\mathrm{er}})\mathbf{I}_d \tag{79}$$

where $c^{\mathrm{er}} := 1 + \max\{d/N, \sqrt{d/N}\}$*, and* $\widetilde{\mathcal{O}}(\cdot)$ *hides the logarithmic factor* $\log(NdT)$*.*
And denote $\|\cdot\|_{\mathrm{op}}$ *as the operator norm. With probability at least* $1 - Td^{-10}$*, it holds that*

$$\Big\| \frac{1}{T} \sum_{\tau=1}^{T} \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} - \mathbb{E}\big[\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N}\big] \Big\|_{\mathrm{op}} \le \widetilde{\mathcal{O}}\Big(c^{\mathrm{er}} \sqrt{\frac{d}{T}} + d^{-4} \Big), \tag{80a}$$

$$\Big\| \frac{1}{T} \sum_{\tau=1}^{T} (\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N})^2 - \mathbb{E}\big[(\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N})^2\big] \Big\|_{\mathrm{op}} \le \widetilde{\mathcal{O}}\Big((c^{\mathrm{er}})^2 \sqrt{\frac{d}{T}} + d^{-4} \Big). \tag{80b}$$

Proof: The proof is similar to Lemma C.4 in Bai et al. (2021), the difference is we do not need Assumption 3, $\mathbf{x}_{\tau,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, but only requires $\mathbf{x}_{\tau,i}$ to be sub-gaussian with parameter K. Recall that $\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} = \frac{1}{N} \mathbf{X}^{\mathrm{all}\top}_{\tau,N} \mathbf{X}^{\mathrm{all}}_{\tau,N} = \hat{\mathbf{Q}}_{\tau,N}$. Applying the sub-gaussian covariance concentration (Vershynin (2018), Exercise 4.7.3), we have with probability at least $1 - d^{-10}$ that

$$\begin{aligned} \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} = \hat{\mathbf{Q}}_{\tau,N} &\preceq \mathbf{Q}_\tau + \Big\| \hat{\mathbf{Q}}_{\tau,N} - \mathbf{Q}_\tau \Big\|_{\mathrm{op}} \mathbf{I}_d \\ &\preceq \Big(\bar{\lambda} + CK^2 \sqrt{\frac{d + \log d}{N}} + CK^2 \frac{d + \log d}{N} \Big) \mathbf{I}_d \preceq K_\tau c^{\mathrm{er}} \mathbf{I}_d \end{aligned} \tag{81}$$

where $K_\tau = \mathcal{O}(1)$ is an absolute constant dependent on $\bar{\lambda}, C, K$. Let $\mathcal{W}_\tau := \big\{ \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \preceq K_\tau c^{\mathrm{er}} \mathbf{I}_d \big\}$ denote this event. We have $\mathbb{P}(\mathcal{W}_\tau) \ge 1 - Td^{-10}$. Let $\mathcal{W} := \bigcup_{t=1}^{T} \mathcal{W}_\tau$ denote the union event. Note that on the event $\mathcal{W}$ we have

$$\frac{1}{T} \sum_{\tau=1}^{T} \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} = \frac{1}{T} \sum_{\tau=1}^{T} \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\{\mathcal{W}_\tau\}.$$

And on the event $\mathcal{W}_\tau, \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N}$ is bounded by: $\mathbf{0} \preceq \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\{\mathcal{W}_\tau\} \preceq K_\tau c^{\mathrm{er}} \mathbf{I}_d$, which means that for any $\mathbf{v} \in \mathbb{R}^d$ and $\|\mathbf{v}\|_2 = 1$, the random variable $\mathbf{v}^\top \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\{\mathcal{W}_\tau\} \mathbf{v} - \mathbf{v}^\top \mathbb{E}\big[\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\{\mathcal{W}_\tau\}\big] \mathbf{v}$ is mean-zero and sub-gaussian with parameter $K_\tau c^{\mathrm{er}}$. Therefore by the standard sub-gaussian concentration, we have

$$\mathbb{P}\Big(\Big| \mathbf{v}^\top \big(\frac{1}{T} \sum_{\tau=1}^{T} \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\{\mathcal{W}_\tau\} \big) \mathbf{v} - \mathbf{v}^\top \mathbb{E}\big[\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\{\mathcal{W}_\tau\}\big] \mathbf{v} \Big| \ge t \Big) \le 2 \exp\Big(- \frac{Tt^2}{2(K_\tau c^{\mathrm{er}})^2} \Big).$$

Using the fact that for any symmetric matrix $\mathbf{M}$, $\|\mathbf{M}\|_{\mathrm{op}} \le 2 \sup_{\mathbf{v} \in N_{1/4}(\mathbb{S}^{d-1})} |\mathbf{v}^\top \mathbf{M} \mathbf{v}|$ where $N_{1/4}(\mathbb{S}^{d-1})$ is a $1/4$-covering set of the $(d-1)$-unit sphere $\mathbb{S}^{d-1}$ with $|N_{1/4}(\mathbb{S}^{d-1})| \le 9^d$ (Vershynin (2018) , Exercise 4.4.3), we have

$$\begin{aligned} &\mathbb{P}\Big(\Big\| \frac{1}{T} \sum_{\tau=1}^{T} \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\Big\{\mathcal{W}_\tau\Big\} - \mathbb{E}\big[\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\Big\{\mathcal{W}_\tau\Big\}\big] \Big\|_{\mathrm{op}} \ge t \Big) \\ \le &\Big| N_{1/4}\Big(\mathbb{S}^{d-1}\Big) \Big| \cdot \sup_{\|\mathbf{v}\|_2 = 1} \mathbb{P}\Big(\Big| \mathbf{v}^\top \Big(\frac{1}{T} \sum_{\tau=1}^{T} \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\Big\{\mathcal{W}_\tau\Big\} \Big) \mathbf{v} - \mathbf{v}^\top \mathbb{E}\big[\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\Big\{\mathcal{W}_\tau\Big\}\big] \mathbf{v} \Big| \ge t \Big) \\ \le &\exp\Big(- Tt^2/2(K_\tau c^{\mathrm{er}})^2 + 3d \Big). \end{aligned}$$

Taking $t = \mathcal{O}\big(K_\tau c^{\mathrm{er}} \sqrt{\frac{6d + 20\log(d)}{T}}\big) = \widetilde{\mathcal{O}}\big(K_\tau c^{\mathrm{er}} \sqrt{\frac{d}{T}}\big)$, the above probability is upper bounded by d^{-10}. In other words, with probability at least $1 - Td^{-10}$, we have

$$\Big\| \frac{1}{T} \sum_{\tau=1}^{T} \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\left\{\mathcal{W}_\tau\right\} - \mathbb{E}\big[\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\left\{\mathcal{W}_\tau\right\}\big] \Big\|_{\mathrm{op}} \le \widetilde{\mathcal{O}}\Big(K_\tau c^{\mathrm{er}} \sqrt{\frac{d}{T}} \Big). \tag{82}$$

To bound the difference between $\mathbb{E}\big[\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N}\big]$ and $\mathbb{E}\big[\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\{\mathcal{W}_\tau\}\big]$, it follows

$$\begin{aligned} \Big\| \mathbb{E}\big[\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N}\big] - \mathbb{E}\big[\hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \mathbf{1}\left\{\mathcal{W}_\tau\right\}\big] \Big\|_{\mathrm{op}} &\le \mathbb{E}\Big[\big\| \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \big\|_{\mathrm{op}} 1\left\{\mathcal{W}^c_\tau\right\} \Big] \le \Big(\mathbb{E}\big[\big\| \hat{\mathbf{W}}^{\mathrm{er}}_{\tau,N} \big\|^2_{\mathrm{op}} \big] \cdot \mathbb{P}\left(\mathcal{W}^c_\tau\right) \Big)^{\frac{1}{2}} \\ &\le \sqrt{\mathbb{E}\big[\max_i \|\mathbf{x}_{\tau,i}\|_2^2 \big] \cdot d^{-10}} \le \sqrt{k^2\left(d + C\log N\right) \cdot d^{-10}} = \widetilde{\mathcal{O}}\left(d^{-4.5}\right) \end{aligned} \tag{83}$$

where $\mathcal{W}_\tau^c$ is the complement of $\mathcal{W}_\tau$, and the last inequality is by sub-gaussian norm concentration. Combining (82) and (83), with probability at least $1 - Td^{-10}$, we have that

$$\Big\|\frac{1}{T}\sum_{\tau=1}^T \hat{\mathbf{W}}_{\tau,N}^{\mathrm{er}} - \mathbb{E}\big[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{er}}\big]\Big\|_{\mathrm{op}} \leq \Big\|\frac{1}{T}\sum_{\tau=1}^T \hat{\mathbf{W}}_{\tau,N}^{\mathrm{er}}\mathbf{1}\left\{\mathcal{W}_\tau\right\} - \mathbb{E}\big[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{er}}\mathbf{1}\left\{\mathcal{W}_\tau\right\}\big]\Big\|_{\mathrm{op}}$$

$$+ \Big\|\mathbb{E}\big[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{er}}\big] - \mathbb{E}\big[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{er}}\mathbf{1}\left\{\mathcal{W}_\tau\right\}\big]\Big\|_{\mathrm{op}} \leq \widetilde{\mathcal{O}}\left(c^{\mathrm{er}}\sqrt{\frac{d}{T}} + d^{-4.5}\right).$$

Similarly we can prove that with probability at least $1 - Td^{-10}$ that

$$\Big\|\frac{1}{T}\sum_{\tau=1}^T (\hat{\mathbf{W}}_{\tau,N}^{\mathrm{er}})^2 - \mathbb{E}\big[(\hat{\mathbf{W}}_{\tau,N}^{\mathrm{er}})^2\big]\Big\|_{\mathrm{op}} \leq \widetilde{\mathcal{O}}\left((c^{\mathrm{er}})^2\sqrt{\frac{d}{T}} + d^{-4}\right).$$

This completes the proof of Lemma 2. Note that, similar results apply to random weight matrices $\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}$ of other methods by replacing $\hat{\mathbf{W}}_{\tau,N}^{\mathrm{er}} = \hat{\mathbf{Q}}_{\tau,N}$ with $\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}} \preceq \hat{\mathbf{Q}}_{\tau,N_2}$ in (81). And Lemma 2 still holds with Assumption 3.

□

Lemma 3 (Hanson-Wright inequality) *(Restatement of Theorem 6.2.1 in (Vershynin, 2018)). Let $\mathbf{z} \in \mathbb{R}^d$ be a random vector with independent, mean-zero, and K-sub-gaussian entries, and let $\mathbf{C} \in \mathbb{R}^{d\times d}$ be a fixed matrix. Then it holds with probability at least $1-\delta$ that*

$$\Big|\mathbf{z}^\top \mathbf{C}\mathbf{z} - \mathbb{E}[\mathbf{z}^\top \mathbf{C}\mathbf{z}]\Big| \leq \mathcal{O}\Big(K^2\|\mathbf{C}\|_{\mathrm{F}}\log\frac{2}{\delta}\Big).$$

Lemma 4 (Linear combination of sub-gaussian) *(Vershynin, 2018) Let $\mathbf{z} \in \mathbb{R}^d$ be a random vector with independent and K-sub-gaussian entries. Then for any $\mathbf{v} \in \mathbb{S}^{d-1}(r), \mathbf{v}^\top\mathbf{z}$ is rK-sub-gaussian. In other words, it holds with probability at least $1-\delta$ that*

$$\Big|\mathbf{v}^\top\mathbf{z} - \mathbb{E}[\mathbf{v}^\top\mathbf{z}]\Big| \leq \mathcal{O}\Big(rK\sqrt{\log\frac{2}{\delta}}\Big).$$

Lemma 5 (Hanson-Wright inequality with non-zero mean) *Let $\mathbf{z} \in \mathbb{R}^d$ be a random vector with independent, and K-sub-gaussian entries, and let $\mathbf{C} \in \mathbb{R}^{d\times d}$ be a fixed matrix. Then it holds with probability at least $1-\delta$ that*

$$\big|\mathbf{z}^\top\mathbf{C}\mathbf{z} - \mathbb{E}\big[\mathbf{z}^\top\mathbf{C}\mathbf{z}\big]\big| \leq \mathcal{O}\Big(K^2\|\mathbf{C}\|_{\mathrm{F}}\log(2/\delta)\Big) + \mathcal{O}\Big(K\|\mathbb{E}[\mathbf{z}]\|\|\mathbf{C}\|_{\mathrm{op}}\sqrt{\log(2/\delta)}\Big).$$

Proof:

$$\begin{aligned}
&\big|\mathbf{z}^\top\mathbf{C}\mathbf{z} - \mathbb{E}\big[\mathbf{z}^\top\mathbf{C}\mathbf{z}\big]\big| = \Big|(\mathbf{z}-\mathbb{E}[\mathbf{z}])^\top\mathbf{C}(\mathbf{z}-\mathbb{E}[\mathbf{z}]) - \mathbb{E}\big[(\mathbf{z}-\mathbb{E}[\mathbf{z}])^\top\mathbf{C}(\mathbf{z}-\mathbb{E}[\mathbf{z}])\big] + 2\mathbb{E}[\mathbf{z}]^\top\mathbf{C}(\mathbf{z}-\mathbb{E}[\mathbf{z}])\Big|\\
&\leq \Big|(\mathbf{z}-\mathbb{E}[\mathbf{z}])^\top\mathbf{C}(\mathbf{z}-\mathbb{E}[\mathbf{z}]) - \mathbb{E}\big[(\mathbf{z}-\mathbb{E}[\mathbf{z}])^\top\mathbf{C}(\mathbf{z}-\mathbb{E}[\mathbf{z}])\big]\Big| + 2\Big|\mathbb{E}[\mathbf{z}]^\top\mathbf{C}(\mathbf{z}-\mathbb{E}[\mathbf{z}])\Big|\\
&\leq \mathcal{O}\Big(K^2\|\mathbf{C}\|_{\mathrm{F}}\log(2/\delta)\Big) + \mathcal{O}\Big(K\|\mathbb{E}[\mathbf{z}]\|\|\mathbf{C}\|_{\mathrm{op}}\sqrt{\log(2/\delta)}\Big)
\end{aligned}$$

where the last inequality follows from Lemma 3 and 4. Note that when $\mathbb{E}[\mathbf{z}] = \mathbf{0}$, this Lemma reduces to the zero-mean version of Hanson-Wright inequality, i.e. Lemma 3. □

Lemma 6 (sub-gaussian random vector concentration) *Let $\mathbf{U}_\tau \in \mathbb{R}^{d\times d}, \mathbf{z}_\tau \in \mathbb{R}^d$. Assume $\big\|\mathbf{U}_\tau\big\| \leq \bar{\lambda}$ and $\mathbf{z}_\tau$ has independent, mean-zero, K-sub-gaussian entries. With probability at least $1-\delta$, it holds that*

$$\Big|\big\|\frac{1}{T}\sum_{\tau=1}^T \mathbf{U}_\tau\mathbf{z}_\tau\big\| - \big\|\mathbb{E}_\tau\left[\mathbf{U}_\tau\mathbf{z}_\tau\right]\big\|\Big| \leq \widetilde{\mathcal{O}}\Big(K\bar{\lambda}\sqrt{\frac{d}{T}}\log\frac{2}{\delta}\Big).$$

Proof: Apply Lemma 3, the Hanson-Wright inequality, and let $\mathbf{z} = \mathbf{z}_\tau$, $\mathbf{C} = \mathbf{U}_\tau^\top \mathbf{U}_\tau$, we obtain that with probability at least $1-\delta$,

$$\left|\mathbf{z}_\tau^\top \mathbf{U}_\tau^\top \mathbf{U}_\tau \mathbf{z}_\tau - \mathbb{E}_{\mathbf{z}_\tau|\mathbf{U}_\tau}\left[\mathbf{z}_\tau^\top \mathbf{U}_\tau^\top \mathbf{U}_\tau \mathbf{z}_\tau\right]\right| \le \mathcal{O}\Big(K^2 \|\mathbf{U}_\tau^\top \mathbf{U}_\tau\|_{\mathrm{F}} \log\frac{2}{\delta}\Big).$$

Since

$$\begin{aligned}\Big|\|\mathbf{U}_\tau \mathbf{z}_\tau\| - \|\mathbb{E}_{\mathbf{z}_\tau|\mathbf{U}_\tau}\left[\mathbf{U}_\tau \mathbf{z}_\tau\right]\|\Big|^2 &\le \Big|\|\mathbf{U}_\tau \mathbf{z}_\tau\|^2 - \|\mathbb{E}_{\mathbf{z}_\tau|\mathbf{U}_\tau}\left[\mathbf{U}_\tau \mathbf{z}_\tau\right]\|^2\Big| \\ = \left|\mathbf{z}_\tau^\top \mathbf{U}_\tau^\top \mathbf{U}_\tau \mathbf{z}_\tau - \mathbb{E}_{\mathbf{z}_\tau|\mathbf{U}_\tau}\left[\mathbf{z}_\tau^\top \mathbf{U}_\tau^\top \mathbf{U}_\tau \mathbf{z}_\tau\right]\right| &\le \mathcal{O}\Big(K^2 d \|\mathbf{U}_\tau\|_{\mathrm{op}}^2 \log\frac{2}{\delta}\Big)\end{aligned}$$

where the last equation holds because $\mathbf{z}_\tau$ has mean-zero entries. Therefore, it holds with probability at least $1-\delta$ that

$$\Big|\|\mathbf{U}_\tau \mathbf{z}_\tau\| - \|\mathbb{E}_{\mathbf{z}_\tau|\mathbf{U}_\tau}\left[\mathbf{U}_\tau \mathbf{z}_\tau\right]\|\Big| \le \mathcal{O}\Big(K\bar{\lambda}\sqrt{d}\log\frac{2}{\delta}\Big). \tag{84}$$

Also, based on Lemma 4, it holds with probability at least $1-\delta$ that

$$\begin{aligned}\Big|\|\mathbb{E}_{\mathbf{z}_\tau,\mathbf{U}_\tau}\left[\mathbf{U}_\tau \mathbf{z}_\tau\right]\| - \|\mathbb{E}_{\mathbf{z}_\tau|\mathbf{U}_\tau}\left[\mathbf{U}_\tau \mathbf{z}_\tau\right]\|\Big| &\le \Big\|\mathbb{E}_{\mathbf{z}_\tau,\mathbf{U}_\tau}\left[\mathbf{U}_\tau \mathbf{z}_\tau\right] - \mathbb{E}_{\mathbf{z}_\tau|\mathbf{U}_\tau}\left[\mathbf{U}_\tau \mathbf{z}_\tau\right]\Big\| \\ &\le \widetilde{\mathcal{O}}\big(K\bar{\lambda}\sqrt{d}\log\frac{2}{\delta}\big).\end{aligned} \tag{85}$$

Combining (84) and (85), it holds with probability at least $1-\delta$ that

$$\Big|\|\frac{1}{T}\sum_{\tau=1}^{T}\mathbf{U}_\tau \mathbf{z}_\tau\| - \|\mathbb{E}_{\mathbf{z}_\tau,\mathbf{U}_\tau}\left[\mathbf{U}_\tau \mathbf{z}_\tau\right]\|\Big| \le \widetilde{\mathcal{O}}\big(K\bar{\lambda}\sqrt{\frac{d}{T}}\log\frac{2}{\delta}\big). \tag{86}$$

□

Lemma 7 (Bound of statistical error not caused by data noise) *Define*

$$\begin{aligned}\mathbf{z}_{\mathcal{A}} &:= \left[(\boldsymbol{\theta}_1^{\mathrm{gt}} - \boldsymbol{\theta}_0^{\mathcal{A}})^\top, \ldots, (\boldsymbol{\theta}_T^{\mathrm{gt}} - \boldsymbol{\theta}_0^{\mathcal{A}})^\top\right]^\top \in \mathbb{R}^{dT}, \\ \mathbf{U}_{\mathcal{A}} &:= \Big[\hat{\mathbf{W}}_{1,N}^{\mathcal{A}}(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}})^{-1}, \ldots, \hat{\mathbf{W}}_{T,N}^{\mathcal{A}}(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}})^{-1}\Big]^\top \in \mathbb{R}^{dT\times d}.\end{aligned}$$

1) Suppose Assumptions 1-2 hold, the statistical error for method $\mathcal{A}$ is computed by

$$\mathcal{E}_{\mathcal{A}}^2(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}) = \|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}}\|_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]}^2 = \underbrace{\|\mathbf{U}_{\mathcal{A}}^\top \mathbf{z}_{\mathcal{A}}\|_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]}^2}_{I_1^{\mathcal{A}}} + \|\Delta_T^{\mathcal{A}}\|_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]}^2 + 2\mathbf{z}_{\mathcal{A}}^\top \mathbf{U}_{\mathcal{A}}\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]\Delta_T^{\mathcal{A}}$$

where with probability at least $1-Td^{-10}$, the first term $I_1^{\mathcal{A}}$ can be bounded above by [2]

$$\begin{aligned}I_1^{\mathcal{A}} \le \frac{R^2}{T}\Big(&\lambda_{\min}(\mathbb{E}[\mathbf{W}_\tau^{\mathcal{A}}])^{-1}\lambda_{\max}(\mathbb{E}[(\mathbf{W}_\tau^{\mathcal{A}})^2]) + \widetilde{\mathcal{O}}(\frac{d}{N}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big) \\ &+ \Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big)M^2.\end{aligned}$$

2) Suppose Assumptions 1,3 hold, the statistical error for method $\mathcal{A}$ can be computed by

$$\mathcal{E}_{\mathcal{A}}^2(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}) = w_{\mathcal{A}}\|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}}\|_2^2 = w_{\mathcal{A}}(\underbrace{\|\mathbf{U}_{\mathcal{A}}^\top \mathbf{z}_{\mathcal{A}}\|^2}_{I_2^{\mathcal{A}}} + \|\Delta_T^{\mathcal{A}}\|^2 + 2\mathbf{z}_{\mathcal{A}}^\top \mathbf{U}_{\mathcal{A}}\Delta_T^{\mathcal{A}})$$

Define $\tilde{C}_0^{\mathcal{A}} := \frac{1}{d}\big\langle\mathbb{E}^{-2}\big[\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big], \mathbb{E}\big[(\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}})^2\big]\big\rangle$. With probability at least $1-Td^{-10}$, $I_2^{\mathcal{A}}$ can be bounded above by

$$I_2^{\mathcal{A}} \le \frac{R^2}{T}\Big(\tilde{C}_0^{\mathcal{A}} + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big).$$

[2] Note that, we provide bound for $I_1^{\mathcal{A}}$ and $I_2^{\mathcal{A}}$ in this lemma since it has the same form for different methods $\mathcal{A}$. And the bound for the rest terms in the statistical error are deferred to later sections for the specific methods.

Proof: The derivations in Section A give the empirical solutions $\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}$ as below

$$\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} = \Big(\sum_{\tau=1}^{T} \hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\Big)^{-1}\Big(\sum_{\tau=1}^{T} \hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}} \boldsymbol{\theta}_{\tau}^{\mathrm{gt}}\Big) + \Delta_T^{\mathcal{A}}. \tag{87}$$

Thus the difference between the estimated model parameter $\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}$ and the population-wise optimal model parameter $\boldsymbol{\theta}_0^{\mathcal{A}}$, which is be used to compute the statistical error, is given by

$$\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}} = \Big(\sum_{\tau=1}^{T} \hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\Big)^{-1}\Big(\sum_{\tau=1}^{T} \hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}} (\boldsymbol{\theta}_{\tau}^{\mathrm{gt}} - \boldsymbol{\theta}_0^{\mathcal{A}})\Big) + \Delta_T^{\mathcal{A}} = \mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}} + \Delta_T^{\mathcal{A}} \tag{88}$$

based on which the statistical error $\mathcal{E}_{\mathcal{A}}^2(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}})$ in (11) can be computed by

$$\mathcal{E}_{\mathcal{A}}^2(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}) = \|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]} = \underbrace{\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]}}_{I_1^{\mathcal{A}}} + \|\Delta_T^{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]} + 2\mathbf{z}_{\mathcal{A}}^{\top}\mathbf{U}_{\mathcal{A}}\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]\Delta_T^{\mathcal{A}} \tag{89}$$

where $I_1^{\mathcal{A}}$ is the only term that does not depend on $\Delta_T^{\mathcal{A}}$, which is caused by the random noise ϵ in the data. In other words, when the variance of ϵ becomes zero, the statistical error $\mathcal{E}_{\mathcal{A}}^2(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}})$ reduces to $I_1^{\mathcal{A}}$, or $I_1^{\mathcal{A}}$ is the statistical error in the noiseless realizable case. We next proceed to bound $I_1^{\mathcal{A}}$ by considering the concentration around its mean as follows

$$\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]} \leq \Big|\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]} - \mathbb{E}_{\mathbf{z}_{\mathcal{A}}|\mathbf{U}_{\mathcal{A}}}\big[\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]}\big]\Big| + \mathbb{E}_{\mathbf{z}_{\mathcal{A}}|\mathbf{U}_{\mathcal{A}}}\big[\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]}\big].$$

Next we will bound the above two terms respectively. We first bound $\Big|\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]} - \mathbb{E}_{\mathbf{z}_{\mathcal{A}}|\mathbf{U}_{\mathcal{A}}}\big[\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]}\big]\Big|$. From Assumption 2, $\boldsymbol{\theta}_{\tau}^{\mathrm{gt}} - \boldsymbol{\theta}_0^{\mathcal{A}}$ are $(R/\sqrt{d})$-sub-gaussian. To bound the absolute error around the expectation, from the Hanson-Wright inequality in Lemma 5, with probability at least $1-\delta$, the following inequality holds

$$\begin{aligned}
&\Big|\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]} - \mathbb{E}_{\mathbf{z}_{\mathcal{A}}|\mathbf{U}_{\mathcal{A}}}[\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]}]\Big| \\
\leq &\widetilde{\mathcal{O}}\Big(\frac{R^2}{d}\Big\|\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}(\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}})^2\big)\Big\|_{\mathrm{F}}\Big) \\
&+ \widetilde{\mathcal{O}}\Big(\frac{R}{\sqrt{d}}M\Big\|\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}(\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}})^2\big)\Big\|_{\mathrm{op}}\Big) \\
\leq &\widetilde{\mathcal{O}}\Big(\frac{R^2+RM}{dT}\Big\|\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\Big\|^{-2}\cdot\sqrt{d}\Big\|\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\Big\|^2_{\mathrm{op}}\Big\|\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]\Big\|_{\mathrm{op}}\Big) = \widetilde{\mathcal{O}}\Big(\frac{R^2+RM}{T\sqrt{d}}\Big).
\end{aligned} \tag{90}$$

Note that in the last equation, we ignore the higher order terms in $\|\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\|_{\mathrm{op}}$, which can be obtained from Lemma 2.

To bound the expected statistical error, $\mathbb{E}_{\mathbf{z}_{\mathcal{A}}|\mathbf{U}_{\mathcal{A}}}[\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]}]$, first note that since for all τ, $\mathbf{W}_{\tau}^{\mathcal{A}}$ is symmetric positive definite (PD) based on Assumption 1, $\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]$ is also symmetric PD, who has a Cholesky decomposition, $\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}] = \mathbb{E}_{\tau}^{\frac{1}{2}}[\mathbf{W}_{\tau}^{\mathcal{A}}]\mathbb{E}_{\tau}^{\frac{1}{2}}[\mathbf{W}_{\tau}^{\mathcal{A}}]^{\top}$ with $\mathbb{E}_{\tau}^{\frac{1}{2}}[\mathbf{W}_{\tau}^{\mathcal{A}}]$ defined as the lower triangular matrix in the decomposition. The statistical error can be rewritten as

$$\begin{aligned}
&\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]} = \mathrm{tr}(\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]}) = \mathrm{tr}\Big((\mathbb{E}_{\tau}^{\frac{1}{2}}[\mathbf{W}_{\tau}^{\mathcal{A}}]\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}})(\mathbb{E}_{\tau}^{\frac{1}{2}}[\mathbf{W}_{\tau}^{\mathcal{A}}]\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}})^{\top}\Big) \\
=&\mathrm{tr}\Bigg(\Big(\mathbb{E}_{\tau}^{\frac{1}{2}}[\mathbf{W}_{\tau}^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_{\tau}^{\mathrm{gt}} - \boldsymbol{\theta}_0^{\mathcal{A}})\big)\Big)\Big(\mathbb{E}_{\tau}^{\frac{1}{2}}[\mathbf{W}_{\tau}^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_{\tau}^{\mathrm{gt}} - \boldsymbol{\theta}_0^{\mathcal{A}})\big)\Big)^{\top}\Bigg)
\end{aligned}$$

whose conditional expectation is given by

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{z}_{\mathcal{A}}|\mathbf{U}_{\mathcal{A}}}[\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|^2_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]}]\\
=&\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}|\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\mathrm{tr}\Big(\Big(\mathbb{E}_\tau^{\frac{1}{2}}[\mathbf{W}_\tau^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\big)\Big)\\
&\qquad\Big(\mathbb{E}_\tau^{\frac{1}{2}}[\mathbf{W}_\tau^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\big)\Big)^{\top}\Big)\Big]\\
=&\underbrace{\mathrm{tr}\Big(\mathrm{Cov}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}|\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\mathbb{E}_\tau^{\frac{1}{2}}[\mathbf{W}_\tau^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\big)\Big]\Big)}_{I_a}\\
&+\underbrace{\Big\|\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}|\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\mathbb{E}_\tau^{\frac{1}{2}}[\mathbf{W}_\tau^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\big)\Big]\Big\|^2}_{I_b}
\end{aligned}
\tag{91}
$$

where the last equation is given by the fact that $\mathbb{E}[\mathbf{z}\mathbf{z}^{\top}] = \mathrm{Cov}[\mathbf{z}] + \mathbb{E}[\mathbf{z}]\mathbb{E}[\mathbf{z}]^{\top}$ for any random vector $\mathbf{z}$, and the linear and cyclic property of trace. From Assumption 2, $\boldsymbol{\theta}_\tau^{\mathrm{gt}} - \boldsymbol{\theta}_0^{\mathcal{A}}$ has independent $(R/\sqrt{d})$-sub-gaussian entries, and Lemma 4, linear combinations of sub-gaussian random variables are still sub-gaussian, $\mathbb{E}_\tau^{\frac{1}{2}}[\mathbf{W}_\tau^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\big)$ has sub-gaussian entries with parameter $\big\|\mathbb{E}_\tau^{\frac{1}{2}}[\mathbf{W}_\tau^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big\|_{\mathrm{op}}^2\big\|\sum_{\tau=1}^{T}(\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}})^2\big\|_{\mathrm{op}}R^2/d$, which is the upper bound of the variance of each entry based on the sub-gaussian property. Since the trace of the covariance is the sum of the variance of all entries, it holds that

$$
\begin{aligned}
I_a\leq& d\Big\|\mathbb{E}_\tau^{\frac{1}{2}}[\mathbf{W}_\tau^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\Big\|_{\mathrm{op}}^2\Big\|\sum_{\tau=1}^{T}(\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}})^2\Big\|_{\mathrm{op}}\frac{R^2}{d}\\
\leq&\Big\|\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]\Big\|_{\mathrm{op}}\Big\|\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\Big\|_{\mathrm{op}}^{-2}\Big\|\sum_{\tau=1}^{T}(\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}})^2\Big\|_{\mathrm{op}}R^2\\
\leq&\frac{R^2}{T}\Big\|\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]\Big\|_{\mathrm{op}}\Big\|\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\Big\|_{\mathrm{op}}^{-2}\Big\|\frac{1}{T}\sum_{\tau=1}^{T}(\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}})^2\Big\|_{\mathrm{op}}.
\end{aligned}
\tag{92}
$$

I_b can be further derived as

$$
\begin{aligned}
I_b =& \Big\|\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}|\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\mathbb{E}_\tau^{\frac{1}{2}}[\mathbf{W}_\tau^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\big)\Big]\Big\|^2\\
=&\underbrace{\begin{aligned}&\Big\langle\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]\Big(\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}|\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\big)\Big]+\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\Big]\Big),\\
&\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}|\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\big)\Big]-\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\Big]\Big\rangle\end{aligned}}_{I_{b1}}\\
&+\underbrace{\begin{aligned}&\Big\langle\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]\Big(\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\Big]+\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\mathbf{W}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}\mathbf{W}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\Big]\Big),\\
&\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\Big]-\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\mathbf{W}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}\mathbf{W}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\Big]\Big\rangle\end{aligned}}_{I_{b2}}\\
&+\underbrace{\Big\langle\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\mathbf{W}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}\mathbf{W}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\Big],\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\mathbf{W}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}\mathbf{W}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}}-\boldsymbol{\theta}_0^{\mathcal{A}})\Big]\Big\rangle}_{I_{b3}}
\end{aligned}
$$

where I_{b1} can be further derived as

$$\begin{aligned}
I_{b1} &= \big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\Big\{\mathbb{E}_{\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}|\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}-\boldsymbol{\theta}_{0}^{\mathcal{A}})\Big]-\mathbb{E}_{\boldsymbol{\theta}_{\tau}^{\mathrm{gt}},\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}-\boldsymbol{\theta}_{0}^{\mathcal{A}})\Big]\Big\}\\
&= \big(\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\Big\{\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\mathbb{E}_{\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}|\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}-\boldsymbol{\theta}_{0}^{\mathcal{A}}\Big]-\mathbb{E}_{\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\mathbb{E}_{\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}|\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}[\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}-\boldsymbol{\theta}_{0}^{\mathcal{A}}]\Big]\Big\}\\
\|I_{b1}\| &\le \widetilde{\mathcal{O}}\Big(\sqrt{\frac{d}{T}}\Big)KM\Big\|\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\Big\|_{\mathrm{op}}^{-1}.
\end{aligned}$$

And for I_{b2}, it holds that

$$\begin{aligned}
I_{b2} =&\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\Big\{\mathbb{E}_{\boldsymbol{\theta}_{\tau}^{\mathrm{gt}},\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}-\boldsymbol{\theta}_{0}^{\mathcal{A}})\Big]-\mathbb{E}_{\boldsymbol{\theta}_{\tau}^{\mathrm{gt}},\mathbf{W}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}\mathbf{W}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}-\boldsymbol{\theta}_{0}^{\mathcal{A}})\Big]\Big\}\\
=&\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\Big\{\mathbb{E}_{\boldsymbol{\theta}_{\tau}^{\mathrm{gt}},\mathbf{W}_{\tau,N}^{\mathcal{A}},\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\sum_{\tau=1}^{T}(\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}-\mathbf{W}_{\tau,N}^{\mathcal{A}})(\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}-\boldsymbol{\theta}_{0}^{\mathcal{A}})\Big]\Big\}\\
=&\big(\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\Big\{\mathbb{E}_{\boldsymbol{\theta}_{\tau}^{\mathrm{gt}},\mathbf{W}_{\tau,N}^{\mathcal{A}},\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\frac{1}{T}\sum_{\tau=1}^{T}(\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}-\mathbb{E}[\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}]+\mathbb{E}[\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}]-\mathbf{W}_{\tau,N}^{\mathcal{A}})(\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}-\boldsymbol{\theta}_{0}^{\mathcal{A}})\Big]\Big\}\\
\|I_{b_2}\| \le&\Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})K+\|\mathbb{E}[\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}-\mathbf{W}_{\tau,N}^{\mathcal{A}}]\|\Big)M\Big\|\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\Big\|_{\mathrm{op}}^{-1}.
\end{aligned}$$

$I_{b3}=0$ since $\mathbb{E}_{\boldsymbol{\theta}_{\tau}^{\mathrm{gt}},\mathbf{W}_{\tau,N}^{\mathcal{A}}}\big[\sum_{\tau=1}^{T}\mathbf{W}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}-\boldsymbol{\theta}_{0}^{\mathcal{A}})\big]=\mathbf{0}$. Combining I_{b1},I_{b2},I_{b3} from above discussions we can bound I_b by

$$\begin{aligned}
I_b &= \Big\|\mathbb{E}_{\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}|\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}}\Big[\mathbb{E}_{\tau}^{\frac{1}{2}}[\mathbf{W}_{\tau}^{\mathcal{A}}]\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\big)^{-1}\big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}(\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}-\boldsymbol{\theta}_{0}^{\mathcal{A}})\big)\Big]\Big\|^{2}\\
&\le \Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})K+\|\mathbb{E}[\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}-\mathbf{W}_{\tau,N}^{\mathcal{A}}]\|\Big)M^{2}\Big\|\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]\Big\|_{\mathrm{op}}\Big\|\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\Big\|_{\mathrm{op}}^{-1}. \qquad (93)
\end{aligned}$$

Combining the bound for I_a and I_b, the expected statistical error conditioned on $\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}$ is bounded by

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}_{\mathcal{A}}|\mathbf{U}_{\mathcal{A}}}[\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]}^{2}] &\le \frac{R^2}{T}\Big\|\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]\Big\|_{\mathrm{op}}\Big\|\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\Big\|_{\mathrm{op}}^{-2}\Big\|\frac{1}{T}\sum_{\tau=1}^{T}(\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}})^{2}\Big\|_{\mathrm{op}}\\
&+\Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})K+\|\mathbb{E}[\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}-\mathbf{W}_{\tau,N}^{\mathcal{A}}]\|\Big)M^{2}\Big\|\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]\Big\|_{\mathrm{op}}\Big\|\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\Big\|_{\mathrm{op}}^{-1}. \qquad (94)
\end{aligned}$$

Finally note that $\big\|\mathbb{E}[\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}-\mathbf{W}_{\tau,N}^{\mathcal{A}}]\big\|\le\widetilde{\mathcal{O}}(\frac{d}{N})$, by combining (90)(94), with probability at least $1-Td^{-10}$, it holds that

$$\begin{aligned}
\|\mathbf{U}_{\mathcal{A}}^{\top}\mathbf{z}_{\mathcal{A}}\|_{\mathbb{E}_{\tau}[\mathbf{W}_{\tau}^{\mathcal{A}}]}^{2} \le& \frac{R^2}{T}\Big(\lambda_{\min}(\mathbb{E}[\mathbf{W}_{\tau}^{\mathcal{A}}])^{-1}\lambda_{\max}(\mathbb{E}[(\mathbf{W}_{\tau}^{\mathcal{A}})^{2}])+\widetilde{\mathcal{O}}(\frac{d}{N})+\widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})+\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big)\\
&+\Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})+\widetilde{\mathcal{O}}(\frac{d}{N})\Big)M^{2}. \qquad (95)
\end{aligned}$$

This completes the proof in the data agnostic case without Assumption 3. Next we proceed to prove the bound under the case with Assumption 3.

From Assumption 3, $\mathbf{W}_{\tau}^{\mathcal{A}}=w_{\mathcal{A}}\mathbf{I}_d$, where $w_{\mathcal{A}}$ is the same for different task τ, therefore $\boldsymbol{\theta}_{0}^{\mathcal{A}}=\mathbb{E}_{\tau}[\boldsymbol{\theta}_{\tau}^{\mathrm{gt}}]$, and $w_{\mathrm{er}}=1$, $w_{\mathrm{ma}}=(1-\alpha)^2$, $w_{\mathrm{im}}=(1+\gamma^{-1})^{-2}$, $w_{\mathrm{ba}}=(1+(\gamma s)^{-1})^{-1}(1+\gamma^{-1})^{-1}$. By using this property in (11), we obtain a simplified meta-test risk decomposition of method $\mathcal{A}$ to the statistical and optimal population risk in the linear centroid model by

$$\lim_{N_a\to\infty}\mathcal{R}^{\mathcal{A}}_{N_a}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}) = \underbrace{w_{\mathcal{A}}\|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}-\boldsymbol{\theta}_0^{\mathcal{A}}\|_2^2}_{\text{statistical error } \mathcal{E}^2_{\mathcal{A}}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}})} + \underbrace{\lim_{N_a\to\infty}\mathcal{R}^{\mathcal{A}}_{N_a}(\boldsymbol{\theta}_0^{\mathcal{A}})}_{\text{optimal population risk}} \quad . \tag{96}$$

Thus the statistical error in (96) can be computed by

$$\mathcal{E}^2_{\mathcal{A}}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}) = w_{\mathcal{A}}\|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}-\boldsymbol{\theta}_0^{\mathcal{A}}\|_2^2 = w_{\mathcal{A}}(\underbrace{\|\mathbf{U}_{\mathcal{A}}^\top\mathbf{z}_{\mathcal{A}}\|^2}_{I_2^{\mathcal{A}}} + \|\Delta_T^{\mathcal{A}}\|_2^2 + 2\mathbf{z}_{\mathcal{A}}^\top\mathbf{U}_{\mathcal{A}}\Delta_T^{\mathcal{A}}). \tag{97}$$

We will bound term $I_2^{\mathcal{A}}$ in the above equation. The only difference of $I_2^{\mathcal{A}}$ from $I_1^{\mathcal{A}}$ is that we can treat $\mathbb{E}_\tau[\mathbf{W}_\tau] = \mathbf{I}_d$ and $M = 0$ when adding Assumption 3. Therefore, with probability at least $1-\delta$, we have

$$\Big|\|\mathbf{U}_{\mathcal{A}}^\top\mathbf{z}_{\mathcal{A}}\|^2 - \mathbb{E}_{\boldsymbol{\theta}_\tau^{\text{gt}},\mathbf{e}_\tau|\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N}}[\|\mathbf{U}_{\mathcal{A}}^\top\mathbf{z}_{\mathcal{A}}\|^2]\Big| = \widetilde{\mathcal{O}}\Big(\frac{R^2}{T\sqrt{d}}\Big). \tag{98}$$

To compute $\mathbb{E}_{\boldsymbol{\theta}_\tau^{\text{gt}},\mathbf{e}_\tau|\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N}}[\mathbf{z}_{\mathcal{A}}^\top\mathbf{U}_{\mathcal{A}}\mathbf{U}_{\mathcal{A}}^\top\mathbf{z}_{\mathcal{A}}]$, first we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_\tau^{\text{gt}},\mathbf{e}_\tau|\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N}}[\mathbf{z}_{\mathcal{A}}^\top\mathbf{U}_{\mathcal{A}}\mathbf{U}_{\mathcal{A}}^\top\mathbf{z}_{\mathcal{A}}]\\
=&\frac{R^2}{Td}\Big\langle\Big(\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N}\Big)^{-2},\frac{1}{T}\sum_{\tau=1}^T(\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N})^2\Big\rangle = \frac{R^2}{T}\Big[\underbrace{\frac{1}{d}\Big\langle\mathbb{E}^{-2}\big[\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N}\big],\mathbb{E}\big[(\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N})^2\big]\Big\rangle}_{=\tilde{C}_0^{\mathcal{A}}}\\
&+\underbrace{\frac{1}{d}\Big\langle\Big(\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N}\Big)^{-2}-\mathbb{E}^{-2}\big[\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N}\big],\mathbb{E}\big[(\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N})^2\big]\Big\rangle}_{I_c}\\
&+\underbrace{\frac{1}{d}\Big\langle\Big(\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N}\Big)^{-2},\frac{1}{T}\sum_{\tau=1}^T(\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N})^2-\mathbb{E}\big[(\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N})^2\big]\Big\rangle}_{I_d}\Big].
\end{aligned} \tag{99}$$

For term I_c and I_d, from Lemma 2, we have with probability at least $1-Td^{-10}$ that

$$|I_c| \leq \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}), |I_d| \leq \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}). \tag{100}$$

Combining (99) and (100), we have with probability at least $1-Td^{-10}$

$$\mathbb{E}_{\boldsymbol{\theta}_\tau^{\text{gt}},\mathbf{e}_\tau|\hat{\mathbf{W}}^{\mathcal{A}}_{\tau,N}}[\mathbf{z}_{\mathcal{A}}^\top\mathbf{U}_{\mathcal{A}}\mathbf{U}_{\mathcal{A}}^\top\mathbf{z}_{\mathcal{A}}] \leq \frac{R^2}{T}\Big(\tilde{C}_0^{\mathcal{A}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big). \tag{101}$$

Then combining (98),(101), it holds with probability at least $1-Td^{-10}$ that

$$I_2^{\mathcal{A}} = \mathbf{z}_{\mathcal{A}}^\top\mathbf{U}_{\mathcal{A}}\mathbf{U}_{\mathcal{A}}^\top\mathbf{z}_{\mathcal{A}} \leq \frac{R^2}{T}\Big(\tilde{C}_0^{\mathcal{A}} + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big). \tag{102}$$

□

Lemma 8 (Dominating constant in statistical error) *Suppose Assumptions 1,3 hold. The dominating constant in the statistical error of meta learning method $\mathcal{A}$ is computed by*

$$\tilde{C}_0^{\mathcal{A}} := \frac{1}{d}\Big\langle\mathbb{E}^{-2}\big[\hat{\mathbf{W}}^{\mathcal{A}}_{\tau}\big],\mathbb{E}\big[(\hat{\mathbf{W}}^{\mathcal{A}}_{\tau})^2\big]\Big\rangle = \frac{1}{d}\mathbb{E}\Big[\text{tr}\big((\hat{\mathbf{W}}^{\mathcal{A}}_{\tau})^2\big)\Big]\Big\{\frac{1}{d}\mathbb{E}\big[\text{tr}(\hat{\mathbf{W}}^{\mathcal{A}}_{\tau})\big]\Big\}^{-2} \geq 1.$$

Proof: We have proved in previous sections that the dominating constant in the statitical error of meta learning method $\mathcal{A}$ adopts the form $\tilde{C}_0^{\mathcal{A}} := \frac{1}{d}\langle \mathbb{E}^{-2}\big[\hat{\mathbf{W}}_\tau^{\mathcal{A}}\big], \mathbb{E}\big[(\hat{\mathbf{W}}_\tau^{\mathcal{A}})^2\big]\rangle$. Next we will prove the equality by showing that $\mathbb{E}[\hat{\mathbf{W}}_\tau^{\mathcal{A}}] = \frac{1}{d}\mathbb{E}[\mathrm{tr}(\hat{\mathbf{W}}_\tau^{\mathcal{A}})]\mathbf{I}_d$.

Let $\mathbf{X}_{\tau,N} = \mathbf{U}_{\tau,N}\mathbf{D}_{\tau,N}\mathbf{V}_{\tau,N}^\top$ be the SVD of $\mathbf{X}_{\tau,N}$, where $\mathbf{U}_{\tau,N} \in \mathbb{R}^{N\times N}, \mathbf{D}_{\tau,N} \in \mathbb{R}^{N\times d}, \mathbf{V}_{\tau,N} \in \mathbb{R}^{d\times d}$. Define $\hat{\mathbf{Q}}_{\tau,N} := \frac{1}{N}\mathbf{X}_{\tau,N}^\top\mathbf{X}_{\tau,N}$, and denote $\lambda_1^{(N)} \geq \cdots \geq \lambda_d^{(N)}$ as the eigenvalues of $\hat{\mathbf{Q}}_{\tau,N}$. Then $\mathbf{D}_{\tau,N}^\top\mathbf{D}_{\tau,N} = N\mathrm{Diag}(\lambda_1^{(N)}, \ldots, \lambda_d^{(N)})$.

For ERM, based on the expression of $\hat{\mathbf{W}}_\tau^{\mathrm{er}}$, it is apparent that

$$\mathbb{E}[\hat{\mathbf{W}}_\tau^{\mathrm{er}}] = \mathbb{E}[\hat{\mathbf{Q}}_{\tau,N}] = \mathbf{Q}_\tau = \mathbf{I}_d = \frac{1}{d}\mathbb{E}[\mathrm{tr}(\hat{\mathbf{W}}_\tau^{\mathrm{er}})]\mathbf{I}_d. \tag{103}$$

For MAML, using the expression of $\hat{\mathbf{W}}_\tau^{\mathrm{ma}}$, we have

$$\begin{aligned}&\mathbb{E}\big[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}\big] = \mathbb{E}\big[\big(\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N_1}\big)\hat{\mathbf{Q}}_{\tau,N_2}\big(\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N_1}\big)\big] = \mathbb{E}\big[\big(\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N_1}\big)\mathbf{Q}_\tau\big(\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N_1}\big)\big]\\ =&\mathbb{E}\Big[\mathbf{V}_{\tau,N_1}\mathrm{Diag}\Big((1-\alpha\lambda_1^{(N_1)})^2, \ldots, (1-\alpha\lambda_d^{(N_1)})^2\Big)\mathbf{V}_{\tau,N_1}^\top\Big].\end{aligned} \tag{104}$$

Then we show that $\mathbb{E}[\hat{\mathbf{W}}_\tau^{\mathrm{ma}}] = \frac{1}{d}\mathbb{E}[\mathrm{tr}(\hat{\mathbf{W}}_\tau^{\mathrm{ma}})]\mathbf{I}_d$ by the permutation trick. We utilize the isotropicity of $\mathbf{X}_\tau$. For notation simplicity, we use $\mathbf{V}_\tau$ to replace $\mathbf{V}_{\tau,N}$ in the following discussion since the value of N does not affect the arguments. Recall that $\mathbf{V}_\tau$ is uniform on all the orthogonal matrices. Let $\mathbf{P} \in \mathbb{R}^{d\times d}$ be any permutation matrix, then $\mathbf{V}_\tau\mathbf{P}$ has the same distribution as $\mathbf{V}_\tau$. For this permuted data matrix $\mathbf{V}_\tau\mathbf{P}$, $\mathbb{E}[\sum_{i=1}^d \lambda_i^{(N)}\mathbf{v}_{\tau,i}\mathbf{v}_{\tau,i}^\top] = \mathbb{E}\big[\sum_{i=1}^d \lambda_i^{(N)}\mathbf{v}_{\tau,t_p(i)}\mathbf{v}_{\tau,t_p(i)}^\top\big]$ with $t_p(i)$ denoting the permutation of the i-th element in $\mathbf{P}$.

Summing over all the permutations $\mathbf{P}$ (and there are totally $d!$ instances), we deduce

$$\begin{aligned}d!\mathbb{E}[\hat{\mathbf{W}}_\tau^{\mathrm{ma}}] =& \sum_{\text{all } t_p}\mathbb{E}\Big[\sum_{i=1}^d(1-\alpha\lambda_i^{(N_1)})^2\mathbf{v}_{\tau,t_p(i)}\mathbf{v}_{\tau,t_p(i)}^\top\Big]\\ =&(d-1)!\mathbb{E}\Big[\sum_{j=1}^d\big(\sum_{i=1}^d(1-\alpha\lambda_i^{(N_1)})^2\big)\mathbf{v}_{\tau,j}\mathbf{v}_{\tau,j}^\top\Big]\\ =&(d-1)!\mathbb{E}\Big[\mathbf{V}_\tau\,\mathrm{Diag}\,\big(\sum_{i=1}^d(1-\alpha\lambda_i^{(N_1)})^2, \ldots, \sum_{i=1}^d(1-\alpha\lambda_i^{(N_1)})^2\big)\mathbf{V}_\tau^\top\Big]\\ =&(d-1)!\mathbb{E}\Big[\sum_{i=1}^d((1-\alpha\lambda_i^{(N_1)})^2)^2\mathbf{V}_\tau\mathbf{V}_\tau^\top\Big] = (d-1)!\mathbb{E}[\mathrm{tr}(\hat{\mathbf{W}}_\tau^{\mathrm{ma}})]\mathbf{I}_d\end{aligned}$$

which gives $\mathbb{E}[\hat{\mathbf{W}}_\tau^{\mathrm{ma}}] = \frac{1}{d}\mathbb{E}[\mathrm{tr}(\hat{\mathbf{W}}_\tau^{\mathrm{ma}})]\mathbf{I}_d$.

Following similar arguments, for iMAML, it also holds that $\mathbb{E}[\hat{\mathbf{W}}_\tau^{\mathrm{im}}] = \frac{1}{d}\mathbb{E}[\mathrm{tr}(\hat{\mathbf{W}}_\tau^{\mathrm{im}})]\mathbf{I}_d$. And for BaMAML, we use the expression $\hat{\mathbf{W}}_\tau^{\mathrm{ba}} = \frac{\gamma s}{1-s}[(\mathbf{I}_d + \gamma^{-1}\hat{\mathbf{Q}}_{\tau,N_1})^{-1} - (\mathbf{I}_d + (\gamma s)^{-1}\hat{\mathbf{Q}}_{\tau,N})^{-1}]$, which apparently gives $\mathbb{E}[\hat{\mathbf{W}}_\tau^{\mathrm{ba}}] = \frac{1}{d}\mathbb{E}[\mathrm{tr}(\hat{\mathbf{W}}_\tau^{\mathrm{ba}})]\mathbf{I}_d$ using the permutation trick.

To summarize, we have proved for all four methods ERM, MAML, iMAML and BaMAML, $\mathbb{E}[\hat{\mathbf{W}}_\tau^{\mathcal{A}}] = \frac{1}{d}\mathbb{E}[\mathrm{tr}(\hat{\mathbf{W}}_\tau^{\mathcal{A}})]\mathbf{I}_d$. Then it is not hard to see that

$$\tilde{C}_0^{\mathcal{A}} := \frac{1}{d}\Big\langle\mathbb{E}^{-2}\big[\hat{\mathbf{W}}_\tau^{\mathcal{A}}\big], \mathbb{E}\big[(\hat{\mathbf{W}}_\tau^{\mathcal{A}})^2\big]\Big\rangle = \frac{1}{d}\mathbb{E}\Big[\mathrm{tr}\big((\hat{\mathbf{W}}_\tau^{\mathcal{A}})^2\big)\Big]\Big\{\frac{1}{d}\mathbb{E}\big[\mathrm{tr}(\hat{\mathbf{W}}_\tau^{\mathcal{A}})\big]\Big\}^{-2}.$$

Finally, applying Jensen's inequality, for any PSD matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, we have $\frac{1}{d}\mathrm{tr}(\mathbf{M}^2) \geq (\frac{1}{d}\mathrm{tr}(\mathbf{M}))^2$, therefore

$$\tilde{C}_0^{\mathcal{A}} := \frac{1}{d}\Big\langle\mathbb{E}^{-2}\big[\hat{\mathbf{W}}_\tau^{\mathcal{A}}\big], \mathbb{E}\big[(\hat{\mathbf{W}}_\tau^{\mathcal{A}})^2\big]\Big\rangle = \frac{1}{d}\mathbb{E}\Big[\mathrm{tr}\big((\hat{\mathbf{W}}_\tau^{\mathcal{A}})^2\big)\Big]\Big\{\frac{1}{d}\mathbb{E}\big[\mathrm{tr}(\hat{\mathbf{W}}_\tau^{\mathcal{A}})\big]\Big\}^{-2} \geq 1.$$

□

Lemma 9 (Constant in statistical error of MAML) *Suppose Assumptions 1,3 hold. The dominating constant in the statistical error of MAML is computed by*

$$\begin{aligned}\tilde{C}_0^{\rm ma} &:= \frac{1}{d}\Big\langle \mathbb{E}\big[\hat{\mathbf{W}}^{\rm ma}_{\tau,N}\big]^{-2}, \mathbb{E}\big[(\hat{\mathbf{W}}^{\rm ma}_{\tau,N})^2\big]\Big\rangle\\ &= \frac{1}{dN_2}\mathbb{E}\Big[\mathrm{tr}^2\Big((\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})^2\Big) + (N_2+1)\mathrm{tr}\Big((\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})^4\Big)\Big]\Big\{\frac{1}{d}\mathbb{E}\big[\mathrm{tr}\big((\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})^2\big)\big]\Big\}^{-2}.\end{aligned}$$

Proof: We reuse the permutation trick to derive $\mathbb{E}\big[(\hat{\mathbf{W}}^{\rm ma}_{\tau,N})^2\big]$ below.

$$\mathbb{E}\big[(\hat{\mathbf{W}}^{\rm ma}_{\tau,N})^2\big] = \mathbb{E}\big[\big(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1}\big)\hat{\mathbf{Q}}_{\tau,N_2}\big(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1}\big)^2\hat{\mathbf{Q}}_{\tau,N_2}\big(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1}\big)\big] \tag{105}$$

We know that $\mathbb{E}\big[(\hat{\mathbf{W}}^{\rm ma}_{\tau,N})^2\big]$ is equal to a scale factor times $\mathbf{I}_d$, the identity matrix. And the scale factor can be derived below

$$\begin{aligned}&\mathrm{tr}\mathbb{E}\big[(\hat{\mathbf{W}}^{\rm ma}_{\tau,N})^2\big] = \mathrm{tr}\mathbb{E}\big[\hat{\mathbf{Q}}_{\tau,N_2}\big(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1}\big)^2\hat{\mathbf{Q}}_{\tau,N_2}\big(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1}\big)^2\big]\\ =&\frac{1}{N_2^2}\mathrm{tr}\mathbb{E}\big[\mathbf{X}^{\rm val}_{\tau,N_2}\big(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1}\big)^2\mathbf{X}^{\rm val\top}_{\tau,N_2}\mathbf{X}^{\rm val}_{\tau,N_2}\big(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1}\big)^2\mathbf{X}^{\rm val\top}_{\tau,N_2}\big]\\ =&\frac{1}{N_2^2}\mathrm{tr}\mathbb{E}\big[\big(\mathbf{X}^{\rm val}_{\tau,N_2}\mathbf{V}^{\rm trn}_{\tau}\mathrm{Diag}\big((1-\alpha\lambda_1^{(N_1)})^2,\dots,(1-\alpha\lambda_d^{(N_1)})^2\big)\mathbf{V}^{\rm trn\top}_{\tau}\mathbf{X}^{\rm val\top}_{\tau,N_2}\big)^2\big]\\ =&\frac{1}{N_2^2}\mathrm{tr}\mathbb{E}\Big[\Big(\sum_{i,j=1}^{N_2}\mathrm{Diag}\big((1-\alpha\lambda_1^{(N_1)})^2,\dots,(1-\alpha\lambda_d^{(N_1)})^2\big)\mathbf{v}_j\mathbf{v}_i^\top\Big)^2\Big]\\ =&\frac{1}{N_2^2}\mathbb{E}\Big[\sum_{i}^{N_2}\mathrm{tr}\Big(\mathrm{Diag}\big((1-\alpha\lambda_1^{(N_1)})^2,\dots,(1-\alpha\lambda_d^{(N_1)})^2\big)\mathbf{v}_i\mathbf{v}_i^\top\Big)^2\\ &+\sum_{i\neq j}(\mathrm{Diag}\big((1-\alpha\lambda_1^{(N_1)})^2,\dots,(1-\alpha\lambda_d^{(N_1)})^2\big)\mathbf{v}_j\mathbf{v}_i^\top)^2\Big]\\ =&\frac{1}{N_2 d}\mathbb{E}\Big[\mathrm{tr}^2\Big(\mathrm{Diag}\big((1-\alpha\lambda_1^{(N_1)})^2,\dots,(1-\alpha\lambda_d^{(N_1)})^2\big)\Big)+2\Big\|\mathrm{Diag}\Big((1-\alpha\lambda_1^{(N_1)})^2,\dots,(1-\alpha\lambda_d^{(N_1)})^2\Big)\Big\|_{\rm F}^2\\ &+(N_2-1)\Big\|\mathrm{Diag}\Big((1-\alpha\lambda_1^{(N_1)})^2,\dots,(1-\alpha\lambda_d^{(N_1)})^2\Big)\Big\|_{\rm F}^2\Big].\end{aligned} \tag{106}$$

By combining Lemma 8, (105), and (106), we arrive at the following

$$\begin{aligned}\tilde{C}_0^{\rm ma} &:= \frac{1}{d}\Big\langle \mathbb{E}\big[\hat{\mathbf{W}}^{\rm ma}_{\tau,N}\big]^{-2}, \mathbb{E}\big[(\hat{\mathbf{W}}^{\rm ma}_{\tau,N})^2\big]\Big\rangle\\ &= \frac{1}{dN_2}\mathbb{E}\Big[\Big(\sum_{i=1}^{d}(1-\alpha\lambda_i^{(N_1)})^2\Big)^2 + (N_2+1)\Big(\sum_{i=1}^{d}(1-\alpha\lambda_i^{(N_1)})^4\Big)\Big]\Big\{\frac{1}{d}\mathbb{E}\big[\sum_{i=1}^{d}(1-\alpha\lambda_i^{(N_1)})^2\big]\Big\}^{-2}\\ &= \frac{1}{dN_2}\mathbb{E}\Big[\mathrm{tr}^2\Big((\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})^2\Big) + (N_2+1)\mathrm{tr}\Big((\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})^4\Big)\Big]\Big\{\frac{1}{d}\mathbb{E}\big[\mathrm{tr}\big((\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})^2\big)\big]\Big\}^{-2}.\end{aligned} \tag{107}$$

□

B.2.2 Bound of statistical errors under Assumptions 1,2

ERM. Following the definition of $\hat{\boldsymbol{\theta}}_0^{\rm er}$ in (29a), we have

$$\hat{\boldsymbol{\theta}}_0^{\rm er}-\boldsymbol{\theta}_0^{\rm er} = \Big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}^{\rm er}_{\tau}\Big)^{-1}\Big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}^{\rm er}_{\tau}(\boldsymbol{\theta}_\tau^{\rm gt}-\boldsymbol{\theta}_0^{\rm er})\Big)+\Big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}^{\rm er}_{\tau}\Big)^{-1}\Big(\sum_{\tau=1}^{T}\frac{1}{N}\mathbf{X}^{\rm all\top}_{\tau,N}\mathbf{e}^{\rm all}_{\tau,N}\Big). \tag{108}$$

To bound the statistical error, we define

$$\mathbf{z}^{\rm all}_{e,\rm er} := \big[\mathbf{e}_1^{\rm all\top},\dots,\mathbf{e}_T^{\rm all\top}\big]^\top \in \mathbb{R}^{NT}, \tag{109a}$$

$$\mathbf{U}_{e,\rm er} := \frac{1}{N}\left[\mathbf{X}^{\rm all\top}_{1,N},\dots,\mathbf{X}^{\rm all\top}_{T,N}\right]^\top\Big(\sum_{\tau=1}^{T}\hat{\mathbf{W}}^{\rm er}_{\tau}\Big)^{-1} \in \mathbb{R}^{NT\times d}. \tag{109b}$$

Under Assumptions 1-2, with $\mathbf{U}_{\rm er}, \mathbf{z}_{\rm er}$ defined in Lemma 7, the ERM statistical error is given by

$$\mathcal{E}_{\rm er}^2(\hat{\boldsymbol{\theta}}_0^{\rm er}) = \|\hat{\boldsymbol{\theta}}_0^{\rm er} - \boldsymbol{\theta}_0^{\rm er}\|^2_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]} \leq 2\Big(\underbrace{\|\mathbf{U}_{\rm er}^\top \mathbf{z}_{\rm er}\|^2_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]}}_{I_1^{\rm er}} + \underbrace{\|\mathbf{U}_{e,{\rm er}}^\top \mathbf{z}_{e,{\rm er}}^{\rm all}\|^2_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]}}_{I_2}\Big). \tag{110}$$

We will then bound terms $I_1^{\rm er}, I_2$ respectively. First the bound for the term $I_1^{\rm er}$ in (110) is provided in Lemma 7, which states that with probability at least $1 - Td^{-10}$, we have

$$\begin{aligned} I_1^{\rm er} \leq& \frac{R^2}{T}\Big(\lambda_{\min}(\mathbb{E}[\mathbf{W}_\tau^{\rm er}])^{-1}\lambda_{\max}(\mathbb{E}[(\mathbf{W}_\tau^{\rm er})^2]) + \widetilde{\mathcal{O}}(\frac{d}{N}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big) \\ &+ \Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big)M^2. \end{aligned} \tag{111}$$

Following similar arguments from Lemma 7, for term I_2, first

$$\begin{aligned} &|\mathbf{z}_{e,{\rm er}}^{{\rm all}\top}\mathbf{U}_{e,{\rm er}}\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]\mathbf{U}_{e,{\rm er}}^\top\mathbf{z}_{e,{\rm er}}^{\rm all} - \mathbb{E}_{\boldsymbol{\theta}_\tau^{\rm gt},\mathbf{e}_\tau|\hat{\mathbf{W}}_\tau^{\rm er}}[\mathbf{z}_{e,{\rm er}}^{{\rm all}\top}\mathbf{U}_{e,{\rm er}}\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]\mathbf{U}_{e,{\rm er}}^\top\mathbf{z}_{e,{\rm er}}^{\rm all}]| \\ \leq& \widetilde{\mathcal{O}}\Big(\big\|\mathbf{U}_{e,{\rm er}}\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]\mathbf{U}_{e,{\rm er}}^\top\big\|_{\rm F}\Big) = \widetilde{\mathcal{O}}\Big(\frac{1}{N}\Big\|\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]\Big(\sum_{\tau=1}^T\hat{\mathbf{W}}_\tau^{\rm er}\Big)^{-2}\Big(\sum_{\tau=1}^T\hat{\mathbf{W}}_\tau^{\rm er}\Big)\Big\|_{\rm F}\Big) \\ =& \widetilde{\mathcal{O}}\Big(\frac{1}{TN}\Big\|\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]\Big(\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}_\tau^{\rm er}\Big)^{-2}\Big(\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}_\tau^{\rm er}\Big)\Big\|_{\rm F}\Big) = \widetilde{\mathcal{O}}\Big(\frac{\sqrt{d}}{TN}\Big) \end{aligned} \tag{112}$$

and the expectation is given by

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\theta}_\tau^{\rm gt},\mathbf{e}_\tau|\hat{\mathbf{W}}_\tau^{\rm er}}[\mathbf{z}_{e,{\rm er}}^{{\rm all}\top}\mathbf{U}_{e,{\rm er}}\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]\mathbf{U}_{e,{\rm er}}^\top\mathbf{z}_{e,{\rm er}}^{\rm all}] = {\rm tr}\Big(\mathbf{U}_{e,{\rm er}}\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]\mathbf{U}_{e,{\rm er}}^\top\Big) \\ =& \frac{d}{TN}\frac{1}{d}\Big\langle\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]\Big(\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}_\tau^{\rm er}\Big)^{-2}, \Big(\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}_\tau^{\rm er}\Big)\Big\rangle = \frac{d}{TN}\frac{1}{d}{\rm tr}\Big(\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]\Big(\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}_\tau^{\rm er}\Big)^{-1}\Big) \\ =& \frac{d}{TN}\Big\{\frac{1}{d}{\rm tr}\Big(\mathbb{E}_\tau[\mathbf{W}_\tau^{\rm er}]\Big(\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}_\tau^{\rm er}\Big)^{-1} - \mathbf{I}\Big) + \underbrace{\frac{1}{d}{\rm tr}\Big(\mathbf{I}_d\Big)}_{=C_1^{\rm er}}\Big\} \end{aligned} \tag{113}$$

Therefore combining (112) and (113), we have with probability at least $1 - Td^{-10}$

$$I_2 \leq \frac{d}{TN}\Big(C_1^{\rm er} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big). \tag{114}$$

Finally, by combining the bound for I_1 and I_2, we conclude that with probability at least $1-Td^{-10}$, the statistical error of ERM is bounded by

$$\begin{aligned} \mathcal{E}_{\rm er}^2(\hat{\boldsymbol{\theta}}_0^{\rm er}) \leq& \frac{R^2}{T}\Big(2C_0^{\rm er} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big) + \frac{d}{TN}\Big(2C_1^{\rm er} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big) \\ &+ \Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big)M^2. \end{aligned} \tag{115}$$

MAML. From the expressions of $\hat{\boldsymbol{\theta}}_0^{\rm ma}$ and $\hat{\mathbf{W}}_\tau^{\rm ma}$, we have

$$\begin{aligned} \hat{\boldsymbol{\theta}}_0^{\rm ma} - \boldsymbol{\theta}_0^{\rm ma} =& \Big(\sum_{\tau=1}^T\hat{\mathbf{W}}_\tau^{\rm ma}\Big)^{-1}\Big(\sum_{\tau=1}^T\hat{\mathbf{W}}_\tau^{\rm ma}(\boldsymbol{\theta}_\tau^{\rm gt} - \boldsymbol{\theta}_0^{\rm ma})\Big) \\ &+ \Big(\sum_{\tau=1}^T\hat{\mathbf{W}}_\tau^{\rm ma}\Big)^{-1}\Big(\sum_{\tau=1}^T(\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N_1})(\frac{1}{N_2}\mathbf{X}_{\tau,N_2}^{{\rm val}\top}\mathbf{e}_{\tau,N_2}^{\rm val} - \frac{\alpha}{N_1}\hat{\mathbf{Q}}_{\tau,N_2}\mathbf{X}_{\tau,N_1}^{{\rm trn}\top}\mathbf{e}_{\tau,N_1}^{\rm trn})\Big). \end{aligned} \tag{116}$$

To bound the statistical error of MAML, we define

$$\mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}} \coloneqq \left[\mathbf{e}_1^{\mathrm{val}\top}, \dots, \mathbf{e}_T^{\mathrm{val}\top}\right]^\top \in \mathbb{R}^{N_2 T}, \tag{117a}$$

$$\mathbf{U}_{e1,\mathrm{ma}}^\top \coloneqq \frac{1}{N_2}\Big(\sum_{\tau=1}^T \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}\Big)^{-1}\Big[\big(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{1,N_1}\big)\mathbf{X}_{1,N_2}^{\mathrm{val}\top}, \dots, \big(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{T,N_1}\big)\mathbf{X}_{T,N_2}^{\mathrm{val}\top}\Big] \in \mathbb{R}^{d\times N_2 T} \tag{117b}$$

$$\mathbf{z}_{e2,\mathrm{ma}}^{\mathrm{trn}} \coloneqq \left[\mathbf{e}_1^{\mathrm{trn}\top}, \dots, \mathbf{e}_T^{\mathrm{trn}\top}\right]^\top \in \mathbb{R}^{N_1 T}, \tag{118a}$$

$$\mathbf{U}_{e2,\mathrm{ma}}^\top \coloneqq \frac{\alpha}{N_1}\Big(\sum_{\tau=1}^T \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}\Big)^{-1}\Big[\big(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{1,N_1}\big)\hat{\mathbf{Q}}_{1,N_2}\mathbf{X}_{1,N_1}^{\mathrm{trn}\top}, \dots, \big(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{T,N_1}\big)\hat{\mathbf{Q}}_{T,N_2}\mathbf{X}_{T,N_1}^{\mathrm{trn}\top}\Big] \in \mathbb{R}^{d\times N_1 T}. \tag{118b}$$

Under Assumptions 1-2, with $\mathbf{U}_{\mathrm{ma}}, \mathbf{z}_{\mathrm{ma}}$ defined in Lemma 7, the MAML statistical error is given by

$$\mathcal{E}_{\mathrm{ma}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{ma}}) \leq 2\Big(\underbrace{\|\mathbf{U}_{\mathrm{ma}}^\top\mathbf{z}_{\mathrm{ma}}\|_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathrm{ma}}]}^2}_{I_1^{\mathrm{ma}}} + \underbrace{\|\mathbf{U}_{e1,\mathrm{ma}}^\top\mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}}\|_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathrm{ma}}]}^2}_{I_2} + \underbrace{\|\mathbf{U}_{e2,\mathrm{ma}}^\top\mathbf{z}_{e2,\mathrm{ma}}^{\mathrm{trn}}\|_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathrm{ma}}]}^2}_{I_3}\Big). \tag{119}$$

We will then bound terms I_1^{ma}, I_2-I_6 respectively. First the bound for the term I_1^{ma} is provided in Lemma 7, which states that with probability at least $1 - Td^{-10}$, we have

$$\begin{aligned} I_1^{\mathrm{ma}} \leq & \frac{R^2}{T}\Big(\lambda_{\min}(\mathbb{E}[\mathbf{W}_\tau^{\mathrm{ma}}])^{-1}\lambda_{\max}(\mathbb{E}[(\mathbf{W}_\tau^{\mathrm{ma}})^2]) + \widetilde{\mathcal{O}}(\frac{d}{N}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big) \\ & + \Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big)M^2. \end{aligned} \tag{120}$$

Following similar arguments as (112), from Lemma 7, for term I_2, first

$$\Big|\|\mathbf{U}_{e1,\mathrm{ma}}^\top\mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}}\|_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathrm{ma}}]}^2 - \mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}, \mathbf{e}_\tau | \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}}[\|\mathbf{U}_{e1,\mathrm{ma}}^\top\mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}}\|_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathrm{ma}}]}^2]\Big| = \widetilde{\mathcal{O}}\Big(\frac{\sqrt{d}}{TN_2}\Big) \tag{121}$$

and the expectation is given by

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}, \mathbf{e}_\tau | \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}}[\mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}\top}\mathbf{U}_{e1,\mathrm{ma}}\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathrm{ma}}]\mathbf{U}_{e1,\mathrm{ma}}^\top\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathrm{ma}}]\mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}}] = \mathrm{tr}\Big(\mathbf{U}_{e1,\mathrm{ma}}\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathrm{ma}}]\mathbf{U}_{e1,\mathrm{ma}}^\top\Big) \\ = & \frac{d}{TN_2}\frac{1}{d}\mathrm{tr}\Big(\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathrm{ma}}](\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}})^{-1}\Big) \\ = & \frac{d}{TN_2}\Big\{\frac{1}{d}\mathrm{tr}\Big(\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathrm{ma}}](\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}})^{-1} - \mathbb{E}^{-1}[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}]\Big) + \underbrace{\frac{1}{d}\mathrm{tr}\Big(\mathbf{I}_d\Big)}_{=C_{1,1}^{\mathrm{ma}}}\Big\}. \end{aligned} \tag{122}$$

Therefore combining (112) and (113), we have

$$I_2 \leq \frac{d}{TN_2}\Big(C_{1,1}^{\mathrm{ma}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big). \tag{123}$$

Following similar arguments from (121)-(123), I_3 satisfies

$$I_3 \leq \frac{d}{TN_1}\Big(C_{1,2}^{\mathrm{ma}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big) \tag{124}$$

with $C_{1,2}^{\mathrm{ma}}$ defined by

$$C_{1,2}^{\mathrm{ma}} \coloneqq \frac{1}{d}\Big\langle\mathbb{E}^{-1}[\mathbf{W}_\tau^{\mathrm{ma}}], \alpha^2\mathbb{E}(\mathbf{I}-\alpha\mathbf{Q}_\tau)\mathbf{Q}_\tau\mathbf{Q}_\tau\mathbf{Q}_\tau(\mathbf{I}-\alpha\mathbf{Q}_\tau)\Big\rangle.$$

Finally, define $C_1^{\mathrm{ma}} := (1-s)^{-1} C_{1,1}^{\mathrm{ma}} + s^{-1} C_{1,2}^{\mathrm{ma}}$. By combining the bound of I_1-I_3, we conclude that with probability at least $1 - Td^{-10}$, the statistical error of MAML is bounded above by

$$\begin{aligned}\mathcal{E}_{\mathrm{ma}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{ma}}) \le & \frac{R^2}{T}\Big(2C_0^{\mathrm{ma}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big) + \frac{d}{TN}\Big(2C_1^{\mathrm{ma}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big)\\ & + \Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big) M^2. \end{aligned} \tag{125}$$

iMAML. Based on $\hat{\boldsymbol{\theta}}_0^{\mathrm{im}}$ in (45a), and $\hat{\mathbf{W}}_\tau^{\mathrm{im}}$ in (45d), we have

$$\begin{aligned}\hat{\boldsymbol{\theta}}_0^{\mathrm{im}} - \boldsymbol{\theta}_0^{\mathrm{im}} =& \Big(\sum_{\tau=1}^T \hat{\mathbf{W}}_\tau^{\mathrm{im}}\Big)^{-1}\Big(\sum_{\tau=1}^T \hat{\mathbf{W}}_\tau^{\mathrm{im}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}} - \boldsymbol{\theta}_0^{\mathrm{im}})\Big)\\ & + \Big(\sum_{\tau=1}^T \hat{\mathbf{W}}_\tau^{\mathrm{im}}\Big)^{-1}\Big(\sum_{\tau=1}^T \gamma \Sigma_{\boldsymbol{\theta}_\tau} \frac{1}{N_2}\mathbf{X}_\tau^{\mathrm{val}\top}\mathbf{e}_\tau^{\mathrm{val}} - \gamma^{-1}\hat{\mathbf{W}}_\tau^{\mathrm{im}}\frac{1}{N_1}\mathbf{X}_\tau^{\mathrm{trn}\top}\mathbf{e}_{\tau,N}\Big). \end{aligned} \tag{126}$$

To bound the iMAML statistical error, define

$$\mathbf{z}_{e1,\mathrm{im}}^{\mathrm{val}} := \left[\mathbf{e}_1^{\mathrm{val}\top}, \dots, \mathbf{e}_T^{\mathrm{val}\top}\right]^\top \in \mathbb{R}^{N_2 T}, \tag{127}$$

$$\mathbf{U}_{e1,\mathrm{im}}^\top := \frac{1}{N_2}\left(\sum_{\tau=1}^T \hat{\mathbf{W}}_{\tau,N}^{\mathrm{im}}\right)^{-1}\left[\gamma\Sigma_{\boldsymbol{\theta}_1,N_1}\mathbf{X}_{1,N_2}^{\mathrm{val}\top}, \dots, \gamma\Sigma_{\boldsymbol{\theta}_T,N_1}\mathbf{X}_{T,N_2}^{\mathrm{val}\top}\right] \in \mathbb{R}^{d\times N_2 T} \tag{128}$$

where $\Sigma_{\boldsymbol{\theta}_\tau,N_1} = (\frac{1}{N_1}\mathbf{X}_{\tau,N_1}^{\mathrm{trn}\top}\mathbf{X}_{\tau,N_1}^{\mathrm{trn}} + \gamma\mathbf{I})^{-1}$, and

$$\mathbf{z}_{e2,\mathrm{im}}^{\mathrm{trn}} := \left[\mathbf{e}_1^{\mathrm{trn}\top}, \dots, \mathbf{e}_T^{\mathrm{trn}\top}\right]^\top \in \mathbb{R}^{N_1 T}, \tag{129}$$

$$\mathbf{U}_{e2,\mathrm{im}}^\top := \frac{1}{N_1}\left(\sum_{\tau=1}^T \hat{\mathbf{W}}_\tau^{\mathrm{im}}\right)^{-1}[\gamma^{-1}\hat{\mathbf{W}}_{1,N}^{\mathrm{im}}\mathbf{X}_1^{\mathrm{trn}\top}, \dots, \gamma^{-1}\hat{\mathbf{W}}_{T,N}^{\mathrm{im}}\mathbf{X}_T^{\mathrm{trn}\top}] \in \mathbb{R}^{d\times N_1 T}. \tag{130}$$

Following similar arguments as the derivation for MAML in (119)-(125), with probability at least $1 - Td^{-10}$, we have

$$\begin{aligned}\mathcal{E}_{\mathrm{im}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{im}}) \le & \frac{R^2}{T}\Big(2C_0^{\mathrm{im}} + \widetilde{\mathcal{O}}(\frac{d}{N}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big) + \frac{d}{TN}\Big(2C_1^{\mathrm{im}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big)\\ & + \Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big) M^2. \end{aligned} \tag{131}$$

with $C_1^{\mathrm{im}} := (1-s)^{-1} C_{1,1}^{\mathrm{im}} + s^{-1} C_{1,2}^{\mathrm{im}}$, and

$$\begin{aligned} & C_{1,1}^{\mathrm{im}} := 1,\\ & C_{1,2}^{\mathrm{im}} := \frac{1}{d}\Big\langle \mathbb{E}[\mathbf{W}_\tau^{\mathrm{im}}]^{-1}, \frac{1}{T}\sum_{\tau=1}^T (\gamma)^{-2}\mathbb{E}[\hat{\mathbf{W}}_\tau^{\mathrm{im}}]\mathbb{E}[\Sigma_{\boldsymbol{\theta}_\tau}^{-1}]\mathbb{E}[\hat{\mathbf{W}}_\tau^{\mathrm{im}}] - \mathbb{E}[\gamma^{-1}(\hat{\mathbf{W}}_\tau^{\mathrm{im}})^2]\Big\rangle. \end{aligned} \tag{132}$$

BaMAML. Based on the expressions of $\hat{\boldsymbol{\theta}}_0^{\mathrm{ba}}$ and $\hat{\mathbf{W}}_\tau^{\mathrm{ba}}$, we have

$$\begin{aligned}\hat{\boldsymbol{\theta}}_0^{\mathrm{ba}} - \boldsymbol{\theta}_0^{\mathrm{ba}} =& \Big(\sum_{\tau=1}^T \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ba}}\Big)^{-1}\Big(\sum_{\tau=1}^T \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ba}}(\boldsymbol{\theta}_\tau^{\mathrm{gt}} - \boldsymbol{\theta}_0^{\mathrm{ba}})\Big)\\ & + \Big(\sum_{\tau=1}^T \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ba}}\Big)^{-1}\Big(\sum_{\tau=1}^T \mathbf{X}_{\tau,N}^{\mathrm{all}\top}\Sigma_{y,N}^{-1}\mathbf{e}_{\tau,N}^{\mathrm{all}} - \mathbf{X}_{\tau,N_1}^{\mathrm{trn}\top}\Sigma_{y,N_1}^{-1}\mathbf{e}_{\tau,N_1}^{\mathrm{trn}}\Big) \end{aligned} \tag{133}$$

where $\Sigma_{y,N}^{-1} = (\mathbf{I}_N + \gamma_b^{-1}\mathbf{X}_{\tau,N}\mathbf{X}_{\tau,N}^\top)^{-1}$.

To bound the statistical error of BaMAML, define

$$\mathbf{z}_{e,\mathrm{ba}}^{\mathrm{all}} \coloneqq \Big[\mathbf{e}_1^{\mathrm{all}\top}, \ldots, \mathbf{e}_T^{\mathrm{all}\top}\Big]^\top \in \mathbb{R}^{NT}, \tag{134}$$

$$\mathbf{U}_{e1,\mathrm{ba}}^\top \coloneqq \frac{1}{N_2}\Big(\sum_{\tau=1}^{T} \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ba}}\Big)^{-1}\Big[\mathbf{X}_{1,N}^\top \Sigma_{y,N}^{-1}, \ldots, \mathbf{X}_{T,N}^\top \Sigma_{y,N}^{-1}\Big] \in \mathbb{R}^{d\times NT}, \tag{135}$$

$$\mathbf{z}_{e2,\mathrm{ba}}^{\mathrm{trn}} \coloneqq \Big[\mathbf{e}_1^{\mathrm{trn}\top}, \ldots, \mathbf{e}_T^{\mathrm{trn}\top}\Big]^\top \in \mathbb{R}^{N_1 T}, \tag{136}$$

$$\mathbf{U}_{e2,\mathrm{ba}}^\top \coloneqq \frac{1}{N_2}\Big(\sum_{\tau=1}^{T} \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ba}}\Big)^{-1}\Big[\mathbf{X}_{1,N_1}^\top \Sigma_{y,N_1}^{-1}, \ldots, \mathbf{X}_{T,N_1}^\top \Sigma_{y,N_1}^{-1}\Big] \in \mathbb{R}^{d\times NT}. \tag{137}$$

Following similar arguments as the derivation for ERM in (110)-(115), with probability at least $1 - Td^{-10}$, we have

$$\begin{aligned}\mathcal{E}_{\mathrm{ba}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{ba}}) \leq & \frac{R^2}{T}\Big(2C_0^{\mathrm{ba}} + \widetilde{\mathcal{O}}(\frac{d}{N}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big) + \frac{d}{TN}\Big(2C_1^{\mathrm{ba}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big) \\ & + \Big(\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{d}{N})\Big)M^2.\end{aligned} \tag{138}$$

with C_1^{ba} derived by

$$\begin{aligned}C_1^{\mathrm{ba}} = 1 \geq & \frac{1}{d}\Big\langle \mathbb{E}[\mathbf{W}_{\tau,N}^{\mathrm{ba}}]^{-1}, (1-s)^{-1}\mathbb{E}[(\mathbf{I}_d + (\gamma s)^{-1}\mathbf{Q}_\tau)^{-1}\mathbf{Q}_\tau(\mathbf{I}_d + (\gamma s)^{-1}\mathbf{Q}_\tau)^{-1} \\ & - s(\mathbf{I}_d + \gamma^{-1}\mathbf{Q}_\tau)^{-1}\mathbf{Q}_\tau(\mathbf{I}_d + \gamma^{-1}\mathbf{Q}_\tau)^{-1}]\Big\rangle.\end{aligned} \tag{139}$$

B.2.3 Bound of statistical error under Assumptions 1,3

ERM. Then under Assumptions 1,3, with $\mathbf{U}_{\mathrm{er}}, \mathbf{z}_{\mathrm{er}}$ defined in Lemma 7, we have

$$\mathcal{E}_{\mathrm{er}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{er}}) = w_{\mathrm{er}}\|\hat{\boldsymbol{\theta}}_0^{\mathrm{er}} - \boldsymbol{\theta}_0^{\mathrm{er}}\|_2^2 = w_{\mathrm{er}}(\underbrace{\|\mathbf{U}_{\mathrm{er}}^\top \mathbf{z}_{\mathrm{er}}\|^2}_{I_1^{\mathrm{er}}} + \underbrace{\|\mathbf{U}_{e,\mathrm{er}}^\top \mathbf{z}_{e,\mathrm{er}}^{\mathrm{all}}\|^2}_{I_2} + \underbrace{2\,\mathbf{z}_{\mathrm{er}}^\top \mathbf{U}_{\mathrm{er}} \mathbf{U}_{e,\mathrm{er}}^\top \mathbf{z}_{e,\mathrm{er}}^{\mathrm{all}}}_{I_3}). \tag{140}$$

We will then bound the redefined terms $I_1^{\mathrm{er}}, I_2, I_3$ in (140) respectively. The bound for the term I_1^{er} in (140) is provided in Lemma 7, which states that with probability at least $1 - Td^{-10}$, the following holds

$$I_1 = \mathbf{z}_{\mathrm{er}}^\top \mathbf{U}_{\mathrm{er}} \mathbf{U}_{\mathrm{er}}^\top \mathbf{z}_{\mathrm{er}} \leq \frac{R^2}{T}\Big(\tilde{C}_0^{\mathrm{er}} + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big). \tag{141}$$

For $\tilde{C}_0^{\mathrm{er}}$, since $\mathbb{E}[\hat{\mathbf{W}}_\tau^{\mathrm{er}}] = \mathbf{Q}_\tau$, and by Lemma 8, we have

$$\tilde{C}_0^{\mathrm{er}} = \frac{1}{d}\mathbb{E}\Big[\mathrm{tr}(\hat{\mathbf{Q}}_{\tau,N}^2)\Big]\Big(\frac{1}{d}\mathbb{E}\big[\mathrm{tr}(\mathbf{Q}_\tau)\big]\Big)^{-2} \tag{142}$$

Following similar arguments from Lemma 7, for term I_2, first

$$|\mathbf{z}_{e,\mathrm{er}}^{\mathrm{all}\top} \mathbf{U}_{e,\mathrm{er}} \mathbf{U}_{e,\mathrm{er}}^\top \mathbf{z}_{e,\mathrm{er}}^{\mathrm{all}} - \mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}, \mathbf{e}_\tau | \hat{\mathbf{W}}_\tau^{\mathrm{er}}}[\mathbf{z}_{e,\mathrm{er}}^{\mathrm{all}\top} \mathbf{U}_{e,\mathrm{er}} \mathbf{U}_{e,\mathrm{er}}^\top \mathbf{z}_{e,\mathrm{er}}^{\mathrm{all}}]| = \widetilde{\mathcal{O}}\Big(\frac{\sqrt{d}}{TN}\Big) \tag{143}$$

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}, \mathbf{e}_\tau | \hat{\mathbf{W}}_\tau^{\mathrm{er}}}[\mathbf{z}_{e,\mathrm{er}}^{\mathrm{all}\top} \mathbf{U}_{e,\mathrm{er}} \mathbf{U}_{e,\mathrm{er}}^\top \mathbf{z}_{e,\mathrm{er}}^{\mathrm{all}}] = \mathrm{tr}\Big(\mathbf{U}_{e,\mathrm{er}} \mathbf{U}_{e,\mathrm{er}}^\top\Big) = \frac{d}{TN}\frac{1}{d}\Big\langle\Big(\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_\tau^{\mathrm{er}}\Big)^{-2}, \Big(\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_\tau^{\mathrm{er}}\Big)\Big\rangle \\ =&\frac{d}{TN}\frac{1}{d}\mathrm{tr}\Big(\Big(\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_\tau^{\mathrm{er}}\Big)^{-1}\Big) = \frac{d}{TN}\Big\{\frac{1}{d}\mathrm{tr}\Big(\Big(\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_\tau^{\mathrm{er}}\Big)^{-1} - \mathbb{E}^{-1}[\hat{\mathbf{W}}_\tau^{\mathrm{er}}]\Big) + \underbrace{\frac{1}{d}\mathrm{tr}\Big(\mathbb{E}^{-1}[\hat{\mathbf{W}}_\tau^{\mathrm{er}}]\Big)}_{=\tilde{C}_1^{\mathrm{er}}}\Big\}\end{aligned} \tag{144}$$

Therefore combining (143) and (144), we have

$$I_2 = \mathbf{z}_{e,\mathrm{er}}^{\mathrm{all}\top} \mathbf{U}_{e,\mathrm{er}} \mathbf{U}_{e,\mathrm{er}}^\top \mathbf{z}_{e,\mathrm{er}}^{\mathrm{all}} \leq \frac{d}{TN}\Big(\tilde{C}_1^{\mathrm{er}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big). \tag{145}$$

For term I_3, note that $\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\mathbf{e}_\tau|\hat{\mathbf{W}}_\tau^{\mathrm{er}}}[\mathbf{z}_{\mathrm{er}}^\top \mathbf{U}_{\mathrm{er}} \mathbf{U}_{e,\mathrm{er}}^\top \mathbf{z}_{e,\mathrm{er}}^{\mathrm{all}}] = 0$. Following a similar argument from (143) to (145), with probability at least $1-\delta$, $|I_3| = |\mathbf{z}_{\mathrm{er}}^\top \mathbf{U}_{\mathrm{er}} \mathbf{U}_{e,\mathrm{er}}^\top \mathbf{z}_{e,\mathrm{er}}^{\mathrm{all}}| \leq \widetilde{\mathcal{O}}(\frac{R}{T\sqrt{N}})$. Finally, by combining (141)-(145), and applying the weight w_{er}, we conclude that with probability at least $1 - Td^{-10}$, the statistical error of ERM is bounded by

$$\begin{aligned}\mathcal{E}_{\mathrm{er}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{er}}) =& w_{\mathrm{er}}\|\hat{\boldsymbol{\theta}}_0^{\mathrm{er}}(\gamma) - \boldsymbol{\theta}_0^{\mathrm{er}}(\gamma)\|_2^2 = \frac{R^2}{T}\Big(w_{\mathrm{er}}\tilde{C}_0^{\mathrm{er}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big)\\ &+ \frac{d}{TN}\Big(w_{\mathrm{er}}\tilde{C}_1^{\mathrm{er}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big) + \widetilde{\mathcal{O}}\Big(\frac{R}{T\sqrt{N}}\Big).\end{aligned} \tag{146}$$

MAML. Under Assumptions 1,3 and $\mathbf{U}_{\mathrm{ma}}, \mathbf{z}_{\mathrm{ma}}$ defined in Lemma 7, we have

$$\begin{aligned}\mathcal{E}_{\mathrm{ma}}^2 = w_{\mathrm{ma}}\|\hat{\boldsymbol{\theta}}_0^{\mathrm{ma}} - \boldsymbol{\theta}_0^{\mathrm{ma}}\|_2^2 = w_{\mathrm{ma}}(&\underbrace{\|\mathbf{U}_{\mathrm{ma}}^\top \mathbf{z}_{\mathrm{ma}}\|^2}_{I_1} + \underbrace{\|\mathbf{U}_{e1,\mathrm{ma}}^\top \mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}}\|^2}_{I_2} + \underbrace{\|\mathbf{U}_{e2,\mathrm{ma}}^\top \mathbf{z}_{e2,\mathrm{ma}}^{\mathrm{trn}}\|^2}_{I_3}\\ + 2\underbrace{\mathbf{z}_{\mathrm{ma}}^\top \mathbf{U}_{\mathrm{ma}} \mathbf{U}_{e1,\mathrm{ma}}^\top \mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}}}_{I_4} &- 2\underbrace{\mathbf{z}_{\mathrm{ma}}^\top \mathbf{U}_{\mathrm{ma}} \mathbf{U}_{e2,\mathrm{ma}}^\top \mathbf{z}_{e2,\mathrm{ma}}^{\mathrm{trn}}}_{I_5} - 2\underbrace{\mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}\top} \mathbf{U}_{e1,\mathrm{ma}} \mathbf{U}_{e2,\mathrm{ma}}^\top \mathbf{z}_{e2,\mathrm{ma}}^{\mathrm{trn}}}_{I_6}).\end{aligned} \tag{147}$$

We will then bound these terms I_1-I_6 in (147) one by one.

To bound term I_1 in (147), from Lemma 7, we have with probability at least $1 - Td^{-10}$

$$I_1 = \mathbf{z}_{\mathrm{ma}}^\top \mathbf{U}_{\mathrm{ma}} \mathbf{U}_{\mathrm{ma}}^\top \mathbf{z}_{\mathrm{ma}} \leq \frac{R^2}{T}\left(\tilde{C}_0^{\mathrm{ma}} + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\right). \tag{148}$$

For $\tilde{C}_0^{\mathrm{ma}}$, by Lemma 9

$$\tilde{C}_0^{\mathrm{ma}} = \frac{1}{dN_2}\mathbb{E}\Big[\mathrm{tr}^2\Big((\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N_1})^2\Big) + (N_2+1)\mathrm{tr}\Big((\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N_1})^4\Big)\Big]\Big[\frac{1}{d}\mathbb{E}\big[\mathrm{tr}\big((\mathbf{I} - \alpha\hat{\mathbf{Q}}_{\tau,N_1})^2\big)\big]\Big]^{-2}. \tag{149}$$

For I_2, from Lemma 3, we have the absolute error around the expectation is given by

$$|\mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}\top} \mathbf{U}_{e1,\mathrm{ma}} \mathbf{U}_{e1,\mathrm{ma}}^\top \mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}} - \mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\mathbf{e}_\tau|\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}}[\mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}\top} \mathbf{U}_{e1,\mathrm{ma}} \mathbf{U}_{e1,\mathrm{ma}}^\top \mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}}]| = \widetilde{\mathcal{O}}\Big(\frac{\sqrt{d}}{TN_2}\Big) \tag{150}$$

and the expectation is given by

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\mathbf{e}_\tau|\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}}[\mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}\top} \mathbf{U}_{e1,\mathrm{ma}} \mathbf{U}_{e1,\mathrm{ma}}^\top \mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}}] = \mathrm{tr}\Big(\mathbf{U}_{e1,\mathrm{ma}} \mathbf{U}_{e1,\mathrm{ma}}^\top\Big) = \frac{d}{TN_2}\frac{1}{d}\mathrm{tr}\Big(\big(\frac{1}{T}\sum_{\tau=1}^T \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}\big)^{-1}\Big)\\ = \frac{d}{TN_2}\Big\{\frac{1}{d}\mathrm{tr}\Big(\big(\frac{1}{T}\sum_{\tau=1}^T \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}\big)^{-1} - \mathbb{E}^{-1}[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}]\Big) + \underbrace{\frac{1}{d}\mathrm{tr}\Big(\mathbb{E}^{-1}[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}]\Big)}_{=\tilde{C}_{1,1}^{\mathrm{ma}}}\Big\}.\end{aligned} \tag{151}$$

And by combining (150) and (151) , with probability at least $1-\delta$, we have

$$I_2 = \mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}\top} \mathbf{U}_{e1,\mathrm{ma}} \mathbf{U}_{e1,\mathrm{ma}}^\top \mathbf{z}_{e1,\mathrm{ma}}^{\mathrm{val}} = \frac{d}{TN_2}\Big(\tilde{C}_{1,1}^{\mathrm{ma}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big). \tag{152}$$

For I_3, the absolute error around the expectation is given by

$$|\|\mathbf{U}_{e2,\mathrm{ma}}^\top \mathbf{z}_{e2,\mathrm{ma}}^{\mathrm{trn}}\|^2 - \mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\mathbf{e}_\tau|\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}}[\|\mathbf{U}_{e2,\mathrm{ma}}^\top \mathbf{z}_{e2,\mathrm{ma}}^{\mathrm{trn}}\|^2]| = \widetilde{\mathcal{O}}\Big(\frac{\sqrt{d}}{TN_1}\Big) \tag{153}$$

and the expectation is given by

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}},\mathbf{e}_\tau|\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}}[\mathbf{z}_{e2,\mathrm{ma}}^{\mathrm{trn}\top}\mathbf{U}_{e2,\mathrm{ma}}\mathbf{U}_{e2,\mathrm{ma}}^\top\mathbf{z}_{e2,\mathrm{ma}}^{\mathrm{trn}}] = \mathrm{tr}\Big(\mathbf{U}_{e2,\mathrm{ma}}\mathbf{U}_{e2,\mathrm{ma}}^\top\Big)\\
=&\frac{d}{TN_1}\frac{1}{d}\mathrm{tr}\Big(\Big(\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}\Big)^{-2}\Big(\frac{1}{T}\sum_{\tau=1}^T\alpha^2(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})\hat{\mathbf{Q}}_{\tau,N_2}\hat{\mathbf{Q}}_{\tau,N_1}\hat{\mathbf{Q}}_{\tau,N_2}(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})\Big)\Big)\\
=&\frac{d}{TN_1}\Big\{\frac{1}{d}\Big\langle\Big(\frac{1}{T}\sum_{\tau=1}^T\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}\Big)^{-2}-\mathbb{E}^{-2}[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}],\frac{1}{T}\sum_{\tau=1}^T\alpha^2(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})\hat{\mathbf{Q}}_{\tau,N_2}\hat{\mathbf{Q}}_{\tau,N_1}\hat{\mathbf{Q}}_{\tau,N_2}(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})\Big\rangle\\
&+\frac{1}{d}\Big\langle\mathbb{E}^{-2}[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}],\frac{1}{T}\sum_{\tau=1}^T\alpha^2(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})\hat{\mathbf{Q}}_{\tau,N_2}\hat{\mathbf{Q}}_{\tau,N_1}\hat{\mathbf{Q}}_{\tau,N_2}(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})\\
&-\mathbb{E}[\alpha^2(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})\hat{\mathbf{Q}}_{\tau,N_2}\hat{\mathbf{Q}}_{\tau,N_1}\hat{\mathbf{Q}}_{\tau,N_2}(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})]\Big\rangle\\
&+\underbrace{\frac{1}{d}\Big\langle\mathbb{E}^{-2}[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ma}}],\mathbb{E}[\alpha^2(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})\hat{\mathbf{Q}}_{\tau,N_2}\hat{\mathbf{Q}}_{\tau,N_1}\hat{\mathbf{Q}}_{\tau,N_2}(\mathbf{I}-\alpha\hat{\mathbf{Q}}_{\tau,N_1})]\Big\rangle}_{=\tilde{C}_{1,2}^{\mathrm{ma}}}\Big\}.
\end{aligned}\tag{154}$$

For I_3, with probability at least $1-\delta$

$$I_3=\mathbf{z}_{e2,\mathrm{ma}}^{\mathrm{trn}\top}\mathbf{U}_{e2,\mathrm{ma}}\mathbf{U}_{e2,\mathrm{ma}}^\top\mathbf{z}_{e2,\mathrm{ma}}^{\mathrm{trn}}=\frac{d}{TN_1}\Big(\tilde{C}_{1,2}^{\mathrm{ma}}+\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})+\widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big).\tag{155}$$

For I_4, I_5, I_6, with probability at least $1-\delta$

$$|I_4|\leq\widetilde{\mathcal{O}}(\frac{R}{T\sqrt{N_2}}),|I_5|\leq\widetilde{\mathcal{O}}(\frac{R}{T\sqrt{N_1}}),|I_6|\leq\widetilde{\mathcal{O}}(\frac{\sqrt{d}}{T\sqrt{N_1N_2}}).\tag{156}$$

Finally, applying the weight w_{ma}, we have with probability $1-Td^{-10}$ the statistical error of MAML is bounded by

$$\begin{aligned}
\mathcal{E}_{\mathrm{ma}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{ma}})=w_{\mathrm{ma}}\|\hat{\boldsymbol{\theta}}_0^{\mathrm{ma}}(\gamma)-\boldsymbol{\theta}_0^{\mathrm{ma}}(\gamma)\|_2^2=\frac{R^2}{T}\Big(w_{\mathrm{ma}}\tilde{C}_0^{\mathrm{ma}}+\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})+\widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big)\\
+\frac{d}{TN}\Big(w_{\mathrm{ma}}\tilde{C}_1^{\mathrm{ma}}+\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})+\widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big)+\widetilde{\mathcal{O}}\Big(\frac{R}{T\sqrt{N}}\Big).
\end{aligned}\tag{157}$$

iMAML. With $\mathbf{U}_{\mathrm{im}}, \mathbf{z}_{\mathrm{im}}$ defined in Lemma 7, we can rewrite (126) as

$$\hat{\boldsymbol{\theta}}_0^{\mathrm{im}}-\boldsymbol{\theta}_0^{\mathrm{im}}=\mathbf{U}_{\mathrm{im}}^\top\mathbf{z}_{\mathrm{im}}+\mathbf{U}_{e1,\mathrm{im}}^\top\mathbf{z}_{e1,\mathrm{im}}^{\mathrm{val}}-\mathbf{U}_{e2,\mathrm{im}}^\top\mathbf{z}_{e2,\mathrm{im}}^{\mathrm{trn}}.\tag{158}$$

Thus the squared error can be computed by

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}_0^{\mathrm{im}}-\boldsymbol{\theta}_0^{\mathrm{im}}\|_2^2=&\underbrace{\|\mathbf{U}_{\mathrm{im}}^\top\mathbf{z}_{\mathrm{im}}\|^2}_{I_1}+\underbrace{\|\mathbf{U}_{e1,\mathrm{im}}^\top\mathbf{z}_{e1,\mathrm{im}}^{\mathrm{val}}\|^2}_{I_2}+\underbrace{\|\mathbf{U}_{e2,\mathrm{im}}^\top\mathbf{z}_{e2,\mathrm{im}}^{\mathrm{trn}}\|^2}_{I_3}\\
&+2\underbrace{\mathbf{z}_{\mathrm{im}}^\top\mathbf{U}_{\mathrm{im}}\mathbf{U}_{e1,\mathrm{im}}^\top\mathbf{z}_{e1,\mathrm{im}}^{\mathrm{val}}}_{I_4}-2\underbrace{\mathbf{z}_{\mathrm{im}}^\top\mathbf{U}_{\mathrm{im}}\mathbf{U}_{e2,\mathrm{im}}^\top\mathbf{z}_{e2,\mathrm{im}}^{\mathrm{trn}}}_{I_5}-2\underbrace{\mathbf{z}_{e1,\mathrm{im}}^{\mathrm{val}\top}\mathbf{U}_{e1,\mathrm{im}}\mathbf{U}_{e2,\mathrm{im}}^\top\mathbf{z}_{e2,\mathrm{im}}^{\mathrm{trn}}}_{I_6}.
\end{aligned}\tag{159}$$

To bound term I_1 in (159), from Lemma 7, we have with probability at least $1-Td^{-10}$

$$I_1=\mathbf{z}_{\mathrm{im}}^\top\mathbf{U}_{\mathrm{im}}\mathbf{U}_{\mathrm{im}}^\top\mathbf{z}_{\mathrm{im}}\leq\frac{R^2}{T}\Big(\tilde{C}_0^{\mathrm{im}}+\widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})+\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big).\tag{160}$$

For $\tilde{C}_0^{\mathrm{im}}$, by Lemma C.2 in (Bai et al., 2021),

$$\tilde{C}_0^{\mathrm{im}}=\frac{\frac{1}{dN_2}\mathbb{E}\big[\,\mathrm{tr}\,\big(\gamma^2\big(\hat{\mathbf{Q}}_{\tau,N_1}+\gamma\mathbf{I}_d\big)^{-2}\big)^2+\big(N_2+1\big)\,\mathrm{tr}\,\big(\gamma^4\big(\hat{\mathbf{Q}}_{\tau,N_1}+\gamma\mathbf{I}_d\big)^{-4}\big)\big]}{\big(\frac{1}{d}\mathbb{E}\big[\,\mathrm{tr}\,\big(\gamma^2\big(\hat{\mathbf{Q}}_{\tau,N_1}+\gamma\mathbf{I}_d\big)^{-2}\big)\big]\big)^2}\tag{161}$$

For I_2, first from Lemma 3

$$\Big|\mathbf{z}_{e1,\mathrm{im}}^{\mathrm{val}\top}\mathbf{U}_{e1,\mathrm{im}}\mathbf{U}_{e1,\mathrm{im}}^{\top}\mathbf{z}_{e1,\mathrm{im}}^{\mathrm{val}} - \mathbb{E}_{\boldsymbol{\theta}_{\tau}^{\mathrm{gt}},\mathbf{e}_{\tau}|\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}}[\mathbf{z}_{e1,\mathrm{im}}^{\mathrm{val}\top}\mathbf{U}_{e1,\mathrm{im}}\mathbf{U}_{e1,\mathrm{im}}^{\top}\mathbf{z}_{e1,\mathrm{im}}^{\mathrm{val}}]\Big| = \widetilde{\mathcal{O}}\Big(\frac{\sqrt{d}}{TN_2}\Big) \tag{162}$$

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_{\tau}^{\mathrm{gt}},\mathbf{e}_{\tau}|\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}}[\mathbf{z}_{e1,\mathrm{im}}^{\mathrm{val}\top}\mathbf{U}_{e1,\mathrm{im}}\mathbf{U}_{e1,\mathrm{im}}^{\top}\mathbf{z}_{e1,\mathrm{im}}^{\mathrm{val}}] = \frac{1}{TN_2}\mathrm{tr}\Big(\Big(\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}\Big)^{-1}\Big)\\
=&\frac{d}{TN_2}\Bigg(\underbrace{\frac{1}{d}\mathrm{tr}\Big(\mathbb{E}^{-1}[\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}]\Big)}_{=C_{1,1}^{\mathrm{im}}} + \underbrace{\frac{1}{d}\mathrm{tr}\Big(\Big(\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}\Big)^{-1} - \mathbb{E}^{-1}[\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}]\Big)}_{I_9}\Bigg),
\end{aligned} \tag{163}$$

and by Lemma 2, with probability at least $1 - Td^{-10}$

$$|I_9| \leq \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}). \tag{164}$$

Therefore, combining (162),(163),(164) with probability at least $1 - Td^{-10}$

$$I_2 \leq \frac{d}{TN_2}\Big(\tilde{C}_{1,1}^{\mathrm{im}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big) \tag{165}$$

$$\tilde{C}_{1,1}^{\mathrm{im}} \coloneqq \frac{1}{d}\mathrm{tr}\Big(\mathbb{E}^{-1}[\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}]\Big). \tag{166}$$

Similarly, for I_3, based on Lemma 3 we have

$$|\|\mathbf{U}_{e2,\mathrm{im}}^{\top}\mathbf{z}_{e2,\mathrm{im}}^{\mathrm{trn}}\|^2 - \mathbb{E}_{\mathbf{e}_{\tau}|\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}}[\|\mathbf{U}_{e2,\mathrm{im}}^{\top}\mathbf{z}_{e2,\mathrm{im}}^{\mathrm{trn}}\|^2]| = \widetilde{\mathcal{O}}\Big(\frac{\sqrt{d}}{TN_1}\Big). \tag{167}$$

And similar to the derivations in ERM and MAML, with Lemma 2, it holds that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{e}_{\tau}|\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}}[\|\mathbf{U}_{e2,\mathrm{im}}^{\top}\mathbf{z}_{e2,\mathrm{im}}^{\mathrm{trn}}\|^2] = \frac{1}{TN_1}\Big\langle\Big(\frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}\Big)^{-2}, \Big(\frac{1}{T}\sum_{\tau=1}^{T}(\gamma)^{-2}\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}\frac{1}{N_1}\mathbf{X}_{\tau}^{\mathrm{trn}\top}\mathbf{X}_{\tau}^{\mathrm{trn}}\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}\Big)\Big\rangle\\
=&\frac{d}{TN_1}\Bigg\{\underbrace{\frac{1}{d}\Big\langle\mathbb{E}[\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}]^{-2}, \frac{1}{T}\sum_{\tau=1}^{T}\gamma^{-2}\mathbb{E}[\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}]\mathbb{E}[\Sigma_{\boldsymbol{\theta}_{\tau}}^{-1}]\mathbb{E}[\hat{\mathbf{W}}_{\tau}^{\mathrm{im}}] - \mathbb{E}[\gamma^{-1}(\hat{\mathbf{W}}_{\tau}^{\mathrm{im}})^2]\Big\rangle}_{=\tilde{C}_{1,2}^{\mathrm{im}}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Bigg\}
\end{aligned} \tag{168}$$

Combining (167) and (168), with probability at least $1 - Td^{-10}$, we have

$$I_3 = \mathbf{z}_{e2,\mathrm{im}}^{\mathrm{trn}\top}\mathbf{U}_{e2,\mathrm{im}}\mathbf{U}_{e2,\mathrm{im}}^{\top}\mathbf{z}_{e2,\mathrm{im}}^{\mathrm{trn}} \leq \frac{d}{TN_1}\Big(\tilde{C}_{1,2}^{\mathrm{im}} + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big) \tag{169}$$

Following a similar argument, for I_4, I_5, I_6, with probability at least $1 - \delta$

$$|I_4| \leq \widetilde{\mathcal{O}}(\frac{R}{T\sqrt{N_2}}), |I_5| \leq \widetilde{\mathcal{O}}(\frac{R}{T\sqrt{N_1}}), |I_6| \leq \widetilde{\mathcal{O}}(\frac{\sqrt{d}}{T\sqrt{N_1N_2}}). \tag{170}$$

Finally, define $\tilde{C}_1^{\mathrm{im}} \coloneqq (1-s)^{-1}\tilde{C}_{1,1}^{\mathrm{im}} + s^{-1}\tilde{C}_{1,2}^{\mathrm{im}}$, applying the weight w_{im} we have with probability at least $1-Td^{-10}$, the statistical error of iMAML is bounded by

$$\begin{aligned}
\mathcal{E}_{\mathrm{im}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{im}}) = w_{\mathrm{im}}\|\hat{\boldsymbol{\theta}}_0^{\mathrm{im}}(\gamma) - \boldsymbol{\theta}_0^{\mathrm{im}}(\gamma)\|_2^2 = \frac{R^2}{T}\Big(w_{\mathrm{im}}\tilde{C}_0^{\mathrm{im}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big)\\
+ \frac{d}{TN}\Big(w_{\mathrm{im}}\tilde{C}_1^{\mathrm{im}} + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big) + \widetilde{\mathcal{O}}\Big(\frac{R}{T\sqrt{N}}\Big).
\end{aligned} \tag{171}$$

BaMAML. With $\mathbf{U}_{\mathrm{ba}}, \mathbf{z}_{\mathrm{ba}}$ defined in Lemma 7

$$\hat{\boldsymbol{\theta}}_0^{\mathrm{ba}} - \boldsymbol{\theta}_0^{\mathrm{ba}} = \mathbf{U}_{\mathrm{ba}}^\top \mathbf{z}_{\mathrm{ba}} + \mathbf{U}_{e1,\mathrm{ba}}^\top \mathbf{z}_{e1,\mathrm{ba}} - \mathbf{U}_{e2,\mathrm{ba}}^\top \mathbf{z}_{e2,\mathrm{ba}} \tag{172}$$

$$\begin{aligned}\|\hat{\boldsymbol{\theta}}_0^{\mathrm{ba}} - \boldsymbol{\theta}_0^{\mathrm{ba}}\|_2^2 =& \underbrace{\|\mathbf{U}_{\mathrm{ba}}^\top \mathbf{z}_{\mathrm{ba}}\|^2}_{I_1} + \underbrace{\|\mathbf{U}_{e1,\mathrm{ba}}^\top \mathbf{z}_{e1,\mathrm{ba}}^{\mathrm{all}}\|^2}_{I_2} + \underbrace{\|\mathbf{U}_{e2,\mathrm{ba}}^\top \mathbf{z}_{e2,\mathrm{ba}}^{\mathrm{trn}}\|^2}_{I_3} \\ &+ 2\underbrace{\mathbf{z}_{\mathrm{ba}}^\top \mathbf{U}_{\mathrm{ba}} \mathbf{U}_{e1,\mathrm{ba}}^\top \mathbf{z}_{e1,\mathrm{ba}}^{\mathrm{all}}}_{I_4} - 2\underbrace{\mathbf{z}_{\mathrm{ba}}^\top \mathbf{U}_{\mathrm{ba}} \mathbf{U}_{e2,\mathrm{ba}}^\top \mathbf{z}_{e2,\mathrm{ba}}^{\mathrm{trn}}}_{I_5} - 2\underbrace{\mathbf{z}_{e1,\mathrm{ba}}^{\mathrm{all}\top} \mathbf{U}_{e1,\mathrm{ba}} \mathbf{U}_{e2,\mathrm{ba}}^\top \mathbf{z}_{e2,\mathrm{ba}}^{\mathrm{trn}}}_{I_6}.\end{aligned} \tag{173}$$

To bound term I_1 in (173), from Lemma 7, we have with probability at least $1 - Td^{-10}$

$$I_1 = \mathbf{z}_{\mathrm{ba}}^\top \mathbf{U}_{\mathrm{ba}} \mathbf{U}_{\mathrm{ba}}^\top \mathbf{z}_{\mathrm{ba}} \leq \frac{R^2}{T}\Big(\tilde{C}_0^{\mathrm{ba}} + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big). \tag{174}$$

To compute $\tilde{C}_0^{\mathrm{ba}}$, by Lemma 8

$$\tilde{C}_0^{\mathrm{ba}} = \frac{1}{d}\mathbb{E}\Big[\mathrm{tr}\Big((\mathbf{W}_{\tau,N_a}^{\hat{\mathrm{ba}}})^2\Big)\Big] \cdot \Big\{\frac{1}{d}\mathbb{E}\Big[\mathrm{tr}\Big(\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ba}}\Big)\Big]\Big\}^{-2} \tag{175}$$

For term I_2, based on Lemma 3, the absolute error around the expectation is given by

$$|\|\mathbf{U}_{e,\mathrm{ba}}^\top \mathbf{z}_{e,\mathrm{ba}}^{\mathrm{all}}\|^2 - \mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}, \mathbf{e}_\tau | \hat{\mathbf{W}}_\tau^{\mathrm{ba}}}[\|\mathbf{U}_{e,\mathrm{ba}}^\top \mathbf{z}_{e,\mathrm{ba}}^{\mathrm{all}}\|^2]| = \widetilde{\mathcal{O}}\Big(\frac{\sqrt{d}}{TN}\Big) \tag{176}$$

where the expectation is given by

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}, \mathbf{e}_\tau | \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ba}}}[\mathbf{z}_{e1,\mathrm{ba}}^{\mathrm{all}\top} \mathbf{U}_{e1,\mathrm{ba}} \mathbf{U}_{e1,\mathrm{ba}}^\top \mathbf{z}_{e1,\mathrm{ba}}^{\mathrm{all}}] = \mathrm{tr}\Big(\mathbf{U}_{e1,\mathrm{ba}} \mathbf{U}_{e1,\mathrm{ba}}^\top\Big) \\ =& \frac{d}{TN}\frac{1}{d}\Big\langle\Big(\frac{1}{T}\sum_{\tau=1}^T \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ba}}\Big)^{-2}, (1-s)^{-1}\Big(\frac{1}{T}\sum_{\tau=1}^T (\mathbf{I}_d + (\gamma s)^{-1}\hat{\mathbf{Q}}_{\tau,N})^{-1}\hat{\mathbf{Q}}_{\tau,N}(\mathbf{I}_d + (\gamma s)^{-1}\hat{\mathbf{Q}}_{\tau,N})^{-1}\Big)\Big\rangle \\ \leq& \frac{d}{TN}\Bigg\{\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \underbrace{\frac{1}{d}\Big\langle\mathbb{E}[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ba}}]^{-2}, (1-s)^{-1}\mathbb{E}[(\mathbf{I}_d + (\gamma s)^{-1}\hat{\mathbf{Q}}_{\tau,N})^{-1}\hat{\mathbf{Q}}_{\tau,N}(\mathbf{I}_d + (\gamma s)^{-1}\hat{\mathbf{Q}}_{\tau,N})^{-1}]\Big\rangle}_{=\tilde{C}_{1,1}^{\mathrm{ba}}}\Bigg\}.\end{aligned} \tag{177}$$

Similarly,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}, \mathbf{e}_\tau | \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ba}}}[\mathbf{z}_{e2,\mathrm{ba}}^\top \mathbf{U}_{e2,\mathrm{ba}} \mathbf{U}_{e2,\mathrm{ba}}^\top \mathbf{z}_{e2,\mathrm{ba}}] \\ \leq& \frac{d}{TN}\Bigg\{\widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \underbrace{\frac{1}{d}\Big\langle\mathbb{E}[\hat{\mathbf{W}}_{\tau,N}^{\mathrm{ba}}]^{-2}, s(1-s)^{-1}\mathbb{E}[(\mathbf{I}_d + \gamma^{-1}\mathbf{Q}_\tau)^{-1}\mathbf{Q}_\tau(\mathbf{I}_d + \gamma^{-1}\mathbf{Q}_\tau)^{-1}]\Big\rangle}_{=\tilde{C}_{1,2}^{\mathrm{ba}}}\Bigg\}.\end{aligned} \tag{178}$$

$\tilde{C}_1^{\mathrm{ba}} := \tilde{C}_{1,1}^{\mathrm{ba}} - \tilde{C}_{1,2}^{\mathrm{ba}}$, combining the above derivations with Lemma 2 gives the higher order terms in (177), which leads to

$$\mathbb{E}_{\boldsymbol{\theta}_\tau^{\mathrm{gt}}, \mathbf{e}_\tau | \hat{\mathbf{W}}_{\tau,N}^{\mathrm{ba}}}[\mathbf{z}_{e,\mathrm{ba}}^{\mathrm{all}\top} \mathbf{U}_{e,\mathrm{ba}} \mathbf{U}_{e,\mathrm{ba}}^\top \mathbf{z}_{e,\mathrm{ba}}^{\mathrm{all}}] \leq \frac{d}{TN}(\tilde{C}_1^{\mathrm{ba}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})) \tag{179}$$

Combining (176) and (179), with probability at least $1 - Td^{-10}$, we have

$$I_2 = \mathbf{z}_{e,\mathrm{ba}}^{\mathrm{all}\top} \mathbf{U}_{e,\mathrm{ba}} \mathbf{U}_{e,\mathrm{ba}}^\top \mathbf{z}_{e,\mathrm{ba}}^{\mathrm{all}} \leq \frac{d}{TN}\Big(\tilde{C}_1^{\mathrm{ba}} + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}}) + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}})\Big) \tag{180}$$

Following a similar argument, with probability $1 - \delta$, $|I_3| \leq \widetilde{\mathcal{O}}(\frac{R}{T\sqrt{N}})$.

Finally, applying the weight w_{ba}, we have with probability $1 - Td^{-10}$, the statistical error of BaMAML is bounded by

$$\begin{aligned}\mathcal{E}_{\mathrm{ba}}^2(\hat{\boldsymbol{\theta}}_0^{\mathrm{ba}}) =& w_{\mathrm{ba}}\|\hat{\boldsymbol{\theta}}_0^{\mathrm{ba}} - \boldsymbol{\theta}_0^{\mathrm{ba}}\|_2^2 = \frac{R^2}{T}\Big(w_{\mathrm{ba}}\tilde{C}_0^{\mathrm{ba}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big) \\ &+ \frac{d}{TN}\Big(w_{\mathrm{ba}}\tilde{C}_1^{\mathrm{ba}} + \widetilde{\mathcal{O}}(\sqrt{\frac{d}{T}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{d}})\Big) + \widetilde{\mathcal{O}}\Big(\frac{R}{T\sqrt{N}}\Big).\end{aligned} \tag{181}$$

B.2.4 Asymptotic dominating constant under Assumptions 1,3

Theorem 9 (Asymptotic ERM constant) *As $d, N \to \infty$, $d/N \to \eta$, the optimal constant of the ERM method $\tilde{C}_0^{\mathrm{er}}$ satisfies*

$$\lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}} \tilde{C}_0^{\mathrm{er}} = 1+\eta.$$

Proof: Recall that $\tilde{C}_0^{\mathrm{er}} = \frac{1}{d}\mathbb{E}\Big[\mathrm{tr}(\hat{\mathbf{Q}}_{\tau,N}^2)\Big]\Big(\frac{1}{d}\mathbb{E}\big[\mathrm{tr}(\mathbf{Q}_\tau)\big]\Big)^{-2}$. Based on Assumption 3, $\mathbb{E}\big[\mathrm{tr}(\mathbf{Q}_\tau)\big] = \mathbb{E}\big[\mathrm{tr}(\mathbf{I}_d)\big] = d$, And $\mathbb{E}[(\hat{\mathbf{Q}}_{\tau,N}^2)]$ can be derived by

$$\begin{aligned}
\mathbb{E}\big[\hat{\mathbf{Q}}_{\tau,N}^2\big] &= \mathbb{E}\big[(\frac{1}{N}\mathbf{X}_{\tau,N}^\top\mathbf{X}_{\tau,N})^2\big] = \mathbb{E}\big[(\frac{1}{N}\mathbf{X}_{\tau,N}^\top\mathbf{X}_{\tau,N})^2\big] = \mathbb{E}\big[(\frac{1}{N}\sum_i \mathbf{x}_{\tau,i}\mathbf{x}_{\tau,i}^\top)^2\big]\\
&= \frac{1}{N}\mathbb{E}\big[(\mathbf{x}_{\tau,i}\mathbf{x}_{\tau,i}^\top)^2\big] + \frac{N-1}{N}\mathbf{I}_d
\end{aligned}$$

where $\mathrm{tr}(\mathbb{E}\big[(\mathbf{x}_{\tau,i}\mathbf{x}_{\tau,i}^\top)^2\big]) = \mathbb{E}\big[(\sum_j x_{\tau,ij}^2)^2\big] = d(d+2)$.

Therefore

$$\tilde{C}_0^{\mathrm{er}} = \frac{d+N+1}{N}, \qquad \lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}} \tilde{C}_0^{\mathrm{er}} = 1+\eta.$$

□

Theorem 10 (Asymptotic MAML constant) *As $d, N \to \infty$, $d/N \to \eta$, the optimal constant of the MAML method, $\tilde{C}_0^{\mathrm{ma}}$, by tuning the step size $\alpha \in (0, 1/\bar{\lambda})$ and the train-val split ratio $s \in (0,1)$, satisfies*

$$\inf_{\substack{\alpha\,>\,0\\ s\,\in\,(0,1)}} \lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}} \tilde{C}_0^{\mathrm{ma}} = 1+\eta$$

Proof: We first derive a lower bound for $\inf_{\substack{\alpha\,>\,0\\ s\,\in\,(0,1)}} \lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}}$ by

$$\begin{aligned}
&\inf_{\substack{\alpha\,>\,0\\ s\,\in\,(0,1)}} \lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}} \tilde{C}_0^{\mathrm{ma}}\\
&\geq \lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}} \inf_{\substack{\alpha\,>\,0\\ s\,\in\,(0,1)}} \frac{\frac{1}{dN_2}\mathbb{E}\Big[\Big(\sum_{i=1}^d(1-\alpha\lambda_i^{(N_1)})^2\Big)^2 + (N_2+1)\Big(\sum_{i=1}^d(1-\alpha\lambda_i^{(N_1)})^4\Big)\Big]}{\frac{1}{d^2}\mathbb{E}^2\Big[\sum_{i=1}^d(1-\alpha\lambda_i^{(N_1)})^2\Big]}\\
&\geq \lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}} \inf_{\substack{\alpha\,>\,0\\ s\,\in\,(0,1)}} \frac{d+N_2+1}{N_2} = 1+\eta.
\end{aligned} \tag{182}$$

Next we derive the upper bound for $\inf_{\substack{\alpha\,>\,0\\ s\,\in\,(0,1)}} \lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}} \tilde{C}_0^{\mathrm{ma}}$. As for any PD matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, $\frac{1}{d}\mathrm{tr}(\mathbf{M}^2) \geq (\frac{1}{d}\mathrm{tr}(\mathbf{M}))^2$, then applying this inequality we obtain

$$\begin{aligned}
\lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}} \tilde{C}_0^{\mathrm{ma}} &\leq \lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}} \frac{\frac{1}{N_2}(d+N_2+1)}{\frac{1}{d^2}\mathbb{E}^2\big[\sum_{i=1}^d(1-\alpha\lambda_d^{(N_1)})^2\big]} \leq \lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}} \frac{\frac{1}{N_2}(d+N_2+1)}{\mathbb{E}^2\big[\big(1-\frac{\alpha}{d}\sum_{i=1}^d\lambda_i^{(N_1)}\big)^2\big]}\\
&\leq \lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}} \frac{\frac{1}{N_2}(d+N_2+1)}{\mathbb{E}^2\big[\big(1-\alpha\mathbb{E}[\frac{1}{d}\sum_{i=1}^d\lambda_i^{(N_1)}]\big)^2\big]} = \lim_{\substack{d,\,N\,\to\,\infty\\ d/N\,\to\,\eta}} \frac{\frac{1}{N_2}(d+N_2+1)}{(1-\alpha)^2}.
\end{aligned} \tag{183}$$

Therefore

$$\inf_{\substack{\alpha > 0 \\ s \in (0,1)}} \lim_{\substack{d,\, N \to \infty \\ d/N \to \eta}} \tilde{C}_0^{\mathrm{ma}} \leq \inf_{s \in (0,1)} \lim_{\substack{d,\, N \to \infty \\ d/N \to \eta}} \frac{d + N_2 + 1}{N_2} = 1 + \eta \tag{184}$$

Based on (182) and (184) we arrive at

$$\inf_{\substack{\alpha > 0 \\ s \in (0,1)}} \lim_{\substack{d,\, N \to \infty \\ d/N \to \eta}} \tilde{C}_0^{\mathrm{ma}} = 1 + \eta \tag{185}$$

□

Theorem 11 (Asymptotic dominating constant of iMAML) *(Bai et al., 2021) As $d, N \to \infty$, $d/N \to \eta$, the optimal constant of the iMAML method, $\tilde{C}_0^{im}$, by tuning the regularization $\gamma \in (0, \infty)$ and the train-val split ratio $s \in (0,1)$, satisfies*

$$\inf_{\substack{\gamma > 0 \\ s \in (0,1)}} \lim_{\substack{d,\, N \to \infty \\ d/N \to \eta}} \tilde{C}_0^{\mathrm{im}} = 1 + \eta.$$

Theorem 12 (Asymptotic BaMAML constant) *As $d, N \to \infty$, $d/N \to \eta$, the optimal constant of the BaMAML method, $\tilde{C}_0^{\mathrm{ba}}$, by tuning the regularization $\gamma \in (0, \infty)$ and the train-val split ratio $s \in (0,1)$, satisfies*

$$\inf_{\substack{\gamma > 0 \\ s \in (0,1)}} \lim_{\substack{d,\, N \to \infty \\ d/N \to \eta}} \tilde{C}_0^{\mathrm{ba}} \begin{cases} = 1, & \eta \in (0, 1], \\ \leq \eta, & \eta \in (1, \infty). \end{cases}$$

Proof: Adopt the Stieltjes transform to obtain $\lim_{\substack{d,\, N \to \infty \\ d/N \to \eta}} \tilde{C}_0^{\mathrm{ba}}$ as a function of γ, s, η, given below. For all $\omega_1, \omega_2 > 0, \eta > 0$, define

$$s(\omega_1, \omega_2) \coloneqq \lim_{d,N \to \infty, d/N \to \eta} \frac{1}{d} \mathbb{E}\Big[\operatorname{tr}\Big(\big(\omega_1 \mathbf{I}_d + \omega_2 \hat{\mathbf{Q}}_N \big)^{-1} \Big) \Big]$$

whose closed form solution is given by

$$\begin{aligned} s(\omega_1, \omega_2) &= \frac{\eta - 1 - \omega_1/\omega_2 + \sqrt{(\omega_1/\omega_2 + 1 + \eta)^2 - 4\eta}}{2\eta\omega_1} \\ &= \frac{\sqrt{(\omega_1/\omega_2 + 1 + \eta)^2 - 4\eta} - (\omega_1/\omega_2 + 1 + \eta) + 2\eta}{2\eta\omega_1} \leq \frac{1}{\omega_1}. \end{aligned}$$

$$\begin{aligned} \frac{d}{d\omega_1} s(\omega_1, \omega_2) =& \big[\big((1/\omega_2 + (1+\eta)/\omega_1)^2 - 4\eta/\omega_1^2 \big)^{-\frac{1}{2}} \cdot \\ & \big((1/\omega_2 + (1+\eta)/\omega_1)(1+\eta) - 4\eta/\omega_1 \big)(-\omega_1^{-2}) + (1-\eta)\omega_1^{-2} \big] / 2\eta \\ \frac{d}{d\omega_1} s(\omega_1, \omega_2)\big|_{\omega_1 = 1} =& [-\big((1/\omega_2 + 1 + \eta)^2 - 4\eta \big)^{-\frac{1}{2}} \big((1/\omega_2 + (1+\eta))(1+\eta) - 4\eta \big) + (1-\eta)] / 2\eta. \end{aligned}$$

Therefore

$$\lim_{d,N \to \infty, d/N \to \eta} \frac{1}{d} \mathbb{E}\Big[\operatorname{tr}\Big(\big(\mathbf{I}_d + \gamma^{-1} \hat{\mathbf{Q}}_N \big)^{-1} \Big) \Big] = s(1, \gamma^{-1}) \leq 1$$

where by L'Hospital's rule,

$$\begin{aligned} \lim_{\gamma \to \infty} s(1, \gamma^{-1}) &= \lim_{\gamma \to \infty} \frac{\sqrt{(\gamma + 1 + \eta)^2 - 4\eta} - (\gamma + 1 + \eta)}{2\eta} + 1 \\ &= \lim_{\gamma \to \infty} \frac{\sqrt{(1/\gamma + 1 + \eta/\gamma)^2 - 4\eta/\gamma^2} - (1/\gamma + 1 + \eta/\gamma)}{2\eta/\gamma} + 1 = 1 \end{aligned}$$

$$\begin{aligned}\lim_{\gamma\to 0} s(1,\gamma^{-1}) &= \lim_{\gamma\to 0}\frac{\sqrt{(\gamma+1+\eta)^2-4\eta}-(\gamma+1-\eta)}{2\eta}\\ &= \lim_{\gamma\to 0}\frac{|\eta-1|-(1-\eta)}{2\eta} = \begin{cases}0, & \eta\in(0,1];\\ 1-\frac{1}{\eta}, & \eta\in(1,\infty).\end{cases}\end{aligned}$$

By the derivative trick,

$$\begin{aligned}&\lim_{d,N\to\infty,d/N\to\eta}\frac{1}{d}\mathbb{E}\Big[\operatorname{tr}\Big(\big(\omega_1\mathbf{I}_d+\omega_2\hat{\mathbf{Q}}_N\big)^{-2}\Big)\Big] = \frac{d}{d\omega_1}s(\omega_1,\omega_2)\\ &\lim_{d,N\to\infty,d/N\to\eta}\frac{1}{d}\mathbb{E}\Big[\operatorname{tr}\Big(\big(\mathbf{I}_d+\gamma^{-1}\hat{\mathbf{Q}}_N\big)^{-2}\Big)\Big]\\ =&[\big((\gamma+1+\eta)^2-4\eta\big)^{-\frac{1}{2}}\big((\gamma+(1+\eta))(1+\eta)-4\eta\big)-(1-\eta)]/2\eta\end{aligned}$$

Therefore

$$\lim_{\gamma\to\infty}\lim_{d,N\to\infty,d/N\to\eta}\frac{1}{d}\mathbb{E}\Big[\operatorname{tr}\Big(\big(\omega_1\mathbf{I}_d+\omega_2\hat{\mathbf{Q}}_N\big)^{-2}\Big)\Big] = \frac{1+\eta-(1-\eta)}{2\eta} = 1$$

$$\lim_{\gamma\to 0}\lim_{d,N\to\infty,d/N\to\eta}\frac{1}{d}\mathbb{E}\Big[\operatorname{tr}\Big(\big(\omega_1\mathbf{I}_d+\omega_2\hat{\mathbf{Q}}_N\big)^{-2}\Big)\Big] = \frac{|\eta-1|-(1-\eta)}{2\eta} = \begin{cases}0, & \eta\in(0,1];\\ 1-\frac{1}{\eta}, & \eta\in(1,\infty).\end{cases}$$

For $\frac{1}{d}\mathbb{E}\Big[\operatorname{tr}\Big(\big(\mathbf{I}_d+\gamma^{-1}\hat{\mathbf{Q}}_{N_1}\big)^{-1}\big(\mathbf{I}_d+(\gamma s)^{-1}\hat{\mathbf{Q}}_N\big)^{-1}\Big)\Big]$, first $\lim_{\gamma\to\infty}\big(\mathbf{I}_d+\gamma^{-1}\hat{\mathbf{Q}}_{N_1}\big)^{-1}=\mathbf{I}_d$, therefore

$$\begin{aligned}&\lim_{\gamma\to\infty}\frac{1}{d}\mathbb{E}\Big[\operatorname{tr}\Big(\big(\mathbf{I}_d+\gamma^{-1}\hat{\mathbf{Q}}_{N_1}\big)^{-1}\big(\mathbf{I}_d+(\gamma s)^{-1}\hat{\mathbf{Q}}_N\big)^{-1}\Big)\Big] = \lim_{\gamma\to\infty}\frac{1}{d}\mathbb{E}\Big[\operatorname{tr}\Big(\big(\mathbf{I}_d+(\gamma s)^{-1}\hat{\mathbf{Q}}_N\big)^{-1}\Big)\Big]\\ &\lim_{\substack{\gamma\to\infty\\ s\gamma\to 0}}\lim_{\substack{d,N\to\infty\\ d/N\to\eta}}\frac{1}{d}\mathbb{E}\Big[\operatorname{tr}\Big(\big(\mathbf{I}_d+\gamma^{-1}\hat{\mathbf{Q}}_{N_1}\big)^{-1}\big(\mathbf{I}_d+(\gamma s)^{-1}\hat{\mathbf{Q}}_N\big)^{-1}\Big)\Big]\\ =&\lim_{\substack{\gamma\to\infty\\ s\gamma\to 0}}\lim_{\substack{d,N\to\infty\\ d/N\to\eta}}\frac{1}{d}\mathbb{E}\Big[\operatorname{tr}\Big(\big(\mathbf{I}_d+(\gamma s)^{-1}\hat{\mathbf{Q}}_N\big)^{-1}\Big)\Big] = \begin{cases}0, & \eta\in(0,1];\\ 1-\frac{1}{\eta}, & \eta\in(1,\infty).\end{cases}\end{aligned}$$

Then

$$\inf_{\substack{\gamma>0\\ s\in(0,1)}}\lim_{\substack{d,N\to\infty\\ d/N\to\eta}}\tilde{C}_0^{\mathrm{ba}} \le \lim_{\substack{\gamma\to\infty\\ s\gamma\to 0}}\lim_{\substack{d,N\to\infty\\ d/N\to\eta}}\tilde{C}_0^{\mathrm{ba}} = \begin{cases}1, & \eta\in(0,1];\\ \eta, & \eta\in(1,\infty).\end{cases}$$

Note $\tilde{C}_0^{\mathrm{ba}}\ge 1$, therefore

$$\inf_{\substack{\gamma>0\\ s\in(0,1)}}\lim_{\substack{d,N\to\infty\\ d/N\to\eta}}\tilde{C}_0^{\mathrm{ba}}\begin{cases}=1, & \eta\in(0,1],\\ \le\eta, & \eta\in(1,\infty).\end{cases} \tag{186}$$

□

B.2.5 Comparison of the dominating constants

Based on Theorems 10, 11, 12, Suppose Assumptions 1,3 hold, we have

$$\lim_{\substack{d,N\to\infty\\ d/N\to\eta}}\tilde{C}_0^{\mathrm{er}} = \inf_{\substack{\alpha\in(0,1/\bar{\lambda})\\ s\in(0,1)}}\lim_{\substack{d,N\to\infty\\ d/N\to\eta}}\tilde{C}_0^{\mathrm{ma}} = \inf_{\substack{\gamma>0\\ s\in(0,1)}}\lim_{\substack{d,N\to\infty\\ d/N\to\eta}}\tilde{C}_0^{\mathrm{im}} > \inf_{\substack{\gamma>0\\ s\in(0,1)}}\lim_{\substack{d,N\to\infty\\ d/N\to\eta}}\tilde{C}_0^{\mathrm{ba}}. \tag{187}$$

Table 2: Comparison of different MAML on image classification (testing loss (NLL) with std., code modified from Nguyen et al. (2020))

	miniImageNet		TieredImageNet	
Method	1-shot 5-way	5-shot 5-way	1-shot 5-way	5-shot 5-way
MAML	1.41 ± 0.04	1.18 ± 0.06	1.36 ± 0.08	0.99 ± 0.02
BaMAML	1.38 ± 0.05	1.15 ± 0.05	1.05 ± 0.06	0.76 ± 0.01

Considering the weighted version, recall that $w_{\mathrm{ma}} = (1-\alpha)^2 > (1-1/\bar{\lambda})^2 > 0$, $w_{\mathrm{im}} = (1+\gamma^{-1})^{-2}$, $\lim_{\gamma\to 0} w_{\mathrm{im}} = 0$, $w_{\mathrm{ba}} = (1+\gamma^{-1})^{-1}(1+(\gamma s)^{-1})^{-1} < w_{\mathrm{im}}$, $\lim_{\gamma s\to 0} w_{\mathrm{ba}} = 0$. Therefore $\inf_{\gamma>0} w_{\mathrm{im}} = \inf_{\gamma>0, s\in(0,1)} w_{\mathrm{ba}} = 0 < \inf_{\alpha\in(0,1/\bar{\lambda})} w_{\mathrm{ma}}$, and

$$\inf_{\substack{\alpha\in(0,1/\bar{\lambda})\\ s\in(0,1)}} \lim_{\substack{d,N\to\infty\\ d/N\to\eta}} w_{\mathrm{ma}}\tilde{C}_0^{\mathrm{ma}} \geq \inf_{\substack{\gamma>0\\ s\in(0,1)}} \lim_{\substack{d,N\to\infty\\ d/N\to\eta}} w_{\mathrm{im}}\tilde{C}_0^{\mathrm{im}} \geq \inf_{\substack{\gamma>0\\ s\in(0,1)}} \lim_{\substack{d,N\to\infty\\ d/N\to\eta}} w_{\mathrm{ba}}\tilde{C}_0^{\mathrm{ba}}. \tag{188}$$

Combining (187) (188) with the comparison in optimal population risk, we can conclude that, under Assumptions 1,3, when γ is sufficiently small, it is guaranteed that iMAML and BaMAML will have smaller meta-test risk than MAML. Furthermore, BaMAML has strictly smaller dominating constant in statistical error compared to iMAML under optimal choice of γ and s, as $d, N \to \infty, d/N \to \eta > 0$.

C Additional experiments and details

C.1 Experimental details

For sinewave regression and real image classification, Adam optimizer is used. The hyperparameters of all experiments are chosen based on grid search. In sinewave regression experiments, the learning rate for ERM is initially 0.0001, while the learning rates for both base-learner and meta-learner in MAML and BaMAML are initially 0.001, except that in the experiments with $N = 1000, T = 100, s = 0.5$, the initial learning rate of BaMAML for both base-learner and meta-learner are initially 0.0001. The learning rate decay is set to be 0.98 for all methods. The number of Monte-Carlo samples of model parameters used for BaMAML is 10. In real image classification experiments, the CNN architecture used is ResNet18. The initial learning rate of MAML and BaMAML are 0.001.

C.2 Real datasets

Experiment settings. We test the performance on the 5-way miniImageNet classification (Vinyals et al., 2016) and TieredImageNet. MiniImageNet consists 100 classes of images, each with 600 examples. The classes are split into 64, 12, and 24 for train, validation and test, respectively, following (Finn et al., 2017). Note that, since in this setting ERM without adaptation to new classes does not have practical meaning, we do not compare with ERM in this setting.

Results. The meta-test classification accuracy under different settings are provided in Table 2, where BaMAML shows comparable testing loss on miniImageNet and higher testing loss on the TieredImageNet dataset.