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

Lisha Chen  
Rensselaer Polytechnic Institute

Tianyi Chen

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, Lin 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:

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

In an attempt to provide an affirmative answer to this
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 over-parameterized 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) optimization 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 Senet (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 centrist 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 also we 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 $\tau$ is drawn from a task distribution, i.e. $\tau \sim \mathcal{D}$, with input features $x \in \mathcal{X} \subset \mathbb{R}^d$ and target labels $y \in \mathcal{Y} \subset \mathbb{R}$. For each task $\tau$, we observe $N$ samples drawn i.i.d. from $p_\tau$ in the dataset $D_\tau = \{(x_{\tau,n}, y_{\tau,n})\}_{n=1}^{N}$, and $D_\tau$ is divided into the train and validation datasets, denoted as $D_\tau^{\text{trn}}$ and $D_\tau^{\text{val}}$, respectively. Here $|D_\tau^{\text{trn}}| = N_1$ and $|D_\tau^{\text{val}}| = N_2$ with $N = N_1 + N_2$. Given the data $D_\tau$, we use the empirical loss $\ell_\tau(h_\tau, D_\tau)$ of per-task hypothesis $h_\tau \in \mathcal{H}$ as a measure of the performance.

And the goal for initialization based meta learning methods, such as MAML [Finn et al., 2017] and BaMAML [Yoon et al., 2018], is to learn an initial parameter $\theta_0 \in \Theta_0$, which, with an adaptation method and the training data, can produce a per-task hypothesis $h_\tau$ that performs well on the validation data for task $\tau$. Formally, for a meta-learning method, $\mathcal{A} : \Theta_0 \times (\mathcal{X}_\tau \times \mathcal{Y}_\tau)^N \rightarrow \mathcal{H}_\tau$, represents the adaptation method or base-learner. Given $T$ tasks with corresponding data, our meta-learning objective is to find $\theta_0$ that minimizes the empirical loss, given by

$$\mathcal{L}^A(\theta_0, D) := \frac{1}{T} \sum_{\tau=1}^{T} \ell_\tau(\mathcal{A}(\theta_0, D_\tau^{\text{trn}}), D_\tau^{\text{val}}). \quad (1)$$

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

$$\mathcal{R}^A(\theta_0) := \mathbb{E}_\tau \left[ \mathbb{E}_{D_\tau} \left[ \ell_\tau(\mathcal{A}(\theta_0, D_\tau^{\text{trn}}), D_\tau^{\text{val}}) \right] \right]. \quad (2)$$

Denote $X_\tau^{\text{all}} := [x_{\tau,1}, \ldots, x_{\tau,N}]^T \in \mathbb{R}^{N \times d}$, $y_\tau^{\text{all}} := [y_{\tau,1}, \ldots, y_{\tau,N}]^T \in \mathbb{R}^N$ 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 $\ell_\tau$ defined as the negative log likelihood, given by

$$\ell_\tau(\mathcal{A}(\theta_0, D_\tau^{\text{trn}}), D_\tau^{\text{val}}) = -\frac{1}{N_2} \log p(y_\tau^{\text{val}} | X_\tau^{\text{val}}, \theta_0, D_\tau^{\text{trn}})$$

$$= -\frac{1}{N_2} \log \int p(y_\tau^{\text{val}} | X_\tau^{\text{val}}, \theta_\tau) p_{\mathcal{A}}(\theta_\tau | \theta_0, D_\tau^{\text{trn}}) d\theta_\tau \quad (3)$$

where $p_{\mathcal{A}}(\theta_\tau | \theta_0, D_\tau^{\text{trn}})$ is the posterior distribution induced by $\mathcal{A}$. And the likelihood $p(y_\tau^{\text{val}} | X_\tau^{\text{val}}, \theta_0, D_\tau^{\text{trn}})$ reduces to a Dirac delta function $\delta(\theta_\tau - \theta_0^\mathcal{A})$. And $\mathcal{A}$ specifies a mapping from the initial parameter $\theta_0$ to the task-specific parameter $\theta_\tau^\mathcal{A}(\theta_0, D_\tau^{\text{trn}})$.

In the meta training stage, we obtain $\theta_0^\mathcal{A}$ by minimizing (1) under each meta learning method $\mathcal{A}$. And in the meta testing stage, we evaluate the test error of $\theta_0^\mathcal{A}$ 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 $\theta_\tau^\mathcal{A}(\theta_0)$ is obtained from the initial parameter $\theta_0$ by taking one step gradient descent with step size $\alpha$ of the per-task loss function $\ell_\tau$. Combined with the empirical loss defined in (1), we have
the empirical loss of MAML is given by
\[
\mathcal{L}_{\text{ma}}(\theta_0, D) = \frac{1}{T} \sum_{\tau=1}^{T} \ell_{\tau}(\hat{\theta}_{\text{ma}}(\theta_0, D_{\tau}^{\text{trn}}), D_{\tau}^{\text{val}}) \\
\text{s.t. } \hat{\theta}_{\text{ma}}(\theta_0, D_{\tau}^{\text{trn}}) = \theta_0 - \frac{\alpha}{2} \nabla_{\theta_{\tau}} \ell_{\tau}(\theta_0, D_{\tau}^{\text{trn}}).
\]

BaMAML obtains an approximation of the posterior distribution \(p(\theta_{\tau} | D_{\tau}^{\text{trn}}, \theta_0)\) instead of a point estimate \(\hat{\theta}_{\text{ma}}(\theta_0, D_{\tau}^{\text{trn}})\). In general, the true posterior distribution can be difficult to compute exactly. Alternatively, the approximate distribution \(\hat{p}(\theta_{\tau} | D_{\tau}^{\text{trn}}, \theta_0)\) can be obtained via variational inference formulation, by minimizing the divergence between the approximate and the true posterior distribution. Define \(D_{\text{KL}}(\cdot | \cdot)\) as the KL-divergence between two distributions, we have
\[
\mathcal{L}_{\text{ba}}(\theta_0, D) = \frac{1}{T} \sum_{\tau=1}^{T} \ell_{\tau}(\hat{p}(\theta_{\tau} | D_{\tau}^{\text{trn}}, \theta_0), D_{\tau}^{\text{val}}) \\
\text{s.t. } \hat{p}(\theta_{\tau} | D_{\tau}^{\text{trn}}, \theta_0) = \arg\min_{q(\theta_{\tau}) \in Q} \mathcal{D}_{\text{KL}}(q(\theta_{\tau}) || p(\theta_{\tau} | D_{\tau}^{\text{trn}}, \theta_0)).
\]

It is worth mentioning that BaMAML formulation in this paper contains iMAML, or iMAML \cite{Rajeswaran2019} 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 \(Q\) is chosen to be the set of Dirac Delta functions and the KL-divergence in \(\mathcal{L}_{\text{ba}}\) is replaced by the cross entropy, then \(\mathcal{L}_{\text{ba}}\) reduces to
\[
\hat{p}(\theta_{\tau} | D_{\tau}^{\text{trn}}, \theta_0) = \delta(\theta_{\tau} - \hat{\theta}_{\text{map}}),
\]
with \(\hat{\theta}_{\text{map}} = \arg\max_{\theta_{\tau}} p(\theta_{\tau} | D_{\tau}^{\text{trn}}, \theta_0).

### 2.2 Meta Linear Regression

**Data model.** Under the meta linear regression setting, with the feature vector \(x_{\tau} \in \mathbb{R}^d\), the target \(y_{\tau} \in \mathbb{R}\), and the ground truth parameter of task \(\tau\), \(\theta_{\tau}^4 \in \mathbb{R}^d\), we assume the data generation model for task \(\tau\) is
\[
y_{\tau} = \theta_{\tau}^4 \top x_{\tau} + \epsilon_{\tau}, \text{ with } \epsilon_{\tau} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_{\tau}^2), \text{ Q. } \tau := \mathbb{E}[x_{\tau} x_{\tau}^\top].
\]

Given the estimate of \(\theta_{\tau}^4\) denoted as \(\hat{\theta}_{\text{ba}}\), then the conditional probability \(p(y_{\tau} | x_{\tau}, \theta_{\tau}^4) = \mathcal{N}((\hat{\theta}_{\text{ba}}^\top x_{\tau}, \sigma_{\tau}^2)\).

Thus, ignoring the constant, the negative log likelihood in \(\mathcal{L}_{\text{ba}}\) becomes the squared error \(\|y_{\tau} - \mathbb{E}[y_{\tau} | x_{\tau}, \theta_{\tau}^4]\|^2\). Note that \(\sigma_{\tau}\) depends on task \(\tau\) generally, but does not pose challenges to analysis, therefore we assume \(\sigma_{\tau} = 1\) in this paper for simplicity.

By plugging \(\hat{\theta}_{\text{ba}}\) into \(\mathcal{L}_{\text{ma}}\), 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 \cite{Gao2020} are also included.

**Proposition 1 (Empirical and population level solutions)** Under data model \(\mathcal{L}_{\text{ba}}\), the meta-test risk of method \(A\) can be computed by
\[
R^A(\theta_0) = \mathbb{E}_\tau[\|\theta_0 - \theta_{\tau}^4\|_{W^A_{\tau}}^2] + 1. 
\]

The optimal solutions to the meta-test risk and empirical loss are given below respectively
\[
\theta_0^A := \arg\min_{\theta_0} R^A(\theta_0) = \mathbb{E}_\tau[\mathbb{W}^A_{\tau}]^{-1}\mathbb{E}_\tau[\mathbb{W}^A_{\tau}]\theta_{\tau}^4 \tag{9a}
\]
\[
\theta_0^A := \arg\min_{\theta_0} \mathcal{L}^A(\theta_0, D) \\
= (\sum_{\tau=1}^{T} \mathbb{W}^A_{\tau})^{-1}(\sum_{\tau=1}^{T} \mathbb{W}^A_{\tau} \theta_{\tau}^4) + \Delta^A_T. \tag{9b}
\]

where the error term \(\Delta^A_T\) is a polynomial function of \(T, N, d\) caused by the noise \(\epsilon\), and specified in the appendix. And \(Q_{\tau,N} := \frac{1}{T}X_{\tau}^\top X_{\tau}^\top, s = N_1 / N\). The weight matrices of different methods, \(W^A_{\tau}\) and \(W^\theta_{\tau}\), are given in Table 1.

Note that, in the meta linear regression case in Proposition 1, BaMAML further assumes the prior distribution \(\theta_{\tau} \sim \mathcal{N}(\theta_0, 1/\gamma_0)\) with \(\gamma_0 = \gamma N_1\), resulting in the weight matrices \(W^\text{ba}_{\tau}, W^\text{ma}_{\tau}\) in Table 1 depending on \(\gamma\) and \(s\). The posterior follows a Gaussian distribution, \(p(\theta_{\tau} | D_{\tau}^{\text{trn}}, \theta_0) = \mathcal{N}(\mu_{\theta_{\tau}}, \Sigma_{\theta_{\tau}})\), where the parameters \(\mu_{\theta_{\tau}}\) and \(\mu_{\theta_{\tau}}\) are given by
\[
\Sigma_{\theta_{\tau}} = (N_1 Q_{\tau,N} + \gamma \mathbb{I})^{-1}, \tag{10a}
\]
\[
\mu_{\theta_{\tau}} = \Sigma_{\theta_{\tau}}(X_{\tau}^\top y_{\tau} + \gamma \theta_0). \tag{10b}
\]

If \(p(\theta_{\tau} | D_{\tau}^{\text{trn}}, \theta_0) \in Q\), then \(\hat{p}(\theta_{\tau} | D_{\tau}^{\text{trn}}, \theta_0) = p(\theta_{\tau} | D_{\tau}^{\text{trn}}, \theta_0)\), which holds for the meta linear regression case analyzed in this paper, with \(Q\) 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.
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 $R^A$ 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 $A$ can be decomposed into optimal population risks and statistical errors, given by

$$R^A(\theta^A_0) = R^A(\theta^A_0) + \|\theta^A_0 - \theta^*\|^2_{E_{x}[W^A]}.$$  (11)

Invoking the definition of $\theta^A_0$ in (2) as the optimal solution for $\min_{\theta_0} R^A(\theta_0)$, the optimal population risk $R^A(\theta^A_0)$ 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 $\tau$, $0 < \lambda \leq \lambda(Q_\tau) \leq \overline{\lambda}$, where $\lambda(Q_\tau)$ represents the eigenvalues of $Q_\tau$.

**Assumption 2 (Sub-gaussian task parameter and bounded features)** The ground truth parameter $\theta^*_\tau$ is independent of $X_\tau$ and satisfies that the individual entries $\{\theta^*_\tau - \theta^A_0\}_{i \in [d], \tau \in [T]}$ are independent and $O(R/\sqrt{d})$-sub-gaussian. In addition, $||E[\theta^*_\tau - \theta^A_0]|| \leq M$. The inputs $||x_{\tau,i}|| \leq K$. $R, K$ are constants.

Note that, these assumptions can be easily satisfied in data generation model (7) by controlling the hyper-parameters. 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 $\gamma$ 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 $\theta^{ma}_0$ and $\theta^{ba}_0$ are the minimizers of $R^{ma}(\theta; \alpha)$ and $R^{ba}(\theta; \gamma)$, respectively. Define $r^{ma} := \min_{\alpha} R(\theta^{ma}_0; \alpha) - 1 > 0$, $C_0 := \max\{(M + ||\theta^{ma}_0||)^2 + R^2\gamma^2, (M + ||\theta^{ma}_0||)^2 + R^2\gamma^2\}$. Under Assumptions 2 when $\gamma$ satisfies

$$0 < \gamma < \left((r^{ma})^{\frac{1}{2}}C_0\lambda^{\frac{1}{2}} - 1\right)^{-1}$$  (12)

BaMAML has smaller optimal population risk, i.e.

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

Theorem 2 states that regardless of the choice of $\alpha$, 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 $\gamma$ represents trade-off between adaptation speed and optimal population risk, because $\theta^{ba}_0(\theta_0)$ is a weighted average of the prior $\theta_0$ and the ground truth parameter $\theta^*_\tau$. The larger $\gamma$, the higher weight for the prior $\theta_0$, then the closer the initial parameter $\theta_0$ is to the optimal $\theta^{ma}_0(\theta_0)$, and the faster the adaptation speed. On the other hand, the larger $\gamma$, 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.
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: \(x_{\tau,i} \overset{iid}{\sim} \mathcal{N}(0, I_d)\). Then \(Q_{\tau,i} = I_d\), therefore \(W^*_\tau = w_{A,I_d}^t\). This implies that for different methods, the optimal initial parameters are the same, that is, \(\theta^*_0 = E[\theta^*]\). 2) The ground truth parameter \(\theta^*_0\) is independent of \(X^*\) and satisfies

\[
\mathbb{E}_{\theta^*_0} \left[ (\theta^*_0 - \theta^* \theta^*_0)^T \right] = \frac{R^2}{d} I_d
\]

where \(R\) is a constant, and the individual entries \(\{\theta^*_0 - \theta^*_0\}_{i \in [d], \tau \in [T]}\) are i.i.d. mean-zero and \(O(R/\sqrt{d})\)-subgaussian.

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 3 and 6.

**Theorem 5 (Statistical error of MAML)** Suppose Assumptions 2,3 hold, \(T = \Omega(d)\), \(d/N = \eta > 0\), and \(\alpha > 0\). Define

\[
w_{A} := \frac{1}{d} \text{tr}(E[W^A])\), \quad \tilde{C}_A := \frac{1}{d} \left( E[2[W^A], E[(W^A)^2]] \right)
\]

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

\[
E^2_{ba}(\tilde{\theta}^*_{0}) = \frac{R^2}{d} w_{ba} \tilde{C}^*_{0} + \frac{d}{T N} w_{ma} \tilde{C}^*_{1} + \varphi
\]

where \(\varphi\) is given by (15). The dominating constant \(C^*_{ma}\) satisfies

\[
\inf_{\alpha > 0} \lim_{d, N \to \infty} \tilde{C}^*_{0} = 1 + \eta.
\]

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

**Theorem 6 (Statistical error of BaMAML)** Suppose Assumptions 2,3 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

\[
E^2_{ba}(\tilde{\theta}^*_{0}) = \frac{R^2}{d} w_{ba} \tilde{C}^*_{0} + \frac{d}{T N} w_{ma} \tilde{C}^*_{1} + \varphi
\]

where \(\varphi\) is given by (15). In addition, the dominating constant \(C^*_{ba}\) satisfies

\[
\inf_{\gamma > 0} \lim_{d, N \to \infty} \tilde{C}^*_{0} = 1, \quad \eta \in (0,1),
\]

\[
\leq \eta, \quad \eta \in (1, \infty).
\]
Theorems 5 and 6 state that when $T,d$ are large and $T = \Omega(d)$, the statistical errors of MAML and BaMAML are dominated by $R^2/T$ times $\tilde{C}_0^{\text{ma}}$ and $\tilde{C}_0^{\text{ba}}$, respectively. Therefore we can compare the statistical errors of MAML and BaMAML based on the optimal hyperparameters $\alpha,\gamma$ and split ratio $s$ below.

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

$$\inf_{\alpha > 0} \lim_{d,N \to \infty} \tilde{C}_0^{\text{ma}} > \inf_{\gamma > 0} \lim_{d,N \to \infty} \tilde{C}_0^{\text{ba}}. \quad (26)$$

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.

**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 $\alpha,\gamma$ are given in subfigure titles.

**Figure 3:** Optimal population risks $R(\theta_0^A)$ v.s. $\alpha,\gamma$, and meta-test risks $R(\hat{\theta}_0^A)$ v.s. train validation split ratio $s$ for ERM, MAML and BaMAML. In Figure 3a, 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 $V \sim U(SO(d)), \theta_{\tau} \sim U([0,2]^d), \lambda_\tau \sim U([0,1,2]^d), Q_\tau = V\text{diag}(\lambda_\tau) V^T, x_\tau \sim N(0,Q_\tau)$, and data model in (7), where $SO(d)$ 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 test loss than MAML in Figure 2 where darker blue represents higher probability that BaMAML is better than MAML (Gao and Sener 2020).

In Figure 3a, we report the meta-test risks for MAML and BaMAML under different hyperparameters $\alpha,\gamma$. Figure 3a shows that with proper choice of hyperparameter $\gamma$, 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
approaches that of ERM. This further demonstrates the trade-off between fast adaptation and optimal population risk (meta-test risk) that depends on $\gamma$. On the other hand, MAML is relatively more sensitive to the step size. Too large step size $\alpha$ 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 [21], 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 3a 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 [18], 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 $\epsilon$. For each task, the parameters are sampled from $A \sim U([0.1, 5.0]), b \sim U([0.0, 2\pi]), w \sim U([0.5, 2.0]), \epsilon \sim N(0, (0.01A)^2)$. 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, Charles 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 WHoffman, 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 Connell. 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 Miçaili, 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...
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.

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.

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

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*, volume 29, pages 3630–3638, Barcelona, Spain, 2016.

Sebastian Thrun and Lorien Pratt. *Learning to Learn: Introduction and Overview*. Kluwer Academic Publishers, USA, 1998. ISBN 0792380479.
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. \( \hat{\theta}^{er}(\theta_0, D^{trn}) = \theta_0 \), and based on the definition in (1), the empirical loss of ERM is given by

\[
L^{er}(\theta_0, D) = \frac{1}{TN} \sum_{\tau=1}^{T} \| y_{\tau,N}^{all} - X_{\tau,N}^{all} \theta_0 \|^2. \tag{27}
\]

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

\[
\hat{\theta}^{er}_0 = \arg \min_{\theta_0} L^{er}(\theta_0, D) = \arg \min_{\theta_0} \frac{1}{TN} \sum_{\tau=1}^{T} \| X_{\tau,N}^{all} \theta^{gt}_\tau + e_{\tau,N}^{all} - X_{\tau,N}^{all} \theta_0 \|^2. \tag{28}
\]

Using the optimality condition, we have

\[
\hat{\theta}^{er}_0 = \left( \sum_{\tau=1}^{T} \hat{W}^{er}_{\tau,N} \right)^{-1} \left( \sum_{\tau=1}^{T} \hat{W}^{er}_{\tau} \theta^{gt}_\tau \right) + \Delta^{er}_T \tag{29a}
\]

\[
\Delta^{er}_T = \left( \sum_{\tau=1}^{T} \hat{W}^{er}_{\tau,N} \right)^{-1} \left( \sum_{\tau=1}^{T} \frac{1}{N} X_{\tau,N}^{all} e_{\tau,N}^{all} \right) \tag{29b}
\]

\[
\hat{W}^{er}_{\tau} = \frac{1}{N} X_{\tau,N}^{all} X_{\tau,N}^{all \top}. \tag{29c}
\]

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

\[
R^{er}_{N_a}(\theta_0) := \mathbb{E}_\tau \left[ \mathbb{E}_{D_{\tau,N_a}} \left[ \mathbb{E}_{p(x, y|\tau)} [(y_\tau - \hat{\theta}^{er}_\tau(\theta_0, D_{\tau,N_a})^\top x_\tau)^2] \right] \right] \tag{30}
\]

where \( \hat{\theta}^{er}_\tau(\theta_0, D_\tau) = \theta_0 \), plugging which into (30), we have

\[
R^{er}_{N_a}(\theta_0) = \mathbb{E}_\tau [(y_\tau - \theta_0^\top x_\tau)^2] = \mathbb{E}_\tau [\| \theta_0 - \theta^{gt}_\tau \|^2_{\hat{W}^{er}_{\tau,N_a}}] + 1 \tag{31a}
\]

\[
\hat{W}^{er}_{\tau,N_a} = \mathbb{E}[x_\tau x_\tau^\top | \tau] = Q_\tau. \tag{31b}
\]
By the general definition of optimal population risk in [2], the ERM optimal population risk is given by

\[
R^e_r(\theta_0) := \lim_{N_a \to \infty} R^e_{N_a}(\theta_0) = E_r[\|\theta_0 - \theta_t^g\|^2_{W^*_r}] + 1
\]  \hspace{1cm} (32a)

\[
W^*_r = Q_r.
\]  \hspace{1cm} (32b)

Define \( \theta^e_0 \) as the minimizer of the ERM optimal population risk, given by

\[
\theta^e_0 = \arg\min_{\theta_0} R^e_r(\theta_0) = \arg\min_{\theta_0} E_r[\|\theta_0 - \theta_t^g\|^2_{W^*_r}] = E_r[\theta_t^g]^{-1}E_r[W^*_r\theta_t^g].
\]  \hspace{1cm} (33)

From [29c] [31b] [32b], we have the property \( E[W^*_{r,N}] = W^*_{r,N} = W^*_{r} \), 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 \( \hat{\theta}_r^m \) is computed from the initial parameter \( \theta_0 \) by taking one-step gradient descent of the empirical loss function as shown below

\[
\hat{\theta}_r^m(\theta_0, D_r) = \theta_0 - \frac{\alpha}{2} \nabla_{\theta_0} \ell_r(\theta_0, D_r) = (I - \alpha \hat{Q}_{r,N})\theta_0 + \frac{\alpha}{N} \theta_0 \tau, N
\]  \hspace{1cm} (34)

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 \( \theta_t^e \) in [34], the empirical loss of MAML is given by

\[
L^m(\theta_0, D) = \frac{1}{TN_2} \sum_{r=1}^{T} \|X^\tau_{r,N_2} - X^\tau_{r,N_2} \theta^e_r(\theta_0, D^\tau_{r,N})\|^2.
\]  \hspace{1cm} (35)

The minimizer of the MAML empirical loss is defined as

\[
\hat{\theta}^m_0 = \arg\min_{\theta_0} L^m(\theta_0, D) = \arg\min_{\theta_0} \frac{1}{TN_2} \sum_{r=1}^{T} \|X^\tau_{r,N_2} \theta^g_r + e^\tau_{r,N_2} - X^\tau_{r,N_2} \hat{\theta}^m_r(\theta_0, D^\tau_{r,N})\|^2.
\]  \hspace{1cm} (36)

Using the optimality condition, we have

\[
\hat{\theta}^m_0 = \left( \sum_{r=1}^{T} \hat{W}_r^\tau \right)^{-1} \left( \sum_{r=1}^{T} \hat{W}_r^\tau \theta^g_r \right) + \Delta^m_T
\] \hspace{1cm} (37a)

\[
\Delta^m_T = \left( \sum_{r=1}^{T} \hat{W}_r^\tau \right)^{-1} \left( \sum_{r=1}^{T} \left( I - \alpha \hat{Q}_{r,N_1} \right) \frac{1}{N_2} X^\tau_{r,N_2} e^\tau_{r,N_2} - \frac{\alpha}{N_1} \hat{Q}_{r,N_2} X^\tau_{r,N_1} e^\tau_{r,N_1} \right)
\] \hspace{1cm} (37b)

\[
W^*_r = (I - \alpha \hat{Q}_{r,N_1}) \hat{Q}_{r,N_2} (I - \alpha \hat{Q}_{r,N_1}).
\] \hspace{1cm} (37c)

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

\[
R^m_{N_a}(\theta_0) = E[(y_r - \hat{\theta}^m_r(\theta_0, D, N_a))^T x^\tau_r)]^2 = E_r[\|\theta_0 - \theta_t^g\|^2_{W^*_{r,N_a}}] + 1 + \frac{\alpha^2}{N_a} E_r[|r(Q^2_r)|]
\]  \hspace{1cm} (38a)

\[
W^*_{r,N_a} = E_{Q_r,N_a} [\left( I - \alpha \hat{Q}_{r,N} \right) Q_r \left( I - \alpha \hat{Q}_{r,N} \right)]
\]

\[
= \left( I - \alpha \hat{Q}_r \right) Q_r \left( I - \alpha \hat{Q}_r \right) + \frac{\alpha^2}{N_a} \left( E_{x^\tau_r} [x^\tau_r x^\tau_r^T Q_r x^\tau_r x^\tau_r^T] - Q^2_r \right).
\]  \hspace{1cm} (38b)

Note that, \( \lim_{N_a \to \infty} W^*_{r,N_a} \to W^*_{r} \), and \( \lim_{N_a \to \infty} \frac{\alpha^2}{N_a} E_r[|r(Q^2_r)|] = 0. \) Therefore, from the definition of optimal population risk in [2], we have the MAML optimal population risk is

\[
R^m(\theta_0) = \lim_{N_a \to \infty} R^m_{N_a}(\theta_0) = E_r[\|\theta_0 - \theta_t^g\|^2_{W^*_{r,N_a}}] + 1
\]  \hspace{1cm} (39)

\(^1\text{Note that, here we define the learning rate as } \alpha/2 \text{ to cancel the scale factor 2 from the derivative for notation simplicity.} \)
In MAML, define $\theta_0^{ma}$ as the minimizer of the MAML optimal population risk, given by

$$
\theta_0^{ma} = \arg \min_{\theta_0} \mathcal{R}_0^{ma}(\theta_0) = \arg \min_{\theta_0} \mathbb{E}_r \left[\|\theta_0 - \theta^*_r\|^2 \mathbf{W}_r\right] = \mathbb{E}_r \left[\mathbf{W}_r^{ma}\right]^{-1} \mathbb{E}_r \left[\mathbf{W}_r^{ma} \theta^*_r\right].
$$

(40)

It is worth noting that, from Lemma 1, we have the property $\mathbb{E}_x_r[\mathbf{W}_r^{ma},\mathbf{W}_r] = \mathbf{W}_r^{ma}$, $\lim_{N \to \infty} \mathbf{W}_r^{ma} = \mathbf{W}_r^{ma}$, 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 $\hat{\theta}_{\tau}^{im}$ is computed from the initial parameter $\theta_0$ by optimizing the regularized task-specific empirical loss, given by

$$
\hat{\theta}_{\tau}^{im}(\theta_0) = \arg \min_{\theta_\tau} \frac{1}{N}\|y_{\tau,N} - X_{\tau,N} \theta_\tau\|^2 + \gamma\|\theta_\tau - \theta_0\|^2
$$

(41)

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

$$
\hat{\theta}_{\tau,N}(\theta_0, \mathcal{D}_\tau) = (\mathbf{Q}_{\tau,N} + \gamma \mathbf{I})^{-1} \left(\frac{1}{N} X_{\tau,N} y_{\tau,N} + \gamma \theta_0\right).
$$

(42)

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

$$
\mathcal{L}_{T,\tau}^{im}(\theta_0, \mathcal{D}_\tau) = \frac{1}{TN_2} \sum_{\tau=1}^{T} \|y_{\tau,N_2}^{val} - X_{\tau,N_2}^{val} \hat{\theta}_{\tau}^{im}(\theta_0, \mathcal{D}_{\tau,trn})\|^2
$$

(43)

whose minimizer is

$$
\hat{\theta}_0^{im} = \arg \min_{\theta_0} \mathcal{L}_{T,\tau}^{im}(\theta_0, \mathcal{D}_\tau) = \arg \min_{\theta_0} \frac{1}{TN_2} \sum_{\tau=1}^{T} \|X_{\tau,N_2}^{val} \theta^*_\tau + e_{\tau,N_2}^{val} - X_{\tau,N_2}^{val} \hat{\theta}_{\tau}^{im}(\theta_0, \mathcal{D}_{\tau,trn})\|^2.
$$

(44)

To solve for $\theta_0^{im}$ in the above equation, using the optimality condition, we obtain

$$
\hat{\theta}_0^{im} = \left(\sum_{\tau=1}^{T} \mathbf{W}_\tau^{im}\right)^{-1} \left(\sum_{\tau=1}^{T} \mathbf{W}_\tau^{im} \theta^*_\tau\right) + \Delta_\tau^{im}
$$

(45a)

$$
\Delta_\tau^{im} = \left(\sum_{\tau=1}^{T} \mathbf{W}_\tau^{im}\right)^{-1} \left(\sum_{\tau=1}^{T} \gamma \Sigma_{\tau} \frac{1}{N_2} X_{\tau,N_2}^{val} e_{\tau,N_2}^{val} - \gamma^{-1} \mathbf{W}_\tau^{im} \frac{1}{N_1} X_{\tau,N_1}^{trn} e_{\tau,N_1}^{trn}\right)
$$

(45b)

$$
\Sigma_{\tau,N_1} = \left(\frac{1}{N_1} X_{\tau,N_1}^{trn} X_{\tau,N_1}^{trn}\right)^{-1} = (\mathbf{Q}_{\tau,N_1} + \gamma \mathbf{I})^{-1}
$$

(45c)

$$
\mathbf{W}_\tau^{im} = \gamma^2 \Sigma_{\tau,N_1} \frac{1}{N_2} X_{\tau,N_2}^{val} X_{\tau,N_2}^{val} \Sigma_{\tau,N_1} = \gamma^2 \Sigma_{\tau,N_1} \mathbf{Q}_{\tau,N_2} \Sigma_{\tau,N_1}.
$$

(45d)

The iMAML meta-test risk is defined as

$$
\mathcal{R}_N^{im}(\theta_0) = \mathbb{E}_r \left[\|y_r - \hat{\theta}_r^{im}(\theta_0, \mathcal{D}_{\tau,N_1})\|^2\right] = \mathbb{E}_r \left[\|\theta_0 - \theta^*_r\|^2 \mathbf{W}_r^{im}\right] + 1 + \frac{1}{N_a} \mathbb{E}[\gamma^{-2} \text{tr}(\mathbf{W}_r^{im} \hat{\theta}_r^{im})]
$$

(46a)

$$
\mathbf{W}_r^{im} = \mathbb{E}_x_r \left[\mathbf{Q}_r - \hat{\mathbf{Q}}_r^{im}\right] = \mathbf{W}_r^{im} + \mathbb{E}_x_r \left[\Sigma_{\theta_\tau} \left(\mathbf{Q}_r - \hat{\mathbf{Q}}_r^{im}\right) \Sigma_{\theta_\tau}\right] - \mathbf{W}_r^{im} \left(\mathbf{Q}_r - \hat{\mathbf{Q}}_r^{im}\right) \Sigma_{\theta_\tau}.
$$

(46b)

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


Let $\Sigma_{\theta_\tau} = (\hat{Q}_{\tau,N} + \gamma I)^{-1}$, and $W_{\tau,N}^{im} = \gamma^2 \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau}$. And we simplify the notation of $X_{\tau,N}, y_{\tau,N}, \hat{Q}_{\tau,N}$ as $X_{\tau}, y_{\tau}, \hat{Q}_{\tau}$. The derivation of \cite{46} is given below

\[
R_{\tau,\theta_\tau}^{im}(\theta_0) = \mathbb{E}[\|\hat{\theta}_\tau^{(\theta_0, D_{\tau,N})} - \theta_\tau^{\text{gt}}\|^2_{W_{\tau,N}^{im}}] + 1 = \mathbb{E}[\|\hat{Q}_{\tau}^{\text{gt}} - \theta_\tau^{\text{gt}}\|^2_{Q_{\tau}}] + 1
\]

\[
\quad = \mathbb{E}[\theta_0^{\text{T}} W_{\tau,N}^{im} \theta_0 + 2\gamma(\theta_\tau^{\text{T}} Q_{\tau} \Sigma_{\theta_\tau} - \theta_\tau^{\text{T}}) Q_{\tau} \Sigma_{\theta_\tau} \theta_0 + \frac{1}{N_a} y^\top X_{\tau} \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} + \frac{1}{N_a} X_{\tau}^{\top} y_{\tau}
\]

where $\theta_\tau^{\text{T}} Q_{\tau} \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} \theta_\tau^{\text{T}} = \theta_\tau^{\text{T}} Q_{\tau} \Sigma_{\theta_\tau} \theta_\tau^{\text{T}}$ follows from the fact that $y_{\tau} = X_{\tau} \theta_\tau + e_{\tau}$ and $\mathbb{E}_{\tau}[e_{\tau}] = 0$, one can further derive

\[
R_{\tau,\theta_\tau}^{im}(\theta_0) = \mathbb{E}[\theta_0^{\text{T}} W_{\tau,N}^{im} \theta_0 + 2\gamma(\theta_\tau^{\text{T}} Q_{\tau} \Sigma_{\theta_\tau} - \theta_\tau^{\text{T}}) Q_{\tau} \Sigma_{\theta_\tau} \theta_0 + \frac{1}{N_a} y^\top X_{\tau} \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} + \frac{1}{N_a} X_{\tau}^{\top} y_{\tau}]
\]

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

\[
\mathbb{E}_{e_{\tau}}[y_{\tau}^\top X_{\tau} \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} e_{\tau}] = \text{tr}(X_{\tau} \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} X_{\tau}^{\top} e_{\tau}^e) = \text{tr}(X_{\tau} \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} X_{\tau}^{\top} e_{\tau}^e)
\]

\[
= \text{tr}(X_{\tau} \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} X_{\tau}^{\top}) = N_a \text{tr}(\Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} \hat{Q}_{\tau}) = N_a \text{tr}(W_{\tau,N}^{im} \hat{Q}_{\tau})
\]

also, based on the Woodbury matrix identity, $I - \hat{Q}_{\tau} \Sigma_{\theta_\tau} = I - \Sigma_{\theta_\tau} \hat{Q}_{\tau} = \gamma^2 \Sigma_{\theta_\tau}$, therefore ($\theta_\tau^{\text{T}} Q_{\tau} \Sigma_{\theta_\tau} - \theta_\tau^{\text{T}}) = (\hat{Q}_{\tau}^{\text{T}} Q_{\tau} \Sigma_{\theta_\tau} - \hat{Q}_{\tau}^{\text{T}}) = -\gamma^2 \theta_\tau^{\text{T}} \Sigma_{\theta_\tau}$, and

\[
\theta_\tau^{\text{T}} Q_{\tau} \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} \theta_\tau^{\text{T}} - 2\theta_\tau^{\text{T}} Q_{\tau} \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} \theta_\tau^{\text{T}} + \gamma^2 \theta_\tau^{\text{T}} = \gamma^2 \theta_\tau^{\text{T}} Q_{\tau} \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} \theta_\tau^{\text{T}} - 2\theta_\tau^{\text{T}} Q_{\tau} \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} \theta_\tau^{\text{T}} + \gamma^2 \theta_\tau^{\text{T}}
\]

Combining these equalities and rearranging the equations we obtain

\[
R_{\tau,\theta_\tau}^{im}(\theta_0) = \mathbb{E}[\theta_0^{\text{T}} W_{\tau,N}^{im} \theta_0 + 2\gamma(\hat{Q}_{\tau}^{\text{T}} Q_{\tau} \Sigma_{\theta_\tau} - \hat{Q}_{\tau}^{\text{T}}) Q_{\tau} \Sigma_{\theta_\tau} \theta_0 + \frac{1}{N_a} y^\top X_{\tau} \Sigma_{\theta_\tau} Q_{\tau} \Sigma_{\theta_\tau} + \frac{1}{N_a} X_{\tau}^{\top} y_{\tau}]
\]

where \cite{b} follows from rearranging the equations; \cite{c} follows from the fact that $\theta_\tau^{\text{T}} (W_{\tau,N}^{im} \hat{Q}_{\tau}) \theta_\tau^{\text{T}} = (\theta_\tau^{\text{T}} (W_{\tau,N}^{im} \hat{Q}_{\tau}) \theta_\tau^{\text{T}})^{\top} = \theta_\tau^{\text{T}} (\hat{Q}_{\tau} W_{\tau,N}^{im}) \theta_\tau^{\text{T}}$.

Since \lim_{N_a \to \infty} \frac{1}{N_a} \mathbb{E}_{\tau}[\gamma^{-2} \text{tr}(W_{\tau,N}^{im} \hat{Q}_{\tau}^{\text{T}})] = 0, from the definition of optimal population risk in \cite{46}, the optimal population risk of iMAML is given by

\[
R_{\tau,N_a}^{im}(\theta_0) = \lim_{N_a \to \infty} R_{\tau,\theta_\tau}^{im}(\theta_0) = \mathbb{E}_{\tau}[\|\theta_0 - \theta_\tau^{\text{gt}}\|^2_{W_{\tau,N}^{im}}] + 1
\]

\[
W_{\tau}^{im} = (\gamma^{-1} Q_{\tau} + I)^{-1} Q_{\tau} (\gamma^{-1} Q_{\tau} + I)^{-1}
\]

whose minimizer is given by

\[
\theta_0^{im} = \underset{\theta_0}{\arg \min} \ R_{\tau,N_a}^{im}(\theta_0) = \mathbb{E}_{\tau}[W_{\tau}^{im}]^{-1} \mathbb{E}_{\tau}[W_{\tau}^{im} \theta_\tau^{\text{gt}}].
\]

It is worth noting that, from Lemma \cite{46} we have the property $\mathbb{E}_{\tau}[W_{\tau,N}^{im}] = W_{\tau,N}^{im}$, \lim_{N \to \infty} W_{\tau,N}^{im} = W_{\tau}^{im}$, 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

\[
\hat{p}(\theta_\tau | D_\tau, \theta_0) = \arg \min_{q(\theta_\tau) \in Q} \mathbb{E}_{p(x_\tau, y_\tau | \tau)} \left[ \text{KL}(q(\theta_\tau) || p(\theta_\tau | D_\tau, \theta_0)) \right]
\]

(51)

where \( p(\theta_\tau | D_\tau, \theta_0) \) can be computed from Bayes rule via

\[
p(\theta_\tau | D_\tau, \theta_0) \propto p(D_\tau, \theta_0)p(\theta_\tau | \theta_0).
\]

(52)

Assuming \( y_\tau | x_\tau, \theta_\tau \sim \mathcal{N}(\theta_\tau x_\tau, 1) \), the likelihood \( p(D_\tau | \theta_\tau) \) can be expressed by

\[
p(D_\tau | \theta_\tau) = \prod_{n=1}^{N} p(y_{\tau,n} | x_{\tau,n}, \theta_\tau) \propto \exp\left\{ -\frac{1}{2} \| y_{\tau,n} - X_{\tau,n} \theta_\tau \|^2 \right\}.
\]

(53)

Assuming the prior \( \theta_\tau | \theta_0 \sim \mathcal{N}(\theta_0, \gamma_0 I_d) \), with \( \gamma_0 \) being the prespecified weight of the prior or regularizer. The prior can be expressed by

\[
p(\theta_\tau | \theta_0) \propto \exp\left\{ -\frac{\gamma_0}{2} \| \theta_\tau - \theta_0 \|^2 \right\}.
\]

(54)

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

\[
p(\theta_\tau | D_\tau, \theta_0) \propto \exp\left\{ -\frac{1}{2} \| y_{\tau,N} - X_{\tau,N} \theta_\tau \|^2 - \frac{\gamma_0}{2} \| \theta_\tau - \theta_0 \|^2 \right\}
\]

\[
\times \exp\left\{ -\frac{1}{2} \| \theta_\tau - \mu_{\theta_{trn},N} \|^2 \right\} \mathcal{N}(\mu_{\theta_{trn},N}, \Sigma_{\theta_{trn},N})
\]

with \( \Sigma_{\theta_{trn},N} = (X_{\tau,N}^T X_{\tau,N} + \gamma_0 I)^{-1} \), \( \mu_{\theta_{trn},N} = \Sigma_{\theta_{trn},N}^{\top} X_{\tau,N} y_{\tau,N} + \gamma_0 \theta_0 \).

(55) - (57)

If \( p(\theta_\tau | D_\tau, \theta_0) \in Q \), then \( \hat{p}(\theta_\tau | D_\tau, \theta_0) = p(\theta_\tau | D_\tau, \theta_0) \), which holds in our analysis since \( Q \) is defined to be the set of Gaussian distributions. The empirical loss of BaMAML is

\[
L_{T,N}^{ba}(\theta_0) := \frac{1}{TN_2} \sum_{\tau=1}^{T} \left[ - \int \log p(D_{\tau,val} | \theta_\tau) \hat{p}_{\tau}^{ba}(\theta_\tau | D_{\tau, trn}, \theta_0) d\theta_\tau \right]
\]

s.t. \( \hat{p}_{\tau}^{ba}(\theta_\tau | D_{\tau, trn}, \theta_0) = \arg \min_{q(\theta_\tau) \in Q} \text{KL}(q(\theta_\tau) || p(\theta_\tau | D_{\tau, trn}, \theta_0)) \)

(58)

where \( \hat{p}_{\tau}^{ba}(\theta_\tau | D_{\tau, trn}, \theta_0) = p_{\tau}^{ba}(\theta_\tau | D_{\tau, trn}, \theta_0) = \mathcal{N}(\mu_{\theta_{trn},N}, \Sigma_{\theta_{trn},N}) \) is the solution of the inner problem. Therefore

\[
L_{T,N}^{ba}(\theta_0) = \frac{1}{TN_2} \sum_{\tau=1}^{T} \log p(D_{\tau,val} | D_{\tau, trn}, \theta_0)
\]

\[
- \frac{1}{TN_2} \sum_{\tau=1}^{T} \left[ \log p(D_{\tau,val} | D_{\tau, trn}, \theta_0) - \log p(D_{\tau, trn} | \theta_0) \right]
\]

where

\[
\log p(D_{\tau,val} | D_{\tau, trn}, \theta_0) - \log p(D_{\tau, trn} | \theta_0) = \log p(y_{\tau,N} | X_{\tau,N} \theta_0) - \log p(y_{\tau,N} | X_{\tau,N} \theta_0)
\]

\[
= -\frac{1}{2} \| y_{\tau,N} - X_{\tau,N} \theta_0 \|^2 \| \Sigma_{\theta_{trn}}^{-1} \]
with $\Sigma_{y,N}^{-1} = (I_N + \gamma_b^{-1}X_{r,N}X_{r,N}^T)^{-1}$. The last equation is because $p(y_{r,N}^{all} | x_{r,N}, \theta_0)$ can be computed by

$$p(y_{r,N}^{all} | x_{r,N}, \theta_0) = \int p(y_{r,N}^{all} | x_{r,N}^{all}, \theta_r)p(\theta_r | \theta_0)d\theta_r$$

$$\propto \int \exp\{-\frac{1}{2}||y_{r,N}^{all} - x_{r,N}^{all}\theta_r||^2 - \frac{\gamma_b}{2}||\theta_r - \theta_0||^2\}d\theta_r$$

$$= \int \exp\{-\frac{1}{2}(\theta_r - \mu_{\theta_r})^T\Sigma_{\theta_r}^{-1}(\theta_r - \mu_{\theta_r}) + \frac{1}{2}\mu_{\theta_r}^T\Sigma_{\theta_r}^{-1}\mu_{\theta_r} - \frac{\gamma_b}{2}\theta_0^T\theta_0 - \frac{1}{2}y_{r,N}^{all}y_{r,N}^{all}\}d\theta_r$$

$$\propto \exp\{-\frac{1}{2}(-\mu_{\theta_r}^T\Sigma_{\theta_r}^{-1}\theta_r + \gamma_b\theta_0^T\theta_0 + y_{r,N}^{all}y_{r,N}^{all})\} = \exp\{-\frac{1}{2}||y_{r,N}^{all} - x_{r,N}^{all}\theta_0||^2$$

where the last equation follows from the Binomial inverse theorem. Similarly, $p(y_{r,N}^{trn} | x_{r,N}^{trn}, \theta_0) \propto \exp\{-\frac{1}{2}||y_{r,N}^{trn} - x_{r,N}^{trn}\theta_0||^2_{\Sigma_{y,N}}\}$.

To solve for $\theta_0^{ba}$, using the optimality condition, we obtain

$$\theta_0^{ba} = \left(\sum_{\tau=1}^{T} W_{\tau}^{ba} \right)^{-1} \left(\sum_{\tau=1}^{T} W_{\tau}^{ba} \theta_0^{gt}\right) + \Delta_{\tau}^{ba}$$

$$\Delta_{\tau}^{ba} = \left(\sum_{\tau=1}^{T} W_{\tau}^{ba} \right)^{-1} \frac{1}{N_2} \left(\sum_{\tau=1}^{T} x_{r,N}^{all} \Sigma_{y,N}^{-1} e_{\tau,N}^{all} - x_{r,N}^{trn} \Sigma_{y,N}^{-1} e_{r,N}^{trn}\right)$$

$$W_{\tau}^{ba} = \left(\frac{\gamma_b}{N}\right)^{-1} Q_{\tau,N} + I)^{-1} Q_{\tau,N}$$

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

$$R_{N_a}^{ba}(\theta_0) = \frac{1}{N} E\left[-\log p(y_{r,N} | x_{r,N}, D_{r,N}, \theta_0)\right]$$

$$= E_r \left[||\theta_0 - \theta_r^{gt}||^2_{W_{\tau}^{ba}}\right] + \frac{1}{N_a} E_r \left[\text{tr}((I_d + (\gamma s)^{-1}Q_{r,N+N_a})^{-1} - (I_d + \gamma^{-1}Q_{r,N_a})^{-1})\right]$$

$$W_{r,N}^{ba} = E_{x_r} \left[\frac{\gamma_b}{N}((I_d + \gamma_b^{-1}N_2 Q_{r,N_a})^{-1} - (I_d + \gamma^{-1}(N + N_a)Q_{r,N+N_a})^{-1})\right]$$

Taking limits of $R_{N_a}^{ba}$ w.r.t. $N_a$ further leads to

$$R^{ba}(\theta_0) := \lim_{N_a \to \infty} R_{N_a}^{ba}(\theta_0) = \lim_{N_a \to \infty} E_r \left[||\theta_0 - \theta_r^{gt}||^2_{W_{r}^{ba}}\right] + 1 = E_r \left[||\theta_0 - \theta_r^{gt}||^2_{\hat{W}_{r}^{ba}}\right] + 1$$

$\theta_0^{ba}$ is defined to be the minimizer of $R^{ba}(\theta_0)$, given by

$$\theta_0^{ba} = \arg\min_{\theta_0} R^{ba}(\theta_0) = \arg\min_{\theta_0} E_r \left[||\theta_0 - \theta_r^{gt}||^2_{\hat{W}_{r}^{ba}}\right] = E_r \left[W_{r}^{ba}\right]^{-1} E_r \left[W_{r}^{ba} \theta_r^{gt}\right],$$

$$W_{r}^{ba} = ((\gamma s)^{-1}Q_{r,N} + I)^{-1} Q_{r,N} (\gamma^{-1}Q_{r,N_a} + I)^{-1}.$$
B Optimal population risk and statistical error analysis

Meta-test risk decomposition. Recall the meta-test risk function for the method $A$ of $\theta_0$ in (2). For method $A$, plugging $\theta_0^A$ into (2) and taking the limit over $N_a$, the number of data during adaptation, we have

\[
\mathcal{R}^A(\theta_0^A) = \mathbb{E}_T \left[ \|\theta_0^A - \theta^g_t\|_{W_t}^2 \right] + 1 + \mathbb{E}_r \left[ \|\theta_0^A - \theta_0^A + \theta_0^A - \theta^g_t\|_{W_t}^2 \right] + 1
\]

\[
= \mathbb{E}_r \left[ \|\theta_0^A - \theta_0^A\|_{W_t}^2 + \|\theta_0^A - \theta^g_t\|_{W_t}^2 + 2(\theta_0^A - \theta_0^A)\top W_t(\theta_0^A - \theta^g_t) \right] + 1
\]

\[
= \mathbb{E}_r \left[ \|\theta_0^A - \theta_0^A\|_{W_t}^2 + \|\theta_0^A - \theta^g_t\|_{W_t}^2 + 2(\theta_0^A - \theta_0^A)\top W_t(\mathbb{E}_r[W^A]^{-1} \mathbb{E}_r[W^A\theta^g_t] - \theta^g_t) \right] + 1
\]

\[
= \mathbb{E}_r \left[ \|\theta_0^A - \theta_0^A\|_{W_t}^2 + \|\theta_0^A - \theta^g_t\|_{W_t}^2 + 1 \right]
\]

where (a) follows from $W^A_t = \lim_{N_a \to \infty} W^A_{r,N_a}$; (b) is by plugging $\theta_0^A = \mathbb{E}_r[W^A]^{-1} \mathbb{E}_r[W^A\theta^g_t]$, (c) is because $\mathbb{E}_r[W^A_t(\mathbb{E}_r[W^A]^{-1} \mathbb{E}_r[W^A\theta^g_t] - \theta^g_t)] = \mathbb{E}_r[W^A_t\theta^g_t] - \mathbb{E}_r[W^A\theta^g_t] = 0$. From (65) and the definition of $\mathcal{R}^A_{N_a}(\cdot)$ in (2), we can decompose the meta-test risk to the optimal population risk and statistical error as follows

\[
\lim_{N_a \to \infty} \mathcal{R}^A_{N_a}(\theta_0^A) = \lim_{N_a \to \infty} \mathcal{R}^A_{N_a}(\theta_0^A) + \|\theta_0^A - \theta_0^A\|_{\mathbb{E}_r[W^A]}^2 \cdot \mathcal{R}^A_{\alpha}(\theta_0^A)
\]

This completes the proof of Proposition 3 in the main paper. Note that, the statistical error $\mathcal{R}^A_{\alpha}(\theta_0^A)$ is resulted from finite random data samples during meta-training to obtain the estimation of the parameter $\theta_0^A$, 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 $\mathcal{R}^A(\theta_0^A) = \min_{\theta_0} \mathcal{R}^A(\theta_0)$. 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

\[
\mathcal{R}^{er}(\theta_0^{er}) = \mathbb{E}_r \left[ \|\theta_0^{er} - \theta^g_t\|_{W_t}^2 \right] + 1
\]

For MAML, the optimal population risk is computed by

\[
\mathcal{R}^{ma}(\theta_0^{ma}, \alpha) = \mathbb{E}_r \left[ \|\theta_0^{ma} - \theta^g_t\|_{W_t^{ma}(\alpha)}^2 \right] + 1
\]

Note that when $\alpha = 0$, $\mathcal{R}^{ma}(\theta_0^{ma}, \alpha) = \mathcal{R}^{er}(\theta_0^{er})$.

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 $|Q_x| \leq \lambda$, $0 < \alpha \leq 1/\lambda$, $\mathcal{R}^{ma}(\theta_0^{ma}, \alpha)$ is monotonically decreasing. Therefore $\mathcal{R}^{ma}(\theta_0^{ma}, \alpha) < \mathcal{R}^{er}(\theta_0^{er})$ when $0 < \alpha \leq 1/\lambda$.

For iMAML, the optimal population risk is computed by

\[
\mathcal{R}^{im}(\theta_0^{im}, \gamma) = \mathbb{E}_r \left[ \|\theta_0^{im} - \theta^g_t\|_{W_t^{im}(\gamma)}^2 \right] + 1
\]

Comparison of MAML and iMAML optimal population risk. We can see that as $\gamma \to \infty$, $W_t^{im}(\gamma) \to W_t^{er} = Q_x$, $\mathcal{R}^{im}(\theta_0^{im}, \gamma) \to \mathcal{R}^{er}(\theta_0)$; as $\gamma \to 0$, $W_t^{im}(\gamma) \to 0$, $\mathcal{R}^{im}(\theta_0) \to 1$. To explicitly compare the optimal population risk of MAML and iMAML, we will show next that when $\gamma$ takes certain values, the corresponding risks satisfy $\mathcal{R}^{im}(\theta_0^{im}, \gamma) < \mathcal{R}^{ma}(\theta_0^{ma}, \alpha)$.

Corollary 2 Based on Assumption 7, for any $\tau \sim p(T)$, and $0 < \alpha < 1/\lambda$, $\|W_t^{ma}\| > 0$. And generally in a typical multi-task learning setting, the tasks are not all identical, therefore there exist $\tau \sim p(T)$, $\tau \in T, \theta_0^{ma} - \theta^g_t \neq 0$. Therefore there exists $\tau \in T$ such that $(\theta_0^{ma} - \theta^g_t)\top W_t^{ma}(\theta_0^{ma} - \theta^g_t) > 0$. 
First we show that the minimum value of the MAML population risk is larger than 1, i.e., $\mathcal{R}^{ma}(\theta_0^{ma}, \alpha) > 1$. According to Corollary 2 it is apparent that

$$E_r[\|\theta_0^{ma} - \theta^*_r\|^2_{W_{\text{im}}}] = E_r[(\theta_0^{ma} - \theta^*_r)^\top W_{\text{ma}}(\theta_0^{ma} - \theta^*_r)] > 0. \quad (70)$$

Therefore, we have

$$\mathcal{R}^{ma}(\theta_0^{ma}, \alpha) = E_r[\|\theta_0^{ma} - \theta^*_r\|^2_{W_{\text{ma}}}] + 1 > 1. \quad (71)$$

Note that $\mathcal{R}^{ma}(\theta_0^{ma}, \alpha)$ also depends on $\alpha$. Let $r^{ma} = \min_{\alpha} \mathcal{R}^{ma}(\theta_0^{ma}, \alpha) - 1$. From (71) we know that $r^{ma} > 0$. We will then show one can always find certain $\gamma$ such that

$$\mathcal{R}^{im}(\theta_0^{im}, \gamma) < \min_{\alpha} \mathcal{R}^{ma}(\theta_0^{ma}, \alpha) = r^{ma} + 1 \quad (72)$$

or equivalently $E_r[\|\theta_0^{im} - \theta^*_r\|^2_{W_{\text{im}}}] < r^{ma}$.

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

$$\|(\gamma^{-1}Q_r + I)^{-1}\| = 1/\lambda_{\min}(\gamma^{-1}Q_r + I) = 1/(\gamma^{-1}\lambda_{\min}(Q_r) + 1) \leq \frac{1}{\gamma^{-1}\lambda + 1} \quad (73)$$

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

$$E_r[\|\theta_0^{im} - \theta^*_r\|^2_{W_{\text{im}}}] \leq E_r[\|\theta_0^{im} - \theta^*_r\|^2_{W_{\text{im}}} \|W_{\text{im}}\|]$$

where we have discussed the bound for $\|W_{\text{im}}\|$ based on Assumption 1. Thus it suffices to bound $E_r[\|\theta_0^{im} - \theta^*_r\|^2_{W_{\text{im}}} \|W_{\text{im}}\|]$. We then have

$$E_r[\|\theta_0^{im} - \theta^*_r\|^2_{W_{\text{im}}} \|W_{\text{im}}\|] \leq E_r[\|\theta_0^{im} - \theta^*_r\|^2_{W_{\text{im}}} \|W_{\text{im}}\|]$$

$$\leq (M + \|\theta_0^{im}\|^2 + R^2)E_r[\|\gamma^{-1}Q_r + I\|^{-1}\|\gamma^{-1}Q_r + I\|^{-1}]$$

$$\leq (M + \|\theta_0^{im}\|^2 + R^2)E_r[\|\gamma^{-1}Q_r + I\|^{-1}\|\gamma^{-1}Q_r + I\|^{-1}]$$

$$\leq \frac{(a)(M + \|\theta_0^{im}\|^2 + R^2)^{\lambda}}{(\gamma^{-1}\lambda + 1)^2} \quad (74)$$

where $(a)$ holds because $\|Q_r\| \leq \tilde{\lambda}$, $\|(\gamma^{-1}Q_r + I)^{-1}\| \leq (\gamma^{-1}\lambda + 1)^{-1}$ from Assumption 1.

Let $C_\theta = \max\{((M + \|\theta_0^{im}\|^2 + R^2)^{\frac{1}{2}}), ((M + \|\theta_0^{im}\|^2 + R^2)^{\frac{1}{2}})\}$. In order to ensure $((M + \|\theta_0^{im}\|^2 + R^2)^{\frac{1}{2}})\leq C^2_\theta \frac{1}{(\gamma^{-1}\lambda + 1)^2} < r^{ma}$ it suffices to ensure

$$\gamma^{-1}\lambda + 1 > (r^{ma})^{-\frac{1}{2}} C_\theta \tilde{\lambda} \quad (75)$$

Since $0 < r^{ma} = E_r[\|\theta_0^{ma} - \theta^*_r\|^2_{W_{\text{ma}}}] \leq C^2_\theta E_r[\|W_{\text{ma}}\|] < C^2_\theta \tilde{\lambda}$. It follows that

$$(r^{ma})^{-\frac{1}{2}} C_\theta \tilde{\lambda} - 1 > 0. \quad (76)$$

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

$$0 < \gamma < ((r^{ma})^{-\frac{1}{2}} C_\theta \tilde{\lambda} - 1)^{-1} \tilde{\lambda} \quad (77)$$

In other words, by choosing (77), we have $\mathcal{R}^{im}(\theta_0^{im}, \gamma) < \mathcal{R}^{ma}(\theta_0^{ma}, \alpha) < \mathcal{R}^{ct}(\theta_0^{ct}), \forall 0 < \alpha \leq 1/\tilde{\lambda}$. 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 \( \tau \) and arbitrary \( \theta \), the population risks for MAML and iMAML, as functions of \( \theta \), are

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

where \( W_{\tau}^{\text{ma}}(\alpha) = (I - \alpha Q_\tau)(I - \alpha Q_\tau) \) and \( W_{\tau}^{\text{im}}(\gamma) = (\gamma^{-1} Q_\tau + I)^{-1} Q_\tau (\gamma^{-1} Q_\tau + I)^{-1} \). And the two functions are minimized by \( \theta_{0}^{\text{ma}} \) and \( \theta_{0}^{\text{im}} \) respectively. Let \( r^{\text{ma}} = \min_{\alpha} \mathcal{R}^{\text{ma}}(\theta_{0}^{\text{ma}}, \alpha) - c > 0 \), and when \( 0 < \gamma < \left( r^{\text{ma}} \right)^{-1/2} C_\theta \lambda^{\frac{1}{2}} - 1 \), then \( \mathcal{R}^{\text{im}}(\theta_{0}^{\text{im}}, \gamma) < \mathcal{R}^{\text{ma}}(\theta_{0}^{\text{ma}}, \alpha) \).

For BaMAML, the optimal population risk is

\[
\mathcal{R}^{ba}(\theta_{0}^{ba}, \gamma) = \mathbb{E}_\tau \left[ \| \theta_{0}^{ba} - \theta_\tau \|^2_{W_{\tau}^{ba}(\gamma)} \right] + 1
\]

(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 \( \tau \) and arbitrary \( \theta \), the optimal population risk for BaMAML, as functions of \( \theta \), is

\[
\mathcal{R}^{ba}(\theta, \gamma) = \mathbb{E}_\tau \left[ \| \theta - \theta_\tau \|^2_{W_{\tau}^{ba}(\gamma)} \right] + 1.
\]

And when \( 0 < \gamma < \left( r^{\text{ma}} \right)^{-1/2} C_\theta \lambda^{\frac{1}{2}} - 1 \), then \( \mathcal{R}^{ba}(\theta_{0}^{ba}, \gamma) < \mathcal{R}^{ma}(\theta_{0}^{ma}, \alpha) \).

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 \( W_{\tau,N}^A := \mathbb{E}_\tau \left[ W_{\tau,N}^A \right] \), then

\[
\| W_{\tau,N}^A \| \leq \| W_{\tau}^A \| + L^A \left( \tilde{O} \left( \frac{d}{N} \right) \right) + \tilde{O} \left( \sqrt{\frac{d}{N}} \right).
\]

Proof: For ERM, \( W_{\tau,N}^{er} := \mathbb{E}_\tau \left[ W_{\tau,N}^{er} \right] = \mathbb{E}[\hat{W}_{\tau,N}^{er}] = \mathbb{E}[\hat{Q}_{\tau,N}] = Q_{\tau} = W_{\tau}^{er}, L^{er} = 0. \)

For MAML, from [38], we have

\[
W_{\tau,N}^{ma} = \mathbb{E}_{Q_{\tau,N}} \left[ (I - \alpha \hat{Q}_{\tau,N}) Q_{\tau}(I - \alpha \hat{Q}_{\tau,N}) \right]
\]

\[
= W_{\tau}^{ma} + \frac{\alpha^2}{N} \left( \mathbb{E}_{x_{r,i}} [x_{r,i}x_{r,i}^T Q_{\tau}x_{r,i}] - Q_{\tau}^2 \right)
\]

therefore

\[
\| W_{\tau,N}^{ma} \| \leq \| W_{\tau}^{ma} \| + \frac{\alpha^2}{N} \left\| \mathbb{E}_{x_{r,i}} [x_{r,i}x_{r,i}^T Q_{\tau}x_{r,i}] - Q_{\tau}^2 \right\|
\]

\[
= \| W_{\tau}^{ma} \| + L^{ma} \left( \tilde{O} \left( \frac{d}{N} \right) \right) + \tilde{O} \left( \sqrt{\frac{d}{N}} \right).
\]
For iMAML, recall $W_{\tau,N}^{im} = \gamma^2 \Sigma_{\theta_x} Q_{x} \Sigma_{\theta_x}$, let $I_0 = \gamma \Sigma_{\theta_x} - (I + \gamma^{-1}Q_{x})^{-1}$, further derived as

$$I_0 = \gamma \Sigma_{\theta_x} - (I + \gamma^{-1}Q_{x})^{-1} = (I + \gamma^{-1}Q_{x})^{-1} - (I + \gamma^{-1}Q_{x})^{-1}$$

$$= -\gamma^{-1}(I + \gamma^{-1}Q_{x})^{-1}(Q_{x} - \hat{Q}_{x})(I + \gamma^{-1}Q_{x})^{-1}$$

Then we have

$$W_{\tau,N}^{im} = \mathbb{E}_{x}[\hat{W}_{\tau,N}^{im}] = \mathbb{E}_{x}[\gamma\Sigma_{\theta_x} Q_{x} \gamma \Sigma_{\theta_x}]$$

$$= \mathbb{E}_{x}[\gamma \Sigma_{\theta_x} + (I + \gamma^{-1}Q_{x})^{-1} - (I + \gamma^{-1}Q_{x})^{-1}]$$

$$= \mathbb{E}_{x}[I + \gamma^{-1}Q_{x} - (I + \gamma^{-1}Q_{x})^{-1}]$$

$$= \mathbb{E}_{x}[I_0Q_{x}I_0] + \mathbb{E}_{x}[I_0Q_{x}(I + \gamma^{-1}Q_{x})^{-1} + (I + \gamma^{-1}Q_{x})^{-1}Q_{x}I_0]$$

$$= \mathbb{E}_{x}[I_0Q_{x}I_0] + \mathbb{E}_{x}[I_0(I + \gamma^{-1}Q_{x})W_{\tau,N}^{im} + W_{\tau,N}^{im}(I + \gamma^{-1}Q_{x})I_0]$$

$$= \mathbb{E}_{x}[\Sigma_{\theta_x}(Q_{x} - \hat{Q}_{x})W_{\tau}^{im}(Q_{x} - \hat{Q}_{x})\Sigma_{\theta_x} + \Sigma_{\theta_x}(Q_{x} - \hat{Q}_{x})W_{\tau}^{im} + W_{\tau}^{im}(Q_{x} - \hat{Q}_{x})\Sigma_{\theta_x}]$$

where because

$$\|\mathbb{E}_{x}[\Sigma_{\theta_x}(Q_{x} - \hat{Q}_{x})W_{\tau}^{im}]\| \leq \mathbb{E}_{x}[\Sigma_{\theta_x}\|Q_{x} - \hat{Q}_{x}\|]W_{\tau}^{im}\| \leq \mathbb{E}_{x}[\|Q_{x} - \hat{Q}_{x}\|]W_{\tau}^{im}\|$$

and $\mathbb{E}_{x}[\Sigma_{\theta_x}(Q_{x} - \hat{Q}_{x})W_{\tau}^{im}(Q_{x} - \hat{Q}_{x})\Sigma_{\theta_x}] \leq \mathbb{E}_{x}[\|Q_{x} - \hat{Q}_{x}\|^2]\|W_{\tau}^{im}\|$. Based on sub-gaussian concentration inequality, it holds with probability at least $1 - \delta$ that

$$\|Q_{x} - \hat{Q}_{x}\| \leq \tilde{\lambda}CK^2\left(\sqrt{\frac{d + \log \frac{2}{\delta}}{N}} + \frac{d + \log \frac{2}{\delta}}{N}\right).$$

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

$$\|W_{\tau,N}^{im}\| \leq \|W_{\tau}^{im}\| + \tilde{O}(\sqrt{\frac{d}{N}}) + \tilde{O}(\frac{d}{N})L^{im}.$$ 

Similarly, for BaMAML,

$$W_{\tau,N}^{ba} = \mathbb{E}_{x}[\gamma \frac{N_1}{N_1}((I_d + \gamma^{-1}N_1\hat{Q}_{x,N})^{-1} - (I_d + \gamma^{-1}(N + N_1)\hat{Q}_{x,N+N_1})^{-1})]$$

$$= \mathbb{E}_{x}[\gamma \frac{N_1}{N_1}((I_d + \gamma^{-1}N_1\hat{Q}_{x})^{-1} - (I_d + \gamma^{-1}(N + N_1)\hat{Q}_{x})^{-1})$$

$$+ (I_d + \gamma^{-1}N_1\hat{Q}_{x,N})^{-1} - (I_d + \gamma^{-1}N\hat{Q}_{x,N})^{-1} - (I_d + \gamma^{-1}N_1\hat{Q}_{x})^{-1} + (I_d + \gamma^{-1}N\hat{Q}_{x})^{-1}]$$

$$= \mathbb{E}_{x}[\gamma \frac{N_1}{N_1}(I_d + \gamma^{-1}N_1\hat{Q}_{x})^{-1} - (I_d + \gamma^{-1}N\hat{Q}_{x})^{-1} - (I_d + \gamma^{-1}N_1\hat{Q}_{x})^{-1} + (I_d + \gamma^{-1}N\hat{Q}_{x})^{-1}]$$

$$= \mathbb{E}_{x}[\gamma \frac{N_1}{N_1}((I_d + \gamma^{-1}N_1\hat{Q}_{x})^{-1}(Q_{x} - \hat{Q}_{x,N})(I_d + \gamma^{-1}N_1\hat{Q}_{x,N})^{-1}$$

$$- \gamma^{-1}N(I_d + \gamma^{-1}N\hat{Q}_{x,N})^{-1}(Q_{x} - \hat{Q}_{x,N})(I_d + \gamma^{-1}N\hat{Q}_{x,N})^{-1}]$$

$$= \mathbb{E}_{x}[\gamma \frac{N_1}{N_1}(Q_{x} - \hat{Q}_{x,N})(I_d + \gamma^{-1}N\hat{Q}_{x,N})^{-1}$$

$$- (\gamma^{-1}N)(I_d + \gamma^{-1}N\hat{Q}_{x,N})^{-1}(Q_{x} - \hat{Q}_{x,N})(I_d + \gamma^{-1}N\hat{Q}_{x,N})^{-1}]$$

therefore

$$\|W_{\tau,N}^{ba}\| \leq \|W_{\tau}^{ba}\| + \tilde{O}(\sqrt{\frac{d}{N}}) + \tilde{O}(\frac{d}{N})L^{ba}.$$ 

**Lemma 2 (Concentration of $\hat{W}_{\tau,N}^{A}$)** Denote $d$ as the dimension of $\theta_x$, $T$ as the number of tasks. Suppose Assumption 1 holds, and $x_{\tau,i}$ 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

$$0 \leq \hat{W}_{\tau,N}^{cr} \leq \tilde{O}(c^{cr})I_d$$

(79)
where \( c^{er} := 1 + \max\{d/N, \sqrt{d/N}\} \), and \( \tilde{O}(\cdot) \) hides the logarithmic factor \( \log(NdT) \).
And denote \( \| \cdot \|_{op} \) as the operator norm. With probability at least \( 1 - T d^{-10} \), it holds that
\[
\left\| \frac{1}{T} \sum_{\tau=1}^{T} \hat{W}_{\tau,N}^{er} - E[\hat{W}_{\tau,N}^{er}] \right\|_{op} \leq \tilde{O}(c^{er} \sqrt{\frac{d}{T}} + d^{-4}),
\]
\[
\left\| \frac{1}{T} \sum_{\tau=1}^{T} (\hat{W}_{\tau,N}^{er})^2 - E[(\hat{W}_{\tau,N}^{er})^2] \right\|_{op} \leq \tilde{O}(c^{er}) \sqrt{\frac{d}{T}} + d^{-4}).
\]

**Proof:** The proof is similar to Lemma C.4 in Bai et al. (2021), the difference is we do not need Assumption 3, \( x_{\tau,i} \sim N(0, I_d) \), but only requires \( x_{\tau,i} \) to be sub-gaussian with parameter \( K \). Recall that \( \hat{W}_{\tau,N}^{er} = \frac{1}{N} X_{\tau,N}^{all\top} X_{\tau,N}^{all} = Q_{\tau,N} \). Applying the sub-gaussian covariance concentration (Vershynin (2018), Exercise 4.7.3), we have with \( v \in \mathbb{R}^d \) and \( \|v\|_2 = 1 \), the random variable
\[
\| v^\top \hat{W}_{\tau,N}^{er} 1\{W_{\tau}\} v - v^\top E[\hat{W}_{\tau,N}^{er} 1\{W_{\tau}\}] v \|
\]
is bounded by:
\[
\| v^\top (\frac{1}{T} \sum_{\tau=1}^{T} \hat{W}_{\tau,N}^{er} 1\{W_{\tau}\}) v - v^\top E[\hat{W}_{\tau,N}^{er} 1\{W_{\tau}\}] v \| \geq t \leq 2 \exp \left( - \frac{T t^2}{2(K \cdot c^{er})^2} \right).
\]

Using the fact that for any symmetric matrix \( M \), \( \|M\|_{op} \leq 2 \sup_{v \in S^{d-1}} | v^\top M v | \) where \( N_{1/4}(S^{d-1}) \) is a 1/4-covering set of the \((d-1)\)-unit sphere \( S^{d-1} \) with \( |N_{1/4}(S^{d-1})| \leq 9^d \) (Vershynin (2018), Exercise 4.4.3), we have
\[
\| v^\top (\frac{1}{T} \sum_{\tau=1}^{T} \hat{W}_{\tau,N}^{er} 1\{W_{\tau}\}) v - v^\top E[\hat{W}_{\tau,N}^{er} 1\{W_{\tau}\}] v \| \leq \exp \left( - \frac{T t^2}{2(K \cdot c^{er})^2} + 3d \right).
\]

Taking \( t = \tilde{O}(K \cdot c^{er} \sqrt{\frac{3d + 20 \log(d)}{T}}) \), the above probability is upper bounded by \( d^{-10} \). In other words, with probability at least \( 1 - T d^{-10} \), we have
\[
\left\| \frac{1}{T} \sum_{\tau=1}^{T} \hat{W}_{\tau,N}^{er} 1\{W_{\tau}\} - E[\hat{W}_{\tau,N}^{er} 1\{W_{\tau}\}] \right\|_{op} \leq \tilde{O}(K \cdot c^{er} \sqrt{\frac{d}{T}}).
\]

To bound the difference between \( E[\hat{W}_{\tau,N}^{er}] \) and \( E[\hat{W}_{\tau,N}^{er} 1\{W_{\tau}\}] \), it follows
\[
\| E[\hat{W}_{\tau,N}^{er}] - E[\hat{W}_{\tau,N}^{er} 1\{W_{\tau}\}] \|_{op} \leq E \left[ \left\| \hat{W}_{\tau,N}^{er} 1\{W_{\tau}\} \right\|_{op}^2 \right] \leq \left( E \left[ \left\| \hat{W}_{\tau,N}^{er} \right\|_{op}^2 \right] \cdot P(W_{\tau}^c) \right)^{1/2}
\]
\[
\leq \sqrt{E \left[ \max_i \|x_{\tau,i}\|^2 \right]} \cdot d^{-10} \leq \sqrt{k^2 (d + C \log N) \cdot d^{-10}} = \tilde{O}(d^{-4.5})
\]
where $W_{c}^c$ is the complement of $W_c$, and the last inequality is by sub-gaussian norm concentration. Combining (82) and (83), with probability at least $1 - Td^{-10}$, we have that

$$\left\| \frac{1}{T} \sum_{\tau=1}^{T} \tilde{W}_{\tau,N}^{er} - E[\tilde{W}_{\tau,N}^{er}] \right\|_{op} \leq \left\| \frac{1}{T} \sum_{\tau=1}^{T} \tilde{W}_{\tau,N}^{er} 1\{W_{\tau}\} - E[\tilde{W}_{\tau,N}^{er} 1\{W_{\tau}\}] \right\|_{op}$$

$$+ \left\| E[\tilde{W}_{\tau,N}^{er}] - E[\tilde{W}_{\tau,N}^{er} 1\{W_{\tau}\}] \right\|_{op} \leq \tilde{O} \left( \epsilon^{er} \sqrt{\frac{d}{T}} + d^{-4.5} \right).$$

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

$$\left\| \frac{1}{T} \sum_{\tau=1}^{T} (\tilde{W}_{\tau,N}^{er})^2 - E[(\tilde{W}_{\tau,N}^{er})^2] \right\|_{op} \leq \tilde{O} \left( (\epsilon^{er})^2 \sqrt{\frac{d}{T}} + d^{-4} \right).$$

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

$\square$

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

$$|z^\top Cz - E[z^\top Cz]| \leq O \left( K^2 \|C\|_F \log \frac{2}{\delta} \right).$$

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

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

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

$$|z^\top Cz - E[z^\top Cz]| \leq O \left( K^2 \|C\|_F \log(2/\delta) \right) + O \left( K \|E[z]\| \|C\|_{op} \sqrt{\log(2/\delta)} \right).$$

Proof:

$$|z^\top Cz - E[z^\top Cz]| = |(z - E[z])^\top C(z - E[z]) - E[(z - E[z])^\top C(z - E[z])] + 2E[z]^\top C(z - E[z])|$$

$$\leq |(z - E[z])^\top C(z - E[z]) - E[(z - E[z])^\top C(z - E[z])]| + 2|E[z]^\top C(z - E[z])|$$

$$\leq O \left( K^2 \|C\|_F \log(2/\delta) \right) + O \left( K \|E[z]\| \|C\|_{op} \sqrt{\log(2/\delta)} \right)$$

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

$\square$

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

$$\left\| \frac{1}{T} \sum_{\tau=1}^{T} U_{\tau} z_{\tau} \right\| - \|E_{\tau} [U_{\tau} z_{\tau}]\| \leq \tilde{O} \left( K\tilde{\lambda} \sqrt{\frac{d}{T}} \log \frac{2}{\delta} \right).$$
Proof: Apply Lemma 3, the Hanson-Wright inequality, and let \( z = z_\tau, C = U_\tau^T U_\tau \), we obtain that with probability at least \( 1 - \delta \),

\[
|z_\tau^T U_\tau^T U_\tau z_\tau - \mathbb{E}_{z_\tau | U_\tau} [z_\tau^T U_\tau^T U_\tau z_\tau]| \leq O\left( K^2 \|U_\tau^T U_\tau\|_F \log \frac{2}{\delta} \right).
\]

Since

\[
\left| \|U_\tau z_\tau\| - \mathbb{E}_{z_\tau | U_\tau} [U_\tau z_\tau]\right|^2 \leq \left| \|U_\tau z_\tau\|^2 - \|\mathbb{E}_{z_\tau | U_\tau} [U_\tau z_\tau]\|^2 \right|,
\]

\[
= |z_\tau^T U_\tau^T U_\tau z_\tau - \mathbb{E}_{z_\tau | U_\tau} [z_\tau^T U_\tau^T U_\tau z_\tau]| \leq O\left( K^2 d \|U_\tau\|_{op} \log \frac{2}{\delta} \right)
\]

where the last equation holds because \( z_\tau \) has mean-zero entries. Therefore, it holds with probability at least \( 1 - \delta \) that

\[
\left| \|U_\tau z_\tau\| - \mathbb{E}_{z_\tau | U_\tau} [U_\tau z_\tau]\right| \leq O\left( K\lambda \sqrt{d} \log \frac{2}{\delta} \right).
\]

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

\[
\left| \mathbb{E}_{z_\tau, U_\tau} [U_\tau z_\tau] - \mathbb{E}_{z_\tau | U_\tau} [U_\tau z_\tau] \right| \leq \mathbb{O}(K\lambda \sqrt{d} \log \frac{2}{\delta}).
\]

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

\[
\left| \frac{1}{T} \sum_{\tau=1}^T \|U_\tau z_\tau\| - \mathbb{E}_{z_\tau, U_\tau} [U_\tau z_\tau]\right| \leq \mathbb{O}(K\lambda \sqrt{\frac{d}{T}} \log \frac{2}{\delta}).
\]

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

\[
\begin{align*}
z_A := & \left[ (\theta_1^A - \theta_0^A)^\top, \ldots, (\theta_T^A - \theta_0^A)^\top \right] \in \mathbb{R}^{dT}, \\
U_A := & \left[ \bar{W}_1^A (\sum_{r=1}^T \bar{W}_r^A)^{-1}, \ldots, \bar{W}_T^A (\sum_{r=1}^T \bar{W}_r^A)^{-1} \right] \in \mathbb{R}^{dT \times d}.
\end{align*}
\]

1) Suppose Assumptions 1-2 hold, the statistical error for method \( A \) is computed by

\[
\mathcal{E}_A^2(\hat{\theta}_0^A) = \|\hat{\theta}_0^A - \theta_0^A\|_{\mathbb{E}[W^A]}^2 = \|U_A z_A\|_{\mathbb{E}[W^A]}^2 + \|\Delta_t^A\|^2_{\mathbb{E}[W^A]} + 2z_A^T U_A \mathbb{E}[W^A] \Delta_t^A I_t^A
\]

where with probability at least \( 1 - Td^{-10} \), the first term \( I_t^A \) can be bounded above by \( 2 \)

\[
I_t^A \leq \frac{R^2}{T} \left( \min \left( \mathbb{E}[W^A] \right)^{-1} \lambda_{\max} \left( \mathbb{E}[W^A]^2 \right) + \mathbb{O}(\frac{d}{N}) + \mathbb{O}(\frac{1}{\sqrt{d}}) + \mathbb{O}(\sqrt{\frac{d}{T}}) \right)
+ \left( \mathbb{O}(\sqrt{\frac{d}{T}}) + \mathbb{O}(\frac{d}{N}) \right) M^2.
\]

2) Suppose Assumptions 1,3 hold, the statistical error for method \( A \) can be computed by

\[
\mathcal{E}_A^2(\hat{\theta}_0^A) = w_A \|\hat{\theta}_0^A - \theta_0^A\|_{\mathbb{E}[W^A]}^2 = w_A \|U_A z_A\|_{\mathbb{E}[W^A]}^2 + \|\Delta_t^A\|^2_{\mathbb{E}[W^A]} + 2z_A^T U_A \mathbb{E}[W^A] \Delta_t^A I_t^A
\]

Define \( \hat{C}_0^A := \frac{1}{N} \mathbb{E}^{-2} [\bar{W}_r^A, \mathbb{E}(\bar{W}_r^A)^2] \). With probability at least \( 1 - Td^{-10} \), \( I_t^A \) can be bounded above by

\[
I_t^A \leq \frac{R^2}{T} \left( \hat{C}_0^A + \mathbb{O}(\frac{1}{\sqrt{d}}) + \mathbb{O}(\sqrt{\frac{d}{T}}) \right).
\]

\[\text{Note that, we provide bound for } I_t^A \text{ and } I_t^A \text{ in this lemma since it has the same form for different methods } A. \text{ And the bound for the rest terms in the statistical error are deferred to later sections for the specific methods.}\]
Thus the difference between the estimated model parameter $\hat{\theta}_0^A$ and the population-wise optimal model parameter $\theta_0^A$, which is be used to compute the statistical error, is given by

$$\hat{\theta}_0^A - \theta_0^A = \left( \sum_{\tau=1}^{T} \hat{W}_{\tau,N}^A \right)^{-1} \left( \sum_{\tau=1}^{T} \hat{W}_{\tau,N}^A (\theta_{\tau}^{st} - \theta_0^A) \right) + \Delta_\tau^A = U_A^T z_A + \Delta_\tau^A$$

(88)

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

$$E_{\tau}^2(\hat{\theta}_0^A) = \|\hat{\theta}_0^A - \theta_0^A\|_{E_{\tau}[W_{\tau}^A]}^2 = \|U_A^T z_A\|_{E_{\tau}[W_{\tau}^A]}^2 + \|\Delta_\tau^A\|_{E_{\tau}[W_{\tau}^A]}^2$$

(89)

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

$$\|U_A^T z_A\|_{E_{\tau}[W_{\tau}^A]}^2 \leq \|U_A^T z_A\|_{E_{\tau}[W_{\tau}^A]}^2 - E_{\tau,N}[U_A][\|U_A^T z_A\|_{E_{\tau}[W_{\tau}^A]}^2] + E_{\tau,A}[\|U_A^T z_A\|_{E_{\tau}[W_{\tau}^A]}^2]$$

Next we will bound the above two terms respectively. We first bound $\|U_A^T z_A\|_{E_{\tau}[W_{\tau}^A]}^2 - E_{\tau,N}[U_A][\|U_A^T z_A\|_{E_{\tau}[W_{\tau}^A]}^2]$. From Assumption 2, $\theta_{\tau}^{st} - \theta_0^A$ are $(R/\sqrt{d})$-sub-gaussian. To bound the absolute error around the expectation, from the Hanson-Wright inequality in Lemma 3 with probability at least $1 - \delta$, the following inequality holds

$$\|U_A^T z_A\|_{E_{\tau}[W_{\tau}^A]}^2 - E_{\tau,N}[U_A][\|U_A^T z_A\|_{E_{\tau}[W_{\tau}^A]}^2] \leq \widehat{O}(\frac{R^2}{d}) \left( \sum_{\tau=1}^{T} W_{\tau,N}^A \right)^{-1} \left( \sum_{\tau=1}^{T} W_{\tau,N}^A \right)^{-1} \left( \sum_{\tau=1}^{T} (W_{\tau,N}^A)^2 \right)_{T\sigma} + \widehat{O}(\frac{RM}{dT}) \|U_A^T z_A\|_{E_{\tau}[W_{\tau}^A]}^2 \leq \widehat{O}(\frac{R^2 + RM}{dT}) \leq \widehat{O}(\frac{R^2 + RM}{dT}) \leq \widehat{O}(\frac{R^2 + RM}{dT})$$

(90)

Note that in the last equation, we ignore the higher order terms in $\|\sum_{\tau=1}^{T} W_{\tau,N}^A\|_{E_{\tau}[W_{\tau}^A]}^2$, which can be obtained from Lemma 2.

To bound the expected statistical error, $E_{\tau,A}[\|U_A^T z_A\|_{E_{\tau}[W_{\tau}^A]}^2]$, first note that since for all $\tau$, $W_{\tau}^A$ is symmetric positive definite (PD) based on Assumption 4 $E_{\tau}[W_{\tau}^A]$ is also symmetric PD, who has a Cholesky decomposition, $E_{\tau}[W_{\tau}^A] = E_{\tau}^T [W_{\tau}^A]E_{\tau}^T [W_{\tau}^A]^T$ with $E_{\tau}^T [W_{\tau}^A]$ defined as the lower triangular matrix in the decomposition. The statistical error can be rewritten as

$$\|U_A^T z_A\|_{E_{\tau}[W_{\tau}^A]}^2$$

Note that in the last equation, we ignore the higher order terms in $\|\sum_{\tau=1}^{T} W_{\tau,N}^A\|_{E_{\tau}[W_{\tau}^A]}^2$, which can be obtained from Lemma 2.
whose conditional expectation is given by
\[
\mathbb{E}_{z \sim \mathbb{U}_{\mathcal{A}}||z|\mathcal{A}|} \left[ \left\| U_{\mathcal{A}} z \right\|_2^2 \right] = \mathbb{E}_{\theta^g \sim \mathbb{W}^\tau_{\mathcal{A}}} \left[ \mathrm{tr} \left( \left( \mathbb{E}_{\tau \sim \mathbb{U}_{\mathcal{A}}} \left[ W^\tau_{\mathcal{A}} \right] \left( \sum_{\tau=1}^T W^\tau_{\mathcal{A},N} \right)^{-1} \left( \sum_{\tau=1}^T \hat{W}^\tau_{\mathcal{A},N} \left( \theta^g_\tau - \theta^0_\tau \right) \right) \right) \right] = \mathbb{E}_{\tau \sim \mathbb{U}_{\mathcal{A}}} \left[ \left\| \mathbb{E}_{\theta^g \sim \mathbb{W}^\tau_{\mathcal{A}}} \left[ W^\tau_{\mathcal{A}} \right] \left( \sum_{\tau=1}^T \hat{W}^\tau_{\mathcal{A},N} \left( \theta^g_\tau - \theta^0_\tau \right) \right) \right\|_2^2 \right]
\]

where the last equation is given by the fact that $\mathbb{E}[zz^T] = \text{Cov}[z] + \mathbb{E}[z] \mathbb{E}[z]^T$ for any random vector $z$, and the linear and cyclic property of trace. From Assumption $\theta^g - \theta^0_\tau$ has independent $(R/\sqrt{T})$-subgaussian entries, and Lemma 4 linear combinations of sub-gaussian random variables are still sub-gaussian, $\mathbb{E}_{\tau \sim \mathbb{U}_{\mathcal{A}}} \left[ W^\tau_{\mathcal{A}} \right] \left( \sum_{\tau=1}^T \hat{W}^\tau_{\mathcal{A},N} \left( \theta^g_\tau - \theta^0_\tau \right) \right)$ has sub-gaussian entries with parameter $\left\| \mathbb{E}_{\tau \sim \mathbb{U}_{\mathcal{A}}} \left[ W^\tau_{\mathcal{A}} \right] \left( \sum_{\tau=1}^T \hat{W}^\tau_{\mathcal{A},N} \right) \right\|_2^2 \leq \frac{1}{T} \sum_{\tau=1}^T \left\| \hat{W}^\tau_{\mathcal{A},N} \right\|_2^2 \leq \frac{R^2}{T}$, 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

\[
I_a \leq \left\| \mathbb{E}_{\tau \sim \mathbb{U}_{\mathcal{A}}} \left[ W^\tau_{\mathcal{A}} \right] \left( \sum_{\tau=1}^T \hat{W}^\tau_{\mathcal{A},N} \right)^{-1} \right\|_2^2 \leq \left\| \mathbb{E}_{\tau \sim \mathbb{U}_{\mathcal{A}}} \left[ W^\tau_{\mathcal{A}} \right] \right\|_2^2 \left\| \sum_{\tau=1}^T \hat{W}^\tau_{\mathcal{A},N} \right\|_2^2 \leq \frac{R^2}{T} \left\| \mathbb{E}_{\tau \sim \mathbb{U}_{\mathcal{A}}} \left[ W^\tau_{\mathcal{A}} \right] \right\|_2^2 \left\| \sum_{\tau=1}^T \hat{W}^\tau_{\mathcal{A},N} \right\|_2^2 \leq \frac{R^2}{T} \left\| \mathbb{E}_{\tau \sim \mathbb{U}_{\mathcal{A}}} \left[ W^\tau_{\mathcal{A}} \right] \right\|_2^2 \left\| \sum_{\tau=1}^T \hat{W}^\tau_{\mathcal{A},N} \right\|_2^2 .
\]

$I_b$ can be further derived as

\[
I_b = \left\| \mathbb{E}_{\theta^g \sim \mathbb{W}^\tau_{\mathcal{A}}} \left[ W^\tau_{\mathcal{A}} \right] \left( \sum_{\tau=1}^T \hat{W}^\tau_{\mathcal{A},N} \right)^{-1} \left( \sum_{\tau=1}^T \hat{W}^\tau_{\mathcal{A},N} \left( \theta^g_\tau - \theta^0_\tau \right) \right) \right\|_2^2 = \left\langle \mathbb{E}_{\tau \sim \mathbb{U}_{\mathcal{A}}} \left[ W^\tau_{\mathcal{A}} \right] \left( \sum_{\tau=1}^T \hat{W}^\tau_{\mathcal{A},N} \right)^{-1} \left( \sum_{\tau=1}^T \hat{W}^\tau_{\mathcal{A},N} \left( \theta^g_\tau - \theta^0_\tau \right) \right) \right\rangle.
\]
where $I_{b1}$ can be further derived as

$$I_{b1} = (\sum_{\tau=1}^{T} \mathbf{W}_{\tau,N}^{A})^{-1} \left\{ \mathbb{E}_{\theta^{a}_{\tau}} [\mathbf{W}_{\tau,N}^{A}] \left[ \sum_{\tau=1}^{T} \mathbb{W}_{\tau,N}^{A} (\theta^{a}_{\tau} - \theta^{a}_0) \right] - \mathbb{E}_{\theta^{a}_{\tau}}[\mathbf{W}_{\tau,N}^{A}] \left[ \sum_{\tau=1}^{T} \mathbb{W}_{\tau,N}^{A} (\theta^{a}_{\tau} - \theta^{a}_0) \right] \right\}$$

$$= \left( \frac{1}{T} \sum_{\tau=1}^{T} \mathbf{W}_{\tau,N}^{A} \right)^{-1} \left\{ \frac{1}{T} \sum_{\tau=1}^{T} \mathbf{W}_{\tau,N}^{A} \mathbb{E}_{\theta^{a}_{\tau}} \left[ \theta^{a}_{\tau} - \theta^{a}_0 \right] - \mathbb{E}_{\theta^{a}_{\tau}} \left[ \mathbf{W}_{\tau,N}^{A} \mathbb{E}_{\theta^{a}_{\tau}} \left[ \theta^{a}_{\tau} - \theta^{a}_0 \right] \right] \right\}$$

$$\|I_{b1}\| \leq \tilde{O} \left( \sqrt{\frac{d}{T}} \right) K M \left\| \frac{1}{T} \sum_{\tau=1}^{T} \mathbf{W}_{\tau,N}^{A} \right\|^{-1}.$$ 

And for $I_{b2}$, it holds that

$$I_{b2} = (\sum_{\tau=1}^{T} \mathbf{W}_{\tau,N}^{A})^{-1} \left\{ \mathbb{E}_{\theta^{a}_{\tau}} \left[ \mathbf{W}_{\tau,N}^{A} \right] \left[ \sum_{\tau=1}^{T} \mathbb{W}_{\tau,N}^{A} (\theta^{a}_{\tau} - \theta^{a}_0) \right] - \mathbb{E}_{\theta^{a}_{\tau}}[\mathbf{W}_{\tau,N}^{A}] \left[ \sum_{\tau=1}^{T} \mathbb{W}_{\tau,N}^{A} (\theta^{a}_{\tau} - \theta^{a}_0) \right] \right\}$$

$$= \left( \frac{1}{T} \sum_{\tau=1}^{T} \mathbf{W}_{\tau,N}^{A} \right)^{-1} \left\{ \mathbb{E}_{\theta^{a}_{\tau}} \left[ \mathbf{W}_{\tau,N}^{A} \right] \sum_{\tau=1}^{T} \left( \mathbf{W}_{\tau,N}^{A} - \mathbf{W}_{\tau,N}^{A} \right) (\theta^{a}_{\tau} - \theta^{a}_0) \right\}$$

$$= \left( \frac{1}{T} \sum_{\tau=1}^{T} \mathbf{W}_{\tau,N}^{A} \right)^{-1} \left\{ \mathbb{E}_{\theta^{a}_{\tau}} \left[ \mathbf{W}_{\tau,N}^{A} \right] \sum_{\tau=1}^{T} \left( \mathbb{W}_{\tau,N}^{A} - \mathbb{W}_{\tau,N}^{A} \right) \left( \theta^{a}_{\tau} - \theta^{a}_0 \right) \right\}$$

$$\|I_{b2}\| \leq \tilde{O} \left( \sqrt{\frac{d}{T}} K + \|\mathbb{E}[\mathbf{W}_{\tau,N}^{A} - \mathbf{W}_{\tau,N}^{A}] \| \right) M \left\| \frac{1}{T} \sum_{\tau=1}^{T} \mathbf{W}_{\tau,N}^{A} \right\|^{-1}.$$ 

$I_{b3} = 0$ since $\mathbb{E}_{\theta^{a}_{\tau}} \left[ \mathbf{W}_{\tau,N}^{A} (\theta^{a}_{\tau} - \theta^{a}_0) \right] = 0$. Combining $I_{b1}, I_{b2}, I_{b3}$ from above discussions we can bound $I_b$ by

$$I_b = \left\| \mathbb{E}_{\theta^{a}_{\tau}} \left[ \mathbf{W}_{\tau,N}^{A} \right] \left[ \sum_{\tau=1}^{T} \mathbb{W}_{\tau,N}^{A} (\theta^{a}_{\tau} - \theta^{a}_0) \right] \right\|^2$$

$$\leq \left( \tilde{O} \left( \sqrt{\frac{d}{T}} K + \|\mathbb{E}[\mathbf{W}_{\tau,N}^{A} - \mathbf{W}_{\tau,N}^{A}] \| \right) M \right) \left\| \mathbb{E}_{\theta^{a}_{\tau}} [\mathbf{W}_{\tau,N}^{A}] \right\| \left\| \frac{1}{T} \sum_{\tau=1}^{T} \mathbf{W}_{\tau,N}^{A} \right\|^{-1}. \quad (93)$$

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

$$\mathbb{E}_{\mathbf{U}_A} \left[ \left\| \mathbf{U}_A \mathbf{z}_A \right\|_2^2 \right] \leq \frac{R^2}{T} \left\| \mathbb{E}_{\tau} [\mathbf{W}_{\tau}^{A}] \right\| \left\| \frac{1}{T} \sum_{\tau=1}^{T} \mathbf{W}_{\tau,N}^{A} \right\|^{-2} \left\| \frac{1}{T} \sum_{\tau=1}^{T} (\mathbf{W}_{\tau,N}^{A})^2 \right\|$$

$$+ \left( \tilde{O} \left( \sqrt{\frac{d}{T}} K + \|\mathbb{E}[\mathbf{W}_{\tau,N}^{A} - \mathbf{W}_{\tau,N}^{A}] \| \right) M \right) \left\| \mathbb{E}_{\tau} [\mathbf{W}_{\tau}^{A}] \right\| \left\| \frac{1}{T} \sum_{\tau=1}^{T} \mathbf{W}_{\tau,N}^{A} \right\|^{-1}. \quad (94)$$

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

$$\left\| \mathbf{U}_A \mathbf{z}_A \right\|_2^2 \leq \frac{R^2}{T} \left( \lambda_{\min} (\mathbb{E} [\mathbf{W}_{\tau}^{A}])^{-1} \lambda_{\max} (\mathbb{E} [\mathbf{W}_{\tau}^{A}]) \right) + \tilde{O}(\frac{d}{N}) + \tilde{O}(\sqrt{\frac{d}{N}})$$

$$+ \left( \tilde{O} \left( \sqrt{\frac{d}{T}} K + \tilde{O}\left( \frac{d}{N} \right)^2 \right) \right) M. \quad (95)$$

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, $\mathbf{W}_{\tau,A} = w_A I_d$, where $w_A$ is the same for different task $\tau$, therefore $\theta^{a}_0 = \mathbb{E}_{\tau} [\theta^{a}_\tau]$, and $w_{ma} = 1, w_{ma} = (1 - \alpha)^2, w_{im} = (1 + \gamma^{-1})^{-2}, w_{as} = (1 + (\gamma\alpha)^{-1})^{-1}(1 + \gamma^{-1})^{-1}$. 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

$$\lim_{N_a \to \infty} \mathcal{R}_{\mathcal{A}}^A(\hat{\theta}_0^A) = \frac{w_A \|\hat{\theta}_0^A - \theta_0^A\|^2}{I_2^d} + \lim_{N_a \to \infty} \mathcal{R}_{\mathcal{A}}^A(\theta_0^A).$$

(96)

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

$$E_{\mathcal{A}}^2(\theta_0^A) = \frac{\mathcal{R}_{\mathcal{A}}^A(\hat{\theta}_0^A) - \mathcal{R}_{\mathcal{A}}^A(\theta_0^A)}{I_2^d} = w_A \frac{\|\hat{\theta}_0^A - \theta_0^A\|^2}{I_2^d} + \|\hat{\Delta}_A^2\|^2 + 2\hat{\Delta}_A^2 \Delta^2_A.$$

(97)

We will bound term $I_2^d$ in the above equation. The only difference of $I_2^d$ from $I_1^d$ is that we can treat $E_{\tau}[W_\tau] = I_d$ and $M = 0$ when adding Assumption 3. Therefore, with probability at least $1 - \delta$, we have

$$\left\|E_{\mathcal{A}}\|U_A^\top z_A\|^2 - E_{\theta^*, \epsilon, \tau}[U_A^\top z_A]\right\| = \tilde{O}(\frac{R^2}{T\sqrt{d}}).$$

(98)

To compute $E_{\theta^*, \epsilon, \tau}[U_A^\top z_A]$, first we have

$$E_{\theta^*, \epsilon, \tau}[U_A^\top z_A] = R^2 \frac{1}{Td} \left\{ \left(\frac{1}{T} \sum_{\tau=1}^{T} \hat{W}_A^{\tau}, \sum_{\tau=1}^{T} \hat{W}_A^{\tau} \right) - \frac{1}{d} \mathbb{E} \left[ \hat{W}_A^{\tau} \right] \mathbb{E} \left[ \hat{W}_A^{\tau} \right] \right\} = C_0^d$$

(99)

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

$$|I_c| \leq \tilde{O}(\sqrt{d} \sqrt{T}), |I_d| \leq \tilde{O}(\sqrt{d} \sqrt{T}).$$

(100)

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

$$E_{\theta^*, \epsilon, \tau}[U_A^\top z_A] \leq \frac{R^2}{T} \left( C_0^d + \tilde{O}(\sqrt{d} \sqrt{T}) \right).$$

(101)

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

$$I_2^d = z_A^\top U_A^\top z_A \leq \frac{R^2}{T} \left( C_0^d + \tilde{O}(\sqrt{d} \sqrt{T}) \right).$$

(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

$$C_0^d := \frac{1}{d} \mathbb{E} \left[ \mathbb{E} \left[ (\hat{W}_A^2)^2 \right] \right] = \frac{1}{d} \mathbb{E} \left[ \text{tr}((\hat{W}_A^2)^2) \right] \left\{ \frac{1}{d} \mathbb{E} \left[ \text{tr}(\hat{W}_A^2) \right] \right\}^{-2} \geq 1.$$
which gives

\[ E \left[ W^2 \right] = \frac{1}{d} \mathbb{E} \left[ \lambda_{max}^2 \right] \]

Following similar arguments, for iMAML, it also holds that

\[ E \left[ W_{im}^2 \right] = \frac{1}{d} \mathbb{E} \left[ \lambda_{max}^2 \right] \]

and for BaMAML, we use the expression

\[ W_{ba} = \frac{1}{d} \left( I_d + \gamma^{-1} Q_{r,N} \right)^{-1} - \left( I_d + (\gamma s)^{-1} Q_{r,N} \right)^{-1} \]

which apparently gives 

\[ E \left[ W_{ba}^2 \right] = \frac{1}{d} \mathbb{E} \left[ \lambda_{max}^2 \right] \]

To summarize, we have proved for all four methods ERM, MAML, iMAML and BaMAML, \( E \left[ W^2 \right] = \frac{1}{d} \mathbb{E} \left[ \lambda_{max}^2 \right] \). Then it is not hard to see that

\[ C_0^d := \frac{1}{d} \mathbb{E} \left[ \lambda_{max}^2 \right] \]

Finally, applying Jensen’s inequality, for any PSD matrix \( M \in \mathbb{R}^{d \times d} \), we have \( \frac{1}{d} \mathbb{E} \left[ \lambda_{max}^2 \right] \geq \left( \frac{1}{d} \mathbb{E} \left[ \lambda_{max}^2 \right] \right)^2 \), therefore

\[ C_0^d := \frac{1}{d} \mathbb{E} \left[ \lambda_{max}^2 \right] \geq 1. \]
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

\[
C_0^{\text{ma}} := \frac{1}{d} \mathbb{E} \left[ \mathbb{E} \left[ \mathbb{E} \left[ \left( \mathbb{W}_{\tau,N}^{\text{ma}} \right)^2 \right] \right] \right] \\
= \frac{1}{d N_2} \mathbb{E} \left[ \text{tr}^2 \left( (I - \alpha \hat{Q}_{\tau,N_1})^2 \right) + (N_2 + 1) \text{tr} \left( (I - \alpha \hat{Q}_{\tau,N_1})^4 \right) \right] \left\{ \frac{1}{d} \mathbb{E} \left[ \text{tr} \left( (I - \alpha \hat{Q}_{\tau,N_1})^2 \right) \right] \right\}^{-2}.
\]

**Proof:** We reuse the permutation trick to derive \( \mathbb{E} \left[ (\mathbb{W}_{\tau,N}^{\text{ma}})^2 \right] \) below.

\[
\mathbb{E} \left[ (\mathbb{W}_{\tau,N}^{\text{ma}})^2 \right] = \mathbb{E} \left[ (I - \alpha \hat{Q}_{\tau,N_1}) \hat{Q}_{\tau,N_2} (I - \alpha \hat{Q}_{\tau,N_1})^2 \hat{Q}_{\tau,N_2} (I - \alpha \hat{Q}_{\tau,N_1}) \right] \tag{105}
\]

We know that \( \mathbb{E} \left[ (\mathbb{W}_{\tau,N}^{\text{ma}})^2 \right] \) is equal to a scale factor times \( I_d \), the identity matrix. And the scale factor can be derived below

\[
\text{tr} \mathbb{E} \left[ (\mathbb{W}_{\tau,N}^{\text{ma}})^2 \right] = \text{tr} \mathbb{E} \left[ \hat{Q}_{\tau,N_2} (I - \alpha \hat{Q}_{\tau,N_1})^2 \hat{Q}_{\tau,N_2} (I - \alpha \hat{Q}_{\tau,N_1})^2 \right] \\
= \frac{1}{N_2^2} \text{tr} \mathbb{E} \left[ \mathbf{X}_{\tau,N_2}^{\text{val}} (I - \alpha \hat{Q}_{\tau,N_1})^2 \mathbf{X}_{\tau,N_2}^{\text{val}} (I - \alpha \hat{Q}_{\tau,N_1})^2 \right] \\
= \frac{1}{N_2^2} \text{tr} \mathbb{E} \left[ \sum_{i,j=1}^{N_2} \text{Diag} \left( (1 - \alpha \lambda^{(N_1)}_i)^2, \ldots, (1 - \alpha \lambda^{(N_1)}_d)^2 \right) \mathbf{v}_i \mathbf{v}_j^\top \right] \\
= \frac{1}{N_2^2} \mathbb{E} \left[ \sum_{i=1}^{N_2} \text{tr} \left( \text{Diag} \left( (1 - \alpha \lambda^{(N_1)}_i)^2, \ldots, (1 - \alpha \lambda^{(N_1)}_d)^2 \right) \mathbf{v}_i \mathbf{v}_i^\top \right)^2 \right. \\
\left. + \sum_{i \neq j} \left( \text{Diag} \left( (1 - \alpha \lambda^{(N_1)}_i)^2, \ldots, (1 - \alpha \lambda^{(N_1)}_d)^2 \right) \mathbf{v}_i \mathbf{v}_j^\top \right)^2 \right] \\
= \frac{1}{N_2 d} \mathbb{E} \left[ \text{tr}^2 \left( \text{Diag} \left( (1 - \alpha \lambda^{(N_1)}_1)^2, \ldots, (1 - \alpha \lambda^{(N_1)}_d)^2 \right) \right) + 2 \left\| \text{Diag} \left( (1 - \alpha \lambda^{(N_1)}_1)^2, \ldots, (1 - \alpha \lambda^{(N_1)}_d)^2 \right) \right\|_F^2 \\ + (N_2 - 1) \left\| \text{Diag} \left( (1 - \alpha \lambda^{(N_1)}_1)^2, \ldots, (1 - \alpha \lambda^{(N_1)}_d)^2 \right) \right\|_F^2 \right] \tag{106}
\]

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

\[
C_0^{\text{ma}} := \frac{1}{d} \mathbb{E} \left[ \mathbb{E} \left[ \mathbb{E} \left[ \left( \mathbb{W}_{\tau,N}^{\text{ma}} \right)^2 \right] \right] \right] \\
= \frac{1}{d N_2} \mathbb{E} \left[ \left( \sum_{i=1}^{d} (1 - \alpha \lambda^{(N_1)}_i)^2 \right)^2 + (N_2 + 1) \left( \sum_{i=1}^{d} (1 - \alpha \lambda^{(N_1)}_i)^4 \right) \right] \left\{ \frac{1}{d} \mathbb{E} \left[ \sum_{i=1}^{d} (1 - \alpha \lambda^{(N_1)}_i)^2 \right] \right\}^{-2} \\
= \frac{1}{d N_2} \mathbb{E} \left[ \text{tr}^2 \left( (I - \alpha \hat{Q}_{\tau,N_1})^2 \right) + (N_2 + 1) \text{tr} \left( (I - \alpha \hat{Q}_{\tau,N_1})^4 \right) \right] \left\{ \frac{1}{d} \mathbb{E} \left[ \text{tr} \left( (I - \alpha \hat{Q}_{\tau,N_1})^2 \right) \right] \right\}^{-2}. \tag{107}
\]

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

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

\[
\hat{\theta}_0^{er} - \theta_0^{er} = \left( \sum_{\tau=1}^{T} \hat{\mathbf{W}}_{\tau}^{er} \right)^{-1} \left( \sum_{\tau=1}^{T} \hat{\mathbf{W}}_{\tau}^{er} (\hat{\theta}_0^{gt} - \theta_0^{er}) \right) + \left( \sum_{\tau=1}^{T} \hat{\mathbf{W}}_{\tau}^{er} \right)^{-1} \left( \sum_{\tau=1}^{T} \frac{1}{N} \mathbf{X}_{\tau,N}^{all} \mathbf{e}_{\tau,N}^{all} \right). \tag{108}
\]
To bound the statistical error, we define

\[ Z^{all}_{e,er} := [e^{allT}_{1}, \ldots, e^{allT}_{T}]^T \in \mathbb{R}^{NT}, \quad (109a) \]

\[ U^{\prime}_{e,er} := \frac{1}{N} [X^{allT}_{1,N}, \ldots, X^{allT}_{T,N}]^T \left( \sum_{\tau=1}^{T} \hat{W}_{\tau}^{er} \right)^{-1} \in \mathbb{R}^{NT \times d}. \quad (109b) \]

Under Assumptions 1-2, with \( U_{er}, z_{er} \) defined in Lemma 7, the ERM statistical error is given by

\[ E_{er}^{2}(\theta_{0}^{er}) = \| \hat{\theta}_{0}^{er} - \theta_{0}^{er} \|_{(E_{\tau}[W^{er}_{\tau}])}^2 \leq 2 \left( \frac{\| U^{\prime}_{er} z_{er} \|_{(E_{\tau}[W^{er}_{\tau}])}^2}{I_{1}^{er}} + \frac{\| U^{\prime}_{e,er} z^{all}_{e,er} \|_{(E_{\tau}[W^{er}_{\tau}])}^2}{I_{2}} \right). \]

We will then bound terms \( I_{1}^{er}, I_{2} \) respectively. First the bound for the term \( I_{1}^{er} \) in (109) is provided in Lemma 7 which states that with probability at least \( 1 - Td^{-10} \), we have

\[ I_{1}^{er} \leq \frac{R^2}{T} \left( \lambda_{\min}(E_{\tau}[W^{er}_{\tau}])^{-1} \lambda_{\max}(E_{\tau}[(W^{er}_{\tau})^2]) + \tilde{O}(\frac{d}{N}) + \tilde{O}(\frac{1}{\sqrt{d}}) + \tilde{O}(\frac{\sqrt{d}}{T}) \right) \]

\[ + \left( \tilde{O}(\frac{\sqrt{d}}{T}) \right) M^2. \]

Following similar arguments from Lemma 7 for term \( I_{2} \), first

\[ |z^{all}_{e,er} U^{\prime}_{e,er} E_{\tau}[W^{er}_{\tau}] U^{T}_{e,er} z^{all}_{e,er} - E_{\theta_{0}^{er},e,er} W^{er}_{\tau} U^{T}_{e,er} z^{all}_{e,er} \mid \leq \tilde{O} \left( \| U^{\prime}_{e,er} E_{\tau}[W^{er}_{\tau}] U^{T}_{e,er} z^{all}_{e,er} \|_{F} \right) \]

\[ = \tilde{O} \left( \frac{1}{T} \| E_{\tau}[W^{er}_{\tau}] (\sum_{\tau=1}^{T} \hat{W}_{\tau}^{er})^{-2} (\sum_{\tau=1}^{T} \hat{W}_{\tau}^{er}) \|_{F} \right) \]

\[ = \tilde{O} \left( \frac{1}{T N} \| E_{\tau}[W^{er}_{\tau}] (\sum_{\tau=1}^{T} \hat{W}_{\tau}^{er})^{-2} (\sum_{\tau=1}^{T} \hat{W}_{\tau}^{er}) \|_{F} \right) = \tilde{O} \left( \frac{\sqrt{d}}{T N} \right) \]

and the expectation is given by

\[ E_{\theta_{0}^{er},e,er} W^{er}_{\tau} U^{T}_{e,er} z^{all}_{e,er} = \text{tr} \left( U^{\prime}_{e,er} E_{\tau}[W^{er}_{\tau}] U^{T}_{e,er} z^{all}_{e,er} \right) \]

\[ = \frac{d}{TN} \frac{1}{d} \langle E_{\tau}[W^{er}_{\tau}] (\sum_{\tau=1}^{T} \hat{W}_{\tau}^{er})^{-2} (\sum_{\tau=1}^{T} \hat{W}_{\tau}^{er}) \rangle = \frac{d}{TN} \frac{1}{d} \text{tr} \left( E_{\tau}[W^{er}_{\tau}] (\sum_{\tau=1}^{T} \hat{W}_{\tau}^{er})^{-1} \right) \]

\[ = \frac{d}{TN} \left\{ \frac{1}{d} \text{tr} \left( E_{\tau}[W^{er}_{\tau}] (\sum_{\tau=1}^{T} \hat{W}_{\tau}^{er})^{-1} - I \right) + \frac{1}{d} \text{tr} (I) \right\} \]

\[ = C_{I_{1}}^{er} \]

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

\[ I_{2} \leq \frac{d}{TN} \left( C_{I_{1}}^{er} + \tilde{O}(\frac{\sqrt{d}}{T}) + \tilde{O}(\frac{1}{\sqrt{d}}) + \tilde{O}(\frac{d}{N}) \right). \]

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

\[ E_{er}^{2}(\theta_{0}^{er}) \leq \frac{R^2}{T} \left( 2C_{0}^{er} + \tilde{O}(\frac{\sqrt{d}}{T}) + \tilde{O}(\frac{1}{\sqrt{d}}) + \tilde{O}(\frac{d}{N}) \right) + \frac{d}{TN} \left( 2C_{I_{1}}^{er} + \tilde{O}(\frac{\sqrt{d}}{T}) + \tilde{O}(\frac{1}{\sqrt{d}}) + \tilde{O}(\frac{d}{N}) \right) \]

\[ + \left( \tilde{O}(\frac{\sqrt{d}}{T}) + \tilde{O}(\frac{d}{N}) \right) M^2. \]
MAML. From the expressions of $\hat{\theta}_0^{ma}$ and $\hat{W}_r^{ma}$, we have

$$
\hat{\theta}_0^{ma} - \theta_0^{ma} = \left( \sum_{\tau=1}^{T} \hat{W}_r^{ma} \right)^{-1} \left( \sum_{\tau=1}^{T} \hat{W}_r^{ma} (\theta_r^t - \theta_0^{ma}) \right) \\
+ \left( \sum_{\tau=1}^{T} \hat{W}_r^{ma} \right)^{-1} \left( \sum_{\tau=1}^{T} (I - \alpha Q_r,N_l) \left( \frac{1}{N_2} X_r \tau,N_2 e_{\tau,N_2} - \frac{\alpha}{N_1} Q_r,N_2 X_r \tau,N_2 e_{\tau,N_1} \right) \right).
$$

To bound the statistical error of MAML, we define

$$
z_{c1,ma}^{val} := [e_{1}^{valT}, \ldots, e_{T}^{valT}]^T \in \mathbb{R}^{N_2T},
$$

$$
U_{c1,ma}^{T} := \frac{1}{N_2} \left( \sum_{\tau=1}^{T} \hat{W}_r^{ma} \right)^{-1} \left[ (I - \alpha \bar{Q}_l,N_l) X_{l1,ma}^{valT}, \ldots, (I - \alpha \bar{Q}_l,N_l) X_{l2,ma}^{valT} \right] \in \mathbb{R}^{d \times N_2T},
$$

$$
z_{e2,ma}^{trn} := [e_{1}^{trnT}, \ldots, e_{T}^{trnT}]^T \in \mathbb{R}^{N_1T},
$$

$$
U_{e2,ma}^{T} := \frac{\alpha}{N_1} \left( \sum_{\tau=1}^{T} \hat{W}_r^{ma} \right)^{-1} \left[ (I - \alpha \bar{Q}_l,N_l) \bar{Q}_l,N_1 X_{l1,ma}^{trnT}, \ldots, (I - \alpha \bar{Q}_l,N_l) \bar{Q}_l,N_1 X_{l2,ma}^{trnT} \right] \in \mathbb{R}^{d \times N_1T}.
$$

Under Assumptions 1-2, with $U_{ma}, z_{ma}$ defined in Lemma 7, the MAML statistical error is given by

$$
\mathcal{E}_2^{ma}(\hat{\theta}_0^{ma}) \leq 2 \left( \| U_{ma} z_{ma}^{val} \|_{\mathbb{E}_{|W_{ma}}^{val}}^2 + \| U_{c1,ma}^{T} z_{c1,ma}^{val} \|_{\mathbb{E}_{r,N}^{val}} + \| U_{e2,ma}^{T} z_{e2,ma}^{trn} \|_{\mathbb{E}_{r,N}^{trn}} \right).
$$

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

$$
I_1^{ma} \leq R^2 \left( \lambda_{\min}(\mathbb{E}[W_{ma}])^{-1} \lambda_{\max}(\mathbb{E}[(W_{ma}^{val})^2]) + \tilde{O}(\frac{d}{N}) + \tilde{O}(\frac{1}{\sqrt{d}}) + \tilde{O}(\frac{d}{T}) \right)
\leq \left( \tilde{O}(\sqrt{\frac{d}{T}}) + \tilde{O}(\frac{d}{N}) \right) M^2.
$$

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

$$
\left\| U_{c1,ma}^{T} z_{c1,ma}^{val} \|_{\mathbb{E}_{r,N}^{val}}^2 - \mathbb{E}_{\theta_r^*, e_r} W_{ma} \| U_{c1,ma}^{T} z_{c1,ma}^{val} \|_{\mathbb{E}_{r,N}^{val}} \right\| = \tilde{O}(\frac{\sqrt{d}}{TN_2})
$$

and the expectation is given by

$$
\mathbb{E}_{\theta_r^*, e_r} W_{ma} \| U_{c1,ma}^{T} z_{c1,ma}^{val} \|_{\mathbb{E}_{r,N}^{val}} = \text{tr} \left( U_{c1,ma}^{T} E_r W_{ma} U_{c1,ma}^{T} z_{c1,ma}^{val} \right) = \text{tr}(U_{c1,ma}^{T} E_r W_{ma} U_{c1,ma}^{T} z_{c1,ma}^{val})
$$

$$
= \frac{d}{TN_2} \frac{1}{d} \text{tr} \left( E_r W_{ma} \right) \left( \frac{1}{T} \sum_{\tau=1}^{T} W_{r,N_{\tau}}^{-1} \right)
$$

$$
= \frac{d}{TN_2} \left\{ \frac{1}{d} \text{tr}(E_r W_{ma}) \left( \frac{1}{T} \sum_{\tau=1}^{T} W_{r,N_{\tau}}^{-1} \right) - \text{tr}^{-1}(E_r W_{ma}) \right\}.
$$

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

$$
I_2 \leq \frac{d}{TN_2} \left( c_{1,1}^{ma} + \tilde{O}(\sqrt{\frac{d}{T}}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right).$$
To bound the iMAML statistical error, define $C$ with $\Sigma$ defined by

$$C_{1,2} := \frac{1}{d} \left( E^{-1}[W^\text{ma}], \alpha^2 E(I - \alpha Q_r)Q_r, Q_r(I - \alpha Q_r) \right).$$

Following similar arguments as the derivation for MAML in (119)-(125), with probability at least $1 - Td^{-10}$, the statistical error of MAML is bounded above by

$$\mathcal{E}^2_{\text{ma}}(\hat{\theta}_{0}^\text{ma}) \leq \frac{R^2}{T} \left( 2C_{0}^\text{ma} + \tilde{O}(\sqrt{d/T}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right) + \frac{d}{T N} \left( 2C_{1}^\text{ma} + \tilde{O}(\sqrt{d/T}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right) + \left( \tilde{O}(\frac{d}{T}) + \tilde{O}(\frac{d}{N}) \right) M^2. \tag{125}$$

**iMAML.** Based on $\hat{\theta}_{0}^\text{im}$ in (14a), and $\hat{W}_{\tau}^\text{im}$ in (14d), we have

$$\hat{\theta}_{0}^\text{im} - \theta_0^\text{im} = \left( \sum_{\tau=1}^{T} \hat{W}_{\tau}^\text{im} \right)^{-1} \left( \sum_{\tau=1}^{T} \hat{W}_{\tau}^\text{im} (\theta_{\tau}^{gt} - \theta_0^\text{im}) \right) \tag{126}$$

Finally, define $C_{0}^\text{im} := (1 - s)^{-1} C_{1,1}^\text{im} + s^{-1} C_{1,2}^\text{im}$. 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

$$\mathcal{E}^2_{\text{im}}(\hat{\theta}_{0}^\text{im}) \leq \frac{R^2}{T} \left( 2C_{0}^\text{im} + \tilde{O}(\sqrt{d/T}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right) + \frac{d}{T N} \left( 2C_{1}^\text{im} + \tilde{O}(\sqrt{d/T}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right) + \left( \tilde{O}(\frac{d}{T}) + \tilde{O}(\frac{d}{N}) \right) M^2. \tag{131}$$

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

$$\mathcal{E}^2_{\text{im}}(\hat{\theta}_{0}^\text{im}) \leq \frac{R^2}{T} \left( 2C_{0}^\text{im} + \tilde{O}(\frac{d}{N}) + \tilde{O}(\sqrt{d/T}) + \tilde{O}(\sqrt{d/T}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right) + \frac{d}{T N} \left( 2C_{1}^\text{im} + \tilde{O}(\sqrt{d/T}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right) + \left( \tilde{O}(\frac{d}{T}) + \tilde{O}(\frac{d}{N}) \right) M^2. \tag{131}$$

with $C_{0}^\text{im} := (1 - s)^{-1} C_{1,1}^\text{im} + s^{-1} C_{1,2}^\text{im}$, and

$$C_{1,1}^\text{im} := 1,$$

$$C_{1,2}^\text{im} := \frac{1}{d} \left( E[W_{\tau}^\text{im}]^{-1}, \frac{1}{T} \sum_{\tau=1}^{T} (\gamma)^{-2} E[W_{\tau}^\text{im}] E[\Sigma_{\theta_r}] E[W_{\tau}^\text{im}] - E[\gamma^{-1}(\hat{W}_{\tau}^\text{im})^2] \right). \tag{132}$$
We will then bound the redefined terms with $\Sigma$.

To bound the statistical error of BaMAML, define $\tilde{\theta}_0 = (\sum_{\tau=1}^{T} \tilde{W}_{\tau,N})^{-1} (\sum_{\tau=1}^{T} \tilde{W}_{\tau,N} (\theta_t - \hat{\theta}_0))$

For $\tilde{\theta}_0$, derived by $\tilde{W}_{\tau,N} = \sum_{\tau=1}^{T} X_{\tau,N} - X_{\tau,N} e^{\tau}_{\tau,N} - X_{\tau,N} \Sigma^{-1} e^{\tau}_{\tau,N}$, (133)

where $\Sigma^{-1} = (I_N + \gamma^{-1} X_{\tau,N} X_{\tau,N}^T)^{-1}$.

To bound the statistical error of BaMAML, define

$z_{e,ba} := [e_{\text{all}}^T, \ldots, e_T^T] \in \mathbb{R}^{NT}$,

$U_{e,1,ba} := \frac{1}{N_2} (\sum_{\tau=1}^{T} \tilde{W}_{\tau,N})^{-1} [X_{\tau,N} \Sigma^{-1} N, \ldots, X_{\tau,N} \Sigma^{-1} N] \in \mathbb{R}^{d \times NT}$,

$z_{e,2,ba} := [e_{\text{trn}}^T, \ldots, e_T^T] \in \mathbb{R}^{NT}$,

$U_{e,2,ba} := \frac{1}{N_2} (\sum_{\tau=1}^{T} \tilde{W}_{\tau,N})^{-1} [X_{\tau,N} \Sigma^{-1} N, \ldots, X_{\tau,N} \Sigma^{-1} N] \in \mathbb{R}^{d \times NT}$.

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

$E_2(\hat{\theta}_0) \leq \frac{R^2}{T} \left( 2C_{1,ba}^0 + \tilde{O}(\frac{d}{N}) + \tilde{O}(\sqrt{\frac{d}{T}}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right)$

with $C_{1,ba}^0$ derived by

$C_{1,ba}^0 = 1 \geq \frac{1}{d} \mathbb{E}[(W_{\tau,N})^{-1}, (1-s)^{-1} \mathbb{E}[(I_d + (\gamma s)^{-1} Q) \tau^{-1} Q_d (I_d + (\gamma s)^{-1} Q) 

- s(I_d + (\gamma^{-1} Q) \tau^{-1} Q_d (I_d + (\gamma^{-1} Q) 

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

ERM. Then under Assumptions 1,3, with $U_{er}, z_{er}$ defined in Lemma 7, we have

$E_2(\hat{\theta}_0) = w_{er}\|\hat{\theta}_0^\tau - \theta_0^\tau\|^2 = w_{er}(\|U_{er} z_{er}^T\|^2 + \|U_{er} z_{er}^T\|^2 + 2z_{er}^T U_{er} U_{er}^{-1} z_{er}^T)$.

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

$I_1 = z_{er}^T U_{er} U_{er}^T z_{er} \leq \frac{R^2}{T} \left( \tilde{O}(\frac{1}{\sqrt{d}}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right)$.

For $\hat{\theta}_0$ since $\mathbb{E}[\hat{W}_{\tau}^T] = Q_d$, and by Lemma 8, we have

$C_{er}^0 = \frac{1}{d} \mathbb{E}[(Q_d^2)^2] \left( \frac{1}{d} \mathbb{E}[\text{tr}(Q_d)] \right)^{-2}$

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

$|z_{e,all}^T U_{e,all} U_{e,all}^T z_{e,all}^T - E_{\theta_t, e, \tau} [\hat{W}_{\tau} z_{e,all} U_{e,all} U_{e,all}^T z_{e,all}]| = \tilde{O}(\frac{\sqrt{d}}{TN}$)
\[ E_{\theta^*_{w}, e, r} [z_{e, r}^{\text{all}} U_{e, r} U_{e, r}^T z_{e, r}^{\text{all}}] = \text{tr} \left( U_{e, r} U_{e, r}^T \right) = \frac{d}{TN} \left( \frac{1}{T} \sum_{\tau=1}^{T} \tilde{W}_{\tau}^r \right)^{-1} \left( \frac{1}{T} \sum_{\tau=1}^{T} \tilde{W}_{\tau}^r \right) \]
\[ = \frac{d}{TN} \text{tr} \left( \left( \frac{1}{T} \sum_{\tau=1}^{T} \tilde{W}_{\tau}^r \right)^{-1} \right) = \frac{d}{TN} \left( \frac{1}{d} \text{tr} \left( \left( \frac{1}{T} \sum_{\tau=1}^{T} \tilde{W}_{\tau}^r \right)^{-1} - \mathbb{E}^{-1} [\tilde{W}_{\tau}^r] \right) \right) + \frac{1}{d} \text{tr} \left( \mathbb{E}^{-1} [\tilde{W}_{\tau}^r] \right). \] (144)

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

\[ I_2 = z_{e, r}^{\text{all}} U_{e, r} U_{e, r}^T z_{e, r}^{\text{all}} \leq \frac{d}{TN} \left( \hat{C}_0^* + \tilde{O}(\sqrt{\frac{d}{T}}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right). \] (145)

For term \( I_3 \), note that \( E_{\theta^*_{w}, e, r} [z_{e, r}^{\text{all}} U_{e, r} U_{e, r}^T z_{e, r}^{\text{all}}] = 0 \). Following a similar argument from (143) to (145), with probability at least \( 1 - \delta \), \( |I_3| = |z_{e, r}^{\text{all}} U_{e, r} U_{e, r}^T z_{e, r}^{\text{all}}| \leq \tilde{O}(\frac{R}{T \sqrt{N}}) \). Finally, by combining (141) - (145), and applying the weight \( w_{e, r} \), we conclude that with probability at least \( 1 - T d^{-10} \), the statistical error of ERM is bounded by

\[ E_{\theta^*_{w}}^2 (\hat{\theta}_0^*) = w_{e, r} \| \theta^*_{0} (\gamma) - \theta^*_{0} (\gamma) \|_2^2 = \frac{R^2}{T} \left( w_{e, r} \hat{C}_0^* + \tilde{O}(\sqrt{\frac{d}{T}}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right) \]
\[ + \frac{d}{TN} \left( w_{e, r} \hat{C}_0^* + \tilde{O}(\sqrt{\frac{d}{T}}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right) + \tilde{O}(\frac{R}{T \sqrt{N}}). \] (146)

**MAML.** Under Assumptions 1,3 and \( U_{ma}, z_{ma} \) defined in Lemma 7, we have

\[ E_{\theta^*_{w}}^2 = w_{ma} \| \theta^*_{0} - \theta^*_{0} \|_2^2 = w_{ma} \left( \| U_{ma} z_{ma} \|_2^2 + \| U_{e1, ma} z_{e1, ma} \|_2^2 + \| U_{e2, ma} z_{e2, ma} \|_2^2 \right) \]
\[ + 2 z_{ma}^T U_{ma} U_{e1, ma} z_{e1, ma} - 2 z_{ma}^T U_{ma} U_{e2, ma} z_{e2, ma} - 2 z_{ma}^T U_{e1, ma} U_{e2, ma} z_{e2, ma}. \] (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 \( 1 - T d^{-10} \)

\[ I_1 = z_{e, r}^T U_{e, r} U_{e, r}^T z_{e, r} \leq \frac{R^2}{T} \left( \hat{C}_0^* + \tilde{O}(\frac{1}{\sqrt{d}}) + \tilde{O}(\sqrt{\frac{d}{T}}) \right). \] (148)

For \( \hat{C}_0^* \), by Lemma 9

\[ \hat{C}_0^* = \frac{1}{d N^2} \mathbb{E} \left[ \text{tr}^2 \left( I - \alpha Q_{r, N} \right)^2 \right] + (N^2 + 1) \text{tr} \left( (I - \alpha Q_{r, N} \alpha)^4 \right) \] \[ \left[ \frac{1}{d} \mathbb{E} \left[ \text{tr} \left( (I - \alpha Q_{r, N} \alpha)^2 \right) \right] \right]^{-2}. \] (149)

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

\[ |z_{e1, ma}^T U_{e1, ma} z_{e1, ma} - E_{\theta^*_{w}, e, r} [z_{e1, ma}^T U_{e1, ma} U_{e1, ma}^T z_{e1, ma}^{\text{all}}]| = \tilde{O}(\frac{\sqrt{d}}{T N^2}). \] (150)

and the expectation is given by

\[ E_{\theta^*_{w}, e, r} [z_{e1, ma}^T U_{e1, ma} U_{e1, ma}^T z_{e1, ma}^{\text{all}}] = \text{tr} \left( U_{e1, ma} U_{e1, ma}^T \right) = \frac{d}{TN} \text{tr} \left( \left( \frac{1}{T} \sum_{t=1}^{T} W_{\tau, N} \right)^{-1} \right) \]
\[ = \frac{d}{TN} \left( \frac{1}{d} \text{tr} \left( \left( \frac{1}{T} \sum_{t=1}^{T} W_{\tau, N} \right)^{-1} - \mathbb{E}^{-1} [W_{\tau, N}] \right) \right) + \frac{1}{d} \text{tr} \left( \mathbb{E}^{-1} [W_{\tau, N}] \right). \] (151)
And by combining (150) and (151), with probability at least $1 - \delta$, we have

$$I_2 = z_{e1,ma}^T U_{e1,ma} U_{e1,ma}^T z_{e1,ma} = \frac{d}{TN_2} \left( \tilde{C}_{1,1}^{ma} + \tilde{O}(\sqrt{\frac{d}{T}}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right). \quad (152)$$

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

$$\|U_{e2,ma}^T z_{e2,ma}^{trn}\|^2 - \mathbb{E}_{\theta^{1,2}_{\gamma}, \gamma} \|U_{e2,ma}^T z_{e2,ma}^{trn}\| = \tilde{O} \left( \frac{\sqrt{d}}{TN_1} \right) \quad (153)$$

and the expectation is given by

$$\mathbb{E}_{\theta^{1,2}_{\gamma}, \gamma} \|U_{e2,ma}^T z_{e2,ma}^{trn}\| = \frac{d}{TN_1} \text{tr} \left( \left( \frac{1}{T} \sum_{\tau=1}^T W_{\tau,N}^{ma} \right)^2 - \mathbb{E} \left\{ \left( \frac{1}{T} \sum_{\tau=1}^T \alpha^2 (I - \alpha \hat{Q}_{r,N_1}) \hat{Q}_{r,N_2} \hat{Q}_{r,N_1} (I - \alpha \hat{Q}_{r,N_1}) \right) \right\} \right) \quad (154)$$

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

$$I_3 = z_{e2,ma}^T U_{e2,ma} U_{e2,ma}^T z_{e2,ma}^{trn} = \frac{d}{TN_1} \left( \tilde{C}_{1,2}^{ma} + \tilde{O}(\sqrt{\frac{d}{T}}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right). \quad (155)$$

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

$$|I_4| \leq \tilde{O} \left( \frac{R}{T \sqrt{N_2}} \right), |I_5| \leq \tilde{O} \left( \frac{R}{T \sqrt{N_1}} \right), |I_6| \leq \tilde{O} \left( \frac{\sqrt{d}}{T \sqrt{N_1N_2}} \right). \quad (156)$$

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

$$\mathcal{L}_{ma}^2(\hat{\theta}_0^{ma}) = w_{ma}^2 ||\hat{\theta}_0^{ma}(\gamma) - \theta_0^{ma}(\gamma)||_2^2 = \frac{R^2}{T} \left( w_{ma} \tilde{C}_{1,1}^{ma} + \tilde{O}(\sqrt{\frac{d}{T}}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right)$$

$$+ \frac{d}{TN} \left( w_{ma} \tilde{C}_{1,2}^{ma} + \tilde{O}(\sqrt{\frac{d}{T}}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right) + \tilde{O} \left( \frac{R}{T \sqrt{N}} \right). \quad (157)$$

**iMAML.** With $U_{im}, z_{im}$ defined in Lemma 7 we can rewrite (126) as

$$\hat{\theta}_0^{im} - \theta_0^{im} = U_{im}^T z_{im} + U_{e1,im}^T z_{e1,im} - U_{e2,im}^T z_{e2,im}. \quad (158)$$

Thus the squared error can be computed by

$$||\hat{\theta}_0^{im} - \theta_0^{im}||_2^2 = ||U_{im}^T z_{im}||^2 + ||U_{e1,im}^T z_{e1,im}||^2 + ||U_{e2,im}^T z_{e2,im}||^2$$

$$+ 2 z_{im}^T U_{im}^T z_{e1,im} - 2 z_{im}^T U_{im}^T z_{e2,im} - 2 z_{e1,im}^T U_{e1,im} U_{e2,im}^T z_{e2,im}. \quad (159)$$
To bound term $I_1$ in (159), from Lemma 7 we have with probability at least $1 - Td^{-10}$

$$I_1 = z_{e1,im}^T U_{e1,im} U_{e1,im}^T z_{e1,im} \leq \frac{R^2}{T} (\tilde{C}_0^{im} + \tilde{O}(\frac{1}{\sqrt{d}}) + \tilde{O}(\frac{1}{\sqrt{T}})).$$ (160)

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

$$\tilde{C}_0^{im} = \frac{1}{dN_2} E \left[ \left( \gamma^2 \left( Q_{r,N_1} + \gamma I_d \right)^{-2} \right)^2 + \left( N_2 + 1 \right) \text{tr} \left( \gamma^4 \left( Q_{r,N_1} + \gamma I_d \right)^{-4} \right) \right]$$

(161)

For $I_2$, first from Lemma 3

$$\left| \varepsilon_{e1,im}^T U_{e1,im} U_{e1,im}^T z_{e1,im}^T - E_{\theta^T_x,y,x} [ \varepsilon_{e1,im}^T U_{e1,im} U_{e1,im}^T z_{e1,im}^T ] \right| = \tilde{O}(\sqrt{dN_2})$$ (162)

$$\mathbb{E}_{\theta^T_x,y,x} [ \varepsilon_{e1,im}^T U_{e1,im} U_{e1,im}^T z_{e1,im}^T ] = \frac{1}{T} \text{tr} \left( \frac{1}{T} \sum_{\tau=1}^T \tilde{W}_\tau^{im} \right)^{-1}$$

(163)

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

$$|I_2| \leq \tilde{O}(\sqrt{d/T}).$$ (164)

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

$$I_2 \leq \frac{d}{T} \tilde{C}_{1,1}^{im} + \tilde{O}(\sqrt{d/T}) + \tilde{O}(\frac{1}{\sqrt{d}})$$

(165)

$$\tilde{C}_{1,1}^{im} := \frac{1}{d} \text{tr} \left( \tilde{E}^{-1}[\tilde{W}_{\tau}^{im}] \right).$$ (166)

Similarly, for $I_3$, based on Lemma 3 we have

$$\| U_{e2,im}^T z_{e2,im}^{trn} \|^2 - E_{\varepsilon_{e1,im}}[\| U_{e2,im}^T z_{e2,im}^{trn} \|^2] = \tilde{O}(\frac{\sqrt{d}}{T N_2}).$$ (167)

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

$$\mathbb{E}_{\varepsilon_{e1,im}} [\| U_{e2,im}^{trn} z_{e2,im}^{trn} \|^2] = \frac{1}{T N_1} \left( \frac{1}{T} \sum_{\tau=1}^T \tilde{W}_\tau^{im} \right)^{-2} \left( \frac{1}{T} \sum_{\tau=1}^T (\gamma)^{-2} \tilde{W}_\tau^{im} \tilde{X}_{\tau}^{trn} \tilde{X}_{\tau}^{trn} \tilde{W}_\tau^{im} \right)$$

(168)

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

$$I_3 = z_{e2,im}^{trn} U_{e2,im}^{trn} U_{e2,im}^{trn} z_{e2,im}^{trn} \leq \frac{d}{T N_2} \left( \tilde{C}_{1,2}^{im} + \tilde{O}(\frac{1}{\sqrt{d}}) + \tilde{O}(\frac{1}{\sqrt{T}}) \right)$$

(169)

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

$$|I_4| \leq \tilde{O}(\frac{R}{T \sqrt{N_2}}), |I_5| \leq \tilde{O}(\frac{R}{T \sqrt{N_1}}), |I_6| \leq \tilde{O}(\frac{\sqrt{d}}{T \sqrt{N_1 N_2}}).$$ (170)
Finally, define $\tilde{C}^0_{1,1} := (1 - s)^{-1}\tilde{C}^0_{1,1} + s^{-1}\tilde{C}^0_{1,2}$, applying the weight $w_{im}$ we have with probability at least $1 - Td^{-10}$, the statistical error of iMAML is bounded by

$$\mathcal{E}^2_{im}(\hat{\theta}^0_{im}) = w_{im}\|(\hat{\theta}^0_{im}(\gamma) - \theta^0_{im}(\gamma))\|^2 = \frac{R^2}{T}\left(w_{im}\tilde{C}^0_{1,1} + \tilde{O}(\sqrt{d}) + \tilde{O}\left(\frac{1}{\sqrt{d}}\right)\right)$$

$$+ \frac{d}{TN}\left(w_{im}\tilde{C}^0_{1,1} + \tilde{O}(\frac{1}{\sqrt{d}}) + \tilde{O}(\sqrt{\frac{d}{T}}) + \tilde{O}\left(\frac{R}{T\sqrt{N}}\right)\right).$$

(171)

**BaMAML.** With $U_{ba}, z_{ba}$ defined in Lemma [7]

$$\theta^0_{ba} - \theta^0_{0} = U_{ba}^T z_{ba} + U_{e1,ba}^{T} z_{e1,ba} - U_{e2,ba}^{T} z_{e2,ba}$$

$$\|\theta^0 - \theta^0_{ba}\|^2 = \left\|U_{ba}^T z_{ba}\right\|^2 + \left\|U_{e1,ba}^{T} z_{e1,ba}\right\|^2 + \left\|U_{e2,ba}^{T} z_{e2,ba}\right\|^2$$

$$+ 2 z_{ba} U_{ba}^T I_{e1,ba} U_{e1,ba}^{T} z_{e1,ba} - 2 z_{ba} U_{ba}^T I_{e2,ba} U_{e2,ba}^{T} z_{e2,ba} - 2 z_{e1,ba}^T U_{e1,ba}^T U_{e2,ba}^{T} z_{e2,ba}.\quad (172)$$

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

$$I_1 = z_{ba} U_{ba} U_{ba}^T z_{ba} \leq \frac{R^2}{T}\left(C^0_{ba} + \tilde{O}(\frac{1}{\sqrt{d}}) + \tilde{O}(\sqrt{\frac{d}{T}})\right).$$

(174)

To compute $\tilde{C}^0_{ba}$, by Lemma [8]

$$\tilde{C}^0_{ba} = \frac{1}{d}\mathbb{E}\left[\text{tr}\left((W_{r,N}^{ba})^2\right)\right] \cdot \left\{\frac{1}{d}\mathbb{E}\left[\text{tr}(\tilde{W}_{r,N}^{ba})\right]\right\}^{-2}$$

(175)

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

$$||U_{e,ba}^{T} z_{ba}^\text{all}||^2 - \mathbb{E}_{\theta^0_{e,ba}, \tilde{z}_{e,ba}}[||U_{e,ba}^{T} z_{ba}^\text{all}||^2] = \tilde{O}(\frac{\sqrt{d}}{TN})$$

(176)

where the expectation is given by

$$\mathbb{E}_{\theta^0_{e,ba}, \tilde{z}_{e,ba}}[\tilde{W}_{r,N}^{ba}] = \text{tr}\left(U_{e1,ba} U_{e1,ba}^T z_{e1,ba}\right)$$

$$= \frac{1}{TN} \left(\frac{1}{d} \mathbb{E}[\tilde{W}_{r,N}^{ba}]\right)^{-2}, (1 - s)^{-1}\left(\frac{1}{d} \mathbb{E}[\tilde{W}_{r,N}^{ba}]\right)^{-2} \leq \frac{d}{TN} \left(\tilde{O}(\frac{\sqrt{d}}{T}) + \frac{1}{d} \mathbb{E}[\tilde{W}_{r,N}^{ba}]\right)^{-2}, (1 - s)^{-1}\mathbb{E}[\tilde{W}_{r,N}^{ba}]^{-1} \mathbb{E}[\tilde{Q}_{r,N}^{-1} \tilde{Q}_{r,N}^{-1} \tilde{Q}_{r,N}^{-1}]^{-1}\right)$$

$$= \tilde{C}^0_{1,1}.$$ (177)

Similarly,

$$\mathbb{E}_{\theta^0_{e,ba}, \tilde{z}_{e,ba}}[\tilde{W}_{r,N}^{ba}] = \text{tr}\left(U_{e2,ba} U_{e2,ba}^T z_{e2,ba}\right)$$

$$\leq \frac{d}{TN} \left(\tilde{O}(\frac{\sqrt{d}}{T}) + \frac{1}{d} \mathbb{E}[\tilde{W}_{r,N}^{ba}]^{-2}, s(1 - s)^{-1}\mathbb{E}[\tilde{Q}_{r,N}^{-1} \tilde{Q}_{r,N}^{-1} \tilde{Q}_{r,N}^{-1}]^{-1}\right)$$

$$= \tilde{C}^0_{1,2}.$$ (178)

$\tilde{C}^0_{1,1} - \tilde{C}^0_{1,2}$, combining the above derivations with Lemma [2] gives the higher order terms in (177), which leads to

$$\mathbb{E}_{\theta^0_{e,ba}, \tilde{z}_{e,ba}}[\tilde{W}_{r,N}^{ba}] = \text{tr}\left(U_{e2,ba} U_{e2,ba}^T z_{e2,ba}\right)$$

$$\leq \frac{d}{TN} \left(\tilde{O}(\frac{\sqrt{d}}{T}) + \frac{1}{d} \mathbb{E}[\tilde{W}_{r,N}^{ba}]^{-2}, s(1 - s)^{-1}\mathbb{E}[\tilde{Q}_{r,N}^{-1} \tilde{Q}_{r,N}^{-1} \tilde{Q}_{r,N}^{-1}]^{-1}\right)$$

$$= \tilde{C}^0_{1,2}.$$ (179)
Combining (176) and (179), with probability at least $1 - Td^{-10}$, we have

$$I_2 = z_{e, ba}^\top U_{e, ba} U_{e, ba}^\top z_{e, ba} \leq \frac{d}{TN} \left( \tilde{C}^\text{ba} + \tilde{O}(\frac{1}{\sqrt{d}}) + \tilde{O}(\frac{d}{T}) \right)$$  \hspace{1cm} (180)

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

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

$$\epsilon^2_{ba}(\hat{\theta}_{ba}^0) = w_{ba} \| \hat{\theta}_{ba}^0 - \theta_{ba}^0 \|^2 = \frac{R^2}{T} \left( w_{ba} \tilde{C}^\text{ba}_0 + \tilde{O}(\sqrt{\frac{d}{T}}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right)$$

$$+ \frac{d}{TN} \left( w_{ba} \tilde{C}^\text{ba}_1 + \tilde{O}(\frac{d}{T}) + \tilde{O}(\frac{1}{\sqrt{d}}) \right) + \tilde{O}(\frac{R}{T\sqrt{N}}).$$  \hspace{1cm} (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 $\tilde{C}^\text{er}$ satisfies

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

**Proof:** Recall that $\tilde{C}^\text{er}_0 = \frac{1}{N} \mathbb{E} \left[ \text{tr}(Q_{r,N}^2) \right] \left( \frac{1}{5} \mathbb{E} \left[ \text{tr}(Q_r) \right] \right)^{-2}$. Based on Assumption 3 $\mathbb{E} \left[ \text{tr}(Q_r) \right] = \mathbb{E} \left[ \text{tr}(I_d) \right] = d$, and $\mathbb{E}(\tilde{Q}^2_{r,N})$ can be derived by

$$\mathbb{E}[\tilde{Q}^2_{r,N}] = \mathbb{E}\left[ \frac{1}{N} X_{r,N} X_{r,N}^\top \right] = \mathbb{E}\left[ \frac{1}{N} X_{r,N}^\top X_{r,N} \right] = \mathbb{E}\left[ \frac{1}{N} \sum_i x_{r,i} x_{r,i}^\top \right]$$

$$= \frac{1}{N} \mathbb{E}\left[ (x_{r,i} x_{r,i}^\top)^2 \right] + \frac{N-1}{N} I_d$$

where $\text{tr}(\mathbb{E}[x_{r,i} x_{r,i}^\top]) = \mathbb{E}[\sum_j x_{r,i,j}^2] = d(d + 2)$. Therefore

$$\tilde{C}^\text{er}_0 = \frac{d + N + 1}{N}, \quad \lim_{d,N \to \infty \atop d/N \to \eta} \tilde{C}^\text{er}_0 = 1 + \eta. \quad \square$$

#### Theorem 10 (Asymptotic MAML constant)

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

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

**Proof:** We first derive a lower bound for $\inf_{\alpha > 0 \atop s \in (0, 1)} \lim_{d, N \to \infty \atop d/N \to \eta} \tilde{C}^\text{ma}_0$

$$\geq \lim_{d, N \to \infty \atop d/N \to \eta} \inf_{\alpha > 0 \atop s \in (0, 1)} \frac{1}{d^2} \mathbb{E} \left[ \left( \sum_{i=1}^d (1 - \alpha \lambda_i(N_1))^2 \right)^2 + (N_2 + 1) \left( \sum_{i=1}^d (1 - \alpha \lambda_i(N_1))^2 \right) \right]$$

$$\geq \lim_{d, N \to \infty \atop d/N \to \eta} \inf_{\alpha > 0 \atop s \in (0, 1)} \frac{d + N_2 + 1}{N_2} = 1 + \eta. \quad (182)$$
Next we derive the upper bound for $\inf_{s \in (0, 1)} \lim_{d/N \to \eta} \hat{C}^{\text{ma}}_0$. As for any PD matrix $M \in \mathbb{R}^{d \times d}$, $\frac{1}{d} \text{tr}(M^2) \geq \left(\frac{d}{2} \text{tr}(M)\right)^2$, then applying this inequality we obtain

$$
\lim_{d/N \to \eta} \hat{C}^{\text{ma}}_0 \leq \lim_{d/N \to \eta} \frac{\frac{1}{N_2}(d + N_2 + 1)}{2 \mathbb{E}^2 \left[ \sum_{i=1}^{d} (1 - \alpha \lambda_i^{(N_1)})^2 \right]} \leq \lim_{d/N \to \eta} \frac{\frac{1}{N_2}(d + N_2 + 1)}{(1 - \alpha)^2}.
$$

Therefore

$$
\inf_{\alpha > 0} \lim_{d/N \to \eta} \hat{C}^{\text{ma}}_0 \leq \inf_{s \in (0, 1)} \lim_{d/N \to \eta} \frac{d + N_2 + 1}{N_2} = 1 + \eta
$$

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

$$
\inf_{\alpha > 0} \lim_{d/N \to \eta} \hat{C}^{\text{ma}}_0 = 1 + \eta
$$

**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, $\hat{C}^{\text{im}}_0$, by tuning the regularization $\gamma \in (0, \infty)$ and the train-val split ratio $s \in (0, 1)$, satisfies

$$
\inf_{\gamma > 0} \lim_{d/N \to \eta} \hat{C}^{\text{im}}_0 = 1 + \eta.
$$

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

$$
\inf_{\gamma > 0} \lim_{d/N \to \eta} \hat{C}^{\text{ba}}_0 = \begin{cases} 1, & \gamma \in (0, 1), \\ \eta, & \gamma \in (1, \infty). \end{cases}
$$

**Proof:** Adopt the Stieltjes transform to obtain $\lim_{d/N \to \eta} \hat{C}^{\text{ba}}_0$ as a function of $\gamma, s, \eta$, given below. For all $\omega_1, \omega_2 > 0, \eta > 0$, define

$$
s(\omega_1, \omega_2) := \lim_{d/N \to \eta} \frac{1}{d} \mathbb{E} \left[ \text{tr} \left( (\omega_1 I_d + \omega_2 \hat{Q}_N)^{-1} \right) \right]
$$

whose closed form solution is given by

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

$$
\frac{d}{d\omega_1} s(\omega_1, \omega_2) = \left((1/\omega_2 + (1 + \eta)/\omega_1)^2 - 4\eta/\omega_1^2\right)^{-\frac{1}{2}} \cdot
$$

$$
\left((1/\omega_2 + (1 + \eta)/\omega_1)(1 + \eta) - 4\eta/\omega_1(\omega_1^2 - 1 + (1 - \eta)\omega_1^{-2})/2\eta
$$

$$
\frac{d}{d\omega_1} s(\omega_1, \omega_2)\big|_{\omega_1 = 1} = -\left((1/\omega_2 + (1 + \eta)^2 - 4\eta)^{-\frac{1}{2}} ((1/\omega_2 + (1 + \eta))(1 + \eta) - 4\eta) + (1 - \eta)\right)/2\eta.
$$
Therefore
\[
\lim_{d,N \to \infty, d/N \to \eta} \frac{1}{d} \mathbb{E} \left[ \text{tr} \left( (\mathbf{I}_d + \gamma^{-1} \hat{\mathbf{Q}}_N)^{-1} \right) \right] = s(1, \gamma^{-1}) \leq 1
\]

where by L'Hospital's rule,
\[
\lim_{\gamma \to \infty} s(1, \gamma^{-1}) = \lim_{\gamma \to \infty} \frac{\sqrt{(\gamma + 1 + \eta)^2 - 4\eta} - (\gamma + 1 + \eta)}{2\eta} + 1 = \lim_{\gamma \to \infty} \frac{\sqrt{(1/\gamma + 1 + \eta/\gamma)^2 - 4\eta/\gamma^2} - (1/\gamma + 1 + \eta/\gamma)}{2\eta/\gamma} + 1 = 1
\]

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

By the derivative trick,
\[
\lim_{d,N \to \infty, d/N \to \eta} \frac{1}{d} \mathbb{E} \left[ \text{tr} \left( (\omega_1 \mathbf{I}_d + \omega_2 \hat{\mathbf{Q}}_N)^{-2} \right) \right] = \frac{d}{d\omega_1} s(\omega_1, \omega_2)
\]

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

For \( \frac{1}{d} \mathbb{E} \left[ \text{tr} \left( (\mathbf{I}_d + \gamma^{-1} \hat{\mathbf{Q}}_N)^{-1} (\mathbf{I}_d + (\gamma s)^{-1} \hat{\mathbf{Q}}_N)^{-1} \right) \right] \), first limit \( s(\gamma \to \infty, \mathbf{I}_d + \gamma^{-1} \hat{\mathbf{Q}}_N)^{-1} = \mathbf{I}_d \), therefore
\[
\lim_{\gamma \to \infty} \frac{1}{d} \mathbb{E} \left[ \text{tr} \left( (\mathbf{I}_d + \gamma^{-1} \hat{\mathbf{Q}}_N)^{-1} (\mathbf{I}_d + (\gamma s)^{-1} \hat{\mathbf{Q}}_N)^{-1} \right) \right] = \lim_{\gamma \to \infty} \frac{1}{d} \mathbb{E} \left[ \text{tr} \left( (\mathbf{I}_d + (\gamma s)^{-1} \hat{\mathbf{Q}}_N)^{-1} \right) \right]
\]

Then
\[
\inf_{\gamma > 0} \lim_{d,N \to \infty, d/N \to \eta} \tilde{C}_{0}^{ba} \leq \lim_{\gamma \to \infty} \lim_{d,N \to \infty, d/N \to \eta} \tilde{C}_{0}^{ba} = \begin{cases} 1, & \eta \in (0, 1]; \\ \eta, & \eta \in (1, \infty). \end{cases}
\]

Note \( \tilde{C}_{0}^{ba} \geq 1 \), therefore
\[
\inf_{\gamma > 0} \lim_{d,N \to \infty, d/N \to \eta} \tilde{C}_{0}^{ba} \leq \lim_{\gamma \to \infty} \lim_{d,N \to \infty, d/N \to \eta} \tilde{C}_{0}^{ba} = \begin{cases} 1, & \eta \in (0, 1]; \\ \eta, & \eta \in (1, \infty). \end{cases}
\]
Table 2: Comparison of different MAML on image classification (testing loss (NLL) with std., code modified from [Nguyen et al., 2020])

| Method  | miniImageNet | TieredImageNet |
|---------|--------------|----------------|
| MAML    | 1.41 ± 0.04  | 1.86 ± 0.06    |
| BaMAML  | 1.38 ± 0.05  | 1.51 ± 0.06    |

### B.2.5 Comparison of the dominating constants

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

\[
\lim_{d, N \to \infty} \tilde{C}_0^{\text{err}} = \inf_{\alpha \in (0, 1/\lambda)} \lim_{d/N \to \eta} \tilde{C}_0^{\text{ma}} = \inf_{\gamma > 0} \lim_{d, N \to \infty} \tilde{C}_0^{\text{im}} > \inf_{\gamma > 0} \lim_{d, N \to \infty} \tilde{C}_0^{\text{ba}}.
\]  

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

\[
\inf_{\alpha \in (0, 1/\lambda)} \lim_{d/N \to \eta} \tilde{C}_0^{\text{ma}} \geq \inf_{\gamma > 0} \lim_{d, N \to \infty} \tilde{C}_0^{\text{im}} \geq \inf_{\gamma > 0} \lim_{d, N \to \infty} \tilde{C}_0^{\text{ba}}.
\]

Combining (187) with the comparison in optimal population risk, we can conclude that, under Assumptions 1, 3, when \( \gamma \) 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 \( \gamma \) 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 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.