Journal of Machine Learning Research 23 (2022) 1-41 Submitted 7/20; Revised 5/21; Published 2/22

Theoretical Convergence of Multi-Step Model-Agnostic Meta-Learning

Kaiyi Ji JI.367@OSU.EDU
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, OH 98195-4322, USA

Junjie Yang YANG.4972@OSU.EDU
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, OH 98195-4322, USA

Yingbin Liang LIANG.889@OSU.EDU
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, OH 98195-4322, USA

Editor: Suvrit Sra

Abstract

As a popular meta-learning approach, the model-agnostic meta-learning (MAML) algorithm has been widely used due to its simplicity and effectiveness. However, the convergence of the general multi-step MAML still remains unexplored. In this paper, we develop a new theoretical framework to provide such convergence guarantee for two types of objective functions that are of interest in practice: (a) resampling case (e.g., reinforcement learning), where loss functions take the form in expectation and new data are sampled as the algorithm runs; and (b) finite-sum case (e.g., supervised learning), where loss functions take the finite-sum form with given samples. For both cases, we characterize the convergence rate and the computational complexity to attain an ϵ-accurate solution for multi-step MAML in the general nonconvex setting. In particular, our results suggest that an inner-stage stepsize needs to be chosen inversely proportional to the number N of inner-stage steps in order for N-step MAML to have guaranteed convergence. From the technical perspective, we develop novel techniques to deal with the nested structure of the meta gradient for multi-step MAML, which can be of independent interest.

Keywords: Computational complexity, convergence rate, finite-sum, meta-learning, multi-step MAML, nonconvex, resampling.

1. Introduction

Meta-learning or learning to learn (Thrun and Pratt, 2012; Naik and Mammone, 1992; Bengio et al., 1991; Schmidhuber, 1987) is a powerful tool for quickly learning new tasks by using the prior experience from related tasks. Recent works have empowered this idea with neural networks, and their proposed meta-learning algorithms have been shown to enable fast learning over unseen tasks using only a few samples by efficiently extracting the knowledge from a range of observed tasks (Santoro et al., 2016; Vinyals et al., 2016;

Finn et al., 2017a). Current meta-learning algorithms can be generally categorized into metric-learning based (Koch et al., 2015; Snell et al., 2017), model-based (Vinyals et al., 2016; Munkhdalai and Yu, 2017), and optimization-based (Finn et al., 2017a; Nichol and Schulman, 2018; Rajeswaran et al., 2019) approaches. Among them, optimization-based meta-learning is a simple and effective approach used in a wide range of domains including classification/regression (Rajeswaran et al., 2019), reinforcement learning (Finn et al., 2017a), robotics (Al-Shedivat et al., 2018), federated learning (Chen et al., 2018), and imitation learning (Finn et al., 2017b).

Model-agnostic meta-learning (MAML) (Finn et al., 2017a) is a popular optimization-based method, which is simple and compatible generally with models trained with gradient descents. MAML consists of two nested stages, where the inner stage runs a few steps of (stochastic) gradient descent for each individual task, and the outer stage updates the meta parameter over all the sampled tasks. The goal of MAML is to find a good meta initialization w^* based on the observed tasks such that for a new task, starting from this w^*, a few (stochastic) gradient steps suffice to find a good model parameter. Such an algorithm has been demonstrated to have superior empirical performance (Antoniou et al., 2019; Grant et al., 2018; Zintgraf et al., 2018; Nichol et al., 2018). Recently, the theoretical convergence of MAML has also been studied. Specifically, Finn et al. (2019) extended MAML to the online setting, and analyzed the regret for the strongly convex objective function. Fallah et al. (2020a) provided an analysis for one-step MAML for general nonconvex functions, where each inner stage takes a single stochastic gradient descent (SGD) step.

In practice, the MAML training often takes *multiple* SGD steps at the inner stage, for example in Finn et al. (2017a); Antoniou et al. (2019) for supervised learning and in Finn et al. (2017a); Fallah et al. (2020b) for reinforcement learning, in order to attain a higher test accuracy (i.e., better generalization performance) even at a price of higher computational cost. Compared to the single-step MAML, the multi-step MAML has been shown to achieve better test performance. For example, as shown in Fig. 5 of Finn et al. (2017a) and Table 2 of Antoniou et al. (2019), the test accuracy is improved as the number of inner-loop steps increases. In particular, in the original MAML work (Finn et al., 2017a), 5 inner-loop steps are taken in the training of a 20-way convolutional MAML model. In addition, some important variants of MAML also take multiple inner-loop steps, which include but not limited to ANIL (Almost No Inner Loop) (Raghu et al., 2020) and BOIL (Body Only update in Inner Loop) (Oh et al., 2021). For these reasons, it is important and meaningful to analyze the convergence of multi-step MAML, and the resulting analysis can be helpful for studying other MAML-type of variants.

However, the theoretical convergence of such *multi-step* MAML algorithms has not been established yet. In fact, several mathematical challenges will arise in the theoretical analysis if the inner stage of MAML takes multiple steps. First, the meta gradient of multi-step MAML has a nested and recursive structure, which requires the performance analysis of an optimization path over a nested structure. In addition, multi-step update also yields a complicated bias error in the Hessian estimation as well as the statistical correlation between the Hessian and gradient estimators, both of which cause further difficulty in the analysis of the meta gradient. *The main contribution of this paper lies in the development of a new theoretical framework for analyzing the general* multi-step *MAML with techniques for handling the above challenges.*

1.1 Main Contributions

We develop a new theoretical framework, under which we characterize the convergence rate and the computational complexity to attain an ϵ-accurate solution for *multi-step* MAML in the general nonconvex setting. Specifically, for the resampling case where each iteration needs sampling of fresh data (e.g., in reinforcement learning), our analysis enables to decouple the Hessian approximation error from the gradient approximation error based on a novel bound on the distance between two different inner optimization paths, which facilitates the analysis of the overall convergence of MAML. For the finite-sum case where the objective function is based on pre-assigned samples (e.g., supervised learning), we develop novel techniques to handle the difference between two losses over the training and test sets in the analysis.

Our analysis provides a guideline for choosing the inner-stage stepsize at the order of $\mathcal{O}(1/N)$ and shows that N-step MAML is guaranteed to converge with the gradient and Hessian computation complexites growing only linearly with N, which is consistent with the empirical observations in Antoniou et al. 2019. In addition, for problems where Hessians are small, e.g., most classification/regression meta-learning problems (Finn et al., 2017a), we show that the inner stepsize α can be set larger while still maintaining the convergence, which explains the empirical findings for MAML training in Finn et al. 2017a; Rajeswaran et al. 2019.

1.2 Related Work

Optimization-based meta-learning. Optimization-based meta-learning approaches have been widely used due to its simplicity and efficiency (Li et al., 2017; Ravi and Larochelle, 2016; Finn et al., 2017a). As a pioneer along this line, MAML (Finn et al., 2017a) aims to find an initialization such that gradient descent from it achieves fast adaptation. Many follow-up studies (Grant et al., 2018; Finn et al., 2019; Jerfel et al., 2018; Finn and Levine, 2018; Finn et al., 2018; Mi et al., 2019; Liu et al., 2019; Rothfuss et al., 2019; Foerster et al., 2018; Fallah et al., 2020a; Raghu et al., 2020; Collins et al., 2020) have extended MAML from different perspectives. For example, Finn et al. (2019) provided a follow-the-meta-leader extension of MAML for online learning. Alternatively to meta-initialization algorithms such as MAML, meta-regularization approaches aim to learn a good bias for a regularized empirical risk minimization problem for intra-task learning (Alquier et al., 2017; Denevi et al., 2018b,a, 2019; Rajeswaran et al., 2019; Balcan et al., 2019; Zhou et al., 2019). Balcan et al. (2019) formalized a connection between meta-initialization and meta-regularization from an online learning perspective. Zhou et al. (2019) proposed an efficient meta-learning approach based on a minibatch proximal update. Raghu et al. (2020) proposed an efficient variant of MAML named ANIL (Almost No Inner Loop) by adapting only a small subset (e.g., head) of neural network parameters in the inner loop. Ji and Liang (2021); Ji et al. (2020b) proposed efficient bilevel optimization algorithms for meta-learning with performance guarantee.

Various Hessian-free MAML algorithms have been proposed to avoid the costly computation of second-order derivatives, which include but not limited to FOMAML (Finn et al., 2017a), Reptile (Nichol and Schulman, 2018), ES-MAML (Song et al., 2020), and HF-MAML (Fallah et al., 2020a). In particular, FOMAML (Finn et al., 2017a) omits all

second-order derivatives in its meta-gradient computation, HF-MAML (Fallah et al., 2020a) estimates the meta gradient in one-step MAML using Hessian-vector product approximation. This paper focuses on the first MAML algorithms, but the techniques here can be extended to analyze the Hessian-free multi-step MAML.

Optimization theory for meta-learning. Theoretical property of MAML was initially established in Finn and Levine (2018), which showed that MAML is a universal learning algorithm approximator under certain conditions. Then *MAML-type algorithms* have been studied recently from the optimization perspective, where the convergence rate and computation complexity is typically characterized. Finn et al. (2019) analyzed online MAML for a strongly convex objective function under a bounded-gradient assumption. Fallah et al. (2020a) developed a convergence analysis for one-step MAML for a general nonconvex objective in the resampling case. Our study here provides a new convergence analysis for *multi-step* MAML in the *nonconvex* setting for both the resampling and finite-sum cases.

Since the initial version of this manuscript was posted in arXiv, there have been a few studies on multi-step MAML more recently. Wang et al. (2020b,a) studied the global optimality of MAML under the over-parameterized neural networks, while our analysis focus on general nonconvex functions. Kim et al. (2020) proposed an efficient extension of multi-step MAML by gradient reuse in the inner loop, while our analysis focuses on the most basic MAML algorithm. Ji et al. (2020a) analyzed the convergence and complexity performance of multi-step ANIL algorithm, which is an efficient simplification of MAML by adapting only partial parameters in the inner loop. We emphasize that the study here is the first along the line of studies on multi-step MAML.

We note that a concurrent work Fallah et al. (2020b) also studies multi-step MAML for reinforcement learning setting, where they design an unbiased multi-step estimator. As a comparison, our estimator is biased due to the data sampling in the inner loop, and hence we need extra developments to control this bias, e.g., by bounding the difference between batch-gradient and the stochastic-gradient parameter updates in the inner loop.

Another type of meta-learning algorithms has also been studied as a bi-level optimization problem. Rajeswaran et al. (2019) proposed a meta-regularization variant of MAML named iMAML via bilevel optimization, and analyzed its convergence by assuming that the regularized empirical risk minimization problem in the inner optimization stage is strongly convex. Likhosherstov et al. (2020) studied the convergence properties of a class of first-order bilevel optimization algorithms.

Statistical theory for meta-learning. Zhou et al. (2019) statistically demonstrated the importance of prior hypothesis in reducing the excess risk via a regularization approach. Du et al. (2020) studied few-shot learning from a representation learning perspective, and showed that representation learning can provide a sufficient rate improvement in both linear regression and learning neural networks. Tripuraneni et al. (2020) studied a multi-task linear regression problem with shared low-dimensional representation, and proposed a sample-efficient algorithm with performance guarantee. Arora et al. (2020) proposed a representation learning approach for imitation learning via bilevel optimization, and demonstrated the improved sample complexity brought by representation learning.

2. Problem Setup

In this paper, we study the convergence of the multi-step MAML algorithm. We consider two types of objective functions that are commonly used in practice: (a) **resampling case** (Finn et al., 2017a; Fallah et al., 2020a), where loss functions take the form in expectation and new data are sampled as the algorithm runs; and (b) **finite-sum case** (Antoniou et al., 2019), where loss functions take the finite-sum form with given samples. The resampling case occurs often in reinforcement learning where data are continuously sampled as the algorithm iterates, whereas the finite-sum case typically occurs in classification problems where the datasets are already sampled in advance. In Appendix A, we provide examples for these two types of problems.

2.1 Resampling Case: Problem Setup and Multi-Step MAML

Suppose a set $\mathcal{T} = \{\mathcal{T}_i, i \in \mathcal{I}\}$ of tasks are available for learning and tasks are sampled based on a probability distribution $p(\mathcal{T})$ over the task set. Assume that each task $\mathcal{T}_i$ is associated with a loss $l_i(w) : \mathbb{R}^d \to \mathbb{R}$ parameterized by w.

The goal of multi-step MAML is to find a good initial parameter w^* such that after observing a new task, a few gradient descend steps starting from such a point w^* can efficiently approach the optimizer (or a stationary point) of the corresponding loss function. Towards this end, multi-step MAML consists of two nested stages, where the inner stage consists of *multiple* steps of (stochastic) gradient descent for each individual tasks, and the outer stage updates the meta parameter over all the sampled tasks. More specifically, at each inner stage, each $\mathcal{T}_i$ initializes at the meta parameter, i.e., $\widetilde{w}_0^i := w$, and runs N *gradient descent* steps as

$$\widetilde{w}_{j+1}^i = \widetilde{w}_j^i - \alpha \nabla l_i(\widetilde{w}_j^i), \quad j = 0, 1, ..., N-1. \tag{1}$$

Thus, the loss of task $\mathcal{T}_i$ after the N-step inner stage iteration is given by $l_i(\widetilde{w}_N^i)$, where $\widetilde{w}_N^i$ depends on the meta parameter w through the iteration updates in (1), and can hence be written as $\widetilde{w}_N^i(w)$. We further define $\mathcal{L}_i(w) := l_i(\widetilde{w}_N^i(w))$, and hence the overall meta objective is given by

$$\min_{w \in \mathbb{R}^d} \mathcal{L}(w) := \mathbb{E}_{i \sim p(\mathcal{T})}[\mathcal{L}_i(w)] := \mathbb{E}_{i \sim p(\mathcal{T})}[l_i(\widetilde{w}_N^i(w))]. \tag{2}$$

Then the outer stage of meta update is a gradient decent step to optimize the above objective function. Using the chain rule, we provide a simplified form (see Appendix B for its derivations) of gradient $\nabla \mathcal{L}_i(w)$ by

$$\nabla \mathcal{L}_i(w) = \left[\prod_{j=0}^{N-1} (I - \alpha \nabla^2 l_i(\widetilde{w}_j^i)) \right] \nabla l_i(\widetilde{w}_N^i), \tag{3}$$

where $\widetilde{w}_0^i = w$ for all tasks. Hence, the *full gradient descent* step of the outer stage for (2) can be written as

$$w_{k+1} = w_k - \beta_k \mathbb{E}_{i \sim p(\mathcal{T})} \left[\prod_{j=0}^{N-1} (I - \alpha \nabla^2 l_i(\widetilde{w}_{k,j}^i)) \right] \nabla l_i(\widetilde{w}_{k,N}^i), \tag{4}$$

Algorithm 1 Multi-step MAML in the resampling case

1: **Input:** Initial parameter w_0, inner stepsize $\alpha > 0$
2: **for** $k = 1, ..., K$ **do**
3: Sample $B_k \subset \mathcal{I}$ of i.i.d. tasks by distribution $p(\mathcal{T})$
4: **for** all tasks $\mathcal{T}_i$ in B_k **do**
5: **for** $j = 0, 1, ..., N-1$ **do**
6: Sample a training set $S_{k,j}^i$
Update $w_{k,j+1}^i = w_{k,j}^i - \alpha \nabla l_i(w_{k,j}^i; S_{k,j}^i)$
7: **end for**
8: **end for**
9: Sample T_k^i and $D_{k,j}^i$ and compute $\widehat{G}_i(w_k)$ through (7).
10: update $w_{k+1} = w_k - \beta_k \frac{\sum_{i \in B_k} \widehat{G}_i(w_k)}{|B_k|}$.
11: **end for**

where the index k is added to $\widetilde{w}_j^i$ in (3) to denote that these parameters are at the k^{th} iteration of the meta parameter w.

The inner- and outer-stage updates of MAML given in (1) and (4) involve the gradient $\nabla l_i(\cdot)$ and the Hessian $\nabla^2 l_i(\cdot)$ of the loss function $l_i(\cdot)$, which takes the form of the expectation over the distribution of data samples as given by

$$l_i(\cdot) = \mathbb{E}_\tau l_i(\cdot\,; \tau), \tag{5}$$

where τ represents the data sample. In practice, these two quantities based on the population loss function are estimated by samples. In specific, each task $\mathcal{T}_i$ samples a batch Ω of data under the current parameter w, and uses $\nabla l_i(\cdot\,; \Omega) := \frac{\sum_{\tau \in \Omega} \nabla l_i(\cdot\,; \tau)}{|\Omega|}$ and $\nabla^2 l_i(\cdot\,; \Omega) := \frac{\sum_{\tau \in \Omega} \nabla^2 l_i(\cdot\,; \tau)}{|\Omega|}$ as *unbiased* estimates of the gradient $\nabla l_i(\cdot)$ and the Hessian $\nabla^2 l_i(\cdot)$, respectively.

For practical multi-step MAML as shown in Algorithm 1, at the k^{th} outer stage, we sample a set B_k of tasks. Then, at the inner stage, each task $\mathcal{T}_i \in B_k$ samples a training set $S_{k,j}^i$ for each iteration j in the inner stage, uses $\nabla l_i(w_{k,j}^i; S_{k,j}^i)$ as an estimate of $\nabla l_i(\widetilde{w}_{k,j}^i)$ in (1), and runs a SGD update as

$$w_{k,j+1}^i = w_{k,j}^i - \alpha \nabla l_i(w_{k,j}^i; S_{k,j}^i), \quad j = 0, .., N-1, \tag{6}$$

where the initialization parameter $w_{k,0}^i = w_k$ for all $i \in B_k$.

At the k^{th} outer stage, we draw a batch T_k^i and $D_{k,j}^i$ of data samples independent from each other and both independent from $S_{k,j}^i$ and use $\nabla l_i(w_{k,N}^i; T_k^i)$ and $\nabla^2 l_i(w_{k,j}^i; D_{k,j}^i)$ to estimate $\nabla l_i(\widetilde{w}_{k,N}^i)$ and $\nabla^2 l_i(\widetilde{w}_{k,j}^i)$ in (4), respectively. Then, the meta parameter w_{k+1} at the outer stage is updated by a SGD step as shown in line 10 of Algorithm 1, where the estimated gradient $\widehat{G}_i(w_k)$ has a form of

$$\widehat{G}_i(w_k) = \prod_{j=0}^{N-1} \big(I - \alpha \nabla^2 l_i\big(w_{k,j}^i; D_{k,j}^i\big)\big) \nabla l_i(w_{k,N}^i; T_k^i). \tag{7}$$

For simplicity, we suppose the sizes of $S_{k,j}^i$, $D_{k,j}^i$ and T_k^i are S, D and T in this paper.

Algorithm 2 Multi-step MAML in the finite-sum case

Input: Initial parameter w_0, inner stepsize $\alpha > 0$
for $k = 1, ..., K$ **do**
  Sample $B_k \subset \mathcal{I}$ of i.i.d. tasks by distribution $p(\mathcal{T})$
  for all tasks $\mathcal{T}_i$ in B_k **do**
    for $j = 0, 1, ..., N-1$ **do**
      Update $w^i_{k,j+1} = w^i_{k,j} - \alpha \nabla l_{S_i}(w^i_{k,j})$
    end for
  end for
  Update $w_{k+1} = w_k - \frac{\beta_k}{|B_k|} \sum_{i \in B_k} \widehat{G}_i(w_k)$
end for

2.2 Finite-Sum Case: Problem Setup and Multi-Step MAML

In the finite-sum case, each task $\mathcal{T}_i$ is *pre-assigned* with a support/training sample set S_i and a query/test sample set T_i. Differently from the resampling case, these sample sets are fixed and no additional fresh data are sampled as the algorithm runs. The goal here is to learn an initial parameter w such that for each task i, after N *gradient descent* steps on data from S_i starting from this w, we can find a parameter w_N that performs well on the test data set T_i. Thus, each task $\mathcal{T}_i$ is associated with two fixed loss functions $l_{S_i}(w) := \frac{1}{|S_i|} \sum_{\tau \in S_i} l_i(w; \tau)$ and $l_{T_i}(w) := \frac{1}{|T_i|} \sum_{\tau \in T_i} l_i(w; \tau)$ with a finite-sum structure, where $l_i(w; \tau)$ is the loss on a single sample point τ and a parameter w. Then, the meta objective function takes the form of

$$\min_{w \in \mathbb{R}^d} \mathcal{L}(w) := \mathbb{E}_{i \sim p(\mathcal{T})}[\mathcal{L}_i(w)] = \mathbb{E}_{i \sim p(\mathcal{T})}[l_{T_i}(\widetilde{w}^i_N)], \tag{8}$$

where $\widetilde{w}^i_N$ is obtained by

$$\widetilde{w}^i_{j+1} = \widetilde{w}^i_j - \alpha \nabla l_{S_i}(\widetilde{w}^i_j), \quad j = 0, 1, ..., N-1 \text{ with } \widetilde{w}^i_0 := w. \tag{9}$$

We want to emphasize that S_i and T_i are both training datasets (they together form into **meta-training datasets**), and (8) is the meta-training loss, i.e., the empirical loss for estimating the test time expected loss. (8) does not involve anything correlated with test error. During the test period, MAML will be evaluated over different **meta-test datasets** that are separate from meta-training datasets S_i and T_i.

Similarly to the resampling case, we define the expected losses $l_S(w) = \mathbb{E}_i l_{S_i}(w)$ and $l_T(w) = \mathbb{E}_i l_{T_i}(w)$, and the meta gradient step of the outer stage for (8) can be written as

$$w_{k+1} = w_k - \beta_k \mathbb{E}_{i \sim p(\mathcal{T})} \prod_{j=0}^{N-1} (I - \alpha \nabla^2 l_{S_i}(\widetilde{w}^i_{k,j})) \nabla l_{T_i}(\widetilde{w}^i_{k,N}), \tag{10}$$

where the index k is added to $\widetilde{w}^i_j$ in (9) to denote that these parameters are at the k^{th} iteration of the meta parameter w.

As shown in Algorithm 2, MAML in the finite-sum case has a nested structure similar to that in the resampling case except that it does not sample fresh data at each iteration.

In the inner stage, MAML performs a sequence of *full gradient descent steps* (instead of stochastic gradient steps as in the resampling case) for each task $i \in B_k$ given by

$$w_{k,j+1}^i = w_{k,j}^i - \alpha \nabla l_{S_i}(w_{k,j}^i), \text{ for } j = 0, ..., N-1 \tag{11}$$

where $w_{k,0}^i = w_k$ for all $i \in B_k$. As a result, the parameter $w_{k,j}$ (which denotes the parameter due to the full gradient update) in the update step (11) is equal to $\widetilde{w}_{k,j}$ in (10) for all $j = 0, ..., N$.

At the outer-stage iteration, the meta optimization of MAML performs a SGD step as shown in line 9 of Algorithm 2, where $\widehat{G}_i(w_k)$ is given by

$$\widehat{G}_i(w_k) = \prod_{j=0}^{N-1} (I - \alpha \nabla^2 l_{S_i}(w_{k,j}^i)) \nabla l_{T_i}(w_{k,N}^i). \tag{12}$$

Compared with the resampling case, the biggest difference for analyzing Algorithm 2 in the finite-sum case is that the losses $l_{S_i}(\cdot)$ and $l_{T_i}(\cdot)$ used in the inner and outer stages respectively are different from each other, whereas in the resampling case, they both are equal to $l_i(\cdot)$ which takes the expectation over the corresponding samples. Thus, the convergence analysis for the finite-sum case requires to develop different techniques. For simplicity, we assume that the sizes of all B_k are B.

3. Convergence of Multi-Step MAML in Resampling Case

In this section, we first make some basic assumptions for the meta loss functions in Section 3.1, and then describe several challenges in analyzing the multi-step MAML in Section 3.2, and then present several properties of the meta gradient in Section 3.3, and finally provide the convergence and complexity results for multi-step MAML in Section 3.4.

3.1 Basic Assumptions

We adopt the following standard assumptions (Fallah et al., 2020a; Rajeswaran et al., 2019). Let $\|\cdot\|$ denote the ℓ_2-norm or spectrum norm for a vector or matrix, respectively.

Assumption 1 *The loss $l_i(\cdot)$ of task $\mathcal{T}_i$ given by (5) satisfies*

1. *The loss $l_i(\cdot)$ is bounded below, i.e., $\inf_{w\in\mathbb{R}^d} l_i(w) > -\infty$.*
2. *$\nabla l_i(\cdot)$ is L_i-Lipschitz, i.e., for any $w, u \in \mathbb{R}^d$, $\|\nabla l_i(w) - \nabla l_i(u)\| \le L_i \|w - u\|$.*
3. *$\nabla^2 l_i(\cdot)$ is ρ_i-Lipschitz, i.e., for any $w, u \in \mathbb{R}^d$, $\|\nabla^2 l_i(w) - \nabla^2 l_i(u)\| \le \rho_i \|w - u\|$.*

By the definition of the objective function $\mathcal{L}(\cdot)$ in (2), item 1 of Assumption 1 implies that $\mathcal{L}(\cdot)$ is bounded below. In addition, item 2 implies $\|\nabla^2 l_i(w)\| \le L_i$ for any $w \in \mathbb{R}^d$.

For notational convenience, we take $L = \max_i L_i$ and $\rho = \max_i \rho_i$. The following assumptions impose the bounded-variance conditions on $\nabla l_i(w)$, $\nabla l_i(w;\tau)$ and $\nabla^2 l_i(w;\tau)$.

Assumption 2 *The stochastic gradient $\nabla l_i(\cdot)$ (with i uniformly randomly chosen from set $\mathcal{I}$) has bounded variance, i.e., there exists a constant $\sigma > 0$ such that, for any $w \in \mathbb{R}^d$,*

$$\mathbb{E}_i \|\nabla l_i(w) - \nabla l(w)\|^2 \le \sigma^2,$$

where the expected loss function $l(w) := \mathbb{E}_i l_i(w)$.

Assumption 3 *For any $w \in \mathbb{R}^d$ and $i \in \mathcal{I}$, there exist constants $\sigma_g, \sigma_H > 0$ such that*

$$\mathbb{E}_\tau \|\nabla l_i(w;\tau) - \nabla l_i(w)\|^2 \le \sigma_g^2 \;\; \text{and} \;\; \mathbb{E}_\tau \|\nabla^2 l_i(w;\tau) - \nabla^2 l_i(w)\|^2 \le \sigma_H^2.$$

Note that the above assumptions are made only on individual loss functions $l_i(\cdot)$ rather than on the total loss $\mathcal{L}(\cdot)$, because some conditions do not hold for $\mathcal{L}(\cdot)$, as shown later.

3.2 Challenges of Analyzing Multi-Step MAML

Several new challenges arise when we analyze the convergence of *multi*-step MAML (with $N \ge 2$) compared to the one-step case (with $N = 1$).

First, each iteration of the meta parameter affects the overall objective function via a nested structure of N-step SGD optimization paths over all tasks. Hence, our analysis of the convergence of such a meta parameter needs to characterize the nested structure and the recursive updates.

Second, the meta gradient estimator $\widehat{G}_i(w_k)$ given in (7) involves $\nabla^2 l_i(w^i_{k,j}; D^i_{k,j})$ for $j = 1, ..., N-1$, which are all *biased* estimators of $\nabla^2 l_i(\widetilde{w}^i_{k,j})$ in terms of the randomness over $D^i_{k,j}$. This is because $w^i_{k,j}$ is a stochastic estimator of $\widetilde{w}^i_{k,j}$ obtained via random training sets $S^i_{k,t}, t = 0, ..., j-1$ along an N-step SGD optimization path in the inner stage. In fact, such a bias error occurs only for multi-step MAML with $N \ge 2$ (which equals zero for $N = 1$), and requires additional efforts to handle.

Third, both the Hessian term $\nabla^2 l_i(w^i_{k,j}; D^i_{k,j})$ for $j = 2, ..., N-1$ and the gradient term $\nabla l_i(w^i_{k,N}; T^i_k)$ in the meta gradient estimator $\widehat{G}_i(w_k)$ given in (7) depend on the sample sets $S^i_{k,i}$ used for inner stage iteration to obtain $w^i_{k,N}$, and hence they are statistically *correlated* even conditioned on w_k. Such complication also occurs only for multi-step MAML with $N \ge 2$ and requires new treatment (the two terms are independent for $N = 1$).

Solutions to address the above challenges. The first challenge is mainly caused by the recursive structure of the meta gradient $\nabla\mathcal{L}(w)$ in (4) and the meta gradient estimator $\widehat{G}_i(w_k)$ given in (7). For example, when analyzing the smoothness of the meta gradient $\nabla\mathcal{L}(w)$, we need to characterize the gap Δ_p between two quantities $\prod_{j=0}^{N-1}(I - \alpha\nabla^2 l_i(\widetilde{w}^i_j))$ and $\prod_{j=0}^{N-1}(I - \alpha\nabla^2 l_i(\widetilde{u}^i_j))$, where w^i_j and u^i_j are the j^{th} iterates of two different inner-loop updating paths. Then, using the error decomposition strategy that $\|f_1 f_2 - f_1' f_2'\| \le \|f_1 - f_1'\|\|f_2\| + \|f_1'\|\|f_2 - f_2'\|$, we can decompose the error Δ_p into N parts, where each one corresponds to the distance $\|w^i_j - u^i_j\|$. The remaining step is to bound the distances $\|w^i_j - u^i_j\|, j = 0, ..., N-1$ by finding the relationship between $\|w^i_{j+1} - u^i_{j+1}\|$ and $\|w^i_j - u^i_j\|$ based on the inner-loop gradient descent updates.

To address the second and third challenges, we first use the strategy we propose in the first challenge to decompose the error into N components with each one taking the form of $\|w^i_{k,j} - \widetilde{w}^i_{k,j}\|$, where $w^i_{k,j}$ and $\widetilde{w}^i_{k,j}$ are the j^{th} stochastic gradient step and true gradient step of the inner loop at iteration k. The remaining step is to upper-bound the first- and second-moment distances between $w^i_{k,j}$ and $\widetilde{w}^i_{k,j}$ for all $j = 0, ..., N$ by finding the relationship between $\|w^i_{k,j+1} - \widetilde{w}^i_{k,j+1}\|$ and $\|w^i_{k,j} - \widetilde{w}^i_{k,j}\|$ based on the inner-loop *stochastic* gradient updates.

3.3 Properties of Meta Gradient

Differently from the conventional gradient whose corresponding loss is evaluated directly at the current parameter w, the meta gradient has a more complicated nested structure with respect to w, because its loss is evaluated at the final output of the inner optimization stage, which is N-step SGD updates. As a result, analyzing the meta gradient is very different and more challenging compared to analyzing the conventional gradient. In this subsection, we establish some important properties of the meta gradient which are useful for characterizing the convergence of multi-step MAML.

Recall that $\nabla\mathcal{L}(w) = \mathbb{E}_{i\sim p(\mathcal{T})}[\nabla\mathcal{L}_i(w)]$ with $\nabla\mathcal{L}_i(w)$ given by (3). The following proposition characterizes the Lipschitz property of the gradient $\nabla\mathcal{L}(\cdot)$.

Proposition 1 *Suppose that Assumptions 1, 2 and 3 hold. For any $w, u \in \mathbb{R}^d$, we have*

$$\|\nabla\mathcal{L}(w) - \nabla\mathcal{L}(u)\| \le \big((1+\alpha L)^{2N} L + C_{\mathcal{L}}\mathbb{E}_i\|\nabla l_i(w)\|\big)\|w-u\|,$$

where $C_{\mathcal{L}}$ is a positive constant given by

$$C_{\mathcal{L}} = \big((1+\alpha L)^{N-1}\alpha\rho + \frac{\rho}{L}(1+\alpha L)^N((1+\alpha L)^{N-1}-1)\big)(1+\alpha L)^N. \tag{13}$$

The proof of Proposition 1 handles the first challenge described in Section 3.2. More specifically, we bound the differences between $\widetilde{w}_j^i$ and $\widetilde{u}_j^i$ along two separate paths $(\widetilde{w}_j^i, j = 0, ..., N)$ and $(\widetilde{u}_j^i, j = 0, ..., N)$, and then connect these differences to the distance $\|w-u\|$. Proposition 1 shows that the objective $\mathcal{L}(\cdot)$ has a gradient-Lipschitz parameter

$$L_w = (1+\alpha L)^{2N} L + C_{\mathcal{L}}\mathbb{E}_i\|\nabla l_i(w)\|,$$

which can be unbounded due to the fact that $\nabla l_i(w)$ may be unbounded. Similarly to Fallah et al. (2020a), we use

$$\widehat{L}_{w_k} = (1+\alpha L)^{2N} L + \frac{C_{\mathcal{L}}\sum_{i\in B_k'}\|\nabla l_i(w_k; D_{L_k}^i)\|}{|B_k'|} \tag{14}$$

to estimate L_{w_k} at the meta parameter w_k, where we *independently* sample the data sets B_k' and $D_{L_k}^i$. As will be shown in Theorem 5, we set the meta stepsize β_k to be inversely proportional to $\widehat{L}_{w_k}$ to handle the possibly unboundedness.

We next characterize several estimation properties of the meta gradient estimator $\widehat{G}_i(w_k)$ in (7). Here, we address the second and third challenges described in Section 3.2. We first quantify how far the stochastic gradient iterate $w_{k,j}^i$ is away from the true gradient iterate $\widetilde{w}_{k,j}^i$, and then provide upper bounds on the first- and second-moment distances between $w_{k,j}^i$ and $\widetilde{w}_{k,j}^i$ for all $j = 0, ..., N$ as below.

Proposition 2 *Suppose that Assumptions 1, 2 and 3 hold. Then, for any $j = 0, ..., N$ and $i \in B_k$, we have*

- **First-moment :** $\mathbb{E}(\|w_{k,j}^i - \widetilde{w}_{k,j}^i\|\,|w_k) \le \big((1+\alpha L)^j - 1\big)\frac{\sigma_g}{L\sqrt{S}}$.
- **Second-moment:** $\mathbb{E}(\|w_{k,j}^i - \widetilde{w}_{k,j}^i\|^2\,|w_k) \le \big((1+\alpha L+2\alpha^2L^2)^j - 1\big)\frac{\alpha\sigma_g^2}{(1+\alpha L)LS}$.

Proposition 2 shows that we can effectively upper-bound the point-wise distance between two paths by choosing α and S properly. Based on Proposition 2, we provide an upper bound on the first-moment estimation error of meta gradient estimator $\widehat{G}_i(w_k)$.

Proposition 3 *Suppose Assumptions 1, 2 and 3 hold, and define constants*

$$C_{\mathrm{err}_1} = (1+\alpha L)^{2N}\sigma_g, \;\; C_{\mathrm{err}_2} = \frac{(1+\alpha L)^{4N}\rho\sigma_g}{\big(2-(1+\alpha L)^{2N}\big)L^2}. \tag{15}$$

Let $e_k := \mathbb{E}[\widehat{G}_i(w_k)] - \nabla\mathcal{L}(w_k)$ be the estimation error. If the inner stepsize $\alpha < (2^{\frac{1}{2N}}-1)/L$, then conditioning on w_k, we have

$$\|e_k\| \leq \frac{C_{\mathrm{err}_1}}{\sqrt{S}} + \frac{C_{\mathrm{err}_2}}{\sqrt{S}}(\|\nabla\mathcal{L}(w_k)\| + \sigma). \tag{16}$$

Note that the estimation error for the multi-step case shown in Proposition 3 involves a term $\mathcal{O}\big(\frac{\|\nabla\mathcal{L}(w_k)\|}{\sqrt{S}}\big)$, which cannot be avoided due to the Hessian approximation error caused by the randomness over the inner-loop samples sets $S^i_{k,j}$. Somewhat interestingly, our later analysis shows that this term does not affect the final convergence rate if we choose the size S properly. The following proposition provides an upper-bound on the second moment of the meta gradient estimator $\widehat{G}_i(w_k)$.

Proposition 4 *Suppose that Assumptions 1, 2 and 3 hold. Define constants*

$$\begin{aligned} C_{\mathrm{squ}_1} &= 3\Big(\frac{\alpha^2\sigma_H^2}{D} + (1+\alpha L)^2\Big)^N\sigma_g^2, \;\; C_{\mathrm{squ}_3} = \frac{2C_{\mathrm{squ}_1}(1+\alpha L)^{2N}}{(2-(1+\alpha L)^{2N})^2\sigma_g^2}, \\ C_{\mathrm{squ}_2} &= C_{\mathrm{squ}_1}\big((1+2\alpha L+2\alpha^2L^2)^N - 1\big)\alpha L(1+\alpha L)^{-1}. \end{aligned} \tag{17}$$

If the inner stepsize $\alpha < (2^{\frac{1}{2N}}-1)/L$, then conditioning on w_k, we have

$$\mathbb{E}\|\widehat{G}_i(w_k)\|^2 \leq \frac{C_{\mathrm{squ}_1}}{T} + \frac{C_{\mathrm{squ}_2}}{S} + C_{\mathrm{squ}_3}\left(\|\nabla\mathcal{L}(w_k)\|^2 + \sigma^2\right). \tag{18}$$

By choosing set sizes D, T, S and the inner stepsize α properly, the factor C_{squ_3} in the second-moment error bound in (18) can be made at a constant level and the first two error terms $\frac{C_{\mathrm{squ}_1}}{T}$ and $\frac{C_{\mathrm{squ}_2}}{S}$ can be made sufficiently small so that the variance of the meta gradient estimator can be well controlled in the convergence analysis, as shown later.

3.4 Main Convergence Result

By using the properties of the meta gradient established in Section 3.3, we provide the convergence rate for multi-step MAML of Algorithm 1 in the following theorem.

Theorem 5 *Suppose that Assumptions 1, 2 and 3 hold. Set the meta stepsize $\beta_k = \frac{1}{C_\beta\widehat{L}_{w_k}}$, where $C_\beta > 0$ is a positive constant and $\widehat{L}_{w_k}$ is the approximated smoothness parameter*

given by (14). *For* $\widehat{L}_{w_k}$ *in* (14), *we choose* $|B_k'| > \frac{4C_{\mathcal{L}}^2\sigma^2}{3(1+\alpha L)^{4N}L^2}$ *and* $|D_{L_k}^i| > \frac{64\sigma_g^2C_{\mathcal{L}}^2}{(1+\alpha L)^{4N}L^2}$ *for all* $i \in B_k'$, *where* $C_{\mathcal{L}}$ *is given by* (13). *Define* $\chi = \frac{(2-(1+\alpha L)^{2N})(1+\alpha L)^{2N}L}{C_{\mathcal{L}}} + \sigma$ *and*

$$\xi = \frac{6}{C_\beta L}\big(\frac{1}{5} + \frac{2}{C_\beta}\big)\big(C_{\mathrm{err}_1}^2 + C_{\mathrm{err}_2}^2\sigma^2\big), \quad \phi = \frac{2}{C_\beta^2 L}\Big(\frac{C_{\mathrm{squ}_1}}{T} + \frac{C_{\mathrm{squ}_2}}{S} + C_{\mathrm{squ}_3}\sigma^2\Big)$$
$$\theta = \frac{2\big(2-(1+\alpha L)^{2N}\big)}{C_\beta C_{\mathcal{L}}}\Big(\frac{1}{5} - \big(\frac{3}{5} + \frac{6}{C_\beta}\big)\frac{C_{\mathrm{err}_2}^2}{S} - \frac{C_{\mathrm{squ}_3}}{C_\beta B} - \frac{2}{C_\beta}\Big) \tag{19}$$

where $C_{\mathrm{err}_1}, C_{\mathrm{err}_2}$ *are given in* (15) *and* $C_{\mathrm{squ}_1}, C_{\mathrm{squ}_2}, C_{\mathrm{squ}_3}$ *are given in* (17). *Choose the inner stepsize* $\alpha < (2^{\frac{1}{2N}}-1)/L$, *and choose* C_β, S *and* B *such that* $\theta > 0$. *Then, Algorithm 1 finds a solution* w_ζ *such that*

$$\mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\| \leq \frac{\Delta}{\theta}\frac{1}{K} + \frac{\xi}{\theta}\frac{1}{S} + \frac{\phi}{\theta}\frac{1}{B} + \sqrt{\frac{\chi}{2}}\sqrt{\frac{\Delta}{\theta}\frac{1}{K} + \frac{\xi}{\theta}\frac{1}{S} + \frac{\phi}{\theta}\frac{1}{B}}, \tag{20}$$

where $\Delta = \mathcal{L}(w_0) - \mathcal{L}^*$ *with* $\mathcal{L}^* = \inf_{w\in\mathbb{R}^d}\mathcal{L}(w)$.

Note that for χ in Theorem 5, we replace the notation C_l by $(1+\alpha L)^{2N} - 1$ based on its definition. The proof of Theorem 5 (see Section 5.1 for details) consists of four main steps: step 1 of bounding an iterative meta update by the meta-gradient smoothness established by Proposition 1; step 2 of characterizing first-moment estimation error of the meta-gradient estimator $\widehat{G}_i(w_k)$ by Proposition 3; step 3 of characterizing second-moment estimation error of the meta-gradient estimator $\widehat{G}_i(w_k)$ by Proposition 4; and step 4 of combining steps 1-3, and telescoping to yield the convergence.

In Theorem 5, the convergence rate given by (20) mainly contains three parts: the first term $\frac{\Delta}{\theta}\frac{1}{K}$ indicates that the meta parameter converges sublinearly with the number K of meta iterations, the second term $\frac{\xi}{\theta}\frac{1}{S}$ captures the estimation error of $\nabla l_i(w_{k,j}^i; S_{k,j}^i)$ for approximating the full gradient $\nabla l_i(w_{k,j}^i)$ which can be made sufficiently small by choosing a large sample size S, and the third term $\frac{\phi}{\theta}\frac{1}{B}$ captures the estimation error and variance of the stochastic meta gradient, which can be made small by choosing large B, T and D (note that ϕ is proportional to both $\frac{1}{T}$ and $\frac{1}{D}$).

It is worthwhile mentioning that our results here focus on our resampling case, where fresh data are resampled as the algorithm runs. This resampling case often happens in bandit or reinforcement learning settings, where batch sizes S, B, D, T can be chosen to be large and the resulting convergence errors will be small. However, for the cases where S, B, D, T are small, our results in Theorem 5 will contain large convergence errors. It is possible to use some techniques such as variance reduction to reduce or even remove such errors. However, this is not the focus of this paper, and require future efforts to address.

Our analysis reveals several insights for the convergence of multi-step MAML as follows. (a) To guarantee convergence, we require $\alpha L < 2^{\frac{1}{2N}} - 1$ (e.g., $\alpha = \Theta(\frac{1}{NL})$). Hence, if the number N of inner gradient steps is large and L is not small (e.g., for some RL problems), we need to choose a small inner stepsize α so that the last output of the inner stage has a *strong dependence* on the initialization (i.e., meta parameter). This is also explained in Rajeswaran et al. (2019), where they add a regularizer $\lambda\|w' - w\|^2$ to make sure the inner-loop output w' has a close connection to the initialization w. (b) For problems with small

Hessians such as many classification/regression problems (Finn et al., 2017a), L (which is an upper bound on the spectral norm of Hessian matrices) is small, and hence we can choose a larger α. This explains the empirical findings in Finn et al. (2017a); Antoniou et al. (2019), where their experiments tend to set a larger stepsize for the regression problems with smaller Hessians.

We next specify the selection of parameters to simplify the convergence result in Theorem 5 and derive the complexity of Algorithm 1 for finding an ϵ-accurate stationary point.

Corollary 6 *Under the setting of Theorem 5, choose* $\alpha = \frac{1}{8NL}, C_\beta = 100$ *and let batch sizes* $S \geq \frac{15\rho^2\sigma_g^2}{L^4}$ *and* $D \geq \sigma_H^2 L^2$. *Then we have*

$$\mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\| \leq \mathcal{O}\Big(\frac{1}{K} + \frac{\sigma_g^2(\sigma^2+1)}{S} + \frac{\sigma_g^2+\sigma^2}{B} + \frac{\sigma_g^2}{TB} + \sqrt{\sigma+1}\sqrt{\frac{1}{K} + \frac{\sigma_g^2(\sigma^2+1)}{S} + \frac{\sigma_g^2+\sigma^2}{B} + \frac{\sigma_g^2}{TB}}\Big).$$

To achieve $\mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\| < \epsilon$, *Algorithm 1 requires at most* $\mathcal{O}\big(\frac{1}{\epsilon^2}\big)$ *iterations, and* $\mathcal{O}(\frac{N}{\epsilon^4} + \frac{1}{\epsilon^2})$ *gradient computations and* $\mathcal{O}\big(\frac{N}{\epsilon^2}\big)$ *Hessian computations per meta iteration.*

Differently from the conventional SGD that requires a gradient complexity of $\mathcal{O}(\frac{1}{\epsilon^4})$, MAML requires a higher gradient complexity by a factor of $\mathcal{O}(\frac{1}{\epsilon^2})$, which is unavoidable because MAML requires $\mathcal{O}(\frac{1}{\epsilon^2})$ tasks to achieve an ϵ-accurate meta point, whereas SGD runs only over one task.

Corollary 6 shows that given a properly chosen inner stepsize, e.g., $\alpha = \Theta(\frac{1}{NL})$, MAML is guaranteed to converge with both the gradient and the Hessian computation complexities growing only *linearly* with N. These results explain some empirical findings for MAML training in Rajeswaran et al. (2019). The above results can also be obtained by using a larger stepsize such as $\alpha = \Theta(c^{\frac{1}{N}} - 1)/L > \Theta\big(\frac{1}{NL}\big)$with a certain constant $c > 1$.

4. Convergence of Multi-Step MAML in Finite-Sum Case

In this section, we provide several properties of the meta gradient for the finite-sum case, and then analyze the convergence and complexity of Algorithm 2. Differently from the resampling case, we develop novel techniques to handle the difference between two losses over the training and test sets (i.e., inner- and outer-loop losses) in the analysis, whereas these two losses are the same for the resampling case.

4.1 Basic Assumptions

We state several standard assumptions for the analysis in the finite-sum case.

Assumption 4 *For each task* $\mathcal{T}_i$, *the loss functions* $l_{S_i}(\cdot)$ *and* $l_{T_i}(\cdot)$ *in* (8) *satisfy*

1. $l_{S_i}(\cdot), l_{T_i}(\cdot)$ *are bounded below, i.e.,* $\inf_{w\in\mathbb{R}^d} l_{S_i}(w) > -\infty$ *and* $\inf_{w\in\mathbb{R}^d} l_{T_i}(w) > -\infty$.
2. *Gradients* $\nabla l_{S_i}(\cdot)$ *and* $\nabla l_{T_i}(\cdot)$ *are* L*-Lipschitz continuous, i.e., for any* $w, u \in \mathbb{R}^d$

$$\|\nabla l_{S_i}(w) - \nabla l_{S_i}(u)\| \leq L\|w-u\| \text{ and } \|\nabla l_{T_i}(w) - \nabla l_{T_i}(u)\| \leq L\|w-u\|.$$

3. *Hessians* $\nabla^2 l_{S_i}(\cdot)$ *and* $\nabla^2 l_{T_i}(\cdot)$ *are* ρ*-Lipschitz continuous, i.e., for any* $w, u \in \mathbb{R}^d$

$$\|\nabla^2 l_{S_i}(w) - \nabla^2 l_{S_i}(u)\| \leq \rho\|w-u\| \text{ and } \|\nabla^2 l_{T_i}(w) - \nabla^2 l_{T_i}(u)\| \leq \rho\|w-u\|.$$

The following assumption provides two conditions $\nabla l_{S_i}(\cdot)$ and $\nabla l_{T_i}(\cdot)$.

Assumption 5 *For all* $w \in \mathbb{R}^d$*, gradients* $\nabla l_{S_i}(w)$ *and* $\nabla l_{T_i}(w)$ *satisfy*

1. $\nabla l_{T_i}(\cdot)$ *has a bounded variance, i.e., there exists a constant* $\sigma > 0$ *such that*

$$\mathbb{E}_i\|\nabla l_{T_i}(w) - \nabla l_T(w)\|^2 \leq \sigma^2,$$

where $\nabla l_T(\cdot) = \mathbb{E}_i\left[\nabla l_{T_i}(\cdot)\right]$.

2. *For each* $i \in \mathcal{I}$*, there exists a constant* $b_i > 0$ *such that* $\|\nabla l_{S_i}(w) - \nabla l_{T_i}(w)\| \leq b_i$.

Instead of imposing a bounded variance condition on the stochastic gradient $\nabla l_{S_i}(w)$, we alternatively assume the difference $\|\nabla l_{S_i}(w) - \nabla l_{T_i}(w)\|$ to be upper-bounded by a constant, which is more reasonable because sample sets S_i and T_i are often sampled from the same distribution and share certain statistical similarity. We note that the second condition also implies $\|\nabla l_{S_i}(w)\| \leq \|\nabla l_{T_i}(w)\| + b_i$, which is weaker than the bounded gradient assumption made in papers such as Finn et al. (2019). It is worthwhile mentioning that the second condition can be relaxed to $\|\nabla l_{S_i}(w)\| \leq c_i\|\nabla l_{T_i}(w)\| + b_i$ for a constant $c_i > 0$. Without the loss of generality, we consider $c_i = 1$ for simplicity.

4.2 Properties of Meta Gradient

We develop several important properties of the meta gradient. The following proposition characterizes a Lipschitz property of the gradient of the objective function

$$\nabla\mathcal{L}(w) = \mathbb{E}_{i\sim p(\mathcal{T})}\prod_{j=0}^{N-1}(I - \alpha\nabla^2 l_{S_i}(\widetilde{w}_j^i))\nabla l_{T_i}(\widetilde{w}_N^i),$$

where the weights $\widetilde{w}_j^i, i \in \mathcal{I}, j = 0, ..., N$ are given by the gradient descent steps in (9).

Proposition 7 *Suppose that Assumptions 4 and 5 hold. Then, for any* $w, u \in \mathbb{R}^d$*, we have*

$$\|\nabla\mathcal{L}(w) - \nabla\mathcal{L}(u)\| \leq L_w\|w-u\|,\ L_w = (1+\alpha L)^{2N}L + C_b b + C_{\mathcal{L}}\mathbb{E}_i\|\nabla l_{T_i}(w)\|$$

where $b = \mathbb{E}_i[b_i]$ *and* $C_b, C_{\mathcal{L}} > 0$ *are constants given by*

$$C_b = \big(\alpha\rho + \frac{\rho}{L}(1+\alpha L)^{N-1}\big)(1+\alpha L)^{2N},\ C_{\mathcal{L}} = \big(\alpha\rho + \frac{\rho}{L}(1+\alpha L)^{N-1}\big)(1+\alpha L)^{2N}. \tag{21}$$

Proposition 7 shows that $\nabla\mathcal{L}(w)$ has a Lipschitz parameter L_w. Similarly to (14), we use the following construction

$$\hat{L}_{w_k} = (1+\alpha L)^{2N}L + C_b b + \frac{C_{\mathcal{L}}}{|B_k'|}\sum_{i\in B_k'}\|\nabla l_{T_i}(w_k)\|, \tag{22}$$

at the k^{th} outer-stage iteration to approximate L_{w_k}, where $B_k' \subset \mathcal{I}$ is chosen independently from B_k. It can be verified that the gradient estimator $\widehat{G}_i(w_k)$ given in (12) is an *unbiased* estimate of $\nabla\mathcal{L}(w_k)$. Thus, our next step is to upper-bound the second moment of $\widehat{G}_i(w_k)$.

Proposition 8 *Suppose Assumptions 4 and 5 are hold, and define constants*

$$A_{\text{squ}_1} = \frac{4(1+\alpha L)^{4N}}{(2-(1+\alpha L)^{2N})^2}, \quad A_{\text{squ}_2} = \frac{4(1+\alpha L)^{8N}}{(2-(1+\alpha L)^{2N})^2}(\sigma+b)^2 + 2(1+\alpha)^{4N}(\sigma^2+\widetilde{b}), \tag{23}$$

where $\widetilde{b} = \mathbb{E}_{i\sim p(\mathcal{T})}[b_i^2]$. *Then, if* $\alpha < (2^{\frac{1}{2N}}-1)/L$, *then conditioning on* w_k, *we have*

$$\mathbb{E}\|\widehat{G}_i(w_k)\|^2 \le A_{\text{squ}_1}\|\nabla\mathcal{L}(w_k)\|^2 + A_{\text{squ}_2}.$$

Based on the above properties, we next characterize the convergence of multi-step MAML.

4.3 Main Convergence Results

In this subsection, we provide the convergence and complexity analysis for Algorithm 2 based on the properties established in the previous subsection.

Theorem 9 *Let Assumptions 4 and 5 hold, and apply Algorithm 2 to solve the objective function* (8). *Choose the meta stepsize* $\beta_k = \frac{1}{C_\beta \widehat{L}_{w_k}}$ *with* $\widehat{L}_{w_k}$ *given by* (22), *where* $C_\beta > 0$ *is a constant. For* $\widehat{L}_{w_k}$ *in* (22), *we choose the batch size* $|B_k'|$ *such that* $|B_k'| \ge \frac{2C_{\mathcal{L}}^2\sigma^2}{(C_b b+(1+\alpha L)^{2N}L)^2}$, *where* C_b *and* $C_{\mathcal{L}}$ *are given by* (21). *Define constants*

$$\begin{aligned}
\xi =& \frac{2-(1+\alpha L)^{2N}}{C_{\mathcal{L}}}(1+\alpha L)^{2N}L + \frac{\big(2-(1+\alpha L)^{2N}\big)C_b b}{C_{\mathcal{L}}} + (1+\alpha L)^{3N}b,\\
\theta =& \frac{2-(1+\alpha L)^{2N}}{C_{\mathcal{L}}}\Big(\frac{1}{C_\beta} - \frac{1}{C_\beta^2}\Big(\frac{A_{\text{squ}_1}}{B}+1\Big)\Big),\ \phi = \frac{A_{\text{squ}_2}}{LC_\beta^2}
\end{aligned} \tag{24}$$

where $C_b, C_{\mathcal{L}}, A_{\text{squ}_1}$ *and* A_{squ_1} *are given by* (21) *and* (23). *Choose* $\alpha < (2^{\frac{1}{2N}}-1)/L$, *and choose* C_β *and* B *such that* $\theta > 0$. *Then, Algorithm 2 attains a solution* w_ζ *such that*

$$\mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\| \le \frac{\Delta}{2\theta K} + \frac{\phi}{2\theta B} + \sqrt{\xi\Big(\frac{\Delta}{\theta K}+\frac{\phi}{\theta B}\Big) + \Big(\frac{\Delta}{2\theta K}+\frac{\phi}{2\theta B}\Big)^2}. \tag{25}$$

The parameters θ, ϕ and ξ in Theorem 9 take complicate forms. The following corollary specifies the parameters C_β, α in Theorem 9 and provides a simplified result for Algorithm 2.

Corollary 10 *Under the same setting of Theorem 9, choose* $\alpha = \frac{1}{8NL}, C_\beta = 80$. *We have*

$$\mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\| \le \mathcal{O}\Big(\frac{1}{K} + \frac{\sigma^2}{B} + \sqrt{\frac{1}{K}+\frac{\sigma^2}{B}}\Big).$$

In addition, suppose the batch size B *further satisfies* $B \ge C_B\sigma^2\epsilon^{-2}$, *where* C_B *is a sufficiently large constant. Then, to achieve an* ϵ*-approximate stationary point, Algorithm 2 requires at most* $K = \mathcal{O}(\epsilon^{-2})$ *iterations, and a total number* $\mathcal{O}\big((T+NS)\epsilon^{-2}\big)$ *of gradient computations and a number* $\mathcal{O}\big(NS\epsilon^{-2}\big)$ *of Hessian computations per iteration, where* T *and* S *correspond to the sample sizes of the pre-assigned sets* $T_i, i\in\mathcal{I}$ *and* $S_i, i\in\mathcal{I}$.

5. Proofs of Main Results

In this section, we provide the proofs the main results for MAML in the resampling case and the finite-sum case, respectively. This section is organized as follows.

For the resampling case, Section 5.1 provides the proofs for the convergence properties of multi-step MAML in the *resampling case*, which include Propositions 1, 2, 3, 4 on the properties of meta gradient, and Theorem 5 and Corollary 6 on the convergence and complexity performance of multi-step MAML. The proofs of these results require several technical lemmas, which we relegate to the Appendix C.

Next, for the finite-sum case, Section 5.2 provides the proofs for the convergence properties of multi-step MAML in the *finite-sum case*, which include Propositions 7, 8 on the properties of meta gradient, and Theorem 9 and Corollary 10 on the convergence and complexity of multi-step MAML. The proofs of these results rely on several technical lemmas, which we relegate to the Appendix D.

5.1 Proofs for Section 3: Convergence of Multi-Step MAML in Resampling Case

To simplify notations, we let $\bar{S}_j^i$ and $\bar{D}_j^i$ denote the randomness over $S_{k,m}^i, D_{k,m}^i, m = 0, ..., j-1$ and let $\bar{S}_j$ and $\bar{D}_j$ denote all randomness over $\bar{S}_j^i, \bar{D}_j^i, i \in \mathcal{I}$, respectively.

Proof of Proposition 1

First recall that $\nabla \mathcal{L}_i(w) = \prod_{j=0}^{N-1}(I - \alpha \nabla^2 l_i(\widetilde{w}_j^i))\nabla l_i(\widetilde{w}_N^i)$. Then, we have

$$\begin{aligned}
\|\nabla \mathcal{L}_i(w) - \nabla \mathcal{L}_i(u)\| \leq & \Big\| \prod_{j=0}^{N-1}(I - \alpha \nabla^2 l_i(\widetilde{w}_j^i)) - \prod_{j=0}^{N-1}(I - \alpha \nabla^2 l_i(\widetilde{u}_j^i)) \Big\| \big\| \nabla l_i(\widetilde{w}_N^i) \big\| \\
& + (1+\alpha L)^N \|\nabla l_i(\widetilde{w}_N^i) - \nabla l_i(\widetilde{u}_N^i)\| \\
\overset{(i)}{\leq} & \Big\| \prod_{j=0}^{N-1}(I - \alpha \nabla^2 l_i(\widetilde{w}_j^i)) - \prod_{j=0}^{N-1}(I - \alpha \nabla^2 l_i(\widetilde{u}_j^i)) \Big\| (1+\alpha L)^N \big\| \nabla l_i(w) \big\| \\
& + (1+\alpha L)^N L \|\widetilde{w}_N^i - \widetilde{u}_N^i\| \\
\overset{(ii)}{\leq} & \underbrace{\Big\| \prod_{j=0}^{N-1}(I - \alpha \nabla^2 l_i(\widetilde{w}_j^i)) - \prod_{j=0}^{N-1}(I - \alpha \nabla^2 l_i(\widetilde{u}_j^i)) \Big\|}_{V(N)} (1+\alpha L)^N \big\| \nabla l_i(w) \big\| \\
& + (1+\alpha L)^{2N} L \|w - u\|, \qquad (26)
\end{aligned}$$

where (i) follows from Lemma 12, and (ii) follows from Lemma 11. We next upper-bound the term $V(N)$ in the above inequality. Specifically, define a more general quantity $V(m)$

by replacing N in $V(N)$ with m. Then, we have

$$\begin{aligned}
V(m) \leq & \Big\| \prod_{j=0}^{m-2} (I - \alpha\nabla^2 l_i(\widetilde{w}_j^i)) \Big\| \big\| \alpha\nabla^2 l_i(\widetilde{w}_{m-1}^i) - \alpha\nabla^2 l_i(\widetilde{u}_{m-1}^i) \big\| \\
& + \Big\| \prod_{j=0}^{m-2} (I - \alpha\nabla^2 l_i(\widetilde{w}_j^i)) - \prod_{j=0}^{m-2} (I - \alpha\nabla^2 l_i(\widetilde{u}_j^i)) \Big\| \big\| I - \alpha\nabla^2 l_i(\widetilde{u}_{m-1}^i) \big\| \\
\leq & (1+\alpha L)^{m-1} \big\| \alpha\nabla^2 l_i(\widetilde{w}_{m-1}^i) - \alpha\nabla^2 l_i(\widetilde{u}_{m-1}^i) \big\| \\
& + (1+\alpha L) \Big\| \prod_{j=0}^{m-2} (I - \alpha\nabla^2 l_i(\widetilde{w}_j^i)) - \prod_{j=0}^{m-2} (I - \alpha\nabla^2 l_i(\widetilde{u}_j^i)) \Big\| \\
\leq & (1+\alpha L)^{m-1} \alpha\rho \|\widetilde{w}_{m-1}^i - \widetilde{u}_{m-1}^i\| + (1+\alpha L) V(m-1) \\
\leq & (1+\alpha L)^{m-1} \alpha\rho (1+\alpha L)^{m-1} \|w-u\| + (1+\alpha L) V(m-1). \qquad (27)
\end{aligned}$$

Telescoping (27) over m from 1 to N and noting $V(1) \leq \alpha\rho\|w-u\|$, we have

$$\begin{aligned}
V(N) &\leq (1+\alpha L)^{N-1} V(1) + \sum_{m=0}^{N-2} \alpha\rho (1+\alpha L)^{2(N-m)-2} \|w-u\| (1+\alpha L)^m \\
&= (1+\alpha L)^{N-1} \alpha\rho \|w-u\| + \alpha\rho (1+\alpha L)^N \sum_{m=0}^{N-2} (1+\alpha L)^m \|w-u\| \\
&\leq \Big((1+\alpha L)^{N-1} \alpha\rho + \frac{\rho}{L} (1+\alpha L)^N ((1+\alpha L)^{N-1} - 1) \Big) \|w-u\|. \qquad (28)
\end{aligned}$$

Recalling the definition of $C_{\mathcal{L}}$ and Combining (26), (28), we have

$$\|\nabla\mathcal{L}_i(w) - \nabla\mathcal{L}_i(u)\| \leq \big(C_{\mathcal{L}} \|\nabla l_i(w)\| + (1+\alpha L)^{2N} L \big) \|w-u\|.$$

Based on the above inequality, we have

$$\begin{aligned}
\|\nabla\mathcal{L}(w) - \nabla\mathcal{L}(u)\| &= \|\mathbb{E}_{i\sim p(\mathcal{T})} (\nabla\mathcal{L}_i(w) - \nabla\mathcal{L}_i(u))\| \\
&\leq \mathbb{E}_{i\sim p(\mathcal{T})} \|(\nabla\mathcal{L}_i(w) - \nabla\mathcal{L}_i(u))\| \\
&\leq \big(C_{\mathcal{L}} \mathbb{E}_{i\sim p(\mathcal{T})} \|\nabla l_i(w)\| + (1+\alpha L)^{2N} L \big) \|w-u\|,
\end{aligned}$$

which finishes the proof.

Proof of Proposition 2

We first prove the first-moment bound. Conditioning on w_k, we have

$$\begin{aligned}
\mathbb{E}_{\bar{S}_m^i}\|w_{k,m}^i-\widetilde{w}_{k,m}^i\| \overset{(i)}{=}&\mathbb{E}_{\bar{S}_m^i}\left\|w_{k,m-1}^i-\alpha\nabla l_i(w_{k,m-1}^i;S_{k,m-1}^i)-(\widetilde{w}_{k,m-1}^i-\alpha\nabla l_i(\widetilde{w}_{k,m-1}^i))\right\|\\
\leq&\mathbb{E}_{\bar{S}_m^i}\|w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i\|+\alpha\mathbb{E}_{\bar{S}_m^i}\left\|\nabla l_i(w_{k,m-1}^i;S_{k,m-1}^i)-\nabla l_i(w_{k,m-1}^i)\right\|\\
&+\alpha\mathbb{E}_{\bar{S}_m^i}\left\|\nabla l_i(w_{k,m-1}^i)-\nabla l_i(\widetilde{w}_{k,m-1}^i)\right\|\\
\leq&\alpha\mathbb{E}_{\bar{S}_{m-1}^i}\Big(\mathbb{E}_{S_{k,m-1}^i}\big(\|\nabla l_i(w_{k,m-1}^i;S_{k,m-1}^i)-\nabla l_i(w_{k,m-1}^i)\|\,\big|\,\bar{S}_{m-2}^i\big)\Big)\\
&+(1+\alpha L)\mathbb{E}_{\bar{S}_{m-1}^i}\|w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i\|\\
\overset{(ii)}{\leq}&(1+\alpha L)\mathbb{E}_{\bar{S}_{m-1}^i}\|w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i\|+\alpha\frac{\sigma_g}{\sqrt{S}},
\end{aligned}$$

where (i) follows from (1) and (6), and (ii) follows from Assumption 3. Telescoping the above inequality over m from 1 to j and using the fact that $w_{k,0}^i=\widetilde{w}_{k,0}^i=w_k$, we have

$$\mathbb{E}_{\bar{S}_j^i}\|w_{k,j}^i-\widetilde{w}_{k,j}^i\|\leq((1+\alpha L)^j-1)\frac{\sigma_g}{L\sqrt{S}},$$

which finishes the proof of the first-moment bound. We next begin to prove the second-moment bound. Conditioning on w_k, we have

$$\begin{aligned}
&\mathbb{E}_{\bar{S}_m^i}\|w_{k,m}^i-\widetilde{w}_{k,m}^i\|^2\\
=&\mathbb{E}_{\bar{S}_{m-1}^i}\|w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i\|^2+\alpha^2\mathbb{E}_{\bar{S}_m^i}\|\nabla l_i(w_{k,m-1}^i;S_{k,m-1}^i)-\nabla l_i(\widetilde{w}_{k,m-1}^i)\|^2\\
&-2\alpha\mathbb{E}_{\bar{S}_{m-1}^i}\Big(\mathbb{E}_{S_{k,m-1}^i}\langle w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i,\nabla l_i(w_{k,m-1}^i;S_{k,m-1}^i)-\nabla l_i(\widetilde{w}_{k,m-1}^i)\rangle\big|\bar{S}_{m-1}^i\Big)\\
\overset{(i)}{\leq}&\mathbb{E}_{\bar{S}_{m-1}^i}\|w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i\|^2-2\alpha\mathbb{E}_{\bar{S}_{m-1}^i}\langle w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i,\nabla l_i(w_{k,m-1}^i)-\nabla l_i(\widetilde{w}_{k,m-1}^i)\rangle\\
&+\alpha^2\mathbb{E}_{\bar{S}_m^i}\left(2\|\nabla l_i(w_{k,m-1}^i;S_{k,m-1}^i)-\nabla l_i(w_{k,m-1}^i)\|^2+2\|\nabla l_i(w_{k,m-1}^i)-\nabla l_i(\widetilde{w}_{k,m-1}^i)\|^2\right)\\
\overset{(ii)}{\leq}&\mathbb{E}_{\bar{S}_{m-1}^i}\|w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i\|^2+2\alpha\mathbb{E}_{\bar{S}_{m-1}^i}\|w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i\|\|\nabla l_i(w_{k,m-1}^i)-\nabla l_i(\widetilde{w}_{k,m-1}^i))\|\\
&+\alpha^2\mathbb{E}_{\bar{S}_m^i}\left(2\|\nabla l_i(w_{k,m-1}^i;S_{k,m-1}^i)-\nabla l_i(w_{k,m-1}^i)\|^2+2\|\nabla l_i(w_{k,m-1}^i)-\nabla l_i(\widetilde{w}_{k,m-1}^i)\|^2\right)\\
\leq&\mathbb{E}_{\bar{S}_{m-1}^i}\|w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i\|^2+2\alpha L\mathbb{E}_{\bar{S}_{m-1}^i}\|w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i\|^2\\
&+2\alpha^2\mathbb{E}_{\bar{S}_{m-1}^i}\Big(\frac{\sigma_g^2}{S}+L^2\|w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i\|^2\Big)\\
\leq&\big(1+2\alpha L+2\alpha^2L^2\big)\mathbb{E}_{\bar{S}_{m-1}^i}\|w_{k,m-1}^i-\widetilde{w}_{k,m-1}^i\|^2+\frac{2\alpha^2\sigma_g^2}{S},
\end{aligned}$$

where (i) follows from $\mathbb{E}_{S_{k,m-1}^i}\nabla l_i(w_{k,m-1}^i;S_{k,m-1}^i)=\nabla l_i(w_{k,m-1}^i)$ and (ii) follows from the inequality that $-\langle a,b\rangle\leq\|a\|\|b\|$ for any vectors a,b. Noting that $w_{k,0}^i=\widetilde{w}_{k,0}^i=w_k$ and telescoping the above inequality over m from 1 to j, we obtain

$$\mathbb{E}_{\bar{S}_j^i}\|w_{k,j}^i-\widetilde{w}_{k,j}^i\|^2\leq\left((1+2\alpha L+2\alpha^2L^2)^j-1\right)\frac{\alpha\sigma_g^2}{L(1+\alpha L)S}.$$

Then,taking the expectation over w_k in the above inequality finishes the proof.

Proof of Proposition 3

Recall the definition that

$$\widehat{G}_i(w_k) = \prod_{j=0}^{N-1} (I - \alpha\nabla^2 l_i(w_{k,j}^i; D_{k,j}^i))\nabla l_i(w_{k,N}^i; T_k^i).$$

Then, conditioning on w_k, we have

$$\begin{aligned}
\mathbb{E}\widehat{G}_i(w_k) =& \mathbb{E}_{\bar{S}_N, i\sim p(\mathcal{T})}\mathbb{E}_{\bar{D}_N}\Big(\prod_{j=0}^{N-1}\big(I - \alpha\nabla^2 l_i(w_{k,j}^i; D_{k,j}^i)\big)\mathbb{E}_{T_k^i}\nabla l_i(w_{k,N}^i; T_k^i)\big|\bar{S}_N, i\Big) \\
=& \mathbb{E}_{\bar{S}_N, i\sim p(\mathcal{T})}\prod_{j=0}^{N-1}\mathbb{E}_{D_{k,j}^i}\big(I - \alpha\nabla^2 l_i(w_{k,j}^i; D_{k,j}^i)\big|\bar{S}_N, i\big)\nabla l_i(w_{k,N}^i) \\
=& \mathbb{E}_{\bar{S}_N, i\sim p(\mathcal{T})}\prod_{j=0}^{N-1}\big(I - \alpha\nabla^2 l_i(w_{k,j}^i)\big)\nabla l_i(w_{k,N}^i), \qquad (29)
\end{aligned}$$

which, combined with $\nabla\mathcal{L}(w_k) = \mathbb{E}_{i\sim p(\mathcal{T})}\prod_{j=0}^{N-1}(I - \alpha\nabla^2 l_i(\widetilde{w}_{k,j}^i))\nabla l_i(\widetilde{w}_{k,N}^i)$, yields

$$\begin{aligned}
&\|\mathbb{E}\widehat{G}_i(w_k) - \nabla\mathcal{L}(w_k)\| \\
&\quad\overset{(i)}{\leq}\mathbb{E}_{\bar{S}_N, i\sim p(\mathcal{T})}\Big\|\prod_{j=0}^{N-1}\big(I - \alpha\nabla^2 l_i(w_{k,j}^i)\big)\nabla l_i(w_{k,N}^i) - \prod_{j=0}^{N-1}(I - \alpha\nabla^2 l_i(\widetilde{w}_{k,j}^i))\nabla l_i(\widetilde{w}_{k,N}^i)\Big\| \\
&\quad\leq\mathbb{E}_{\bar{S}_N, i\sim p(\mathcal{T})}\Big\|\prod_{j=0}^{N-1}\big(I - \alpha\nabla^2 l_i(w_{k,j}^i)\big)\nabla l_i(w_{k,N}^i) - \prod_{j=0}^{N-1}(I - \alpha\nabla^2 l_i(w_{k,j}^i))\nabla l_i(\widetilde{w}_{k,N}^i)\Big\| \\
&\qquad+\mathbb{E}_{\bar{S}_N, i\sim p(\mathcal{T})}\Big\|\prod_{j=0}^{N-1}\big(I - \alpha\nabla^2 l_i(w_{k,j}^i)\big)\nabla l_i(\widetilde{w}_{k,N}^i) - \prod_{j=0}^{N-1}(I - \alpha\nabla^2 l_i(\widetilde{w}_{k,j}^i))\nabla l_i(\widetilde{w}_{k,N}^i)\Big\| \\
&\quad\leq\mathbb{E}_{\bar{S}_N, i}\Big\|\prod_{j=0}^{N-1}\big(I - \alpha\nabla^2 l_i(w_{k,j}^i)\big) - \prod_{j=0}^{N-1}(I - \alpha\nabla^2 l_i(\widetilde{w}_{k,j}^i))\Big\|\|\nabla l_i(\widetilde{w}_{k,N}^i)\| \\
&\qquad+(1+\alpha L)^N\mathbb{E}_{\bar{S}_N, i}\Big\|\nabla l_i(w_{k,N}^i) - \nabla l_i(\widetilde{w}_{k,N}^i)\Big\| \\
&\quad\overset{(ii)}{\leq}(1+\alpha L)^N\mathbb{E}_{\bar{S}_N, i}\|\nabla l_i(w_k)\|\Big\|\prod_{j=0}^{N-1}\big(I - \alpha\nabla^2 l_i(w_{k,j}^i)\big) - \prod_{j=0}^{N-1}(I - \alpha\nabla^2 l_i(\widetilde{w}_{k,j}^i))\Big\| \\
&\qquad+(1+\alpha L)^N L\mathbb{E}_{\bar{S}_N, i}\|w_{k,N}^i - \widetilde{w}_{k,N}^i\| \\
&\quad\overset{(iii)}{\leq}(1+\alpha L)^N\mathbb{E}_i\|\nabla l_i(w_k)\|\underbrace{\mathbb{E}_{\bar{S}_N}\Big(\Big\|\prod_{j=0}^{N-1}\big(I - \alpha\nabla^2 l_i(w_{k,j}^i)\big) - \prod_{j=0}^{N-1}(I - \alpha\nabla^2 l_i(\widetilde{w}_{k,j}^i))\Big\|\,\Big|\,i\Big)}_{R(N)} \\
&\qquad+(1+\alpha L)^N\big((1+\alpha L)^N - 1\big)\frac{\sigma_g}{\sqrt{S}}, \qquad (30)
\end{aligned}$$

where (i) follows from the Jensen's inequality, (ii) follows from Lemma 12 that $\|\nabla l_i(\widetilde{w}^i_{k,N})\| \leq (1+\alpha L)^N \|\nabla l_i(w_k)\|$, and (iii) follows from item 1 in Proposition 2. Our next step is to upper-bound the term $R(N)$. To simplify notations, we define a general quantity $R(m)$ by replacing N in $R(N)$ with m, and we use $\mathbb{E}_{\bar{S}_m|i}(\cdot)$ to denote $\mathbb{E}_{\bar{S}_m}(\cdot|i)$. Then, we have

$$
\begin{aligned}
R(m) \leq & \mathbb{E}_{\bar{S}_m|i}\Big\| \prod_{j=0}^{m-1}\big(I-\alpha\nabla^2 l_i(w^i_{k,j})\big) - \prod_{j=0}^{m-2}(I-\alpha\nabla^2 l_i(w^i_{k,j}))(I-\alpha\nabla^2 l_i(\widetilde{w}^i_{k,m-1})\Big\| \\
& + \mathbb{E}_{\bar{S}_m|i}\Big\| \prod_{j=0}^{m-2}(I-\alpha\nabla^2 l_i(w^i_{k,j}))(I-\alpha\nabla^2 l_i(\widetilde{w}^i_{k,m-1}) - \prod_{j=0}^{m-1}(I-\alpha\nabla^2 l_i(\widetilde{w}^i_{k,j}))\Big\| \\
\leq & (1+\alpha L)^{m-1}\alpha\rho\mathbb{E}_{\bar{S}_m|i}\|w^i_{k,m-1}-\widetilde{w}^i_{k,m-1}\| + (1+\alpha L)R(m-1) \\
\overset{(i)}{\leq} & \alpha\rho(1+\alpha L)^{m-1}((1+\alpha L)^{m-1}-1)\frac{\sigma_g}{L\sqrt{S}} + (1+\alpha L)R(m-1) \\
\leq & \alpha\rho(1+\alpha L)^{N-1}\big((1+\alpha L)^{N-1}-1\big)\frac{\sigma_g}{L\sqrt{S}} + (1+\alpha L)R(m-1), \qquad (31)
\end{aligned}
$$

where (i) follows from Proposition 2. Telescoping the above inequality over m from 2 to N and using $R(1)=0$, we have

$$
R(N) \leq ((1+\alpha L)^{N-1}-1)^2(1+\alpha L)^{N-1}\frac{\rho\sigma_g}{L^2\sqrt{S}}. \qquad (32)
$$

Thus, conditioning on w_k and combining (32) and (30), we have

$$
\begin{aligned}
\|\mathbb{E}\widehat{G}_i(w_k) - \nabla\mathcal{L}(w_k)\| \leq & ((1+\alpha L)^{N-1}-1)^2\frac{\rho}{L}(1+\alpha L)^{2N-1}\frac{\sigma_g}{L\sqrt{S}}\mathbb{E}_{i\sim p(\mathcal{T})}\big(\|\nabla l_i(w_k)\|\big) \\
& + \frac{(1+\alpha L)^N((1+\alpha L)^N-1)\sigma_g}{\sqrt{S}} \\
\leq & ((1+\alpha L)^{N-1}-1)^2\frac{\rho}{L}(1+\alpha L)^{2N-1}\frac{\sigma_g}{L\sqrt{S}}\Big(\frac{\|\nabla\mathcal{L}(w_k)\|}{1-C_l}+\frac{\sigma}{1-C_l}\Big) \\
& + \frac{(1+\alpha L)^N((1+\alpha L)^N-1)\sigma_g}{\sqrt{S}},
\end{aligned}
$$

where the last inequality follows from Lemma 15. Rearranging the above inequality and using C_{err_1} and C_{err_2} defined in Proposition 3 finish the proof.

Proof of Proposition 4

Recall $\widehat{G}_i(w_k) = \prod_{j=0}^{N-1}(I-\alpha\nabla^2 l_i(w^i_{k,j};D^i_{k,j}))\nabla l_i(w^i_{k,N};T^i_k)$. Conditioning on w_k, we have

$$
\begin{aligned}
& \mathbb{E}\|\widehat{G}_i(w_k)\|^2 \\
& \leq \mathbb{E}_{\bar{S}_N,i}\Big(\mathbb{E}_{\bar{D}_N,T^i_k}\Big(\Big\|\prod_{j=0}^{N-1}(I-\alpha\nabla^2 l_i(w^i_{k,j};D^i_{k,j}))\Big\|^2\|\nabla l_i(w^i_{k,N};T^i_k)\|^2\Big|\bar{S}_N,i\Big)\Big) \\
& \leq \underbrace{\mathbb{E}_{\bar{S}_N,i}\Big(\prod_{j=0}^{N-1}\mathbb{E}_{\bar{D}_N}\Big(\Big\|I-\alpha\nabla^2 l_i(w^i_{k,j};D^i_{k,j})\Big\|^2\Big|\bar{S}_N,i\Big)\Big)}_{P}\underbrace{\mathbb{E}_{T^i_k}\Big(\|\nabla l_i(w^i_{k,N};T^i_k)\|^2\Big|\bar{S}_N,i\Big)}_{Q}\Big). \qquad (33)
\end{aligned}
$$

We next upper-bound P and Q in (33). Note that $w_{k,j}^i, j = 0, ..., N-1$ are deterministic when conditioning on S_N, i, and w_k. Thus, conditioning on S_N, i, and w_k, we have

$$\begin{aligned}\mathbb{E}_{\bar{D}_N}\Big\|I - \alpha\nabla^2 l_i(w_{k,j}^i; D_{k,j}^i)\Big\|^2 =& \mathrm{Var}\Big(I - \alpha\nabla^2 l_i(w_{k,j}^i; D_{k,j}^i)\Big) + \big\|I - \alpha\nabla^2 l_i(w_{k,j}^i)\big\|^2 \\ \leq& \frac{\alpha^2\sigma_H^2}{D} + (1+\alpha L)^2. \end{aligned} \tag{34}$$

We next bound Q term. Conditioning on $\bar{S}_N, i$ and w_k, we have

$$\begin{aligned}\mathbb{E}_{T_k^i}\|\nabla l_i(w_{k,N}^i; T_k^i)\|^2 \overset{(i)}{\leq}& 3\mathbb{E}_{T_k^i}\|\nabla l_i(w_{k,N}^i; T_k^i) - \nabla l_i(w_{k,N}^i)\|^2 + 3\mathbb{E}_{T_k^i}\|\nabla l_i(w_{k,N}^i) - \nabla l_i(\widetilde{w}_{k,N}^i)\|^2 \\ &+ 3\mathbb{E}_{T_k^i}\|\nabla l_i(\widetilde{w}_{k,N}^i)\|^2 \\ \overset{(ii)}{\leq}& \frac{3\sigma_g^2}{T} + 3L^2\|w_{k,N}^i - \widetilde{w}_{k,N}^i\|^2 + 3(1+\alpha L)^{2N}\|\nabla l_i(w_k)\|^2, \end{aligned} \tag{35}$$

where (i) follows from the inequality that $\|\sum_{i=1}^n a\|^2 \leq n\sum_{i=1}^n \|a\|^2$, and (ii) follows from Lemma 12. Thus, conditioning on w_k and combining (33), (34) and (35), we have

$$\mathbb{E}\|\widehat{G}_i(w_k)\|^2 \leq 3\Big(\frac{\alpha^2\sigma_H^2}{D} + (1+\alpha L)^2\Big)^N\Big(\frac{\sigma_g^2}{T} + L^2\mathbb{E}\|w_{k,N}^i - \widetilde{w}_{k,N}^i\|^2 + (1+\alpha L)^{2N}\mathbb{E}\|\nabla l_i(w_k)\|^2\Big)$$

which, in conjunction with Proposition 2, yields

$$\mathbb{E}\|\widehat{G}_i(w_k)\|^2 \leq 3(1+\alpha L)^{2N}\Big(\frac{\alpha^2\sigma_H^2}{D} + (1+\alpha L)^2\Big)^N(\|\nabla l(w_k)\|^2 + \sigma^2) + \frac{C_{\mathrm{squ}_1}}{T} + \frac{C_{\mathrm{squ}_2}}{S}. \tag{36}$$

Based on Lemma 15 and conditioning on w_k, we have

$$\|\nabla l(w_k)\|^2 \leq \frac{2}{(1-C_l)^2}\|\nabla\mathcal{L}(w_k)\| + \frac{2C_l^2}{(1-C_l)^2}\sigma^2,$$

which, in conjunction with $\frac{2x^2}{(1-x)^2} + 1 \leq \frac{2}{(1-x)^2}$ and (36), finishes the proof.

Proof of Theorem 5

The proof of Theorem 5 consists of four main steps: step 1 of bounding an iterative meta update by the meta-gradient smoothness established by Proposition 1; step 2 of characterizing first-moment error of the meta-gradient estimator $\widehat{G}_i(w_k)$ by Proposition 3; step 3 of characterizing second-moment error of the meta-gradient estimator $\widehat{G}_i(w_k)$ by Proposition 4; and step 4 of combining steps 1-3, and telescoping to yield the convergence.

To simplify notations, define the smoothness parameter of the meta-gradient as

$$L_{w_k} = (1+\alpha L)^{2N}L + C_{\mathcal{L}}\mathbb{E}_{i\sim p(\mathcal{T})}\|\nabla l_i(w_k)\|,$$

where $C_{\mathcal{L}}$ is given in (13). Based on the smoothness of the gradient $\nabla\mathcal{L}(w)$ given by Proposition 1, we have

$$\mathcal{L}(w_{k+1}) \leq \mathcal{L}(w_k) + \langle\nabla\mathcal{L}(w), w_{k+1} - w_k\rangle + \frac{L_{w_k}}{2}\|w_{k+1} - w_k\|^2$$

Note that the randomness from β_k depends on B'_k and $D^i_{L_k}, i \in B'_k$, and thus is independent of $S^i_{k,j}, D^i_{k,j}$ and T^i_k for $i \in B_k, j = 0, ..., N$. Then, taking expectation over the above inequality, conditioning on w_k, and recalling $e_k := \mathbb{E}\widehat{G}_i(w_k) - \nabla\mathcal{L}(w_k)$, we have

$$\mathbb{E}(\mathcal{L}(w_{k+1})|w_k) \leq \mathcal{L}(w_k) - \mathbb{E}(\beta_k)\langle\nabla\mathcal{L}(w_k), \nabla\mathcal{L}(w_k) + e_k\rangle + \frac{L_{w_k}\mathbb{E}(\beta_k^2)\mathbb{E}\big\|\frac{1}{B}\sum_{i\in B_k}\widehat{G}_i(w_k)\big\|^2}{2}.$$

Then, applying Lemma 16 in the above inequality yields

$$\begin{aligned}
\mathbb{E}(\mathcal{L}(w_{k+1})|w_k) \leq& \mathcal{L}(w_k) - \frac{4}{5C_\beta}\frac{1}{L_{w_k}}\|\nabla\mathcal{L}(w_k)\|^2 + \frac{4}{5C_\beta}\frac{1}{L_{w_k}}|\langle\nabla\mathcal{L}(w_k), e_k\rangle| \\
&+ \frac{2}{C_\beta^2}\frac{1}{L_{w_k}}\Big(\frac{1}{B}\mathbb{E}\big\|\widehat{G}_i(w_k)\big\|^2 + \|\mathbb{E}\widehat{G}_i(w_k)\|^2\Big). \\
\leq& \mathcal{L}(w_k) - \frac{4}{5C_\beta}\frac{1}{L_{w_k}}\|\nabla\mathcal{L}(w_k)\|^2 + \frac{2}{5C_\beta}\frac{1}{L_{w_k}}\|\nabla\mathcal{L}(w_k)\|^2 + \frac{2}{5C_\beta}\frac{1}{L_{w_k}}\|e_k\|^2 \\
&+ \frac{2}{C_\beta^2}\frac{1}{L_{w_k}}\Big(\frac{1}{B}\mathbb{E}\big\|\widehat{G}_i(w_k)\big\|^2 + \|\mathbb{E}\widehat{G}_i(w_k)\|^2\Big). \qquad (37)
\end{aligned}$$

Then, applying Propositions 3 and 4 to the above inequality yields

$$\begin{aligned}
&\mathbb{E}(\mathcal{L}(w_{k+1})|w_k) \\
&\leq \mathcal{L}(w_k) - \frac{2}{5C_\beta}\frac{1}{L_{w_k}}\|\nabla\mathcal{L}(w_k)\|^2 + \frac{2}{C_\beta^2}\frac{1}{L_{w_k}}\frac{1}{B}\mathbb{E}\big\|\widehat{G}_i(w_k)\big\|^2 + \frac{4}{C_\beta^2}\frac{1}{L_{w_k}}\|\nabla\mathcal{L}(w_k)\|^2 \\
&\quad + \Big(\frac{6}{5C_\beta L_{w_k}} + \frac{12}{C_\beta^2 L_{w_k}}\Big)\Big(\frac{C^2_{\mathrm{err}_2}}{S}\|\nabla\mathcal{L}(w_k)\|^2 + \frac{C^2_{\mathrm{err}_1}}{S} + \frac{C^2_{\mathrm{err}_2}\sigma^2}{S}\Big) \\
&\leq \mathcal{L}(w_k) - \frac{2}{C_\beta L_{w_k}}\left(\frac{1}{5} - \left(\frac{3}{5} + \frac{6}{C_\beta}\right)\frac{C^2_{\mathrm{err}_2}}{S} - \frac{C_{\mathrm{squ}_3}}{C_\beta B} - \frac{2}{C_\beta}\right)\|\nabla\mathcal{L}(w_k)\|^2 \\
&\quad + \frac{6}{C_\beta L_{w_k} S}\Big(\frac{1}{5} + \frac{2}{C_\beta}\Big)\Big(C^2_{\mathrm{err}_1} + C^2_{\mathrm{err}_2}\sigma^2\Big) + \frac{2}{C_\beta^2 L_{w_k} B}\Big(\frac{C_{\mathrm{squ}_1}}{T} + \frac{C_{\mathrm{squ}_2}}{S} + C_{\mathrm{squ}_3}\sigma^2\Big). \qquad (38)
\end{aligned}$$

Recalling $L_{w_k} = (1+\alpha L)^{2N}L + C_\mathcal{L}\mathbb{E}_i\|\nabla l_i(w_k)\|$, we have $L_{w_k} \geq L$ and

$$L_{w_k} \overset{(i)}{\leq} (1+\alpha L)^{2N}L + \frac{C_\mathcal{L}\sigma}{1-C_l} + \frac{C_\mathcal{L}}{1-C_l}\|\nabla\mathcal{L}(w_k)\|, \qquad (39)$$

where (i) follows from Assumption 2 and Lemma 15. Combining (38) and (39) yields

$$\begin{aligned}
\mathbb{E}(\mathcal{L}(w_{k+1})|w_k) \leq& \mathcal{L}(w_k) + \frac{6}{C_\beta L}\Big(\frac{1}{5} + \frac{2}{C_\beta}\Big)\Big(C^2_{\mathrm{err}_1} + C^2_{\mathrm{err}_2}\sigma^2\Big)\frac{1}{S} \\
&+ \frac{2}{C_\beta^2 L}\Big(\frac{C_{\mathrm{squ}_1}}{T} + \frac{C_{\mathrm{squ}_2}}{S} + C_{\mathrm{squ}_3}\sigma^2\Big)\frac{1}{B} \\
&- \frac{2}{C_\beta}\frac{\frac{1}{5} - \left(\frac{3}{5} + \frac{6}{C_\beta}\right)\frac{C^2_{\mathrm{err}_2}}{S} - \frac{C_{\mathrm{squ}_3}}{C_\beta B} - \frac{2}{C_\beta}}{(1+\alpha L)^{2N}L + \frac{C_\mathcal{L}\sigma}{1-C_l} + \frac{C_\mathcal{L}}{1-C_l}\|\nabla\mathcal{L}(w_k)\|}\|\nabla\mathcal{L}(w_k)\|^2. \qquad (40)
\end{aligned}$$

Based on the notations in (19), we rewrite (40) as

$$\mathbb{E}(\mathcal{L}(w_{k+1})|w_k) \leq \mathcal{L}(w_k) + \frac{\xi}{S} + \frac{\phi}{B} - \theta \frac{\|\nabla \mathcal{L}(w_k)\|^2}{\chi + \|\nabla \mathcal{L}(w_k)\|}.$$

Unconditioning on w_k in the above inequality and telescoping the above inequality over k from 0 to $K-1$, we have

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left(\frac{\theta \|\nabla \mathcal{L}(w_k)\|^2}{\chi + \|\nabla \mathcal{L}(w_k)\|}\right) \leq \frac{\Delta}{K} + \frac{\xi}{S} + \frac{\phi}{B}, \tag{41}$$

where $\Delta = \mathcal{L}(w_0) - \mathcal{L}^*$. Choosing ζ from $\{0, ..., K-1\}$ uniformly at random, we obtain from (41) that

$$\mathbb{E}\left(\frac{\theta \|\nabla \mathcal{L}(w_\zeta)\|^2}{\chi + \|\nabla \mathcal{L}(w_\zeta)\|}\right) \leq \frac{\Delta}{K} + \frac{\xi}{S} + \frac{\phi}{B}. \tag{42}$$

Consider a function $f(x) = \frac{x^2}{c+x}$, $x > 0$, where $c > 0$ is a constant. Simple computation shows that $f''(x) = \frac{2c^2}{(x+c)^3} > 0$. Thus, using Jensen's inequality in (42), we have

$$\frac{\theta (\mathbb{E}\|\nabla \mathcal{L}(w_\zeta)\|)^2}{\chi + \mathbb{E}\|\nabla \mathcal{L}(w_\zeta)\|} \leq \frac{\Delta}{K} + \frac{\xi}{S} + \frac{\phi}{B}. \tag{43}$$

Rearranging the above inequality yields

$$\begin{aligned}
\mathbb{E}\|\nabla \mathcal{L}(w_\zeta)\| \leq & \frac{\Delta}{2\theta}\frac{1}{K} + \frac{\xi}{2\theta}\frac{1}{S} + \frac{\phi}{2\theta}\frac{1}{B} + \sqrt{\chi\Big(\frac{\Delta}{2\theta}\frac{1}{K} + \frac{\xi}{2\theta}\frac{1}{S} + \frac{\phi}{2\theta}\frac{1}{B}\Big) + \Big(\frac{\Delta}{2\theta}\frac{1}{K} + \frac{\xi}{2\theta}\frac{1}{S} + \frac{\phi}{2\theta}\frac{1}{B}\Big)^2} \\
\leq & \frac{\Delta}{\theta}\frac{1}{K} + \frac{\xi}{\theta}\frac{1}{S} + \frac{\phi}{\theta}\frac{1}{B} + \sqrt{\frac{\chi}{2}}\sqrt{\frac{\Delta}{\theta}\frac{1}{K} + \frac{\xi}{\theta}\frac{1}{S} + \frac{\phi}{\theta}\frac{1}{B}},
\end{aligned} \tag{44}$$

which finishes the proof.

Proof of Corollary 6

Since $\alpha = \frac{1}{8NL}$, we have

$$(1+\alpha L)^N = \big(1 + \frac{1}{8N}\big)^N = e^{N\log(1+\frac{1}{8N})} \leq e^{1/8} < \frac{5}{4}, (1+\alpha L)^{2N} < e^{1/4} < \frac{3}{2},$$

which, in conjunction with (15), implies that

$$C_{\mathrm{err}_1} < \frac{5\sigma_g}{16}, \quad C_{\mathrm{err}_2} < \frac{3\rho\sigma_g}{4L^2}. \tag{45}$$

Furthermore, noting that $D \geq \sigma_H^2/L^2$, we have

$$C_{\mathrm{squ}_1} \leq 3(1 + 2\alpha L + 2\alpha^2 L^2)^N \sigma_g^2 < 3e^{9/32}\sigma_g^2 < 4\sigma_g^2,\ C_{\mathrm{squ}_2} < \frac{1.3\sigma_g^2}{8} < \frac{\sigma_g^2}{5},\ C_{\mathrm{squ}_3} \leq 11. \tag{46}$$

Based on (13), we have

$$C_{\mathcal{L}} < \frac{75}{128}\frac{\rho}{L} < \frac{3}{5}\frac{\rho}{L} \text{ and } C_{\mathcal{L}} \overset{(i)}{>} \frac{\rho}{L}((N-1)\alpha L) > \frac{1}{16}\frac{\rho}{L}, \tag{47}$$

where (i) follows from the inequality that $(1+a)^n > 1+an$. Then, using (45), (46) and (47), we obtain from (19) that

$$\begin{aligned}
&\xi < \frac{7}{500L}\Big(\frac{1}{10} + \frac{9\rho\sigma^2}{16L^4}\Big)\sigma_g^2, \quad \phi \le \frac{1}{5000L}\Big(\frac{3\sigma_g^2}{T} + \frac{\sigma_g^2}{5S} + 11\sigma^2\Big) < \frac{1}{1000L}(\sigma_g^2 + 3\sigma^2) \\
&\theta \ge \frac{L}{60\rho}\Big(\frac{1}{5} - \frac{4}{5}\frac{9}{16}\frac{\rho^2\sigma_g^2}{L^4}\frac{1}{S} - \frac{11}{100B} - \frac{1}{50}\Big) = \frac{L}{1500\rho}, \ \chi \le \frac{24L^2}{\rho} + \sigma.
\end{aligned} \tag{48}$$

Then, treating Δ, ρ, L as constants and using (20), we obtain

$$\mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\| \le \mathcal{O}\Big(\frac{1}{K} + \frac{\sigma_g^2(\sigma^2+1)}{S} + \frac{\sigma_g^2+\sigma^2}{B} + \frac{\sigma_g^2}{TB} + \sqrt{\sigma+1}\sqrt{\frac{1}{K} + \frac{\sigma_g^2(\sigma^2+1)}{S} + \frac{\sigma_g^2+\sigma^2}{B} + \frac{\sigma_g^2}{TB}}\Big).$$

Then, choosing batch sizes $S \ge C_S\sigma_g^2(\sigma^2+1)\max(\sigma,1)\epsilon^{-2}$, $B \ge C_B(\sigma_g^2+\sigma^2)\max(\sigma,1)\epsilon^{-2}$ and $TB > C_T\sigma_g^2\max(\sigma,1)\epsilon^{-2}$, we have

$$\mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\| \le \mathcal{O}\left(\frac{1}{K} + \frac{1}{\epsilon^2}\Big(\frac{1}{C_S} + \frac{1}{C_B} + \frac{1}{C_T}\Big) + \sqrt{\sigma}\sqrt{\frac{1}{K} + \frac{1}{\sigma\epsilon^2}\Big(\frac{1}{C_S} + \frac{1}{C_B} + \frac{1}{C_T}\Big)}\right)$$

After at most $K = C_K\max(\sigma,1)\epsilon^{-2}$ iterations, the above inequality implies, for constants C_S, C_B, C_T and C_K large enough, $\mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\| \le \epsilon$. Recall that we need $|B_k'| > \frac{4C_{\mathcal{L}}^2\sigma^2}{3(1+\alpha L)^{4N}L^2}$ and $|D_{L_k}^i| > \frac{64\sigma_g^2C_{\mathcal{L}}^2}{(1+\alpha L)^{4N}L^2}$ for building stepsize β_k at each iteration k. Based on the selected parameters, we have

$$\frac{4C_{\mathcal{L}}^2\sigma^2}{3(1+\alpha L)^{4N}L^2} \le \frac{4\sigma^2}{3L^2}\frac{3\rho}{5L} \le \Theta(\sigma^2), \quad \frac{64\sigma_g^2C_{\mathcal{L}}^2}{(1+\alpha L)^{4N}L^2} < \Theta(\sigma_g^2),$$

which implies $|B_k'| = \Theta(\sigma^2)$ and $|D_{L_k}^i| = \Theta(\sigma_g^2)$. Then, since the batch size $D = \Theta(\sigma_H^2/L^2)$, the total number of gradient computations at each meta iteration k is given by $B(NS+T) + |B_k'||D_{L_k}^i| \le \mathcal{O}(N\epsilon^{-4} + \epsilon^{-2})$. Furthermore, the total number of Hessian computations at each meta iteration is given by $BND \le \mathcal{O}(N\epsilon^{-2})$. This completes the proof.

5.2 Proofs for Section 4: Convergence of Multi-Step MAML in Finite-Sum Case

In this subsection, we provide proofs for the convergence properties of multi-step MAML in the finite-sum case.

Proof of Proposition 7

By the definition of $\nabla\mathcal{L}_i(\cdot)$, we have

$$\begin{aligned}\|\nabla\mathcal{L}_i(w)-\nabla\mathcal{L}_i(u)\| \leq& \Big\| \prod_{j=0}^{N-1}(I-\alpha\nabla^2 l_{S_i}(\widetilde{w}_j^i))\nabla l_{T_i}(\widetilde{w}_N^i) - \prod_{j=0}^{N-1}(I-\alpha\nabla^2 l_{S_i}(\widetilde{u}_j^i))\nabla l_{T_i}(\widetilde{w}_N^i)\Big\| \\ &+ \Big\| \prod_{j=0}^{N-1}(I-\alpha\nabla^2 l_{S_i}(\widetilde{u}_j^i))\nabla l_{T_i}(\widetilde{w}_N^i) - \prod_{j=0}^{N-1}(I-\alpha\nabla^2 l_{S_i}(\widetilde{u}_j^i))\nabla l_{T_i}(\widetilde{u}_N^i)\Big\| \\ \leq& \underbrace{\Big\| \prod_{j=0}^{N-1}(I-\alpha\nabla^2 l_{S_i}(\widetilde{w}_j^i)) - \prod_{j=0}^{N-1}(I-\alpha\nabla^2 l_{S_i}(\widetilde{u}_j^i))\Big\|}_{A} \|\nabla l_{T_i}(\widetilde{w}_N^i)\| \\ &+ (1+\alpha L)^N \|\nabla l_{T_i}(\widetilde{w}_N^i) - \nabla l_{T_i}(\widetilde{u}_N^i)\|. \end{aligned} \tag{49}$$

We next upper-bound A in the above inequality. Specifically, we have

$$\begin{aligned} A \leq& \Big\| \prod_{j=0}^{N-1}(I-\alpha\nabla^2 l_{S_i}(\widetilde{w}_j^i)) - \prod_{j=0}^{N-2}(I-\alpha\nabla^2 l_{S_i}(\widetilde{w}_j^i))(I-\alpha\nabla^2 l_{S_i}(\widetilde{u}_{N-1}^i))\Big\| \\ &+ \Big\| \prod_{j=0}^{N-2}(I-\alpha\nabla^2 l_{S_i}(\widetilde{w}_j^i))(I-\alpha\nabla^2 l_{S_i}(\widetilde{u}_{N-1}^i)) - \prod_{j=0}^{N-1}(I-\alpha\nabla^2 l_{S_i}(\widetilde{u}_j^i))\Big\| \\ \leq& \Big((1+\alpha L)^{N-1}\alpha\rho + \frac{\rho}{L}(1+\alpha L)^N\big((1+\alpha L)^{N-1}-1\big)\Big)\|w-u\|, \end{aligned} \tag{50}$$

where the last inequality uses an approach similar to (28). Combining (49) and (50) yields

$$\begin{aligned} &\|\nabla\mathcal{L}_i(w)-\nabla\mathcal{L}_i(u)\| \\ &\quad\leq\big((1+\alpha L)^{N-1}\alpha\rho + \frac{\rho}{L}(1+\alpha L)^N\big((1+\alpha L)^{N-1}-1\big)\big)\|w-u\|\|\nabla l_{T_i}(\widetilde{w}_N^i)\| \\ &\qquad+(1+\alpha L)^N L\|\widetilde{w}_N^i-\widetilde{u}_N^i\|. \end{aligned} \tag{51}$$

To upper-bound $\|\nabla l_{T_i}(\widetilde{w}_N^i)\|$ in (51), using the mean value theorem, we have

$$\begin{aligned} \|\nabla l_{T_i}(\widetilde{w}_N^i)\| =& \Big\|\nabla l_{T_i}(w-\sum_{j=0}^{N-1}\alpha\nabla l_{S_i}(\widetilde{w}_j^i))\Big\| \\ \overset{(i)}{\leq}& \|\nabla l_{T_i}(w)\| + \alpha L\sum_{j=0}^{N-1}(1+\alpha L)^j\big\|\nabla l_{S_i}(w)\big\| \\ \overset{(ii)}{\leq}& (1+\alpha L)^N\|\nabla l_{T_i}(w)\| + \big((1+\alpha L)^N-1\big)b_i, \end{aligned} \tag{52}$$

where (i) follows from Lemma 17, and (ii) follows from Assumption 5. In addition, using an approach similar to Lemma 11, we have

$$\|\widetilde{w}_N^i-\widetilde{u}_N^i\| \leq (1+\alpha L)^N\|w-u\|. \tag{53}$$

Combining (51), (52) and (53) yields

$$\begin{aligned}
&\|\nabla\mathcal{L}_i(w)-\nabla\mathcal{L}_i(u)\|\\
&\quad\le\Big((1+\alpha L)^{N-1}\alpha\rho+\frac{\rho}{L}(1+\alpha L)^N\big((1+\alpha L)^{N-1}-1\big)\Big)(1+\alpha L)^N\|\nabla l_{T_i}(w)\|\|w-u\|\\
&\quad\quad+\Big((1+\alpha L)^{N-1}\alpha\rho+\frac{\rho}{L}(1+\alpha L)^N\big((1+\alpha L)^{N-1}-1\big)\Big)\big((1+\alpha L)^N-1\big)b_i\|w-u\|\\
&\quad\quad+(1+\alpha L)^{2N}L\|w-u\|,
\end{aligned}$$

which, in conjunction with C_b and $C_{\mathcal{L}}$ given in (21), yields

$$\|\nabla\mathcal{L}_i(w)-\nabla\mathcal{L}_i(u)\|\le\big((1+\alpha L)^{2N}L+C_bb_i+C_{\mathcal{L}}\|\nabla l_{T_i}(w)\|\big)\|w-u\|.$$

Based on the above inequality and Jensen's inequality, we finish the proof.

Proof of Proposition 8

Conditioning on w_k, we have

$$\mathbb{E}\|\widehat{G}_i(w_k)\|^2=\mathbb{E}\Big\|\prod_{j=0}^{N-1}(I-\alpha\nabla^2 l_{S_i}(w_{k,j}^i))\nabla l_{T_i}(w_{k,N}^i)\Big\|^2\le(1+\alpha L)^{2N}\mathbb{E}\|\nabla l_{T_i}(w_{k,N}^i)\|^2,$$

which, using an approach similar to (52), yields

$$\begin{aligned}
\mathbb{E}\|\widehat{G}_i(w_k)\|^2&\le(1+\alpha L)^{2N}2(1+\alpha L)^{2N}\mathbb{E}\|\nabla l_{T_i}(w_k)\|^2+2(1+\alpha L)^{2N}\big((1+\alpha L)^N-1\big)^2\mathbb{E}_ib_i^2\\
&\le2(1+\alpha L)^{4N}(\|\nabla l_T(w_k)\|^2+\sigma^2)+2(1+\alpha L)^{2N}\big((1+\alpha L)^N-1\big)^2\widetilde{b}\\
&\overset{(i)}{\le}2(1+\alpha L)^{4N}\Big(\frac{2}{C_1^2}\|\nabla l_T(w_k)\|^2+\frac{2C_2^2}{C_1^2}+\sigma^2\Big)+2(1+\alpha L)^{2N}\big((1+\alpha L)^N-1\big)^2\widetilde{b}\\
&\le\frac{4(1+\alpha L)^{4N}}{C_1^2}\|\nabla l_T(w_k)\|^2+\frac{4(1+\alpha L)^{4N}C_2^2}{C_1^2}+2(1+\alpha L)^{4N}(\sigma^2+\widetilde{b}),
\end{aligned}\tag{54}$$

where (i) follows from Lemma 19, and constants C_1 and C_2 are given by (74). Noting that $C_2=\big((1+\alpha L)^{2N}-1\big)\sigma+(1+\alpha L)^N\big((1+\alpha L)^N-1\big)b<\big((1+\alpha L)^{2N}-1\big)(\sigma+b)$ and using the definitions of $A_{\mathrm{squ}_1},A_{\mathrm{squ}_2}$ in (23), we finish the proof.

Proof of Theorem 9

Based on the smoothness of $\nabla\mathcal{L}(\cdot)$ established in Proposition 7, we have

$$\mathcal{L}(w_{k+1})\le\mathcal{L}(w_k)-\beta_k\Big\langle\nabla\mathcal{L}(w_k),\frac{1}{B}\sum_{i\in B_k}\widehat{G}_i(w_k)\Big\rangle+\frac{L_{w_k}\beta_k^2}{2}\Big\|\frac{1}{B}\sum_{i\in B_k}\widehat{G}_i(w_k)\Big\|^2$$

Taking the conditional expectation given w_k over the above inequality and noting that the randomness over β_k is independent of the randomness over $\widehat{G}_i(w_k)$, we have

$$\begin{aligned}
&\mathbb{E}(\mathcal{L}(w_{k+1})|w_k)\\
&\quad\le\mathcal{L}(w_k)-\frac{1}{C_\beta}\mathbb{E}\Big(\frac{1}{\hat{L}_{w_k}}\,\Big|\,w_k\Big)\|\nabla\mathcal{L}(w_k)\|^2+\frac{L_{w_k}}{2C_\beta^2}\mathbb{E}\Big(\frac{1}{\hat{L}_{w_k}^2}\,\Big|\,w_k\Big)\mathbb{E}\Big(\Big\|\frac{1}{B}\sum_{i\in B_k}\widehat{G}_i(w_k)\Big\|^2\Big|w_k\Big).
\end{aligned}\tag{55}$$

Note that, conditioning on w_k,

$$\mathbb{E}\Big\|\frac{1}{B}\sum_{i\in B_k}\widehat{G}_i(w_k)\Big\|^2 \leq \frac{1}{B}\big(A_{\mathrm{squ}_1}\|\nabla\mathcal{L}(w_k)\|^2 + A_{\mathrm{squ}_2}\big) + \|\nabla\mathcal{L}(w_k)\|^2 \tag{56}$$

where the inequality follows from Proposition 8. Then, combining (56), (55) and applying Lemma 20, we have

$$\mathbb{E}(\mathcal{L}(w_{k+1})|w_k) \leq \mathcal{L}(w_k) - \Big(\frac{1}{L_{w_k}C_\beta} - \frac{1}{L_{w_k}C_\beta^2}\Big(\frac{A_{\mathrm{squ}_1}}{B}+1\Big)\Big)\|\nabla\mathcal{L}(w_k)\|^2 + \frac{A_{\mathrm{squ}_2}}{L_{w_k}C_\beta^2 b}. \tag{57}$$

Recalling that $L_{w_k} = (1+\alpha L)^{2N}L + C_b b + C_{\mathcal{L}}\mathbb{E}_{i\sim p(\mathcal{T})}\|\nabla l_{T_i}(w_k)\|$ and conditioning on w_k, we have $L_{w_k} \geq L$ and

$$\begin{aligned} L_{w_k} \leq& (1+\alpha L)^{2N}L + C_b b + C_{\mathcal{L}}(\|\nabla l_T(w_k)\| + \sigma) \\ \overset{(i)}{\leq}& (1+\alpha L)^{2N}L + C_b b + C_{\mathcal{L}}\Big(\frac{C_2}{C_1}+\sigma\Big) + \frac{C_{\mathcal{L}}}{C_1}\|\nabla\mathcal{L}(w_k)\|, \end{aligned} \tag{58}$$

where (i) follows from Lemma 19. Combining (58) and (57) yields

$$\begin{aligned} &\mathbb{E}(\mathcal{L}(w_{k+1})|w_k) \\ \leq& \mathcal{L}(w_k) - \frac{\Big(\frac{1}{C_\beta} - \frac{1}{C_\beta^2}\Big(\frac{A_{\mathrm{squ}_1}}{B}+1\Big)\Big)\|\nabla\mathcal{L}(w_k)\|^2}{(1+\alpha L)^{2N}L + C_b b + C_{\mathcal{L}}\Big(\frac{C_2}{C_1}+\sigma\Big) + \frac{C_{\mathcal{L}}}{C_1}\|\nabla\mathcal{L}(w_k)\|} + \frac{1}{LC_\beta^2}\frac{A_{\mathrm{squ}_2}}{B} \\ =& \mathcal{L}(w_k) - \frac{\frac{C_1}{C_{\mathcal{L}}}\Big(\frac{1}{C_\beta} - \frac{1}{C_\beta^2}\Big(\frac{A_{\mathrm{squ}_1}}{B}+1\Big)\Big)\|\nabla\mathcal{L}(w_k)\|^2}{\frac{C_1}{C_{\mathcal{L}}}(1+\alpha L)^{2N}L + \frac{bC_1C_b}{C_{\mathcal{L}}} + C_2 + C_1\sigma + \|\nabla\mathcal{L}(w_k)\|} + \frac{1}{LC_\beta^2}\frac{A_{\mathrm{squ}_2}}{B} \\ =& \mathcal{L}(w_k) - \frac{\frac{C_1}{C_{\mathcal{L}}}\Big(\frac{1}{C_\beta} - \frac{1}{C_\beta^2}\Big(\frac{A_{\mathrm{squ}_1}}{B}+1\Big)\Big)\|\nabla\mathcal{L}(w_k)\|^2}{\frac{C_1}{C_{\mathcal{L}}}(1+\alpha L)^{2N}L + \frac{bC_1C_b}{C_{\mathcal{L}}} + (1+\alpha L)^N((1+\alpha L)^{2N}-1)b + \|\nabla\mathcal{L}(w_k)\|} + \frac{A_{\mathrm{squ}_2}}{LC_\beta^2 B}, \end{aligned} \tag{59}$$

where the last equality follows from the definitions of C_1, C_2 in (74). Combining the definitions in (24) with (59) and taking the expectation over w_k, we have

$$\mathbb{E}\frac{\theta\|\nabla\mathcal{L}(w_k)\|^2}{\xi + \|\nabla\mathcal{L}(w_k)\|} \leq \mathbb{E}(\mathcal{L}(w_k) - \mathcal{L}(w_{k+1})) + \frac{\phi}{B}.$$

Telescoping the above bound over k from 0 to $K-1$ and choosing ζ from $\{0, ..., K-1\}$ uniformly at random, we have

$$\mathbb{E}\frac{\theta\|\nabla\mathcal{L}(w_\zeta)\|^2}{\xi + \|\nabla\mathcal{L}(w_\zeta)\|} \leq \frac{\Delta}{K} + \frac{\phi}{B}. \tag{60}$$

Using an approach similar to (43), we obtain from (60) that

$$\frac{(\mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\|)^2}{\xi + \mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\|} \leq \frac{\Delta}{\theta K} + \frac{\phi}{\theta B},$$

which further implies that

$$\mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\| \leq \frac{\Delta}{2\theta K} + \frac{\phi}{2\theta B} + \sqrt{\xi\Big(\frac{\Delta}{\theta K} + \frac{\phi}{\theta B}\Big) + \Big(\frac{\Delta}{2\theta K} + \frac{\phi}{2\theta B}\Big)^2}, \tag{61}$$

which finishes the proof.

Proof of Corollary 10

Since $\alpha = \frac{1}{8NL}$, we have $(1+\alpha L)^{4N} < e^{0.5} < 2$, and thus

$$\begin{aligned} &A_{\text{squ}_1} < 32,\ A_{\text{squ}_2} < 8(\sigma+b)^2 + 4(\sigma^2+\widetilde{b}), \\ &C_{\mathcal{L}} < \Big(\frac{5\rho}{32NL} + \frac{\rho}{L}\frac{5}{16}\Big)\frac{5}{4} < \frac{5\rho}{8L},\ C_{\mathcal{L}} > \frac{\rho}{L}\big((1+\alpha L)^{N-1} - 1\big) > \frac{\rho}{L}\alpha L(N-1) > \frac{\rho}{16L}, \\ &C_b < \frac{15}{32}\frac{\rho}{L}\frac{1}{4} < \frac{\rho}{8L}, \end{aligned} \tag{62}$$

which, in conjunction with (24), yields

$$\theta \geq \frac{1}{80}\frac{4L}{5\rho}\Big(1 - \frac{33}{80}\Big) \geq \frac{L}{200\rho},\ \phi \leq \frac{2(\sigma+b)^2 + (\sigma^2+\widetilde{b})}{1600L},\ \xi \leq \frac{24L^2}{\rho} + \frac{37b}{16}. \tag{63}$$

Combining (63) and (25) yields

$$\begin{aligned} \mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\| \leq& \frac{\Delta}{2\theta K} + \frac{\phi}{2\theta B} + \sqrt{\xi\Big(\frac{\Delta}{\theta K} + \frac{\phi}{\theta B}\Big) + \Big(\frac{\Delta}{2\theta K} + \frac{\phi}{2\theta B}\Big)^2} \\ \leq& \mathcal{O}\Big(\frac{1}{K} + \frac{\sigma^2}{B} + \sqrt{\frac{1}{K} + \frac{\sigma^2}{B}}\Big). \end{aligned}$$

Then, based on the parameter selection that $B \geq C_B\sigma^2\epsilon^{-2}$ and after at most $K = C_k\epsilon^{-2}$ iterations, we have

$$\mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\| \leq \mathcal{O}\Big(\big(\frac{1}{C_B} + \frac{1}{C_k}\big)\frac{1}{\epsilon^2} + \frac{1}{\epsilon}\sqrt{\big(\frac{1}{C_B} + \frac{1}{C_k}\big)}\Big).$$

Then, for C_B, C_K large enough, we obtain from the above inequality that $\mathbb{E}\|\nabla\mathcal{L}(w_\zeta)\| \leq \epsilon$. Thus, the total number of gradient computations is given by $B(T+NS) = \mathcal{O}(\epsilon^{-2}(T+NS))$. Furthermore, the total number of Hessian computations is given by $BNS = \mathcal{O}(NS\epsilon^{-2})$ at each iteration. Then, the proof is complete.

6. Conclusion and Future Work

In this paper, we provide a new theoretical framework for analyzing the convergence of multi-step MAML algorithm for both the resampling case and the finite-sum case. Our analysis covers most applications including reinforcement learning and supervised learning of interest. Our analysis reveals that a properly chosen inner stepsize is crucial for guaranteeing MAML to converge with the complexity increasing only linearly with N (the number of the inner-stage gradient updates). Moreover, for problems with small Hessians, the inner stepsize can be set larger while maintaining the convergence. Our results also provide justifications for the empirical findings in training MAML.

We expect that our analysis framework can be applied to understand the convergence of MAML in other scenarios such as various RL problems and Hessian-free MAML algorithms.

Acknowledgments

The work was supported in part by the U.S. National Science Foundation under Grants CCF-1761506, ECCS-1818904, and CCF-1900145.

Appendices

Appendix A. Examples for Two Types of Objective Functions

A.1 RL Example for Resampling Case

RL problems are often captured by objective functions in the expectation form. Consider a RL meta learning problem, where each task corresponds to a Markov decision process (MDP) with horizon H. Each RL task $\mathcal{T}_i$ corresponds to an initial state distribution ρ_i, a policy π_w parameterized by w that denotes a distribution over the action set given each state, and a transition distribution kernel $q_i(x_{t+1}|x_t, a_t)$ at time steps $t = 0, ..., H-1$. Then, the loss $l_i(w)$ is defined as negative total reward, i.e.,

$$\text{(RL example)}: \quad l_i(w) := -\mathbb{E}_{\tau\sim p_i(\cdot|w)}[\mathcal{R}(\tau)],$$

where $\tau = (s_0, a_0, s_1, a_1, ..., s_{H-1}, a_{H-1})$ is a trajectory following the distribution $p_i(\cdot|w)$, and the reward

$$\mathcal{R}(\tau) := \sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t, a_t)$$

with $\mathcal{R}(\cdot)$ given as a reward function. The estimated gradient here is

$$\nabla l_i(w; \Omega) := \frac{1}{|\Omega|} \sum_{\tau\in\Omega} g_i(w; \tau),$$

where $g_i(w; \tau)$ is an unbiased policy gradient estimator s.t. $\mathbb{E}_{\tau\sim p_i(\cdot|w)} g_i(w; \tau) = \nabla l_i(w)$, e.g, REINFORCE (Williams, 1992) or G(PO)MDP (Baxter and Bartlett, 2001). In addition, the estimated Hessian is

$$\nabla^2 l_i(w; \Omega) := \frac{1}{|\Omega|} \sum_{\tau\in\Omega} H_i(w; \tau)$$

, where $H_i(w; \tau)$ is an unbiased policy Hessian estimator, e.g., DiCE (Foerster et al., 2018) or LVC (Rothfuss et al., 2019).

A.2 Classification Example for Finite-Sum Case

The risk minimization problem in classification often has a finite-sum objective function. For example, the mean-squared error (MSE) loss takes the form of

$$\text{(Classification example)}: \quad l_{S_i}(w) := \frac{1}{|S_i|} \sum_{(x_j,y_j)\in S_i} \|y_j - \phi(w; x_i)\|^2 \quad \text{(similarly for } l_{T_i}(w)),$$

where x_j, y_j are a feature-label pair and $\phi(w; \cdot)$ can be a deep neural network parameterized by w.

Appendix B. Derivation of Simplified Form of Gradient $\nabla\mathcal{L}_i(w)$ in (3)

First note that $\mathcal{L}_i(w_k) = l_i(\widetilde{w}^i_{k,N})$ and $\widetilde{w}^i_{k,N}$ is obtained by the following gradient descent updates

$$\widetilde{w}^i_{k,j+1} = \widetilde{w}^i_{k,j} - \alpha\nabla l_i(\widetilde{w}^i_{k,j}), \ j = 0, 1, ..., N-1 \ \text{ with } \ \widetilde{w}^i_{k,0} := w_k. \tag{64}$$

Then, by the chain rule, we have

$$\nabla \mathcal{L}_i(w_k) = \nabla_{w_k} l_i(\widetilde{w}_{k,N}^i) = \prod_{j=0}^{N-1} \nabla_{\widetilde{w}_{k,j}^i} \left(\widetilde{w}_{k,j+1}^i\right) \nabla l_i(\widetilde{w}_{k,N}^i),$$

which, in conjunction with (64), implies that

$$\nabla \mathcal{L}_i(w_k) = \prod_{j=0}^{N-1} \nabla_{\widetilde{w}_{k,j}^i} \left(\widetilde{w}_{k,j}^i - \alpha \nabla l_i(\widetilde{w}_{k,j}^i)\right) \nabla l_i(\widetilde{w}_{k,N}^i) = \prod_{j=0}^{N-1} \left(I - \alpha \nabla^2 l_i(\widetilde{w}_{k,j}^i)\right) \nabla l_i(\widetilde{w}_{k,N}^i),$$

which finishes the proof.

Appendix C. Auxiliary Lemmas for MAML in Resampling Case

In this section, we derive some useful lemmas to prove the propositions given in Section 3.3 on the properties of the meta gradient and the main results Theorem 5 and Corollary 6.

The first lemma provides a bound on the difference between $\|\widetilde{w}_j^i - \widetilde{u}_j^i\|$ for $j = 0, ..., N, i \in \mathcal{I}$, where $\widetilde{w}_j^i$, $j = 0, ..., N, i \in \mathcal{I}$ are given through the *gradient descent* updates in (1) and $\widetilde{u}_j^i$, $j = 0, ..., N$ are defined in the same way.

Lemma 11 *For any $i \in \mathcal{I}$, $j = 0, ..., N$ and $w, u \in \mathbb{R}^d$, we have*

$$\left\|\widetilde{w}_j^i - \widetilde{u}_j^i\right\| \leq (1 + \alpha L)^j \|w - u\|.$$

Proof Based on the updates that $\widetilde{w}_m^i = \widetilde{w}_{m-1}^i - \alpha \nabla l_i(\widetilde{w}_{m-1}^i)$ and $\widetilde{u}_m^i = \widetilde{u}_{m-1}^i - \alpha \nabla l_i(\widetilde{u}_{m-1}^i)$, we obtain, for any $i \in \mathcal{I}$,

$$\begin{aligned}
\|\widetilde{w}_m^i - \widetilde{u}_m^i\| =& \|\widetilde{w}_{m-1}^i - \alpha \nabla l_i(\widetilde{w}_{m-1}^i) - \widetilde{u}_{m-1}^i + \alpha \nabla l_i(\widetilde{u}_{m-1}^i)\| \\
\overset{(i)}{\leq}& \|\widetilde{w}_{m-1}^i - \widetilde{u}_{m-1}^i\| + \alpha L \|\widetilde{w}_{m-1}^i - \widetilde{u}_{m-1}^i\| \\
\leq& (1 + \alpha L) \|\widetilde{w}_{m-1}^i - \widetilde{u}_{m-1}^i\|,
\end{aligned}$$

where (i) follows from the triangle inequality. Telescoping the above inequality over m from 1 to j, we obtain

$$\left\|\widetilde{w}_j^i - \widetilde{u}_j^i\right\| \leq (1 + \alpha L)^j \|\widetilde{w}_0^i - \widetilde{u}_0^i\|,$$

which, in conjunction with the fact that $\widetilde{w}_0^i = w$ and $\widetilde{u}_0^i = u$, finishes the proof. ■

The following lemma provides an upper bound on $\|\nabla l_i(\widetilde{w}_j^i)\|$ for all $i \in \mathcal{I}$ and $j = 0, ..., N$, where $\widetilde{w}_j^i$ is defined in the same way as in Lemma 11.

Lemma 12 *For any $i \in \mathcal{I}$, $j = 0, ..., N$ and $w \in \mathbb{R}^d$, we have*

$$\|\nabla l_i(\widetilde{w}_j^i)\| \leq (1 + \alpha L)^j \|\nabla l_i(w)\|.$$

Proof For $m \geq 1$, we have

$$\begin{aligned}\|\nabla l_i(\widetilde{w}_m^i)\| =&\|\nabla l_i(\widetilde{w}_m^i) - \nabla l_i(\widetilde{w}_{m-1}^i) + \nabla l_i(\widetilde{w}_{m-1}^i)\| \\ \leq&\|\nabla l_i(\widetilde{w}_m^i) - \nabla l_i(\widetilde{w}_{m-1}^i)\| + \|\nabla l_i(\widetilde{w}_{m-1}^i)\| \\ \leq&L\|\widetilde{w}_m^i - \widetilde{w}_{m-1}^i\| + \|\nabla l_i(\widetilde{w}_{m-1}^i)\| \leq (1+\alpha L)\|\nabla l_i(\widetilde{w}_{m-1}^i)\|,\end{aligned}$$

where the last inequality follows from the update $\widetilde{w}_m^i = \widetilde{w}_{m-1}^i - \alpha\nabla l_i(\widetilde{w}_{m-1}^i)$. Then, telescoping the above inequality over m from 1 to j yields

$$\|\nabla l_i(\widetilde{w}_j^i)\| \leq (1+\alpha L)^j\|\nabla l_i(\widetilde{w}_0^i)\|,$$

which, combined with the fact that $\widetilde{w}_0^i = w$, finishes the proof. ■

The following lemma gives an upper bound on the quantity $\left\|I - \prod_{j=0}^m (I - \alpha V_j)\right\|$ for all matrices $V_j \in \mathbb{R}^{d\times d}, j = 0, ..., m$ that satisfy $\|V_j\| \leq L$.

Lemma 13 *For all matrices $V_j \in \mathbb{R}^{d\times d}, j = 0, ..., m$ that satisfy $\|V_j\| \leq L$, we have*

$$\left\|I - \prod_{j=0}^m (I - \alpha V_j)\right\| \leq (1+\alpha L)^{m+1} - 1.$$

Proof First note that the product $\prod_{j=0}^m (I - \alpha V_j)$ can be expanded as

$$\prod_{j=0}^m (I - \alpha V_j) = I - \sum_{j=0}^m \alpha V_j + \sum_{0\leq p<q\leq m} \alpha^2 V_p V_q + \cdots + (-1)^{m+1}\alpha^{m+1}\prod_{j=0}^m V_j.$$

Then, by using $\|V_j\| \leq L$ for $j = 0, ..., m$, we have

$$\begin{aligned}\left\|I - \prod_{j=0}^m (I - \alpha V_j)\right\| \leq&\left\|\sum_{j=0}^m \alpha V_j\right\| + \left\|\sum_{0\leq p<q\leq m} \alpha^2 V_p V_q\right\| + \cdots + \left\|\alpha^{m+1}\prod_{j=0}^m V_j\right\| \\ \leq&\mathrm{C}_{m+1}^1\alpha L + \mathrm{C}_{m+1}^2(\alpha L)^2 + \cdots + \mathrm{C}_{m+1}^{m+1}(\alpha L)^{m+1} \\ =&(1+\alpha L)^{m+1} - 1,\end{aligned}$$

where the notion C_n^k denotes the number of k-element subsets of a set of size n. Then, the proof is complete. ■

Recall the gradient $\nabla\mathcal{L}_i(w) = \prod_{j=0}^{N-1}(I - \alpha\nabla^2 l_i(\widetilde{w}_j^i))\nabla l_i(\widetilde{w}_N^i)$, where $\widetilde{w}_j^i, i \in \mathcal{I}, j = 0, ..., N$ are given by the gradient descent steps in (1) and $\widetilde{w}_0^i = w$ for all tasks $i \in \mathcal{I}$. Next, we provide an upper bound on the difference $\|\nabla l_i(w) - \nabla\mathcal{L}_i(w)\|$.

Lemma 14 *For any $i \in \mathcal{I}$ and $w \in \mathbb{R}^d$, we have*

$$\|\nabla l_i(w) - \nabla\mathcal{L}_i(w)\| \leq C_l\|\nabla l_i(w)\|,$$

where C_l is a positive constant given by

$$C_l = (1+\alpha L)^{2N} - 1 > 0. \tag{65}$$

Proof First note that $\widetilde{w}_N^i$ can be rewritten as $\widetilde{w}_N^i = w - \alpha \sum_{j=0}^{N-1} \nabla l_i(\widetilde{w}_j^i)$. Then, based on the mean value theorem (MVT) for vector-valued functions (McLeod, 1965), we have, there exist constants $r_t, t = 1, ..., d$ satisfying $\sum_{t=1}^{d} r_t = 1$ and vectors $w_t' \in \mathbb{R}^d, t = 1, ..., d$ such that

$$\begin{aligned}\nabla l_i(\widetilde{w}_N^i) =& \nabla l_i\Big(w - \alpha \sum_{j=0}^{N-1} \nabla l_i(\widetilde{w}_j^i)\Big) = \nabla l_i(w) + \Big(\sum_{t=1}^{d} r_t \nabla^2 l_i(w_t')\Big)\Big(-\alpha \sum_{j=0}^{N-1} \nabla l_i(\widetilde{w}_j^i)\Big) \\ =& \Big(I - \alpha \sum_{t=1}^{d} r_t \nabla^2 l_i(w_t')\Big)\nabla l_i(w) - \alpha \sum_{t=1}^{d} r_t \nabla^2 l_i(w_t') \sum_{j=1}^{N-1} \nabla l_i(\widetilde{w}_j^i). \end{aligned} \tag{66}$$

For simplicity, we define $K(N) := \prod_{j=0}^{N-1}(I - \alpha \nabla^2 l_i(\widetilde{w}_j^i))$. Then, using (66), we write $\|\nabla l_i(w) - \nabla \mathcal{L}_i(w)\|$ as

$$\begin{aligned}\|\nabla l_i(w) - \nabla \mathcal{L}_i(w)\| =& \|\nabla l_i(w) - K(N)\nabla l_i(\widetilde{w}_N^i)\| \\ =& \Big\|\nabla l_i(w) - K(N)\Big(I - \alpha \sum_{t=1}^{d} r_t \nabla^2 l_i(w_t')\Big)\nabla l_i(w) + \alpha K(N) \sum_{t=1}^{d} r_t \nabla^2 l_i(w_t') \sum_{j=1}^{N-1} \nabla l_i(\widetilde{w}_j^i)\Big\| \\ \le& \Big\|\Big(I - K(N)\Big(I - \alpha \sum_{t=1}^{d} r_t \nabla^2 l_i(w_t')\Big)\Big)\nabla l_i(w)\Big\| + \Big\|\alpha K(N) \sum_{t=1}^{d} r_t \nabla^2 l_i(w_t') \sum_{j=1}^{N-1} \nabla l_i(\widetilde{w}_j^i)\Big\| \\ \overset{(i)}{\le}& \Big\|\Big(I - K(N)\Big(I - \alpha \sum_{t=1}^{d} r_t \nabla^2 l_i(w_t')\Big)\Big)\nabla l_i(w)\Big\| + \alpha L(1 + \alpha L)^N \sum_{j=1}^{N-1} \Big\|\nabla l_i(\widetilde{w}_j^i)\Big\| \\ \overset{(ii)}{\le}& \Big\|I - K(N)\Big(I - \alpha \sum_{t=1}^{d} r_t \nabla^2 l_i(w_t')\Big)\Big\|\|\nabla l_i(w)\| + \alpha L(1 + \alpha L)^N \sum_{j=1}^{N-1} (1 + \alpha L)^j \|\nabla l_i(w)\| \\ \le& \Big\|I - K(N)\Big(I - \alpha \sum_{t=1}^{d} r_t \nabla^2 l_i(w_t')\Big)\Big\|\|\nabla l_i(w)\| + (1 + \alpha L)^{N+1}((1 + \alpha L)^{N-1} - 1)\|\nabla l_i(w)\| \\ \overset{(iii)}{\le}& ((1 + \alpha L)^{N+1} - 1)\|\nabla l_i(w)\| + (1 + \alpha L)^{N+1}((1 + \alpha L)^{N-1} - 1)\|\nabla l_i(w)\| \\ =& ((1 + \alpha L)^{2N} - 1)\|\nabla l_i(w)\|, \end{aligned}$$

where (i) follows from the fact that $\|\nabla^2 l_i(u)\| \le L$ for any $u \in \mathbb{R}^d$ and $\sum_{t=1}^{d} r_t = 1$, and the inequality that $\|\sum_{j=1}^{n} a_j\| \le \sum_{j=1}^{n} \|a_j\|$, (ii) follows from Lemma 12, and (iii) follows from Lemma 13. ■

Recall that the expected value of the gradient of the loss $\nabla l(w) := \mathbb{E}_{i \sim p(\mathcal{T})} \nabla l_i(w)$ and the objective function $\nabla \mathcal{L}(w) := \nabla \mathcal{L}_i(w)$. Based on the above lemmas, we next provide an upper bound on $\|\nabla l(w)\|$ using $\|\nabla \mathcal{L}(w)\|$.

Lemma 15 *For any $w \in \mathbb{R}^d$, we have*

$$\|\nabla l(w)\| \le \frac{1}{1 - C_l}\|\nabla \mathcal{L}(w)\| + \frac{C_l}{1 - C_l}\sigma,$$

where the constant C_l is given by

$$C_l = (1+\alpha L)^{2N} - 1.$$

Proof Based on the definition of $\nabla l(w)$, we have

$$\begin{aligned}
\|\nabla l(w)\| =& \|\mathbb{E}_{i\sim p(\mathcal{T})}(\nabla l_i(w) - \nabla \mathcal{L}_i(w) + \nabla \mathcal{L}_i(w))\| \\
\leq& \|\mathbb{E}_{i\sim p(\mathcal{T})}\nabla \mathcal{L}_i(w)\| + \|\mathbb{E}_{i\sim p(\mathcal{T})}(\nabla l_i(w) - \nabla \mathcal{L}_i(w))\| \\
\leq& \|\nabla \mathcal{L}(w)\| + \mathbb{E}_{i\sim p(\mathcal{T})}\|\nabla l_i(w) - \nabla \mathcal{L}_i(w)\| \\
\overset{(i)}{\leq}& \|\nabla \mathcal{L}(w)\| + C_l \mathbb{E}_{i\sim p(\mathcal{T})}\|\nabla l_i(w)\| \\
\overset{(ii)}{\leq}& \|\nabla \mathcal{L}(w)\| + C_l(\|\nabla l(w)\| + \sigma),
\end{aligned}$$

where (i) follows from Lemma 14, and (ii) follows from Assumption 2. Then, rearranging the above inequality completes the proof. ■

Recall from (14) that we choose the meta stepsize $\beta_k = \frac{1}{C_\beta \widehat{L}_{w_k}}$, where C_β is a positive constant and $\widehat{L}_{w_k} = (1+\alpha L)^{2N} L + C_{\mathcal{L}} \frac{1}{|B_k'|}\sum_{i\in B_k'} \|\nabla l_i(w_k; D^i_{L_k})\|$. Using an approach similar to Lemma 4.11 in Fallah et al. (2020a), we establish the following lemma to provide the first- and second-moment bounds for β_k.

Lemma 16 *Suppose that Assumptions 1, 2 and 3 hold. Set the meta stepsize $\beta_k = \frac{1}{C_\beta \widehat{L}_{w_k}}$ with $\widehat{L}_{w_k}$ given by (14), where $|B_k'| > \frac{4C_{\mathcal{L}}^2\sigma^2}{3(1+\alpha L)^{4N}L^2}$ and $|D^i_{L_k}| > \frac{64\sigma_g^2 C_{\mathcal{L}}^2}{(1+\alpha L)^{4N}L^2}$ for all $i \in B_k'$. Then, conditioning on w_k, we have*

$$\mathbb{E}\beta_k \geq \frac{4}{C_\beta}\frac{1}{5L_{w_k}}, \quad \mathbb{E}\beta_k^2 \leq \frac{4}{C_\beta^2}\frac{1}{L_{w_k}^2},$$

where $L_{w_k} = (1+\alpha L)^{2N} L + C_{\mathcal{L}} \mathbb{E}_{i\sim p(\mathcal{T})}\|\nabla l_i(w_k)\|$ with $C_{\mathcal{L}}$ given in (13).

Proof Let $\widetilde{L}_{w_k} = 4L + \frac{4C_{\mathcal{L}}}{(1+\alpha L)^{2N}}\frac{1}{|B_k'|}\sum_{i\in B_k'}\|\nabla l_i(w_k; D^i_{L_k})\|$. Note that $|B_k'| > \frac{4C_{\mathcal{L}}^2\sigma^2}{3(1+\alpha L)^{4N}L^2}$ and $|D^i_{L_k}| > \frac{64\sigma_g^2 C_{\mathcal{L}}^2}{(1+\alpha L)^{4N}L^2}$, $i \in B_k'$. Then, using an approach similar to (61) in Fallah et al. (2020a) and conditioning on w_k, we have

$$\mathbb{E}\Big(\frac{1}{\widetilde{L}_{w_k}^2}\Big) \leq \frac{\sigma_\beta^2/(4L)^2 + \mu_\beta^2/(\mu_\beta)^2}{\sigma_\beta^2 + \mu_\beta^2}, \tag{67}$$

where σ_β^2 and μ_β are the variance and mean of variable $\frac{4C_{\mathcal{L}}}{(1+\alpha L)^{2N}}\frac{1}{|B_k'|}\sum_{i\in B_k'}\|\nabla l_i(w_k; D^i_{L_k})\|$. Using an approach similar to (62) in Fallah et al. (2020a), conditioning on w_k and using $|D^i_{L_k}| > \frac{64\sigma_g^2 C_{\mathcal{L}}^2}{(1+\alpha L)^{4N}L^2}$, we have

$$\frac{C_{\mathcal{L}}}{(1+\alpha L)^{2N}}\mathbb{E}_i\|\nabla l_i(w_k)\| - L \leq \mu_\beta \leq \frac{C_{\mathcal{L}}}{(1+\alpha L)^{2N}}\mathbb{E}_i\|\nabla l_i(w_k)\| + L, \tag{68}$$

which implies that $\mu_\beta + 5L \geq \frac{4}{(1+\alpha L)^{2N}} L_{w_k}$, and thus using (67) yields

$$\frac{16}{(1+\alpha L)^{4N}} L_{w_k}^2 \mathbb{E}\Big(\frac{1}{\widetilde{L}_{w_k}^2}\Big) \leq \frac{\mu_\beta^2(25/16 + \sigma_\beta^2/(8L^2)) + 25\sigma_\beta^2/8}{\sigma_\beta^2 + \mu_\beta^2}. \tag{69}$$

Furthermore, conditioning on w_k, σ_β is bounded by

$$\begin{aligned}
\sigma_\beta^2 =& \frac{16C_{\mathcal{L}}^2}{(1+\alpha L)^{4N}|B_k'|} \mathrm{Var}(\|\nabla l_i(w_k; D_{L_k}^i)\|) \\
\leq& \frac{16C_{\mathcal{L}}^2}{(1+\alpha L)^{4N}|B_k'|} \Big(\sigma^2 + \frac{\sigma_g^2}{|D_{L_k}^i|}\Big) \\
\overset{(i)}{\leq}& \frac{16C_{\mathcal{L}}^2\sigma^2}{(1+\alpha L)^{4N}|B_k'|} + \frac{L^2}{4|B_k'|} \overset{(ii)}{\leq} 12L^2 + \frac{1}{4}L^2 < \frac{25}{2}L^2,
\end{aligned} \tag{70}$$

where (i) follows from $|D_{L_k}^i| > \frac{64\sigma_g^2 C_{\mathcal{L}}^2}{(1+\alpha L)^{4N}L^2}$, $i \in B_k'$ and (ii) follows from $|B_k'| > \frac{4C_{\mathcal{L}}^2\sigma^2}{3(1+\alpha L)^{4N}L^2}$ and $|B_k'| \geq 1$. Then, plugging (70) in (69) yields $\frac{16}{(1+\alpha L)^{4N}} L_{w_k}^2 \mathbb{E}\Big(\frac{1}{\widetilde{L}_{w_k}^2}\Big) \leq \frac{25}{8}$. Then, noting that $\beta_k = \frac{4}{C_\beta(1+\alpha L)^{2N}\widetilde{L}_{w_k}}$, using the above inequality and conditioning on w_k, we have

$$\mathbb{E}\beta_k^2 = \frac{16}{C_\beta^2(1+\alpha L)^{4N}} \mathbb{E}\left(\frac{1}{\widetilde{L}_{w_k}^2}\right) \leq \frac{25}{8C_\beta^2}\frac{1}{L_{w_k}^2} < \frac{4}{C_\beta^2}\frac{1}{L_{w_k}^2}. \tag{71}$$

In addition, by Jensen's inequality and conditioning on w_k, we have

$$\begin{aligned}
\mathbb{E}\beta_k =& \frac{4}{C_\beta(1+\alpha L)^{2N}} \mathbb{E}\Big(\frac{1}{\widetilde{L}_{w_k}}\Big) \geq \frac{4}{C_\beta(1+\alpha L)^{2N}} \frac{1}{\mathbb{E}\widetilde{L}_{w_k}} = \frac{4}{C_\beta(1+\alpha L)^{2N}} \frac{1}{4L + \mu_\beta} \\
\overset{(i)}{\geq}& \frac{4}{C_\beta} \frac{1}{4L(1+\alpha L)^{2N} + L_{w_k}} \overset{(ii)}{\geq} \frac{4}{C_\beta}\frac{1}{5L_{w_k}},
\end{aligned} \tag{72}$$

where (i) follows from (68) and (ii) follows from the fact $L_{w_k} > (1+\alpha L)^{2N} L$. ■

Appendix D. Auxiliary Lemmas for MAML in Finite-Sum Case

In this section, we provide some useful lemmas to prove the propositions in Section 4.2 on properties of the meta gradient and the main results Theorem 9 and Corollary 10.

The following lemma provides an upper bound on $\|l_{S_i}(\widetilde{w}_j^i)\|$ for all $i \in \mathcal{I}$ and $j = 0, ..., N$, where $\widetilde{w}_j^i$ is defined by (9) with $\widetilde{w}_0^i = w$.

Lemma 17 *For any $i \in \mathcal{I}$, $j = 0, ..., N$ and $w \in \mathbb{R}^d$, we have*

$$\|\nabla l_{S_i}(\widetilde{w}_j^i)\| \leq (1+\alpha L)^j \|\nabla l_{S_i}(w)\|.$$

Proof The proof is similar to that of Lemma 12, and thus omitted. ■

We next provide a bound on $\|\nabla l_{T_i}(w) - \nabla \mathcal{L}_i(w)\|$, where

$$\nabla \mathcal{L}_i(w) = \prod_{j=0}^{N-1} (I - \alpha \nabla^2 l_{S_i}(w_j^i)) \nabla l_{T_i}(w_N^i).$$

Lemma 18 *For any $i \in \mathcal{I}$ and $w \in \mathbb{R}^d$, we have*

$$\|\nabla l_{T_i}(w) - \nabla \mathcal{L}_i(w)\| \leq \big((1+\alpha L)^N - 1\big)\|\nabla l_{T_i}(w)\| + (1+\alpha L)^N \big((1+\alpha L)^N - 1\big)\|\nabla l_{S_i}(w)\|.$$

Proof Using the mean value theorem (MVT), we have, there exist constants $r_t, t = 1, ..., d$ satisfying $\sum_{t=1}^d r_t = 1$ and vectors $w_t' \in \mathbb{R}^d, t = 1, ..., d$ such that

$$\begin{aligned}
\nabla l_{T_i}(\widetilde{w}_N^i) =& \nabla l_{T_i}\Big(w - \alpha \sum_{j=0}^{N-1} \nabla l_{S_i}(\widetilde{w}_j^i)\Big) = \nabla l_{T_i}(w) + \sum_{t=1}^d r_t \nabla^2 l_{T_i}(w_t')\Big(-\alpha \sum_{j=0}^{N-1} \nabla l_{S_i}(\widetilde{w}_j^i)\Big) \\
=& \nabla l_{T_i}(w) - \alpha \sum_{t=1}^d r_t \nabla^2 l_{T_i}(w_t') \sum_{j=0}^{N-1} \nabla l_{S_i}(\widetilde{w}_j^i).
\end{aligned}$$

Based on the above equality, we have

$$\begin{aligned}
&\|\nabla l_{T_i}(w) - \nabla \mathcal{L}_i(w)\| \\
=& \Big\|\nabla l_{T_i}(w) - \prod_{j=0}^{N-1} (I - \alpha \nabla^2 l_{S_i}(\widetilde{w}_j^i)) \nabla l_{T_i}(\widetilde{w}_N^i)\Big\| \\
=& \Big\|\nabla l_{T_i}(w) - \prod_{j=0}^{N-1} (I - \alpha \nabla^2 l_{S_i}(\widetilde{w}_j^i)) \nabla l_{T_i}(w) + \prod_{j=0}^{N-1} (I - \alpha \nabla^2 l_{S_i}(\widetilde{w}_j^i)) \alpha \sum_{t=1}^d r_t \nabla^2 l_{T_i}(w_t') \sum_{j=0}^{N-1} \nabla l_{S_i}(\widetilde{w}_j^i)\Big\| \\
=& \Big\|I - \prod_{j=0}^{N-1} (I - \alpha \nabla^2 l_{S_i}(\widetilde{w}_j^i))\Big\| \|\nabla l_{T_i}(w)\| + \Big\| \prod_{j=0}^{N-1} (I - \alpha \nabla^2 l_{S_i}(\widetilde{w}_j^i)) \alpha \sum_{t=1}^d r_t \nabla^2 l_{T_i}(w_t') \sum_{j=0}^{N-1} \nabla l_{S_i}(\widetilde{w}_j^i)\Big\| \\
\overset{(i)}{\leq}& \big((1+\alpha L)^N - 1\big)\|\nabla l_{T_i}(w)\| + \alpha L (1+\alpha L)^N \sum_{j=0}^{N-1} \|\nabla l_{S_i}(\widetilde{w}_j^i)\| \\
\overset{(ii)}{\leq}& \big((1+\alpha L)^N - 1\big)\|\nabla l_{T_i}(w)\| + \alpha L (1+\alpha L)^N \sum_{j=0}^{N-1} (1+\alpha L)^j \|\nabla l_{S_i}(w)\| \\
=& \big((1+\alpha L)^N - 1\big)\|\nabla l_{T_i}(w)\| + (1+\alpha L)^N \big((1+\alpha L)^N - 1\big)\|\nabla l_{S_i}(w)\|,
\end{aligned}$$

where (i) follows from Lemma 13 and $\|\sum_{t=1}^d r_t \nabla^2 l_{T_i}(w_t')\| \leq \sum_{t=1}^d r_t \|\nabla^2 l_{T_i}(w_t')\| \leq L$, and (ii) follows from Lemma 17. Then, the proof is complete. ■

Recall that $\nabla l_T(w) = \mathbb{E}_{i \sim p(\mathcal{T})} \nabla l_{T_i}(w)$, $\nabla \mathcal{L}(w) = \mathbb{E}_{i \sim p(\mathcal{T})} \nabla \mathcal{L}_i(w)$ and $b = \mathbb{E}_{i \sim p(\mathcal{T})}[b_i]$. The following lemma provides an upper bound on $\|\nabla l_T(w)\|$.

Lemma 19 *For any $i \in \mathcal{I}$ and $w \in \mathbb{R}^d$, we have*

$$\|\nabla l_T(w)\| \leq \frac{1}{C_1}\|\nabla \mathcal{L}(w)\| + \frac{C_2}{C_1}, \tag{73}$$

where constants $C_1, C_2 > 0$ are give by

$$\begin{aligned} C_1 =& 2-(1+\alpha L)^{2N}, \\ C_2 =& \big((1+\alpha L)^{2N}-1\big)\sigma + (1+\alpha L)^N\big((1+\alpha L)^N-1\big)b. \end{aligned} \tag{74}$$

Proof First note that

$$\begin{aligned} \|\nabla l_T(w)\| =& \|\mathbb{E}_i(\nabla l_{T_i}(w)-\nabla\mathcal{L}_i(w))+\nabla\mathcal{L}(w)\| \\ \le& \|\nabla\mathcal{L}(w)\| + \mathbb{E}_i\|\nabla l_{T_i}(w)-\nabla\mathcal{L}_i(w)\| \\ \overset{(i)}{\le}& \|\nabla\mathcal{L}(w)\| + \mathbb{E}_i\Big(\big((1+\alpha L)^N-1\big)\|\nabla l_{T_i}(w)\| + (1+\alpha L)^N\big((1+\alpha L)^N-1\big)\|\nabla l_{S_i}(w)\|\Big) \\ \overset{(ii)}{\le}& \|\nabla\mathcal{L}(w)\| + \big((1+\alpha L)^N-1\big)\big(\|\nabla l_T(w)\|+\sigma\big) \\ & + (1+\alpha L)^N\big((1+\alpha L)^N-1\big)(\mathbb{E}_i\|\nabla l_{T_i}(w)\| + \mathbb{E}_i b_i) \\ \le& \|\nabla\mathcal{L}(w)\| + \big((1+\alpha L)^N - 1 + (1+\alpha L)^N((1+\alpha L)^N-1)\big)\|\nabla l_T(w)\| \\ & + ((1+\alpha L)^N-1)\sigma + (1+\alpha L)^N((1+\alpha L)^N-1)(\sigma+b) \\ \le& \|\nabla\mathcal{L}(w)\| + \big((1+\alpha L)^{2N}-1\big)\|\nabla l_T(w)\| \\ & + ((1+\alpha L)^{2N}-1)\sigma + (1+\alpha L)^N((1+\alpha L)^N-1)b \end{aligned}$$

where (i) follows from Lemma 18, (ii) follows from Assumption 5. Based on the definitions of C_1 and C_2 in (74), the proof is complete. ■

The following lemma provides the first- and second-moment bounds on $1/\hat{L}_{w_k}$, where

$$\hat{L}_{w_k} = (1+\alpha L)^{2N}L + C_b b + C_{\mathcal{L}}\frac{\sum_{i\in B_k'}\|\nabla l_{T_i}(w_k)\|}{|B_k'|}.$$

Lemma 20 *If the batch size $|B_k'| \ge \frac{2C_{\mathcal{L}}^2\sigma^2}{(C_b b+(1+\alpha L)^{2N}L)^2}$, then conditioning on w_k, we have*

$$\mathbb{E}\Big(\frac{1}{\hat{L}_{w_k}}\Big) \ge \frac{1}{L_{w_k}}, \quad \mathbb{E}\Big(\frac{1}{\hat{L}_{w_k}^2}\Big) \le \frac{2}{L_{w_k}^2}$$

where L_{w_k} is given by

$$L_{w_k} = (1+\alpha L)^{2N}L + C_b b + C_{\mathcal{L}}\mathbb{E}_{i\sim p(\mathcal{T})}\|\nabla l_{T_i}(w_k)\|.$$

Proof Conditioning on w_k and using an approach similar to (67), we have

$$\mathbb{E}\Big(\frac{1}{\hat{L}_{w_k}^2}\Big) \le \frac{\sigma_\beta^2/\big(C_b b+(1+\alpha L)^{2N}L\big)^2 + \mu_\beta^2/(\mu_\beta + C_b b + (1+\alpha L)^{2N}L)^2}{\sigma_\beta^2+\mu_\beta^2}, \tag{75}$$

where μ_β and σ_β^2 are the mean and variance of variable $\frac{C_{\mathcal{L}}}{|B_k'|}\sum_{i\in B_k'}\|\nabla l_{T_i}(w_k)\|$. Noting that $\mu_\beta = C_{\mathcal{L}}\mathbb{E}_{i\sim p(\mathcal{T})}\|\nabla l_{T_i}(w_k)\|$, we have $L_{w_k} = (1+\alpha L)^{2N}L + C_b b + \mu_\beta$, and thus

$$L_{w_k}^2\mathbb{E}\Big(\frac{1}{\hat{L}_{w_k}^2}\Big) \le \frac{\sigma_\beta^2\frac{((1+\alpha L)^{2N}L+C_b b+\mu_\beta)^2}{\big(C_b b+(1+\alpha L)^{2N}L\big)^2}+\mu_\beta^2}{\sigma_\beta^2+\mu_\beta^2} \le \frac{2\sigma_\beta^2+\mu_\beta^2+\frac{2\sigma_\beta^2\mu_\beta^2}{\big(C_b b+(1+\alpha L)^{2N}L\big)^2}}{\sigma_\beta^2+\mu_\beta^2}, \tag{76}$$

where the last inequality follows from $(a+b)^2 \leq 2a^2 + 2b^2$. Note that, conditioning on w_k,

$$\sigma_\beta^2 = \frac{C_{\mathcal{L}}^2}{|B_k'|}\mathrm{Var}\|\nabla l_{T_i}(w_k)\| \leq \frac{C_{\mathcal{L}}^2}{|B_k'|}\sigma^2,$$

which, in conjunction with $|B_k'| \geq \frac{2C_{\mathcal{L}}^2\sigma^2}{(C_b b+(1+\alpha L)^{2N}L)^2}$, yields

$$\frac{2\sigma_\beta^2}{\big(C_b b + (1+\alpha L)^{2N}L\big)^2} \leq 1. \tag{77}$$

Combining (77) and (76) yields

$$\mathbb{E}\Big(\frac{1}{\hat{L}_{w_k}^2}\Big) \leq \frac{2}{L_{w_k}^2}.$$

In addition, conditioning on w_k, we have

$$\mathbb{E}\Big(\frac{1}{\hat{L}_{w_k}}\Big) \overset{(i)}{\geq} \frac{1}{\mathbb{E}\hat{L}_{w_k}} = \frac{1}{L_{w_k}}, \tag{78}$$

where (i) follows from Jensen's inequality. Then, the proof is complete. ■

References

Maruan Al-Shedivat, Trapit Bansal, Yuri Burda, Ilya Sutskever, Igor Mordatch, and Pieter Abbeel. Continuous adaptation via meta-learning in nonstationary and competitive environments. In *International Conference on Learning Representations (ICLR)*, 2018.

Pierre Alquier, Massimiliano Pontil, et al. Regret bounds for lifelong learning. In *Artificial Intelligence and Statistics (AISTATS)*, pages 261–269, 2017.

Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your MAML. In *International Conference on Learning Representations (ICLR)*, 2019.

Sanjeev Arora, Simon S Du, Sham Kakade, Yuping Luo, and Nikunj Saunshi. Provable representation learning for imitation learning via bi-level optimization. In *International conference on machine learning (ICML)*, 2020.

Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning (ICML)*, pages 424–433, 2019.

Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Y Bengio, S Bengio, and J Cloutier. Learning a synaptic learning rule. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1991.

Fei Chen, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning for recommendation. *arXiv preprint arXiv:1802.07876*, 2018.

Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Distribution-agnostic model-agnostic meta-learning. *arXiv preprint arXiv:2002.04766*, 2020.

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Incremental learning-to-learn with statistical guarantees. *arXiv preprint arXiv:1803.08089*, 2018a.

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Learning to learn around a common mean. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10169–10179, 2018b.

Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, and Massimiliano Pontil. Learning-to-learn stochastic gradient descent with biased regularization. *arXiv preprint arXiv:1903.10399*, 2019.

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

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1082–1092, 2020a.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Provably convergent policy gradient methods for model-agnostic meta-reinforcement learning. *arXiv preprint arXiv:2002.05135*, 2020b.

Chelsea Finn and Sergey Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. In *International Conference on Learning Representations (ICLR)*, 2018.

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

Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. In *Conference on Robot Learning (CoRL)*, pages 357–368, 2017b.

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9516–9527, 2018.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning (ICML)*, pages 1920–1930, 2019.

Jakob Foerster, Gregory Farquhar, Maruan Al-Shedivat, Tim Rocktäschel, Eric Xing, and Shimon Whiteson. DiCE: The infinitely differentiable monte carlo estimator. In *International Conference on Machine Learning (ICML)*, pages 1529–1538, 2018.

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

Ghassen Jerfel, Erin Grant, Thomas L Griffiths, and Katherine Heller. Online gradient-based mixtures for transfer modulation in meta-learning. *arXiv preprint arXiv:1812.06080*, 2018.

Kaiyi Ji and Yingbin Liang. Lower bounds and accelerated algorithms for bilevel optimization. *arXiv preprint arXiv:2102.03926*, 2021.

Kaiyi Ji, Jason D Lee, Yingbin Liang, and H Vincent Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. *arXiv preprint arXiv:2006.09486*, 2020a.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Nonasymptotic analysis and faster algorithms. *arXiv preprint arXiv:2010.07962*, 2020b.

Jin-Hwa Kim, Junyoung Park, and Yongseok Choi. Multi-step estimation for gradient-based meta-learning. *arXiv preprint arXiv:2006.04298*, 2020.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-SGD: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

Valerii Likhosherstov, Xingyou Song, Krzysztof Choromanski, Jared Davis, and Adrian Weller. UFO-BLO: Unbiased first-order bilevel optimization. *arXiv preprint arXiv:2006.03631*, 2020.

Hao Liu, Richard Socher, and Caiming Xiong. Taming MAML: Efficient unbiased meta-reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 4061–4071, 2019.

Robert M McLeod. Mean value theorems for vector valued functions. *Proceedings of the Edinburgh Mathematical Society*, 14(3):197–209, 1965.

Fei Mi, Minlie Huang, Jiyong Zhang, and Boi Faltings. Meta-learning for low-resource natural language generation in task-oriented dialogue systems. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3151–3157, 2019.

Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *International Conference on Machine Learning (ICML)*, pages 2554–2563, 2017.

Devang K Naik and Richard J Mammone. Meta-neural networks that learn by learning. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 437–442, 1992.

Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2018.

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

Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. BOIL: Towards representation change for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. In *International Conference on Learning Representations (ICLR)*, 2020.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 113–124, 2019.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2016.

Jonas Rothfuss, Dennis Lee, Ignasi Clavera, Tamim Asfour, and Pieter Abbeel. ProMP: Proximal meta-policy search. In *International Conference on Learning Representations (ICLR)*, 2019.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning (ICML)*, pages 1842–1850, 2016.

Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4077–4087, 2017.

Xingyou Song, Wenbo Gao, Yuxiang Yang, Choromanski Krzysztof, Aldo Pacchiano, and Yunhao Tang. ES-MAML: Simple hessian-free meta learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

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

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3630–3638, 2016.

Haoxiang Wang, Ruoyu Sun, and Bo Li. Global convergence and induced kernels of gradient-based meta-learning with neural nets. *arXiv preprint arXiv:2006.14606*, 2020a.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. On the global optimality of model-agnostic meta-learning. In *International conference on machine learning (ICML)*, 2020b.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

Pan Zhou, Xiaotong Yuan, Huan Xu, Shuicheng Yan, and Jiashi Feng. Efficient meta learning via minibatch proximal update. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1532–1542, 2019.

Luisa M Zintgraf, Kyriacos Shiarlis, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. CAML: Fast context adaptation via meta-learning. *arXiv preprint arXiv:1810.03642*, 2018.