On the Convergence Theory of Gradient-Based Model-Agnostic Meta-Learning Algorithms

Alireza Fallah[*], Aryan Mokhtari[†], Asuman Ozdaglar[*]

Abstract

In this paper, we study the convergence theory of a class of gradient-based Model-Agnostic Meta-Learning (MAML) methods and characterize their overall computational complexity as well as their best achievable level of accuracy in terms of gradient norm for *nonconvex* loss functions. In particular, we start with the MAML algorithm and its first order approximation (FO-MAML) and highlight the challenges that emerge in their analysis. By overcoming these challenges not only we provide the first theoretical guarantees for MAML and FO-MAML in nonconvex settings, but also we answer some of the unanswered questions for the implementation of these algorithms including how to choose their learning rate (stepsize) and the batch size for both tasks and datasets corresponding to tasks. In particular, we show that MAML can find an ϵ-first-order stationary point for any ϵ after at most $\mathcal{O}(1/\epsilon^2)$ iterations while the cost of each iteration is $\mathcal{O}(d^2)$, where d is the problem dimension. We further show that FO-MAML reduces the cost per iteration of MAML to $\mathcal{O}(d)$, but, unlike MAML, its solution cannot reach any small desired level of accuracy. We further propose a new variant of the MAML algorithm called Hessian-free MAML (HF-MAML) which preserves all theoretical guarantees of MAML, while reducing its computational cost per iteration from $\mathcal{O}(d^2)$ to $\mathcal{O}(d)$.

1 Introduction

In a large fraction of artificial intelligence problems, ranging from robotics to image classification and pattern recognition, the goal is to design systems that use prior experience and knowledge to learn new skills more efficiently. *Meta-learning* or *learning to learn* formalizes this goal by using data from previous tasks to learn update rules or model parameters that can be fine-tuned to perform well on new tasks with small amount of data [27]. Recent works have integrated this paradigm with neural networks with examples including learning the initial weights of a neural network [15, 25], updating its architecture [5, 31, 32], or learning the parameters of optimization algorithms using recurrent neural networks [26, 3].

A particularly simple and effective approach, proposed by [15], has been gradient-based meta-learning which assumes all tasks are represented by the same class of parametrized model. This approach forms a shared prior w using existing tasks which is used as a meta initialization for a gradient-based method using few data points from a new task. This method is referred to as *model-agnostic meta learning (MAML)* since it can be applied to any learning problem that is trained with gradient descent. A number of recent papers studied the empirical performance of various methods in this setting, including [25, 4, 22, 30, 18, 8, 1]. However, to the best of our knowledge, its convergence properties have not been established for general and possibly non-convex functions.

In this paper, we study the convergence of variants of MAML methods for nonconvex loss functions and establish their computational complexity as well as their best achievable level of accuracy in terms of gradient norm. We consider both MAML and its first order approximation, FO-MAML, and for both, provide convergence results that show the dependence on batch size and the problem dimension. Our results also highlight how to choose the learning rate (stepsize) to achieve the best possible performance for these algorithms. We also propose a new MAML algorithm, called *Hessian-free MAML (HF-MAML)*, which preserves the theoretical guarantees of MAML while achieving the same cheap cost of FO-MAML.

[*] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. {afallah@mit.edu, asuman@mit.edu}.

[†] Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA. mokhtari@austin.utexas.edu.

More formally, let $\mathcal{T} = \{\mathcal{T}_i\}_{i \in \mathcal{I}}$ denote the set of all tasks and let p be the probability distribution over tasks $\mathcal{T}$, i.e, task $\mathcal{T}_i$ is drawn with probability $p_i := p(\mathcal{T}_i)$. We represent the loss function corresponding to task $\mathcal{T}_i$ by $f_i(w) : \mathbb{R}^d \to \mathbb{R}$ which is parameterized by the same $w \in \mathbb{R}^d$ for all tasks. Here, the loss function f_i measures how well an action w performs on task $\mathcal{T}_i$. We also define $f(w)$ as the expectation of f_i over all tasks, i.e.,

$$f(w) := \mathbb{E}_{i \sim p}[f_i(w)]. \tag{1}$$

By minimizing the function f one can find a solution w that in expectation has the best performance (smallest loss) over all possible tasks.

In most learning applications, the loss function f_i corresponding to task $\mathcal{T}_i$ is defined as an expected loss with respect to the probability distribution which generates data for task $\mathcal{T}_i$, i.e,

$$f_i(w) := \mathbb{E}_\theta[f_i(w, \theta)]. \tag{2}$$

In this case, we estimate the gradient and Hessian of f_i over a batch chosen from the dataset associated with task $\mathcal{T}_i$. More specifically, the estimation of gradient of f_i at point w over a batch $\mathcal{D}$ is given by $\nabla f_i(w, \mathcal{D}) := \frac{1}{|\mathcal{D}|} \sum_{\theta \in \mathcal{D}} \nabla f_i(w, \theta)$, which is an unbiased estimator of $\nabla f_i(w)$. Similarly, $\nabla^2 f_i(w, \mathcal{D})$ is defined as the unbiased estimator of the Hessian $\nabla^2 f_i(w)$ over the batch of data $\mathcal{D}$. To better highlight this point, as an example, consider a supervised learning problem in which each task $\mathcal{T}_i$ corresponds to the problem of finding weights w of a predictive model $h(w; .)$ such as a deep neural network for predicting the label y of a sample point x. In this case, the loss f_i is defined as the population risk $f_i(w) = \mathbb{E}_{(x,y)}[\ell(w; x, y)] = \mathbb{E}_{(x,y)}[\|y - h(w; x)\|^2]$, where $h(w; x)$ is the prediction of x's label given the model w and ℓ is the quadratic loss function. Note that for this special case, the stochastic gradient $\tilde{\nabla} f_i(w, \mathcal{D})$ is given by $\frac{1}{|\mathcal{D}|} \sum_{(x_j, y_j) \in \mathcal{D}} \nabla \ell(w; x_j, y_j)$.

In traditional statistical learning, we aim to minimize the loss function f defined in (1), as we expect its solution to be a proper approximation for the optimal solution of a new unseen task $\mathcal{T}_i$. However, in the model-agnostic meta-learning setting, we aim to find the best point that performs well as an initial point for learning a new task $\mathcal{T}_i$ when *we have budget for running a few steps of gradient descent* [15]. For simplicity, we focus on finding an initialization w such that, after observing a new task $\mathcal{T}_i$, one gradient step would lead to a good approximation for the minimizer of $f_i(w)$. We can formulate this goal as the following optimization problem

$$\min_{w \in \mathbb{R}^d} F(w) := \mathbb{E}_{i \sim p}\left[F_i(w)\right] := \mathbb{E}_{i \sim p}\left[f_i(w - \alpha \nabla f_i(w))\right], \tag{3}$$

where $\alpha > 0$ is the stepsize for the update of gradient descent method and $F_i(w)$ denotes $f_i(w - \alpha \nabla f_i(w))$.

The objective function in (3) is defined in a way that the optimal solution would perform well in expectation when we observe a task and look at the output after running a single step of *gradient* descent.[1] However, in most applications, computing the gradient for each task is costly and we only have access to a stochastic approximation of the gradient and therefore we can run steps of the *stochastic gradient* descent method. In this case, our goal is to find a point w such that when a task $\mathcal{T}_i$ is chosen, after running one step of stochastic gradient descent, the resulting solution performs well in expectation. In particular, we assume we have access to the stochastic gradient $\tilde{\nabla} f_i(w, \mathcal{D}^i_{test})$ which is an unbiased estimator of the gradient $\nabla f_i(w)$ evaluated using the batch $\mathcal{D}^i_{test}$ with size D_{test}. In this formulation, our goal would change to solving the following problem

$$\min_{w \in \mathbb{R}^d} \hat{F}(w) := \mathbb{E}_{i \sim p}\left[\mathbb{E}_{\mathcal{D}^i_{test}}\left[f_i(w - \alpha \tilde{\nabla} f_i(w, \mathcal{D}^i_{test}))\right]\right], \tag{4}$$

where the expectation is taken with respect to selection of the task i as well as selection of the random set $\mathcal{D}^i_{test}$ for computing stochastic gradient. Throughout the paper, we will clarify the connection between F and $\hat{F}$, and we report our results for both of these functions.

It is worth emphasizing that the solution of the standard expected risk minimization in (1) gives us the best answer when we are given many tasks and we plan to choose only "one action"

[1] In the above formulation, we only consider the case that one step of gradient is performed for the new task, but, indeed, a more general case is when we perform multiple steps of gradient descent update. However, running more steps of gradient update comes at the cost of computing higher order derivatives (third derivative and higher) and for simplicity of our analysis we only focus on a single iteration of gradient update which at most requires the second derivative of the loss functions f_i.

that performs well in expectation, when we observe a new unseen task. On the other hand, the solution of the expected risk minimization in (3) is designed for the case that we have access to a large number of tasks and we aim to choose "an action that after one or more steps of gradient descent" performs well for an unseen task. In the first case, we naturally choose an action that is closer to the optimal solutions of the tasks that have higher probability, but in the second case we choose an action that is closer to the optimal solutions of the tasks that have higher probability and are harder for gradient descent to solve them. For instance, when the loss functions are strongly convex and smooth, a harder task (minimization problem) for gradient descent is the problem that has a larger condition number. Therefore, the solution of (3) is naturally closer to the solution of those tasks that have larger condition numbers.

To better highlight this point we consider an example where we have access to three equally likely tasks $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ with the optimal solutions $w_*^{\mathcal{T}_1}$, $w_*^{\mathcal{T}_2}$, $w_*^{\mathcal{T}_3}$, respectively; see Figure 1. Here, w is the solution of expected risk minimization defined in (1) and $\hat{w}$ is the solution of the expected risk defined in (3). In this example, task $\mathcal{T}_1$ is the easiest task as we can make a lot of progress with only two steps of gradient descent and task $\mathcal{T}_2$ is the hardest task as we approach the optimal solution slowly by taking gradient steps. As we observe in Figure 1, for task $\mathcal{T}_3$, if we start from w the outcome after running two steps of gradient descent is almost the same as starting from $\hat{w}$. For task $\mathcal{T}_1$, however, w is a better initial point compared to $\hat{w}$, but the error of their resulted solution after two steps of gradient descent are not significantly different. This is due to the fact that $\mathcal{T}_1$ is easy and for both cases we get very close to the optimal solution even after two steps of gradient descent. The difference between starting from w and $\hat{w}$ is substantial when we aim to solve task $\mathcal{T}_2$ which is the hardest task. Because of this difference, the updated variable after running two steps of gradients has a lower expected error when we start from $\hat{w}$ comparing to the case that we start from w. This simple example illustrates the fact that if we know a-priori that after choosing an action we are allowed to run a single (or more) iteration of gradient descent to learn a new task, then it is better to start from the minimizer of (3) rather than the minimizer of the traditional statistical learning problem in (1).

In this work, we focus on gradient-based meta-learning algorithms and in particular the MAML algorithm proposed in [15] for solving Problems (3) and (4). MAML, at each iteration k, first selects a batch of tasks $\mathcal{B}_k$, and then proceeds in two stages: the *inner loop* and the *outer loop*. In the inner loop, for each chosen task $\mathcal{T}_i$ in $\mathcal{B}_k$, MAML computes a mid-point using a step of stochastic gradient on f_i. Then, in the outer loop, MAML runs one step of stochastic gradient descent with an estimate of ∇F computed over $\mathcal{B}_k$ using the the mid-points for each task $\mathcal{T}_i$ evaluated in the inner loop. We provide a detailed description of MAML in Section 2.

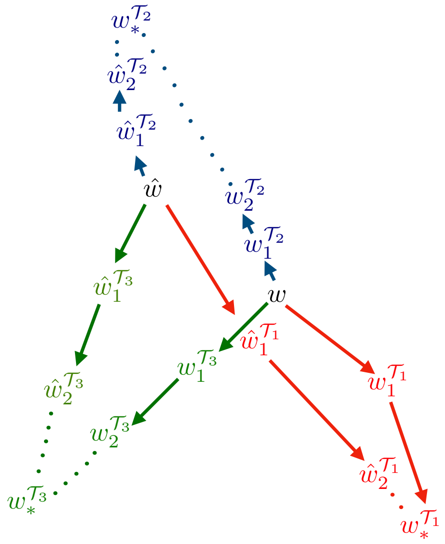

Figure 1: Comparison of the performance of the optimal solution of the statistical learning problem in (1) and the optimal solution of the met-learning problem in (3) when we have budget for two steps of gradient descent update.

1.1 Our contributions

In this paper, we provide theoretical guarantees for the convergence of MAML algorithms to first order stationarity for *non-convex* functions. To the best of our knowledge, this is the first theoretical study of the convergence of MAML algorithms in such a general setting. We build our analysis upon interpreting MAML as a stochastic gradient decent method that solves the optimization problem in (3). As we discuss in Section 4.1 in details, the analysis of MAML-type methods is significantly more challenging than the standard analysis of gradient-type methods for nonconvex problems due to several reasons. First, the function F (and $\hat{F}$) is not necessarily smooth over $\mathbb{R}^d$ since its smoothness parameter could be unbounded. Second, we need to use a stochastic stepsize (learning rate) in the update of MAML to ensure convergence which makes the convergence analysis complicated. Third, the descent direction used in the update of MAML

Algorithm	Having access to sufficient samples				K-shot Learning
	Best accuracy possible	Iteration complexity	# samples/ iteration	Runtime/ iteration	Best accuracy possible
MAML	$\|\nabla F(w)\| \leq \epsilon$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(1/\epsilon^4)$	$\boldsymbol{\mathcal{O}(d^2)}$	$\|\nabla F(w)\| \leq \mathcal{O}(\tilde{\sigma}/\sqrt{K})$
FO-MAML	$\|\nabla F(w)\| \leq \boldsymbol{\mathcal{O}(\alpha\sigma)}$	$\mathcal{O}(1/(\alpha^2\sigma^2))$	$\mathcal{O}(1/(\alpha^4\sigma^2))$	$\mathcal{O}(d)$	$\|\nabla F(w)\| \leq \mathcal{O}(\boldsymbol{\sigma} + \tilde{\sigma}/\sqrt{K})$
HF-MAML	$\|\nabla F(w)\| \leq \epsilon$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(1/\epsilon^4)$	$\mathcal{O}(d)$	$\|\nabla F(w)\| \leq \mathcal{O}(\tilde{\sigma}/\sqrt{K})$

Table 1: Our theoretical results for convergence of MAML, first-order approximation of MAML (FO-MAML), and our proposed Hessian-free MAML (HF-MAML) to a first-order stationary point (FOSP) in nonconvex settings. Here, d is the problem dimension, σ is a bound on the standard deviation of $\nabla f_i(w)$ from its mean $\nabla f(w)$, and $\tilde{\sigma}$ is a bound on the standard deviation of $\nabla f_i(w.\theta)$, an unbiased estimate of $\nabla f_i(w)$, from its mean $\nabla f_i(w)$, for every i. For any $\epsilon > 0$, MAML can find an ϵ-FOSP, while each iteration has a complexity of $\mathcal{O}(d^2)$. FO-MAML has a lower complexity of $\mathcal{O}(d)$, but it cannot reach a point with gradient norm less than $\mathcal{O}(\alpha\sigma)$. HF-MAML has the best of both worlds, i.e., HF-MAML has a cost of $\mathcal{O}(d)$, and it can find an ϵ-FOSP for any $\epsilon > 0$.

is not an unbiased estimator of the true gradient and we need to handle the effect of gradient approximation bias in the analysis of MAML methods.

Overcoming these challenges, we characterize the iteration and sample complexity of MAML method and shed light on the relation of batch sizes and parameters of MAML with its convergence rate and accuracy. Using these results, we provide an explicit approach for tuning the hyper-parameters of MAML and also the required amount of data for reaching a first order stationary point of problem (3). Specifically, we show that when the loss associated to each task f_i is a general smooth nonconvex function, MAML finds an ϵ-first-order stationary point of the stochastic programs defined in (3) and (4) after at most $\mathcal{O}(1/\epsilon^2)$ iterations and $\mathcal{O}(1/\epsilon^4)$ stochastic gradient evaluations per iteration, while the complexity of each iteration is of $\mathcal{O}(d^2)$. Our result also directly applies to the specific case of K-shot learning where for each task in the inner loop we only have access to K samples to estimate the gradient. In this case, MAML finds a solution w_{KS} such that $\|\nabla F(w_{KS})\| \leq \mathcal{O}(\sqrt{\tilde{\sigma}^2/K})$ [2] where $\tilde{\sigma}^2$ denotes the bound on the variance of stochastic gradient approximation of $\nabla f_i(w)$, i.e., an unbiased estimator of $\nabla f_i(w)$ using only one sample (see Assumption 4.6 for the precise definition).

As described in [15], the exact implementation of MAML requires access to second-order information of the loss functions $\nabla^2 f_i$, and, therefore, the computational complexity of MAML per iteration scales poorly with the problem dimension d, i.e., the cost per iteration for MAML is $\mathcal{O}(d^2)$. To resolve this issue, the authors in [15] suggest ignoring the second-order term in the update rule of MAML and show that the proposed first-order approximation does not affect the performance of MAML in practice. In our work, we formally characterize the convergence results for this first-order approximation of MAML (*FO-MAML*) and show that if the learning rate α used for updating each individual task is small or the tasks are statistically close to each other, then the error induced by the first-order approximation is negligible. To be more precise, we show that, when we have access to sufficient samples, FO-MAML finds a point $w^\dagger$ such that $\|\nabla F(w^\dagger)\| \leq \mathcal{O}(\alpha\sigma)$ in $\mathcal{O}(1/(\alpha^2\sigma^2))$ iterations. Hence, when α or σ are small enough such that $\alpha\sigma = \mathcal{O}(\epsilon)$ we do not lose that much compare to MAML by ignoring the second term, i.e., we can find a point with its gradient norm of $\mathcal{O}(\epsilon)$. Nevertheless, in general, in contrast to MAML which can find an ϵ-first order stationary point for any arbitrary ϵ, FO-MAML is limited to $\epsilon \geq \mathcal{O}(\alpha\sigma)$.

To address this issue, we introduce a new method, *Hessian-free MAML (HF-MAML)*, which recovers the complexity bounds for MAML while it does not require computation of the Hessian or a Hessian-vector product and has a computational complexity of $\mathcal{O}(d)$ for each iteration. In fact, we show that HF-MAML enjoys a better convergence rate in comparison to FO-MAML, and, for any positive ϵ, it can find an ϵ-first-order stationary point after at most $\mathcal{O}(1/\epsilon^2)$ iterations while keeping the computational cost at each iteration of $\mathcal{O}(d)$. Hence, HF-MAML has the best of both worlds meaning that it has a low computational complexity of $\mathcal{O}(d)$ as in FO-MAML and it can achieve any arbitrary accuracy for first-order stationarity as in MAML.

A summary of our results is provided in Table 1. For formal definitions of σ and $\tilde{\sigma}$ please check Assumptions 4.5 and 4.6, respectively.

[2] We assume σ and $\tilde{\sigma}$ small enough for the results we state in this section. The general result can be found in Section 4.

Algorithm 1: Model Agnostic Meta Learning (MAML)

```
Input : Initial iterate $w_0$
while not done do
    Choose a batch of i.i.d. tasks $\mathcal{B}_k \subseteq \mathcal{I}$ with distribution $p$ and with size $B = |\mathcal{B}_k|$;
    for all $\mathcal{T}_i$ with $i \in B$ do
        Compute the stochastic gradient $\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i)$ using dataset $\mathcal{D}_{in}^i$;
        Set $w_{k+1}^i = w_k - \alpha \tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i)$;
    end
    Set $w_{k+1} = w_k - \frac{\beta_k}{B} \sum_{i \in \mathcal{B}_k} \left(I - \alpha \tilde{\nabla}^2 f_i(w_k, \mathcal{D}_h^i)\right) \tilde{\nabla} f_i(w_{k+1}^i, \mathcal{D}_o^i)$ using $\mathcal{D}_o^i$ and $\mathcal{D}_h^i$
    $k \leftarrow k + 1$
end
```

2 Model-Agnostic Meta-Learning (MAML) Algorithm

The Model-Agnostic Meta-Learning (MAML) algorithm was proposed in [15] for solving the stochastic optimization problem in (3). In MAML, at each step k, we choose a subset $\mathcal{B}_k$ of the tasks, with each task drawn independently from distribution p. For simplicity assume that the size of $\mathcal{B}_k$ is fixed and equal to B. Then, the update of MAML is implemented at two levels: (i) inner step and (ii) outer step (meta-step).

In the inner step, for each task $\mathcal{T}_i$, we use a subset of the dataset $\mathcal{D}_{in}^i$ corresponding to task $\mathcal{T}_i$ to compute the stochastic gradient $\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i)$ which is an an unbiased estimator of the gradient $\nabla f_i(w_k)$ associated with $\mathcal{T}_i$. The stochastic gradient $\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i)$ is then used to compute a model w_{k+1}^i corresponding to each task $\mathcal{T}_i$ by a single iteration of stochastic gradient descent, i.e.,

$$w_{k+1}^i = w_k - \alpha \tilde{\nabla} f_i(w_k, D_{in}^i). \tag{5}$$

To simplify the notation, we assume the size of dataset $\mathcal{D}_{in}^i$ for all tasks i are equal to each other and we denote it by D_{in}.

In the outer loop, once we have the updated models $\{w_{k+1}^i\}_{i=1}^B$ for all the chosen tasks in $\mathcal{B}_k$, we compute the revised meta-model w_{k+1} by performing the update

$$\begin{aligned} w_{k+1} &= w_k - \beta_k \frac{1}{B} \sum_{i \in \mathcal{B}_k} \left(I - \alpha \tilde{\nabla}^2 f_i(w_k, \mathcal{D}_h^i)\right) \tilde{\nabla} f_i \left(w_k - \alpha \tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i\right) && \text{(6a)} \\ &= w_k - \beta_k \frac{1}{B} \sum_{i \in \mathcal{B}_k} \left(I - \alpha \tilde{\nabla}^2 f_i(w_k, \mathcal{D}_h^i)\right) \tilde{\nabla} f_i(w_{k+1}^i, \mathcal{D}_o^i), && \text{(6b)} \end{aligned}$$

where the stochastic gradient $\tilde{\nabla} f_i(w_{k+1}^i, \mathcal{D}_o^i)$ corresponding to task $\mathcal{T}_i$ is evaluated using the data set $\mathcal{D}_o^i$ and the models $\{w_{k+1}^i\}_{i=1}^B$ computed in the inner loop, and the stochastic Hessian $\tilde{\nabla}^2 f_i(w_k, \mathcal{D}_h^i)$ for each task $\mathcal{T}_i$ is computed using the data set $\mathcal{D}_h^i$. Note that the data sets $\mathcal{D}_{in}^i$ used for the inner update are different from the data sets $\mathcal{D}_o^i$ and $\mathcal{D}_h^i$ used for the outer update. It is also possible to assume that $\mathcal{D}_o^i = \mathcal{D}_h^i$, but in this paper we assume that $\mathcal{D}_o^i$ and $\mathcal{D}_h^i$ are independent from each other that allows us to use a smaller batch for the stochastic Hessian computation which is more costly. Here also we assume that the sizes of $\mathcal{D}_o^i$ and $\mathcal{D}_h^i$ are fixed for all i and equal to D_o and D_h, respectively. The steps of the MAML method are summarized in 1.

MAML as an approximation of SGD. To better highlight the fact that MAML runs a stochastic gradient step for the objective function F in (3), consider the update of gradient descent for minimizing the objective function F with step size β_k which can be written as[3]

$$\begin{aligned} w_{k+1} &= w_k - \beta_k \nabla F(w_k) \\ &= w_k - \beta_k \mathbb{E}_{i \sim p} \left[\left(I - \alpha \nabla^2 f_i(w_k)\right) \nabla f_i(w_k - \alpha \nabla f_i(w_k)) \right]. \end{aligned} \tag{7}$$

As the underlying probability distribution of tasks p is unknown, evaluation of the expectation in the right hand side of (7) is often computationally prohibitive. Therefore, one can use a stochastic

[3] We are allowed to change the order of expectation and derivative as the number of tasks is finite.

Algorithm 2: First-order MAML (FO-MAML)

```
Input : Initial iterate $w_0$
while not done do
    Choose a batch of i.i.d. tasks $\mathcal{B}_k \subseteq \mathcal{I}$ with distribution $p$ and with size $B = |\mathcal{B}_k|$;
    for all $\mathcal{T}_i$ with $i \in B$ do
        Compute the stochastic gradient $\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i)$ using dataset $\mathcal{D}_{in}^i$;
        Set $w_{k+1}^i = w_k - \alpha \tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i)$;
    end
    Set $w_{k+1} = w_k - \frac{\beta_k}{B} \sum_{i \in \mathcal{B}_k} \tilde{\nabla} f_i(w_{k+1}^i, \mathcal{D}_o^i)$ using dataset $\mathcal{D}_o^i$
    $k \leftarrow k+1$
end
```

gradient update for minimizing the function F with a batch $\mathcal{B}_k$ which contains B tasks that are independently drawn. Then, the update can be written as

$$w_{k+1} = w_k - \beta_k \frac{1}{B} \sum_{i \in \mathcal{B}_k} \left(I - \alpha \nabla^2 f_i(w_k)\right) \nabla f_i(w_k - \alpha \nabla f_i(w_k)) \,. \tag{8}$$

In the above update we have assumed that the exact gradients and Hessians of f_i, i.e., ∇f_i and $\nabla^2 f_i$, are available and we only sample the tasks. By comparing the updates in (6a) and (8) we obtain that the update of MAML can be considered as a stochastic gradient descent on F when we do not have access to the gradient ∇f_i and Hessian $\nabla^2 f_i$ of the loss associated with tasks and we approximate them by their stochastic variants. This is due to the fact that in many settings the function f_i is defined either as the expectation over an infinite set of random functions as in (2) or as the sum of a massive number of functions where each individual function is associated with a sample of a very large dataset.

Smaller batch selection for Hessian approximation. The use of first-order methods for solving problem (3) requires computing the gradient of the objective function F which needs evaluating the second derivative of the loss function f_i. Indeed, computation of the Hessians $\nabla^2 f_i$ for all the chosen tasks at each iteration is costly and is of order $\mathcal{O}(d^2 B D_h)$, where d is the dimension of the problem, D_h is the size of batch used for stochastic Hessian approximation, and B is the number of tasks chosen at each iteration for the update of MAML. One approach to lower this cost is to reduce the batch size D_h used for Hessian approximation. Later in our analysis for MAML, we show that we can perform the update in (6a) and have an exact convergent method, while having the size of batch D_h significantly smaller than the size of batches D_{in} and B.

First-order MAML (FO-MAML). To reduce the cost of implementing the update of MAML one might suggest ignoring the second-order term that appears in the update of MAML. In this approach, which is also known as first-order (approximation of) MAML (FO-MAML) [15], we update the model w_k by following the update

$$w_{k+1} = w_k - \beta_k \frac{1}{B} \sum_{i \in \mathcal{B}_k} \tilde{\nabla} f_i(w_{k+1}^i, \mathcal{D}_o^i), \tag{9}$$

where the points w_{k+1}^i are evaluated according to the update in (5). Indeed, this approximation reduces the computational complexity of implementing MAML, but it comes at the cost of inducing an extra error in computation of the stochastic gradient of F. We formally characterize this error in our theoretical results and show under what conditions the error induced by ignoring the second-order term does not impact its convergence. The steps of FO-MAML are outlined in Algorithm 2.

3 Hessian-free MAML (HF-MAML)

The update of MAML for solving the meta-learning problem introduced in (3) requires access to the Hessian of the individual functions $\nabla^2 f_i(w) \in \mathbb{R}^{d \times d}$ and computation of a Hessian-vector product, i.e., $\nabla_i^2 f(w)\nabla f_i(w-\alpha\nabla f_i)$. Computation of the Hessian $\nabla^2 f_i(w)$ and its product with any vector of size d has a computational complexity of $\mathcal{O}(d^2)$. Indeed, the cost of this computation is prohibitive when the dimension of the problem d is very large which is the case in most applications of meta-learning. To resolve this issue we propose an approximate variant of the MAML method that is *Hessian-free*, i.e., only requires evaluation of gradients, and has a computational cost of $\mathcal{O}(d)$ per iteration.

The idea behind our proposed method is that for any function ϕ, the product of its Hessian $\nabla^2\phi(w)$ by any vector v can be approximated by

$$\nabla^2\phi(w)v \approx \left[\frac{\nabla\phi(w+\delta v)-\nabla\phi(w-\delta v)}{2\delta}\right] \tag{10}$$

with an error of at most $\rho\delta\|v\|^2$, where ρ is the parameter for Lipschitz continuity of the Hessian of ϕ. Based on this approximation, we propose a computationally efficient approach for minimizing the expected loss F defined in (3) which we refer to it as Hessian-free MAML (HF-MAML). As the name suggests the HF-MAML is an approximation of the MAML that does not require evaluation of any Hessian, while it provides a very accurate approximation of MAML. To be more precise, the update of HF-MAML is defined as

$$w_{k+1} = w_k - \beta_k \frac{1}{B}\sum_{i\in\mathcal{B}_k}\left[\tilde\nabla f_i\Big(w_k-\alpha\tilde\nabla f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i\Big)-\alpha d_k^i\right] \tag{11}$$

where α is the step size for each task, β_k is the stepsize for the meta update, and the vectors d_k^i are defined as

$$d_k^i := \frac{\tilde\nabla f_i\Big(w_k+\delta_k^i\tilde\nabla f_i(w_k-\alpha\tilde\nabla f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i),\mathcal{D}_h^i\Big)-\tilde\nabla f_i\Big(w_k-\delta_k^i\tilde\nabla f_i(w_k-\alpha\tilde\nabla f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i),\mathcal{D}_h^i\Big)}{2\delta_k^i}. \tag{12}$$

Note that d_k^i is an approximation for $\tilde\nabla^2 f_i(w_k,\mathcal{D}_h^i)\tilde\nabla f_i(w_k-\alpha\tilde\nabla f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)$ which appears in the gradient of the objective function F. In addition, $\delta_k^i > 0$ is a positive scalar which indicates the level of approximation for our proposed Hessian-vector product approximation.

As in MAML, this update can be implemented efficiently in two stages where in the first one we compute the mid-points w_{k+1}^i by performing the following update for all tasks $\mathcal{T}_i \in \mathcal{B}_k$

$$w_{k+1}^i = w_k - \alpha\tilde\nabla f_i(w_k,\mathcal{D}_{in}^i), \tag{13}$$

which is identical to the update in (5). Then we use these models in the outer loop to evaluate the vectors d_k^i based on the update

$$d_k^i = \frac{\tilde\nabla f_i\left(w_k+\delta_k^i\tilde\nabla f_i(w_{k+1}^i,\mathcal{D}_o^i),\mathcal{D}_h^i\right)-\tilde\nabla f_i\left(w_k-\delta_k^i\tilde\nabla f_i(w_{k+1}^i,\mathcal{D}_o^i),\mathcal{D}_h^i\right)}{2\delta_k^i}, \tag{14}$$

where the vector d_k^i approximates the Hessian-vector product $\tilde\nabla^2 f_i(w_k,\mathcal{D}_h^i)\tilde\nabla f_i(w_{k+1}^i,\mathcal{D}_o^i)$. Then we compute the updated meta model w_{k+1} according to the update

$$w_{k+1} = w_k - \frac{\beta_k}{B}\sum_{i\in\mathcal{B}_k}\left(\tilde\nabla f_i(w_{k+1}^i,\mathcal{D}_o^i)-\alpha d_k^i\right). \tag{15}$$

The steps of the HF-MAML method are outlined in Algorithm 3. In our theoretical results we formally characterize the error induced by the Hessian-vector product approximation and show how one should choose all the parameters of HF-MAML including B, D_{in}, D_o, D_h, β_k, and δ_k^i.

4 Theoretical Results

In this section we establish our theoretical results for MAML, FO-MAML, and HF-MAML when the loss function f_i corresponding to task $\mathcal{T}_i$ is a general nonconvex but smooth function. In particular, we characterize the overall computational complexity of these algorithms for finding a first-order stationary point of the global objective function F. A formal definition of a first-order stationary point follows.

Algorithm 3: Hessian-free MAML (HF-MAML)

```
Input : Initial iterate $w_0$
while not done do
    Choose a batch of i.i.d. tasks $\mathcal{B}_k \subseteq \mathcal{I}$ with distribution $p$ and with size $B = |\mathcal{B}_k|$;
    for all $\mathcal{T}_i$ with $i \in B$ do
        Compute the stochastic gradient $\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i)$;
        Set $w_{k+1}^i = w_k - \alpha \tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i)$;
    end
    Compute $d_k^i = \frac{\tilde{\nabla} f_i\left(w_k + \delta_k^i \tilde{\nabla} f_i(w_{k+1}^i, \mathcal{D}_o^i), \mathcal{D}_h^i\right) - \tilde{\nabla} f_i\left(w_k - \delta_k^i \tilde{\nabla} f_i(w_{k+1}^i, \mathcal{D}_o^i), \mathcal{D}_h^i\right)}{2\delta_k^i}$;
    Set $w_{k+1} = w_k - \frac{\beta_k}{B} \sum_{i \in \mathcal{B}_k} \left(\tilde{\nabla} f_i(w_{k+1}^i, \mathcal{D}_o^i) - \alpha d_k^i\right)$;
    $k \leftarrow k + 1$;
end
```

Definition 4.1. *A random vector $w_\epsilon \in \mathbb{R}^d$ is called an ϵ-approximate first order stationary point (FOSP) for problem (3) if it satisfies*

$$\mathbb{E}[\|\nabla F(w_\epsilon)\|] \leq \epsilon. \tag{16}$$

This definition implies that a point w_ϵ is an ϵ-FOSP if its gradient norm for the global loss F is smaller than ϵ in expectation.

Our goal in this section is to answer two fundamental questions for each of the three considered methods. Can they find an ϵ-FOSP for arbitrary $\epsilon > 0$? If yes, how many iterations is needed for achieving such point? Before answering these questions, we first formally state the assumptions that we use for our analysis in the rest of the section.

Assumption 4.2. *Function F is bounded below, i.e., $\min_{w \in \mathbb{R}^d} F(w) > -\infty$.*

Assumption 4.3. *For every $i \in \mathcal{I}$, f_i is twice continuously differentiable and L_i-smooth, i.e., for every $w, u \in \mathbb{R}^d$, we have*

$$\|\nabla f_i(w) - \nabla f_i(u)\| \leq L_i \|w - u\|. \tag{17}$$

Since we also assume that the functions f_i are twice differentiable, the L_i-smoothness assumption also implies that

$$-L_i I_d \preceq \nabla^2 f_i(w) \preceq L_i I_d \quad \forall w \in \mathbb{R}^d, \tag{18a}$$

$$-\frac{L_i}{2}\|w-u\|^2 \leq f_i(w) - f_i(u) - \nabla f_i(u)^\top (w-u) \leq \frac{L_i}{2}\|w-u\|^2 \quad \forall w, u \in \mathbb{R}^d. \tag{18b}$$

For the simplicity of analysis, in the rest of the paper, we mostly work with $L := \max_i L_i$ which can be considered as a parameter for the Lipschitz continuity of the gradients ∇f_i for all $i \in \mathcal{I}$.

As we described in Section 2, the second derivative of the functions f_i appear in the update of MAML. Therefore, we need to impose a regularity condition on the variation of the objective function Hessians expressed in the next Assumption, which is commonly used in the analysis of second-order methods.

Assumption 4.4. *For every $i \in \mathcal{I}$, the Hessian of function f_i is ρ_i-Lipschitz continuous, i.e., for every $w, u \in \mathbb{R}^d$, we have*

$$\|\nabla^2 f_i(w) - \nabla^2 f_i(u)\| \leq \rho_i \|w - u\|. \tag{19}$$

Also, in the rest of the paper, to simplify our notation we use $\rho := \max \rho_i$ as the parameter for the Lipschitz continuity of the Hessians $\nabla^2 f_i$ for all $i \in \mathcal{I}$.

We would like to add that we do not assume any smoothness conditions for the global loss F and all the required conditions are for the individual loss functions f_i. In fact, later in this section, we show that under the conditions in Assumption 4.3, the global loss F may not be gradient-Lipschitz in general.

The goal of Meta-learning is to train a model based on a set of given tasks so that this model can be used for learning a new unseen task. However, this is only possible if the training tasks are somehow related to unseen (test) tasks. In the following assumption, we formalize this condition by assuming that the gradient ∇f_i, which is an unbiased estimator of the expected loss gradient $\nabla f = \mathbb{E}_{i\sim p}[\nabla f_i(w)]$, has bounded variance.

Assumption 4.5. *For any $w \in \mathbb{R}^d$, the variance of gradient $\nabla f_i(w)$ is bounded, i.e., for some non-negative σ, we have*

$$\mathbb{E}_{i\sim p}[\|\nabla f(w) - \nabla f_i(w)\|^2] \leq \sigma^2, \tag{20}$$

for any $w \in \mathbb{R}^d$.

As we stated in the introduction, in many applications, evaluating the exact gradient corresponding to the loss function of each task is computationally costly, and we only have access to a batch of data to compute an unbiased estimator of the exact gradient $\nabla f_i(w)$. In the following assumption we formally state the conditions required for the stochastic approximations of the gradients $\nabla f_i(w, \theta)$ and Hessians $\nabla^2 f_i(w, \theta)$.

Assumption 4.6. *For any i and any $w \in \mathbb{R}^d$, the stochastic gradients $\nabla f_i(w, \theta)$ and Hessians $\nabla^2 f_i(w, \theta)$ have bounded variance, i.e.,*

$$\mathbb{E}_\theta[\|\nabla f_i(w, \theta) - \nabla f_i(w)\|^2] \leq \tilde{\sigma}^2, \tag{21}$$

$$\mathbb{E}_\theta[\|\nabla^2 f_i(w, \theta) - \nabla^2 f_i(w)\|^2] \leq \sigma_H^2, \tag{22}$$

where $\tilde{\sigma}$ and σ_H are non-negative constants.

In particular, it is worth noting that the special case $\tilde{\sigma} = \sigma_H = 0$ represents the case that the exact gradient and Hessian for each task is available.

Finally, to simplify the statement of our results for MAML, FO-MAML, and HF-MAML, we make the following assumption on the relation of parameters. Later in the appendix, we drop this assumption and state the general version of our results.

Assumption 4.7. *Recall the definitions $L := \max L_i$ and $\rho := \max \rho_i$ with L_i and ρ_i defined in Assumptions 4.3 and 4.4, respectively. Then, we assume $\rho\alpha/L = \mathcal{O}(1)$. Also, we assume $\sigma^2 + \tilde{\sigma}^2 = \mathcal{O}(1)$, where σ and $\tilde{\sigma}$ are defined in Assumptions 4.5 and 4.6, respectively.*[4]

Remark 4.8. *We would like to emphasize that all the conditions in Assumption 4.7 are added only to simplify the statements of our main theorems. All our results hold even if the conditions in Assumption 4.7 are not satisfied. In fact, in the appendix, we state our proofs for the general case in which Assumption 4.7 does not hold.*

4.1 Challenges in analyzing MAML algorithms

Before stating our main results for MAML, FO-MAML, and HF-MAML, in this subsection we briefly highlight some of the challenges that emerge in analyzing these algorithms and prove some intermediate lemmas that we will use in the following subsections.

(I) Unbounded smoothness parameter: As we mentioned in the previous section, the global loss function F that we are minimizing in the MAML algorithm by following a stochastic gradient descent step is not necessarily smooth over $\mathbb{R}^d$, and its smoothness parameter could be unbounded. We formally characterize the parameter for the Lipschitz continuity of the gradients ∇F in the following lemma.

Lemma 4.9. *Consider the objective function F defined in (3) for the case that $\alpha \in [0, \frac{1}{L}]$. Suppose that the conditions in Assumptions 4.3-4.4 are satisfied, and recall the definitions $L := \max L_i$ and $\rho := \max \rho_i$. Then, for any $w, u \in \mathbb{R}^d$ we have*

$$\|\nabla F(w) - \nabla F(u)\| \leq (4L + 2\rho\alpha \min\{\mathbb{E}_{i\sim p}\|\nabla f_i(w)\|, \mathbb{E}_{i\sim p}\|\nabla f_i(u)\|\})\, \|w - u\|. \tag{23}$$

Proof. Check Appendix B. □

[4] Note that the constants in $\mathcal{O}(1)$ do not depend on any of the parameters of the problem.

The result in Lemma 4.9 shows that the objective function F is smooth with a parameter that depends on the minimum of the expected norm of gradients. In other words, when we measure the smoothness of gradients between two points w and u, the smoothness parameter depends on $\min\{\mathbb{E}_{i\sim p}\|\nabla f_i(w)\|, \mathbb{E}_{i\sim p}\|\nabla f_i(u)\|\}$. Indeed, this term could be unbounded or arbitrarily large as we have no assumption on the gradients norm. In addition, computation of $\min\{\mathbb{E}_{i\sim p}\|\nabla f_i(w)\|, \mathbb{E}_{i\sim p}\|\nabla f_i(u)\|\}$ could be costly as it requires access to the gradients of all tasks.

Before moving to next challenge in analyzing MAML, we use Lemma 4.9 to show the following result which is analogous to (18b) for F. We skip the proof as it is very similar to the proof of Lemma 1.2.3 in [24].

Corollary 4.10. *Let $\alpha \in [0, \frac{1}{L}]$. Then, for every $w, u \in \mathbb{R}^d$, we have*

$$-\frac{L(w)}{2}\|u-w\|^2 \leq F(u) - F(w) - \nabla F(w)^\top (u-w) \leq \frac{L(w)}{2}\|u-w\|^2, \tag{24}$$

where $L(w) = 4L + 2\rho\alpha \mathbb{E}_{i\sim p}\|\nabla f_i(w)\|$.

(II) Stochastic stepsize: In many optimization algorithms, including gradient descent and stochastic gradient descent, the stepsize is often selected proportional to the inverse of the smoothness parameter. However, as studied above, in our setting, this parameter depends on the norm of gradient of all tasks which is not computationally tractable. To resolve this issue, we propose a method for choosing the stepsize β_k by approximating the expectation in $L(w)$ with an average over a batch of tasks. In other words, we approximate $\mathbb{E}_{i\sim p}\|\nabla f_i(w)\|$ in the definition of $L(w)$ using the estimator $\sum_{j\in\mathcal{B}'}\|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|$ where $\mathcal{D}_\beta^j$ is a dataset corresponding to task j with size D_β. Hence, we estimate $L(w)$ by

$$\tilde{L}(w) := 4L + \frac{2\rho\alpha}{B'}\sum_{j\in\mathcal{B}'}\|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|. \tag{25}$$

Using this estimate, our stepsize β_k is tuned to be a constant times the inverse of $\tilde{L}(w)$ which we denote by $\tilde{\beta}(w) = 1/\tilde{L}(w)$, i.e., $\beta_k = c\tilde{\beta}(w) = c/\tilde{L}(w)$. This simple observation shows that the stepsize that we need to use for MAML algorithms is stochastic as $1/\tilde{L}(w)$ is a random parameter and depends on the choice of $\mathcal{B}'$. Therefore, we need to derive lower and upper bounds on the expectations $E[\beta_k]$ and $\mathbb{E}[\beta_k^2]$, respectively, as they appear in the convergence analysis of gradient-based methods. Considering the defintion $\beta_k = c\tilde{\beta}(w)$, we state these bounds for $\tilde{\beta}(w)$ in the following lemma.

Lemma 4.11. *Consider the objective function F defined in (3) for the case that $\alpha \in [0, \frac{1}{L}]$. Suppose that the conditions in Assumptions 4.3-4.6 are satisfied, and recall the definitions $L := \max L_i$ and $\rho := \max \rho_i$. Further, consider the definition*

$$\tilde{\beta}(w) := \frac{1}{\tilde{L}(w)} := \frac{1}{4L + 2\rho\alpha\sum_{j\in\mathcal{B}'}\|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|/B'}, \tag{26}$$

where $\mathcal{B}'$ is a batch of tasks with size B' which are independently drawn with distribution p, and for every $j \in \mathcal{B}'$, $\mathcal{D}_\beta^j$ is a dataset corresponding to task j with size D_β. Then, if the conditions

$$B' \geq \left\lceil \frac{1}{2}\left(\frac{\rho\alpha\sigma}{L}\right)^2 \right\rceil, \quad D_\beta \geq \left\lceil \left(\frac{2\rho\alpha\tilde{\sigma}}{L}\right)^2 \right\rceil \tag{27}$$

are satisfied, then we have

$$\mathbb{E}\left[\tilde{\beta}(w)\right] \geq \frac{0.8}{L(w)}, \quad \mathbb{E}\left[\tilde{\beta}(w)^2\right] \leq \frac{3.125}{L(w)^2} \tag{28}$$

where $L(w)$ is the Lipschitz parameter of ∇F at w defined in Corollary 4.10.

Proof. Check Appendix C. □

The result in Lemma 4.11 shows that if we choose the stepsize $\beta_k = c\tilde{\beta}(w_k)$, with $\tilde{\beta}(w_k)$ given in (26) and the batch-sizes B' and D_β satisfy the conditions (27), then the first moment of β_k is bounded below by a factor of $1/L(w_k)$ and its second moment is upper bounded by a factor of $1/L(w_k)^2$. We will use these inequalities in the convergence analysis of MAML methods. We would like to add that, throughout the paper, we assume at each iteration k, the batches $\mathcal{B}'_k$, $\{\mathcal{D}^j_\beta\}_{j\in\mathcal{B}'_k}$ are independently drawn from $\mathcal{B}_k$ and $\{\mathcal{D}^i_{in}, \mathcal{D}^i_o, \mathcal{D}^i_h\}_{i\in\mathcal{B}_k}$ used in the updates of MAML methods. Also, it is worth to emphasize that the batch size for the random sets $\mathcal{B}'_k$ and $\{\mathcal{D}^j_\beta\}_{j\in\mathcal{B}'_k}$ are independent of the desired accuracy ϵ and the extra cost for the computation of these batches is of $\mathcal{O}(1)$.

(III) Biased estimator: The statement that MAML performs an update of stochastic gradient descent at each iteration on the objective function F is not quite accurate. To better highlight this point, recall the update of MAML in (6a). According to this update, the descent direction g_k for MAML at step k is given by

$$g_k := \frac{1}{B}\sum_{i\in\mathcal{B}_k}\left(I - \alpha\tilde{\nabla}^2 f_i(w_k, D^i_h)\right)\tilde{\nabla} f_i\Big(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o\Big), \tag{29}$$

while the exact gradient of F at the iterate w_k is given by

$$\nabla F(w_k) = \mathbb{E}_{i\sim p}\left[\left(I - \alpha\nabla^2 f_i(w_k)\right)\nabla f_i(w_k - \alpha\nabla f_i(w_k))\right]. \tag{30}$$

It can be easily verified that, given w_k, g_k is not an unbiased estimator of the gradient $\nabla F(w_k)$ as the stochastic gradient $\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in})$ is within the stochastic gradient $\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o)$. Therefore, the descent direction that we use in the update of MAML for updating models is a biased estimator of the gradient $\nabla F(w_k)$. This is another challenge that we face in analyzing MAML and its variants. To overcome this challenge, we need to characterize the first-order and second-order moments of the expression $\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o)$, as we do in the following lemma.

Lemma 4.12. *Consider the objective function F defined in (3) for the case that $\alpha \in [0, \frac{1}{L}]$. Suppose that the conditions in Assumptions 4.3-4.6 are satisfied, and recall the definitions $L := \max L_i$ and $\rho := \max \rho_i$. Then, we can show that*

$$\mathbb{E}_{\mathcal{D}_{in},\mathcal{D}_o}[\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o) \mid \mathcal{F}_k] = \nabla f_i\left(w_k - \alpha\nabla f_i(w_k)\right) + e_{i,k}, \quad \textit{where} \quad \|e_{i,k}\| \leq \frac{\alpha L\tilde{\sigma}}{\sqrt{D_{in}}}. \tag{31}$$

Moreover, the second moment of $\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o)$ is bounded above by

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_{in},\mathcal{D}_o}\left[\|\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o)\|^2 \mid \mathcal{F}_k\right] \\ &\leq \left(1 + \frac{1}{\phi}\right)\|\nabla f_i(w_k - \alpha\nabla f_i(w_k))\|^2 + \frac{(1+\phi)\alpha^2L^2\tilde{\sigma}^2}{D_{in}} + \frac{\tilde{\sigma}^2}{D_o},\end{aligned} \tag{32}$$

where ϕ is an arbitrary positive constant.

Proof. Check Appendix D. □

The result in Lemma 4.12 clarifies the reason that the descent direction of MAML denoted by g_k is a biased estimator of $\nabla F(w_k)$, as $\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o)$ is not an unbiased estimator of $\nabla f_i(w_k - \alpha\nabla f_i(w_k))$. In particular, it shows that the bias is bounded above by a constant which depends on the variance of the stochastic gradients $\tilde{\nabla} f_i$ as well as the stepsize α for the inner steps. Indeed, by setting $\alpha = 0$, the vector $\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o)$ becomes an unbiased estimator of $\nabla f_i(w_k - \alpha\nabla f_i(w_k))$ as our result in (31) also suggests. Also, the result in (32) shows that the second moment of $\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o)$ is bounded above by the sum of a multiplicand of $\|\nabla f_i(w_k - \alpha\nabla f_i(w_k))\|^2$ and a multiplicand of the variance $\tilde{\sigma}^2$. For the special case of $\alpha = 0$, we would expect the second moment of $\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o)$ to be bounded above by $\|\nabla f_i(w_k - \alpha\nabla f_i(w_k))\|^2 + \tilde{\sigma}^2/D_o$, which can also be obtained from (32) by setting $\phi \to \infty$.

4.2 On the Connection of F and $\hat{F}$

In this subsection, we investigate the connection between F and $\hat{F}$ defined in (3) and (4), respectively. In particular, in the following theorem, we characterize how far the gradient of F and $\hat{F}$ can be from each other. Later, using this result, we show all the methods that we study achieve the same level of gradient norm with respect to both F and $\hat{F}$, up to some constant.

Theorem 4.13. *Consider the objective functions F and $\hat{F}$ defined in (3) and (4), respectively, for the case that $\alpha \in (0, \frac{1}{L}]$. Suppose that the conditions in Assumptions 4.3-4.6 are satisfied, and recall the definitions $L := \max L_i$ and $\rho := \max \rho_i$. Then, for any $w \in \mathbb{R}^d$, we have*

$$\|\nabla \hat{F}(w) - \nabla F(w)\| \leq 2\alpha L \frac{\tilde{\sigma}}{\sqrt{D_{test}}} + \alpha^2 L \frac{\sigma_H \tilde{\sigma}}{D_{test}}. \tag{33}$$

Proof. Check Appendix E. □

In the upcoming results, we mainly focus on characterizing the behavior of MAML, its first order approximation, and HF-MAML with respect to F, and by using the above theorem, we can immediately obtain bounds on the norm of $\nabla \hat{F}$ as well. In fact, the above theorem indicates the difference between ∇F and $\nabla \hat{F}$ is $\mathcal{O}\left(\max\{\frac{\tilde{\sigma}}{\sqrt{D_{test}}}, \frac{\sigma_H \tilde{\sigma}}{D_{test}}\}\right)$.

4.3 Convergence of MAML

In this subsection, we study the overall complexity of MAML for finding an ϵ-first-order stationary point of the loss functions F and $\hat{F}$ defined in (3) and (4), respectively.

Theorem 4.14. *Consider the objective function F defined in (3) for the case that $\alpha \in (0, \frac{1}{6L}]$. Suppose that the conditions in Assumptions 4.3-4.7 are satisfied, and recall the definitions $L := \max L_i$ and $\rho := \max \rho_i$. Consider running MAML with batch sizes satisfying the conditions $D_h \geq \lceil 2\alpha^2 \sigma_H^2 \rceil$ and $B \geq 20$. Let $\beta_k = \tilde{\beta}(w_k)/12$ where $\tilde{\beta}(w)$ is given in defined in (26). Then, for any $1 > \epsilon > 0$, MAML finds a solution w_ϵ such that*

$$\mathbb{E}[\|\nabla F(w_\epsilon)\|] \leq \mathcal{O}\left(\sqrt{\frac{\sigma^2}{B} + \frac{\tilde{\sigma}^2}{BD_o} + \frac{\tilde{\sigma}^2}{D_{in}}}\right) + \epsilon \tag{34}$$

after at most running for

$$\mathcal{O}(1)\Delta \min\left\{\frac{L}{\epsilon^2}, \frac{LB}{\sigma^2} + \frac{L(BD_o + D_{in})}{\tilde{\sigma}^2}\right\} \tag{35}$$

iterations, where $\Delta := (F(w_0) - \min_{w \in \mathbb{R}^d} F(w))$.

Proof. Check Appendix F for the general statement of the theorem and its proof. □

Remark 4.15. *It is worth noting that the condition $B \geq 20$ can be dropped, i.e., B can be any positive integer, at the cost of decreasing the ratio $\beta_k/\tilde{\beta}(w_k)$.*

The result in Theorem 4.14 shows that after running MAML for $\mathcal{O}(\frac{1}{\epsilon^2} + \frac{B}{\sigma^2} + \frac{BD_o + D_{in}}{\tilde{\sigma}^2})$ iterations, we can find a point $w^\dagger$ that its expected gradient norm $\mathbb{E}[\|\nabla F(w^\dagger)\|]$ is at most of $\epsilon + \mathcal{O}(\sqrt{\frac{\sigma^2}{B} + \frac{\tilde{\sigma}^2}{BD_o} + \frac{\tilde{\sigma}^2}{D_{in}}})$. This result implies that if we choose the batch sizes B, D_o, and D_{in} properly (as a function of ϵ), then for any $\epsilon > 0$ it is possible to reach an ϵ-FOSP of problem (3) in a number of iterations which is polynomial in $1/\epsilon$. We formally state this result in the following corollary.

Corollary 4.16. *Suppose the condition in Theorem 4.14 are satisfied. Then, if the batch sizes B, D_o, and D_{in} satisfy the following conditions,*

$$B \geq C_1 \frac{\sigma^2}{\epsilon^2}, \quad D_{in}, BD_o \geq C_2 \frac{\tilde{\sigma}^2}{\epsilon^2}, \tag{36}$$

for some constants C_1 and C_2, then the iterates generated by MAML finds an ϵ-FOSP, i.e., $\mathbb{E}[\|\nabla F(w)\|] \leq \epsilon$, after at most $\Delta \mathcal{O}(L/\epsilon^2)$, where $\Delta := (F(w_0) - \min_{w \in \mathbb{R}^d} F(w))$.

The result in the above corollary shows that if we have sufficient samples for the batch of stochastic gradient evaluations, i.e., D_{in} and D_o, and for the batch of tasks B, then for any $\epsilon > 0$ MAML finds an ϵ-FOSP after at most $\mathcal{O}(1/\epsilon^2)$ iterations.

Remark 4.17. *Recall from Theorem 4.13 that the difference between ∇F and $\nabla \hat{F}$ is $\mathcal{O}\left(\max\{\frac{\tilde{\sigma}}{\sqrt{D_{test}}}, \frac{\sigma_H \tilde{\sigma}}{D_{test}}\}\right)$. Given that, and since in practice, we usually choose D_{test} at least as large as D_{in}, one can see that as long as σ_H is not significantly larger than $\tilde{\sigma}$, the order of norm of gradient for both F and $\hat{F}$ would be similar for all the results, up to some constant. This argument holds for FO-MAML and HF-MAML as well.*

4.4 Convergence of FO-MAML

Now we proceed to characterize the convergence of the first order approximation of MAML presented in Algorithm 2. In particular, we first characterize the overall complexity of this algortihm for finding a first-order stationary point of the function F.

Theorem 4.18. *Consider the objective function F defined in (3) for the case that $\alpha \in (0, \frac{1}{10L}]$. Suppose that the conditions in Assumptions 4.3-4.7 are satisfied, and recall the definitions $L := \max L_i$ and $\rho := \max \rho_i$. Consider running FO-MAML with batch sizes satisfying the conditions $D_h \geq \lceil 2\alpha^2\sigma_H^2 \rceil$ and $B \geq 20$. Let $\beta_k = \tilde{\beta}(w_k)/18$ where $\tilde{\beta}(w)$ is given in defined in (26). Then, for any $1 > \epsilon > 0$, first order MAML finds a solution w_ϵ such that*

$$\mathbb{E}[\|\nabla F(w_\epsilon)\|] \leq \mathcal{O}\left(\sqrt{\sigma^2\left(20\alpha^2L^2 + \frac{1}{B}\right) + \frac{\tilde{\sigma}^2}{BD_o} + \frac{\tilde{\sigma}^2}{D_{in}}}\right) + \epsilon \tag{37}$$

after at most running for

$$\mathcal{O}(1)\Delta \min\left\{\frac{L}{\epsilon^2}, \frac{L}{\sigma^2(20\alpha^2L^2 + 1/B)} + \frac{L(BD_o + D_{in})}{\tilde{\sigma}^2}\right\} \tag{38}$$

iterations, where $\Delta := (F(w_0) - \min_{w\in\mathbb{R}^d} F(w))$.

Proof. Check Appendix G for the general statement of the theorem and its proof. □

Comparing the result in Theorem 4.18 with the one in Theorem 4.14 implies that FO-MAML, in contrast to MAML, might not converge to the exact solution even when we use large batch sizes. To be more specific, even if we choose large batch sizes B, D_{in}, and D_o for FO-MAML, the norm of gradient cannot become smaller than $\mathcal{O}(\alpha\sigma)$. This is because of the $\alpha^2L^2\sigma^2$ term which appears under square root in the upper bound for the gradient norm in (37) and does not decrease by increasing the batch sizes for the tasks and stochastic gradient evaluations.

In the following corollary, we state the results for FO-MAML when, as in corollary 4.16, we use batch sizes of $\mathcal{O}(1/\epsilon^2)$.

Corollary 4.19. *Suppose the condition in Theorem 4.18 are satisfied. Then, if the batch sizes B, D_o, and D_{in} satisfy the following conditions,*

$$B \geq C_1\frac{1}{\alpha^2L^2}, \quad D_{in}, BD_o \geq C_2\frac{\tilde{\sigma}^2}{\alpha^2\sigma^2L^2}, \tag{39}$$

for some constants C_1 and C_2, then FO-MAML finds a point $w^\dagger$ satisfying the condition $\mathbb{E}[\|\nabla F(w^\dagger)\|] \leq \mathcal{O}(\alpha\sigma L)$, after at most $\Delta\mathcal{O}(1/(\alpha^2\sigma^2L))$, iterations, where $\Delta := (F(w_0) - \min_{w\in\mathbb{R}^d} F(w))$.

Indeed, even by making ϵ arbitrary small we cannot reach a point that has a gradient norm less than $\mathcal{O}(\alpha\sigma)$ for small enough α or σ, while in MAML if we choose the batch sizes properly for any $\epsilon > 0$ we can find an ϵ-FOSP after at most $\mathcal{O}(1/\epsilon^2)$ iterations.

Remark 4.20. *The results for FO-MAML can be extended to the case that we aim to achieve first-order optimality for the function $\hat{F}$, and they would be similar to the bounds derived for the function F up to a constant factor.*

4.5 Convergence of HF-MAML

Now we proceed to analyze the overall complexity of our proposed method for finding an ϵ-first-order stationary point of F, i.e., $\mathbb{E}[\|\nabla F(w)\|] \leq \epsilon$.

Theorem 4.21. *Consider the objective function F defined in (3) for the case that $\alpha \in (0, \frac{1}{6L}]$. Suppose that the conditions in Assumptions 4.3-4.7 are satisfied, and recall the definitions $L := \max L_i$ and $\rho := \max \rho_i$. Consider running HF-MAML with batch sizes satisfying the conditions $D_h \geq \lceil 36(\alpha\rho\sigma_H)^2 \rceil$ and $B \geq 20$. Let $\beta_k = \tilde{\beta}(w_k)/25$ where $\tilde{\beta}(w)$ is defined in (26). Also, we choose the approximation parameter δ_k^i in HF-MAML as*

$$\delta_k^i = \frac{1}{6\rho\alpha\|\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i)\|}$$

Then, for any $\epsilon > 0$, HF-MAML finds a solution w_ϵ such that

$$\mathbb{E}[\|\nabla F(w_\epsilon)\|] \leq \mathcal{O}\left(\sqrt{\frac{\sigma^2}{B} + \frac{\tilde{\sigma}^2}{BD_o} + \frac{\tilde{\sigma}^2}{D_{in}}}\right) + \epsilon \tag{40}$$

after at most running for

$$\mathcal{O}(1)\Delta \min\left\{\frac{L}{\epsilon^2}, \frac{LB}{\sigma^2} + \frac{L(BD_o + D_{in})}{\tilde{\sigma}^2}\right\} \tag{41}$$

iterations, where $\Delta := (F(w_0) - \min_{w\in\mathbb{R}^d} F(w))$.

Proof. Check Appendix H for the general statement of the theorem and its proof. □

Comparing the results in Theorem 4.21 for HF-MAML with the result for in Theorem 4.14 for MAML shows that the overall complexity of these methods and the resulted accuracy on gradient are the same, up to a constant factor. Hence, HF-MAML recovers the complexity of MAML *without* computing any second-order information or performing any update that has a complexity of $\mathcal{O}(d^2)$. As the results for HF-MAML are similar to the ones for MAML, we can show that similar results to the ones in Corollary 4.16 also hold for HF-MAML.

5 Related Work

The general problem of learning from prior models and experiences to learn new tasks efficiently has been formulated in various ways in the literature. One of the main approaches in this regard is designing algorithms for updating the parameters of the optimization methods used for training models. This idea goes back to [10, 9], where the authors studied the problem of learning how to update the parameters of a synaptic learning rule used for training neural networks. Recently, many papers have followed this idea in various ways including using random search for optimizing the hyper-parameters of the learner's model [11, 12], learning the optimizer through the lens of reinforcement learning [21], or controlling the parameters of the gradient methods using recurrent neural networks [26] (see Table 1 in [23] for a summary of different approaches and also [28] for a detailed survey).

In one of the first theoretical formalizations, Baxter [27, 7] proposed a statistical framework with tasks drawn from a distribution, and studied the problem of *bias learning* where the goal is to find an automatic way for choosing the inductive bias in learning problems. In a recent work [17], the authors provide a framework for tuning the hyper-parameters of learning algorithms, such as the initialization or the regularization parameter. They investigate the asymptotic behavior of their method under additional assumptions, such as compactness of the hyper-parameter space and uniqueness of the optimal point for any choice of parameter.

In this paper, we focus on the theoretical analysis of gradient-based model-agnostic meta-learning methods. This setting was first introduced in [15], and was followed by a large number of experimental works proposing various algorithms for the model-agnostic setting [25, 4, 22, 30, 8, 18]. For instance, in [18], the authors introduce an adaptation of MAML for learning the parameters of a prior distribution in a hierarchical Bayesian model. Authors in [22], introduce a variant of MAML that replaces the inner loop stepsize with a vector (which is multiplied with the gradient in an element-wise manner) and then learns this step-vector as well. In another recent work [25], the authors introduce a new method named *Reptile*, which samples a batch of tasks, and for each task, runs a few steps of stochastic gradient descent on its corresponding loss function. Reptile, then, takes the average from these resulting points and defines the next iterate to be the convex combination of this average and the current iterate. In particular, for one step of gradient, Reptile simply minimizes (1) using stochastic gradient descent. Also, authors in [16] study MAML and its extension to online setting for the case of strongly convex objective functions and show that, under the bounded gradient assumption, the objective function of MAML is also strongly convex. Nevertheless, this assumption makes the analysis simpler as the challenge of unbounded smoothness parameter, which we discussed in Section 4, will not emerge anymore. In addition, the authors do not elaborate how to deal with the fact that each step of MAML involves a *biased* estimator of gradient as we explained in Section 4.

The online setting in the gradient-based learning to learn problem has gained attention in theoretical papers recently [16, 19, 20]. In particular, authors in [19] consider the case where the agent sees a sequence of tasks, where each task is an online learning setup where the agent chooses its actions sequentially and suffer losses accordingly. For each task, the agent starts with an initial model, and runs a within task online algorithm to reduce the regret. Finally, the agent updates the initialization or regularization parameter using a meta-update online algorithm. The authors study this setting for convex functions and propose a framework using tools from online convex optimization literature. A closely related problem to this setting is the lifelong learning framework [6, 2]. As an example, in [2], the authors consider the case were a series of tasks are presented sequentially where each task itself is associated with a dataset which is also revealed sequentially and processed by a within-task method. The authors focus on convex loss functions and introduce a meta-algorithm which updates a prior distribution on the set of feature maps, and use it to transfer information from the observed tasks to a new one. In a similar line of work, [13, 14] consider the case where each task is a regression problem where its underlying distribution of data points itself comes from a general distribution. They propose an algorithm which incrementally updates the bias regularization parameter using a sequence of observed tasks. Also, in [16], the authors consider the model-agnostic setting and use a stronger comparator that allows adapting to each new task and propose a "follow the meta leader" algorithm that nevertheless achieves a sublinear regret.

6 Conclusion

In this work, we studied the convergence properties of MAML, its first-order approximation (FO-MAML), and our proposed method, Hessian-free MAML (HF-MAML) for non-convex loss functions. In particular, we characterized their best achievable accuracy in terms of gradient norm when we have access to enough samples and further showed their best possible accuracy when the number of available samples is limited. Our results indicate that MAML can find an ϵ-first-order stationary point, for any positive ϵ, while it suffers from an iteration cost of $\mathcal{O}(d^2)$. On the other hand, we illustrated that although the iteration cost of FO-MAML is $\mathcal{O}(d)$, it cannot reach any desired level of accuracy. That said, we next showed that our proposed method, HF-MAML, has the best of both worlds, i.e., it has a cost of $\mathcal{O}(d)$ at each iteration and can find an ϵ-first-order stationary point, for any positive ϵ.

References

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

[2] Pierre Alquier, The Tien Mai, and Massimiliano Pontil. Regret Bounds for Lifelong Learning. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 261–269, Fort Lauderdale, FL, USA, 20–22 Apr 2017.

[3] Marcin Andrychowicz, Misha Denil, Sergio Gómez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems 29*, pages 3981–3989. Curran Associates, Inc., 2016.

[4] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your MAML. In *International Conference on Learning Representations*, 2019.

[5] Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing neural network architectures using reinforcement learning. In *International Conference on Learning Representations*, 2017.

[6] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. Efficient representations for lifelong learning and autoencoding. In *Conference on Learning Theory*, pages 191–210, 2015.

[7] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[8] Harkirat Singh Behl, Atilim Günes Baydin, and Philip H. S. Torr. Alpha MAML: adaptive model-agnostic meta-learning. 2019.

[9] Samy Bengio, Yoshua Bengio, Jocelyn Cloutier, and Jan Gecsei. On the optimization of a synaptic learning rule. In *Preprints Conf. Optimality in Artificial and Biological Neural Networks*, pages 6–8. Univ. of Texas, 1992.

[10] Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier. *Learning a synaptic learning rule*. Université de Montréal, Département d'informatique et de recherche opérationnelle, 1990.

[11] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[12] James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems 24*, pages 2546–2554. Curran Associates, Inc., 2011.

[13] Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, and Massimiliano Pontil. Learning-to-learn stochastic gradient descent with biased regularization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1566–1575, 2019.

[14] Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Learning to learn around a common mean. In *Advances in Neural Information Processing Systems 31*, pages 10169–10179, 2018.

[15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 06–11 Aug 2017.

[16] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1920–1930, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[17] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1568–1577, 2018.

[18] 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*, 2018.

[19] Mikhail Khodak, Maria-Florina Balcan, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[20] Mikhail Khodak, Maria Florina-Balcan, and Ameet Talwalkar. Adaptive gradient-based meta-learning methods. *arXiv preprint arXiv:1906.02717*, 2019.

[21] Ke Li and Jitendra Malik. Learning to optimize. In *International Conference on Learning Representations*, 2017.

[22] 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.

[23] Luke Metz, Niru Maheswaranathan, Brian Cheung, and Jascha Sohl-Dickstein. Learning unsupervised learning rules. In *International Conference on Learning Representations*, 2019.

[24] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer, 2004.

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

[26] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017.

[27] Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 1998.

[28] Joaquin Vanschoren. *Meta-Learning*, pages 35–61. Springer International Publishing, 2019.

[29] David A Wooff. Bounds on reciprocal moments with applications and developments in stein estimation and post-stratification. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(2):362–371, 1985.

[30] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7693–7702, 2019.

[31] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. In *International Conference on Learning Representations*, 2017.

[32] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.

A Intermediate results

In this subsection, we prove some results that will be used in the rest of our proofs.

Theorem A.1. *Let X be random variable with left extremity zero, and let c be a positive constant. Suppose that $\mu_X = \mathbb{E}[X]$ and $\sigma_X^2 = Var\ (X)$ are finite. Then, for every positive integer k,*

$$\frac{1}{(\mu_X + c)^k} \leq \mathbb{E}\left[\frac{1}{(X+c)^k}\right] \leq \frac{\sigma_X^2/c^k + \mu_X^2\gamma^k}{\sigma_X^2 + \mu_X^2} \tag{42}$$

where $\gamma = \mu_X/(\sigma_X^2 + \mu_X(\mu_X + c))$.

Proof. See Theorem 1 in [29]. □

Lemma A.2. *Consider the definitions of f in (1) an F in (3) for the case that $\alpha \in [0, \frac{\sqrt{2}-1}{L})$. Suppose that the conditions in Assumptions 4.3-4.5 are satisfied. Further, recall the definitions $L := \max L_i$ and $\rho := \max \rho_i$. Then, for any $w \in \mathbb{R}^d$ we have*

$$\|\nabla f(w)\| \leq C_1 \|\nabla F(w)\| + C_2\sigma, \tag{43}$$

$$\mathbb{E}_{i\sim p}[\|\nabla F_i(w)\|^2] \leq 2(1+\alpha L)^2 C_1^2 \|\nabla F(w)\|^2 + (1+\alpha L)^2(2C_2^2+1)\sigma^2, \tag{44}$$

where

$$C_1 = \frac{1}{1 - 2\alpha L - \alpha^2 L^2}, \quad C_2 = \frac{2\alpha L + \alpha^2 L^2}{1 - 2\alpha L - \alpha^2 L^2}.$$

Proof. The gradient of the function $F(w)$ is given by

$$\nabla F(w) = \mathbb{E}_{i\sim p}[\nabla F_i(w)], \tag{45a}$$

$$\nabla F_i(w) = A_i(w)\nabla f_i(w - \alpha\nabla f_i(w)) \tag{45b}$$

with $A_i(w) := (I - \alpha\nabla^2 f_i(w))$. Note that using the mean value theorem we can write the gradient $\nabla f_i(w - \alpha\nabla f_i(w))$ as

$$\begin{aligned}\nabla f_i(w - \alpha\nabla f_i(w)) &= \nabla f_i(w) - \alpha\nabla^2 f_i(\tilde{w}_i)\nabla f_i(w)\\ &= (I - \alpha\nabla^2 f_i(\tilde{w}_i))\nabla f_i(w)\end{aligned} \tag{46}$$

for some $\tilde{w}_i$ which can be written as a convex combination of w and $w - \alpha\nabla f_i(w)$. Using (45b) and the result in (46) we can write

$$\nabla F_i(w) = A_i(w)\nabla f_i(w - \alpha\nabla f_i(w)) = A_i(w)A_i(\tilde{w}_i)\nabla f_i(w), \tag{47}$$

where $A_i(\tilde{w}_i) := (I - \alpha\nabla^2 f_i(\tilde{w}_i))$. Now, we have

$$\begin{aligned}\|\nabla f(w)\| = \|\mathbb{E}_{i\sim p}\nabla f_i(w)\| &= \|\mathbb{E}_{i\sim p}\left[\nabla F_i(w) + (\nabla f_i(w) - \nabla F_i(w))\right]\|\\ &\leq \|\mathbb{E}_{i\sim p}\nabla F_i(w)\| + \|\mathbb{E}_{i\sim p}\left[(I - A_i(w)A_i(\tilde{w}_i))\,\nabla f_i(w)\right]\| \qquad (48)\\ &\leq \|\nabla F(w)\| + \mathbb{E}_{i\sim p}\left[\|I - A_i(w)A_i(\tilde{w}_i)\|\|\nabla f_i(w)\|\right], \qquad (49)\end{aligned}$$

where (48) is obtained by substituting $\nabla F_i(w)$ from (47). Next, note that

$$\|I - A_i(w)A_i(\tilde{w}_i)\| = \|\alpha\nabla^2 f_i(w) + \alpha\nabla^2 f_i(\tilde{w}_i) + \alpha^2\nabla^2 f_i(w)\nabla^2 f_i(\tilde{w}_i)\| \leq 2\alpha L + \alpha^2 L^2,$$

where the last inequality can be shown by using (18a) and triangle inequality. Using this bound in (49) yields

$$\begin{aligned}\|\nabla f(w)\| &\leq \|\nabla F(w)\| + (2\alpha L + \alpha^2 L^2)\mathbb{E}_{i\sim p}\|\nabla f_i(w)\|\\ &\leq \|\nabla F(w)\| + (2\alpha L + \alpha^2 L^2)\left(\|\mathbb{E}_{i\sim p}\nabla f_i(w)\| + \mathbb{E}_{i\sim p}\left[\|\nabla f_i(w) - \mathbb{E}_{i\sim p}\nabla f_i(w)\|\right]\right)\\ &\leq \|\nabla F(w)\| + (2\alpha L + \alpha^2 L^2)\left(\|\nabla f(w)\| + \sigma\right), \qquad (50)\end{aligned}$$

where (50) holds since $\mathbb{E}_{i\sim p}\nabla f_i(w) = \nabla f(w)$, and also, by Assumption 4.5,

$$\mathbb{E}_{i\sim p}\left[\|\nabla f_i(w) - \mathbb{E}_{i\sim p}\nabla f_i(w)\|\right] \leq \sqrt{\mathbb{E}_{i\sim p}\left[\|\nabla f_i(w) - \nabla f(w)\|^2\right]} \leq \sigma. \tag{51}$$

Finally, moving the term $\|\nabla f(w)\|$ from the right hand side of (50) to the left hand side and dividing both sides by $1/(1-2\alpha L-\alpha^2L^2)$ completes the proof of (43). To show (44), note that, using (47), and the fact that $\|A_i(w)\| \le (1+\alpha L)$ and $\|A_i(\tilde{w})\| \le (1+\alpha L)$ we can write

$$\begin{aligned}
\mathbb{E}_{i\sim p}[\|\nabla F_i(w)\|^2] &\le \mathbb{E}_{i\sim p}[\|A_i(w)\|^2\|A_i(\tilde{w}_i)\|^2\|\nabla f_i(w)\|^2] \\
&\le (1+\alpha L)^2\mathbb{E}_{i\sim p}[\|\nabla f_i(w)\|^2] \\
&\le (1+\alpha L)^2\left(\|\nabla f(w)\|^2+\sigma^2\right) \\
&\le (1+\alpha L)^2\left(2C_1^2\|\nabla F(w)\|^2+2C_2^2\sigma^2+\sigma^2\right)
\end{aligned}$$

where the last inequality follows from (43) along with the fact that $(a+b)^2 \le 2a^2+2b^2$. □

B Proof of Lemma 4.9

By considering the definition $\nabla F(w) = \mathbb{E}_{i\sim p}[\nabla F_i(w)]$ where $\nabla F_i(w) = (I-\alpha\nabla^2 f_i(w))\nabla f_i(w-\alpha\nabla f_i(w))$ we can show that

$$\begin{aligned}
\|\nabla F(w)-\nabla F(u)\| &\le \sum_{i\in\mathcal{I}} p_i\|\nabla F_i(w)-\nabla F_i(u)\| \\
&\le \sum_{i\in\mathcal{I}} p_i(\|\nabla f_i(w-\alpha\nabla f_i(w))-\nabla f_i(u-\alpha\nabla f_i(u))\| \qquad (52)\\
&\quad +\alpha\|\nabla^2 f_i(w)\nabla f_i(w-\alpha\nabla f_i(w))-\nabla^2 f_i(u)\nabla f_i(u-\alpha\nabla f_i(u))\| \qquad (53)
\end{aligned}$$

To show the desired result, it suffices to bound both terms in (52) and (53). For (52), we have

$$\begin{aligned}
\|\nabla f_i(w-\alpha\nabla f_i(w))-\nabla f_i(u-\alpha\nabla f_i(u))\| &\le L\|w-u+\alpha(\nabla f_i(w)-\nabla f_i(u))\| \\
&\le L(1+\alpha L)\|w-u\|, \qquad (54)
\end{aligned}$$

where we used the smoothness assumption in Assumption 4.3 for both inequalities. To bound (53), note that

$$\begin{aligned}
&\|\nabla^2 f_i(w)\nabla f_i(w-\alpha\nabla f_i(w))-\nabla^2 f_i(u)\nabla f_i(u-\alpha\nabla f_i(u))\| \\
&\quad = \|\nabla^2 f_i(w)\nabla f_i(w-\alpha\nabla f_i(w))-\nabla^2 f_i(w)\nabla f_i(u-\alpha\nabla f_i(u)) \\
&\qquad +\nabla^2 f_i(w)\nabla f_i(u-\alpha\nabla f_i(u))-\nabla^2 f_i(u)\nabla f_i(u-\alpha\nabla f_i(u))\| \\
&\quad \le \|\nabla^2 f_i(w)\|\|\nabla f_i(w-\alpha\nabla f_i(w))-\nabla f_i(u-\alpha\nabla f_i(u))\| \\
&\qquad +\|\nabla^2 f_i(w)-\nabla^2 f_i(u)\|\|\nabla f_i(u-\alpha\nabla f_i(u))\| \\
&\quad \le \left(L^2(1+\alpha L)+\rho\|\nabla f_i(u-\alpha\nabla f_i(u))\|\right)\|w-u\|, \qquad (55)
\end{aligned}$$

where (55) follows from (54), (18a), and Assumption 4.4. To bound the gradient term in (55), we use the mean value theorem which implies that

$$\nabla f_i(u-\alpha\nabla f_i(u)) = \left(I-\alpha\nabla^2 f_i(\tilde{u}_i)\right)\nabla f_i(u)$$

holds for some $\tilde{u}_i$ which can be written as a convex combination of u and $u-\alpha\nabla f_i(u)$. As a result, and by using (18a), we obtain

$$\|\nabla f_i(u-\alpha\nabla f_i(u))\| \le (1+\alpha L)\|\nabla f_i(u)\|. \qquad (56)$$

Next, plugging (56) in (55) leads to

$$\|\nabla^2 f_i(w)\nabla f_i(w-\alpha\nabla f_i(w))-\nabla^2 f_i(u)\nabla f_i(u-\alpha\nabla f_i(u))\| \le \left(L^2+\rho\|\nabla f_i(u)\|\right)(1+\alpha L)\|w-u\|. \qquad (57)$$

Using bounds (54) and (57) in (52) and (53), respectively, along with the fact that $\alpha L \le 1$, yields

$$\|\nabla F(w)-\nabla F(u)\| \le (4L+2\rho\alpha\mathbb{E}_{i\sim p}\|\nabla f_i(u)\|)\|w-u\|.$$

We can show a similar bound with $\nabla f_i(u)$ replaced by $\nabla f_i(w)$ in the right hand side, and these two together complete the proof.

C Proof of Lemma 4.11

First, note that as $\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j) = \frac{1}{D_\beta} \sum_{\theta \in \mathcal{D}_\beta^j} \tilde{\nabla} f_j(w, \theta)$ and each $\tilde{\nabla} f_j(w, \theta)$ is an unbiased estimator of $\nabla f_j(w)$ with a bounded variance of $\tilde{\sigma}^2$, then for each task $\mathcal{T}_j$ we have

$$\mathbb{E}_{\mathcal{D}_\beta^j}[\|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j) - \nabla f_j(w)\|^2] \leq \frac{\tilde{\sigma}^2}{D_\beta}, \tag{58}$$

and, therefore, $\mathbb{E}_{\mathcal{D}_\beta^j}[\|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j) - \nabla f_j(w)\|] \leq \frac{\tilde{\sigma}}{\sqrt{D_\beta}}$ we can write

$$\|\nabla f_j(w)\| - \frac{\tilde{\sigma}}{\sqrt{D_\beta}} \leq \mathbb{E}_{\mathcal{D}_\beta^j}[\|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|] \leq \|\nabla f_j(w)\| + \frac{\tilde{\sigma}}{\sqrt{D_\beta}}. \tag{59}$$

To derive a bound on the second moment of $\tilde{\beta}(w)$, we use the result of Theorem A.1 for $X = 2\rho\alpha \sum_{j \in \mathcal{B}'} \|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|/B'$, $c = 4L$, and $k = 2$ we obtain that

$$\begin{aligned} \mathbb{E}[\tilde{\beta}(w)^2] &= \mathbb{E}\left[\left(\frac{1}{4L + 2\rho\alpha \sum_{j \in \mathcal{B}'} \|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|/B'}\right)^2\right] \\ &\leq \frac{\sigma_b^2 \frac{1}{(4L)^2} + \mu_b^2 (\frac{\mu_b}{(\sigma_b^2 + \mu_b(\mu_b + 4L))})^2}{\sigma_b^2 + \mu_b^2} \end{aligned} \tag{60}$$

where μ_b and σ_b^2 are the mean and variance of random variable $X = 2\rho\alpha \frac{1}{B'} \sum_{j \in \mathcal{B}'} \|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|$. Now replace $\sigma_b^2 + \mu_b(\mu_b + 4L)$ by its lower bound $\mu_b(\mu_b + 4L)$ and simplify the terms to obtain

$$\mathbb{E}[\tilde{\beta}(w)^2] \leq \frac{\sigma_b^2/(4L)^2 + \mu_b^2/(\mu_b + 4L)^2}{\sigma_b^2 + \mu_b^2}. \tag{61}$$

Now recall the result in (59) use the fact that the batch size D_β is larger than

$$D_\beta \geq \left\lceil \left(\frac{2\rho\alpha\tilde{\sigma}}{L}\right)^2 \right\rceil$$

to write that

$$2\rho\alpha \mathbb{E}_{i \sim p} \|\nabla f_i(w)\| - L \leq \mu_b \leq 2\rho\alpha \mathbb{E}_{i \sim p} \|\nabla f_i(w)\| + L \tag{62}$$

Now based on the definition $L(w) = 4L + 2\rho\alpha \mathbb{E}_{i \sim p} \|\nabla f_i(w)\|$ and the first inequality in (62) we can show that

$$\mu_b + 5L \geq L(w). \tag{63}$$

Therefore, using (61), we have

$$\begin{aligned} L(w)^2 \mathbb{E}[\tilde{\beta}(w)^2] &\leq \frac{\sigma_b^2 (\mu_b + 5L)^2/(4L)^2 + \mu_b^2 (\mu_b + 5L)^2/(\mu_b + 4L)^2}{\sigma_b^2 + \mu_b^2} \\ &\leq \frac{\mu_b^2((5/4)^2 + 2\sigma_b^2/(4L)^2) + 2(5/4)^2 \sigma_b^2}{\sigma_b^2 + \mu_b^2} \end{aligned} \tag{64}$$

where for the last inequality we used the fact that $(\mu_b + 5L)^2 \leq 2\mu_b^2 + 2(5L)^2$. Now considering (64), to prove the second result in (28) we only need to show that

$$2\sigma_b^2/(4L)^2 \leq (5/4)^2. \tag{65}$$

Note that

$$\begin{aligned} \sigma_b^2 &= \frac{(2\rho\alpha)^2}{B'} \mathrm{Var}\left(\|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|\right) = \frac{(2\rho\alpha)^2}{B'} \left(\mathbb{E}\left[\|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|^2\right] - (\mathbb{E}_{i \sim p} \|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|)^2\right) \\ &= \frac{(2\rho\alpha)^2}{B'} \left(\mathrm{Var}\left(\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\right) + \|\mathbb{E}\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|^2 - (\mathbb{E}_{i \sim p} \|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|)^2\right) \\ &\leq \frac{(2\rho\alpha)^2}{B'} \left(\sigma^2 + \frac{\tilde{\sigma}^2}{|\mathcal{D}_\beta|} + \|\mathbb{E}\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|^2 - (\mathbb{E}_{i \sim p} \|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|)^2\right) \end{aligned} \tag{66}$$

where the last inequality follows from the law of total variance which states

$$\mathrm{Var}(Y) = \mathbb{E}\left[\mathrm{Var}(Y|X)\right] + \mathrm{Var}\left(\mathbb{E}[Y|X]\right) \tag{67}$$

for any two random variables X and Y (here $X = \nabla f_j(w)$ and $Y = \tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)$). Now, using the fact that $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$ for any random variable X, we obtain the following result from (66)

$$\sigma_b^2 \leq \frac{(2\rho\alpha)^2}{B'} \left(\sigma^2 + \frac{\tilde{\sigma}^2}{|\mathcal{D}_\beta|} \right). \tag{68}$$

Finally, plugging (68) in (64) and using the assumption (27) on size of B', D_β completes the proof and the second result in (28) follows.

To prove the first result in (28) which is a bound on the first moment of $\tilde{\beta}(w)$, note that, using Jensen's inequality we know that $\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$ and hence by replacing X with $\tilde{L}(w)$ which is defined in (25) and can be written as $\tilde{L}(w) := 1/\tilde{\beta}(w)$ we can show that

$$\mathbb{E}[\tilde{\beta}(w)] = \mathbb{E}[\frac{1}{\tilde{L}(w)}] \geq \frac{1}{\mathbb{E}[\tilde{L}(w)]} = \frac{1}{4L + \mu_b}, \tag{69}$$

where μ_b is the mean of $2\rho\alpha \frac{1}{B'} \sum_{j \in \mathcal{B}'} \|\tilde{\nabla} f_j(w, \mathcal{D}_\beta^j)\|$. Now by using this result and the upper bound for μ_b in (62) we obtain that

$$\mathbb{E}[\tilde{\beta}(w)] \geq \frac{1}{5L + 2\rho\alpha \mathbb{E}_{i\sim p}\|\nabla f_i(w)\|}. \tag{70}$$

As $L(w) = 4L + 2\rho\alpha \mathbb{E}_{i\sim p}\|\nabla f_i(w)\|$ we can show that

$$\mathbb{E}[\tilde{\beta}(w)] \geq \frac{1}{L + L(w)} \geq \frac{1}{L(w)/4 + L(w)} = \frac{4/5}{L(w)} \tag{71}$$

and the first claim in (28) follows.

D Proof of Lemma 4.12

Note that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{in},\mathcal{D}_o}[\tilde{\nabla} f_i(w_k - \alpha \tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i) \mid \mathcal{F}_k] \\
&= \mathbb{E}_{\mathcal{D}_{in}}\left[\nabla f_i \Big(w_k - \alpha \tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i) \Big) \mid \mathcal{F}_k \right] \\
&= \mathbb{E}[\nabla f_i\left(w_k - \alpha \nabla f_i(w_k)\right) \mid \mathcal{F}_k] + \mathbb{E}_{\mathcal{D}_{in}}\left[\nabla f_i \Big(w_k - \alpha \tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i) \Big) - \nabla f_i\left(w_k - \alpha \nabla f_i(w_k)\right) \mid \mathcal{F}_k \right] \\
&= \mathbb{E}[\nabla f_i\left(w_k - \alpha \nabla f_i(w_k)\right) \mid \mathcal{F}_k] + e_{i,k}
\end{aligned}$$

where

$$e_{i,k} = \mathbb{E}_{\mathcal{D}_{in}}\left[\nabla f_i \Big(w_k - \alpha \tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i) \Big) - \nabla f_i\left(w_k - \alpha \nabla f_i(w_k)\right) \mid \mathcal{F}_k \right]$$

and its norm is bounded by

$$\begin{aligned}
\|e_{i,k}\| &\leq \mathbb{E}_{\mathcal{D}_{in}^i}\left[\Big\| \nabla f_i \Big(w_k - \alpha \tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i) \Big) - \nabla f_i\left(w_k - \alpha \nabla f_i(w_k)\right) \Big\| \mid \mathcal{F}_k \right] \\
&\leq \alpha L \mathbb{E}_{\mathcal{D}_{in}^i}\left[\Big\| \tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i) - \nabla f_i(w_k) \Big\| \mid \mathcal{F}_k \right] && (72) \\
&\leq \alpha L \frac{\tilde{\sigma}}{\sqrt{D_{in}}} && (73)
\end{aligned}$$

where (72) follows from the Lipschitz property of gradient (Assumption 4.3 and (18a)), and the last line is obtained using Assumption 4.6. To bound the second moment, note that

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{in},\mathcal{D}_o}\left[\|\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\|^2 \mid \mathcal{F}_k\right] \\
&= \mathbb{E}_{\mathcal{D}_{in}^i}\left[\|\nabla f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i))\|^2+\frac{\tilde{\sigma}^2}{D_o} \mid \mathcal{F}_k\right] \\
&\leq (1+\frac{1}{\phi})\|\nabla f_i(w_k-\alpha\nabla f_i(w_k)\|^2 \\
&+(1+\phi)\mathbb{E}_{\mathcal{D}_{in}^i}\left[\|\nabla f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i))-\nabla f_i(w_k-\alpha\nabla f_i(w_k)\|^2 \mid \mathcal{F}_k\right]+\frac{\tilde{\sigma}^2}{D_o} &(74)\\
&\leq (1+\frac{1}{\phi})\|\nabla f_i(w_k-\alpha\nabla f_i(w_k)\|^2+(1+\phi)\alpha^2L^2\frac{\tilde{\sigma}^2}{D_{in}}+\frac{\tilde{\sigma}^2}{D_o} &(75)
\end{aligned}
$$

where (74) follows from the inequality $(a+b)^2 \leq (1+1/\phi)a^2+(1+\phi)b^2$ and (75) is obtained similar to (72).

E Proof of Theorem 4.13

First, note that

$$
\nabla\hat{F}(w)=\mathbb{E}_{i\sim p}\left[\mathbb{E}_{\mathcal{D}_{test}^i}\left[(I-\alpha\tilde{\nabla}^2 f_i(w,\mathcal{D}_{test}^i))\nabla f_i(w-\alpha\tilde{\nabla} f_i(w,\mathcal{D}_{test}^i))\right]\right] \tag{76}
$$

Next, using Assumption 4.6, we have

$$
I-\alpha\tilde{\nabla}^2 f_i(w,\mathcal{D}_{test}^i)=I-\alpha\nabla^2 f_i(w)+e_{H,i} \tag{77}
$$

where

$$
\mathbb{E}_{\mathcal{D}_{test}^i}[e_{H,i}]=0, \quad \mathbb{E}_{\mathcal{D}_{test}^i}[\|e_{H,i}\|^2]\leq\frac{\alpha^2\sigma_H^2}{D_{test}}. \tag{78}
$$

In addition,

$$
\nabla f_i(w-\alpha\tilde{\nabla} f_i(w,\mathcal{D}_{test}^i))=\nabla f_i(w-\alpha\nabla f_i(w))+e_{G,i} \tag{79}
$$

where

$$
e_{G,i}=\nabla f_i(w-\alpha\tilde{\nabla} f_i(w,\mathcal{D}_{test}^i))-\nabla f_i\left(w-\alpha\nabla f_i(w_k)\right)
$$

and the expectation of its norm squared is bounded by

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}_{test}^i}[\|e_{G,i}\|^2] &\leq \mathbb{E}_{\mathcal{D}_{test}^i}\left[\left\|\nabla f_i\left(w-\alpha\tilde{\nabla} f_i(w,\mathcal{D}_{test}^i)\right)-\nabla f_i\left(w-\alpha\nabla f_i(w)\right)\right\|^2\right] \\
&\leq \alpha^2L^2\mathbb{E}_{\mathcal{D}_{test}^i}\left[\left\|\tilde{\nabla} f_i(w,\mathcal{D}_{test}^i)-\nabla f_i(w)\right\|^2\right] &(80)\\
&\leq \alpha^2L^2\frac{\tilde{\sigma}^2}{D_{test}} &(81)
\end{aligned}
$$

where (80) follows from the Lipschitz property of gradient (Assumption 4.3 and (18a)), and the last line is obtained using Assumption 4.6. Now plugging (77) and (79) in (76) implies

$$
\begin{aligned}
\nabla\hat{F}(w) &= \mathbb{E}_{i\sim p}\left[\mathbb{E}_{\mathcal{D}_{test}^i}\left[(I-\alpha\nabla^2 f_i(w)+e_{H,i})(\nabla f_i(w-\alpha\nabla f_i(w))+e_{G,i})\right]\right] &(82)\\
&= \mathbb{E}_{i\sim p}\left[(I-\alpha\nabla^2 f_i(w))\nabla f_i(w-\alpha\nabla f_i(w))\right] \\
&+ \mathbb{E}_{i\sim p}\left[(I-\alpha\nabla^2 f_i(w))\mathbb{E}_{\mathcal{D}_{test}^i}[e_{G,i}]+\nabla f_i(w-\alpha\nabla f_i(w))\mathbb{E}_{\mathcal{D}_{test}^i}[e_{H,i}]\right] \\
&+ \mathbb{E}_{i\sim p}\left[\mathbb{E}_{\mathcal{D}_{test}^i}[e_{G,i}e_{H,i}]\right]. &(83)
\end{aligned}
$$

Using $\nabla F(w)=E_{i\sim p}\left[(I-\alpha\nabla^2 f_i(w))\nabla f_i(w-\alpha\nabla f_i(w))\right]$ along with $\mathbb{E}_{\mathcal{D}_{test}^i}[e_{H,i}]=0$ yields

$$
\nabla\hat{F}(w)=\nabla F(w)+\mathbb{E}_{i\sim p}\left[(I-\alpha\nabla^2 f_i(w))\mathbb{E}_{\mathcal{D}_{test}^i}[e_{G,i}]\right]+\mathbb{E}_{i\sim p}\left[\mathbb{E}_{\mathcal{D}_{test}^i}[e_{G,i}e_{H,i}]\right]. \tag{84}
$$

As a result, using the fact that $\|I - \alpha\nabla^2 f_i(w)\| \leq 1 + \alpha L$ along with Cauchy-Schwarz inequality implies

$$\begin{aligned}
\|\nabla\hat{F}(w) - \nabla F(w)\| &\leq (1+\alpha L)\mathbb{E}_{i\sim p}\left[\mathbb{E}_{\mathcal{D}^i_{test}}[\|e_{G,i}\|]\right] + \mathbb{E}_{i\sim p}\left[\sqrt{\mathbb{E}_{\mathcal{D}^i_{test}}[\|e_{H,i}\|^2]\mathbb{E}_{\mathcal{D}^i_{test}}[\|e_{G,i}\|^2]}\ \right] \\
&\leq (1+\alpha L)\alpha L\frac{\tilde{\sigma}}{\sqrt{D_{test}}} + \alpha^2 L\frac{\sigma_H\tilde{\sigma}}{D_{test}} \\
&\leq 2\alpha L\frac{\tilde{\sigma}}{\sqrt{D_{test}}} + \alpha^2 L\frac{\sigma_H\tilde{\sigma}}{D_{test}} \qquad (85)
\end{aligned}$$

where the last inequality follows from $\alpha \leq \frac{1}{L}$.

F Proof of Theorem 4.14 (General Version)

Theorem F.1. *Consider the objective function F defined in (3) for the case that $\alpha \in (0, \frac{1}{6L}]$. Suppose that the conditions in Assumptions 4.3-4.6 are satisfied, and recall the definitions $L := \max L_i$ and $\rho := \max\rho_i$. Consider running MAML with batch sizes satisfying the conditions $D_h \geq \lceil 2\alpha^2\sigma_H^2\rceil$ and $B \geq 20$. Let $\beta_k = \tilde{\beta}(w_k)/12$ where $\tilde{\beta}(w)$ is given in defined in (26). Then, for any $\epsilon > 0$, MAML finds a solution w_ϵ such that*

$$\mathbb{E}[\|\nabla F(w_\epsilon)\|] \leq \max\left\{\sqrt{61\left(1+\frac{\rho\alpha}{L}\sigma\right)\left(\frac{\sigma^2}{B} + \frac{\tilde{\sigma}^2}{BD_o} + \frac{\tilde{\sigma}^2}{D_{in}}\right)}, \frac{61\rho\alpha}{L}\left(\frac{\sigma^2}{B} + \frac{\tilde{\sigma}^2}{BD_o} + \frac{\tilde{\sigma}^2}{D_{in}}\right), \epsilon\right\} \qquad (86)$$

after at most running for

$$\mathcal{O}(1)\Delta\min\left\{\frac{L+\rho\alpha(\sigma+\epsilon)}{\epsilon^2}, \frac{LB}{\sigma^2} + \frac{L(BD_o + D_{in})}{\tilde{\sigma}^2}\right\} \qquad (87)$$

iterations, where $\Delta := (F(w_0) - \min_{w\in\mathbb{R}^d} F(w))$.

Proof. To simplify the notation, we denote $L(w_k)$ by L_k. Also, let $\mathcal{F}_k$ be the information up to iteration k. Note that, conditioning on $\mathcal{F}_k$, the iterate w_k, and hence, $F(w_k)$ and $\nabla F(w_k)$, are not random variables anymore. Let

$$G_i(w_k) := \left(I - \alpha\tilde{\nabla}^2 f_i(w_k, D^i_h)\right)\tilde{\nabla} f_i\left(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o\right)$$

First, we characterize the first and second moment of $G_i(w_k)$ conditioning on $\mathcal{F}_k$[5] . Note that, since $\mathcal{D}^i_{in}, \mathcal{D}^i_o$, and $\mathcal{D}^i_h$ are drawn independently, we have

$$\begin{aligned}
\mathbb{E}[G_i(w_k)] &= \mathbb{E}_{i\sim p}\left[\mathbb{E}_{\mathcal{D}^i_h}\left[I - \alpha\tilde{\nabla}^2 f_i(w_k, D^i_h)\right]\mathbb{E}_{\mathcal{D}^i_o,\mathcal{D}^i_{in}}\left[\tilde{\nabla} f_i\left(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o\right)\right]\right] \\
&= \mathbb{E}_{i\sim p}\left[\left(I - \alpha\nabla^2 f_i(w_k)\right)\left(\nabla f_i\left(w_k - \alpha\nabla f_i(w_k)\right) + e_{i,k}\right)\right] \qquad (88)
\end{aligned}$$

where $e_{i,k}$ as defined in Lemma (4.12) is given by

$$e_{i,k} := \mathbb{E}_{\mathcal{D}_{in},\mathcal{D}_o}[\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}^i_{in}), \mathcal{D}^i_o)] - \nabla f_i\left(w_k - \alpha\nabla f_i(w_k)\right)$$

. By simplifying the right hand side of (88) we obtain that

$$\begin{aligned}
\mathbb{E}[G_i(w_k)] &= \mathbb{E}_{i\sim p}\left[\left(I - \alpha\nabla^2 f_i(w_k)\right)\nabla f_i\left(w_k - \alpha\nabla f_i(w_k)\right) + \left(I - \alpha\nabla^2 f_i(w_k)\right)e_{i,k}\right] \\
&= \mathbb{E}_{i\sim p}\left[\nabla F_i(w_k) + \left(I - \alpha\nabla^2 f_i(w_k)\right)e_{i,k}\right] \\
&= \nabla F(w_k) + r_k \qquad (89)
\end{aligned}$$

and r_k is given by $r_k = \mathbb{E}_{i\sim p}\left[(I - \alpha\nabla^2 f_i(w_k))e_{i,k}\right]$. Note that the second equality in (89) due to definition $F_i(w) := f_i(w - \alpha\nabla f_i(w))$. Next, we derive an upper bound on the norm of r_k as

$$\begin{aligned}
\|r_k\| &\leq \mathbb{E}_{i\sim p}\left[\|I - \alpha\nabla^2 f_i(w_k)\|\|e_{i,k}\|\right] \\
&\leq (1+\alpha L)\alpha L\frac{\tilde{\sigma}}{\sqrt{D_{in}}} \qquad (90) \\
&\leq 0.2\frac{\tilde{\sigma}}{\sqrt{D_{in}}}, \qquad (91)
\end{aligned}$$

[5] we suppress the conditioning on $\mathcal{F}_k$ to simplify the notation

where (90) follows from Lemma (4.12) along with the Lipschitz property of gradient (Assumption 4.3 and (18a)), and the last line is obtained using the fact that $\alpha L \leq \frac{1}{6}$. Hence, we have

$$\|\mathbb{E}[G_i(w_k)]\| \leq \|\nabla F(w_k)\| + 0.2\frac{\tilde{\sigma}}{\sqrt{D_{in}}}$$

Now, note that this inequality and the fact that $a \leq b + c$ yields $a^2 \leq 2b^2 + 2c^2$ for any positive scalars a, b, c, imply that

$$\|\mathbb{E}[G_i(w_k)]\|^2 \leq 2\|\nabla F(w_k)\|^2 + 0.08\frac{\tilde{\sigma}^2}{D_{in}}. \tag{92}$$

To bound the variance of $G_i(w_k)$, we bound its second moment. A similar argument to what we did above implies

$$\mathbb{E}[\|G_i(w_k)\|^2] = \mathbb{E}_{i\sim p}\left[\mathbb{E}_{\mathcal{D}_h^i}\left\|I - \alpha\tilde{\nabla}^2 f_i(w_k, D_h^i)\right\|^2 \mathbb{E}_{\mathcal{D}_o^i, D_{in}^i}\left\|\tilde{\nabla} f_i\left(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i\right)\right\|^2\right] \tag{93}$$

To simplify the right hand side we first use the fact that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_h^i}\left\|I - \alpha\tilde{\nabla}^2 f_i(w_k, D_h^i)\right\|^2 &= Var\left[I - \alpha\tilde{\nabla}^2 f_i(w_k, D_h^i)\right] + \|I - \alpha\nabla^2 f_i(w_k)\|^2 \\
&= \alpha^2 Var\left[\tilde{\nabla}^2 f_i(w_k, D_h^i)\right] + \|I - \alpha\nabla^2 f_i(w_k)\|^2 \\
&\leq \frac{\alpha^2\sigma_H^2}{D_h} + \|I - \alpha\nabla^2 f_i(w_k)\|^2
\end{aligned} \tag{94}$$

where the last inequality follows from Assumption 4.6. Substitute the upper bound in (94) into (93) to obtain

$$\mathbb{E}[\|G_i(w_k)\|^2] \leq \mathbb{E}_{i\sim p}\left[\left(\|I - \alpha\nabla^2 f_i(w_k)\|^2 + \frac{\alpha^2\sigma_H^2}{D_h}\right)\mathbb{E}_{\mathcal{D}_o^i, D_{in}^i}\left\|\tilde{\nabla} f_i\left(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i\right)\right\|^2\right] \tag{95}$$

Note that using the fact that $\|I - \alpha\nabla^2 f_i(w_k)\| \leq 1 + \alpha L$ and the assumption that $\alpha L \leq \frac{1}{6}$ we can show that $\|I - \alpha\nabla^2 f_i(w_k)\| \leq 7/6$. Further, we know that $D_h \geq 2\alpha^2\sigma_H^2$ which implies that $\alpha^2\sigma_H^2/D_h \leq 1/2$. By combining these two bounds we can show that

$$\|I - \alpha\nabla^2 f_i(w_k)\|^2 + \frac{\alpha^2\sigma_H^2}{D_h} \leq 2 \tag{96}$$

As a result of (96), we can simplify the right hand side of (95) to

$$\mathbb{E}[\|G_i(w_k)\|^2] \leq 2\mathbb{E}_{i\sim p}\left[\mathbb{E}_{\mathcal{D}_o^i, D_{in}^i}\left\|\tilde{\nabla} f_i\left(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i\right)\right\|^2\right]. \tag{97}$$

Note that, using Lemma 4.12 with $\phi = 1$, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_o^i, D_{in}^i}\left\|\tilde{\nabla} f_i\left(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i\right)\right\|^2 &\leq 2\|\nabla f_i(w_k - \alpha\nabla f_i(w_k)\|^2 + 2\alpha^2 L^2\frac{\tilde{\sigma}^2}{D_{in}} + \frac{\tilde{\sigma}^2}{D_o} \\
&\leq 2\frac{\|\nabla F_i(w_k)\|^2}{(1-\alpha L)^2} + 2\alpha^2 L^2\frac{\tilde{\sigma}^2}{D_{in}} + \frac{\tilde{\sigma}^2}{D_o}
\end{aligned} \tag{98}$$

where the last inequality follows from (45b) and the fact that $\|I - \alpha\nabla^2 f_i(w)\| \geq 1 - \alpha L$. Plugging (98) in (97) and using (44) in Lemma A.2 yields

$$\mathbb{E}[\|G_i(w_k)\|^2] \leq 40\|\nabla F(w_k)\|^2 + 14\sigma^2 + \tilde{\sigma}^2\left(\frac{2}{D_o} + \frac{1}{6D_{in}}\right). \tag{99}$$

Now that we have upper bounds on $\mathbb{E}[\|G_i(w_k)\|]$ and $\mathbb{E}[\|G_i(w_k)\|^2]$, we proceed to prove the main result. According to Corollary 4.10, we have

$$\begin{aligned} F(w_{k+1}) &\le F(w_k) + \nabla F(w_k)^\top (w_{k+1} - w_k) + \frac{L_k}{2}\|w_{k+1} - w_k\|^2 \\ &= F(w_k) - \beta_k \nabla F(w_k)^\top \left(\frac{1}{B} \sum_{i \in \mathcal{B}_k} G_i(w_k) \right) + \frac{L_k}{2}\beta_k^2 \left\| \frac{1}{B} \sum_{i \in \mathcal{B}_k} G_i(w_k) \right\|^2 . \end{aligned} \tag{100}$$

By computing the expectation of both sides of (100) conditioning on $\mathcal{F}_k$, we obtain that

$$\begin{aligned} \mathbb{E}[F(w_{k+1})|\mathcal{F}_k] \le F(w_k) &- \mathbb{E}[\beta_k|\mathcal{F}_k] \nabla F(w_k)^\top \mathbb{E}[G_i(w_k)|\mathcal{F}_k] \\ &+ \frac{L_k}{2}\mathbb{E}[\beta_k^2|\mathcal{F}_k] \left(\|\mathbb{E}[G_i(w_k)|\mathcal{F}_k]\|^2 + \frac{1}{B}\mathbb{E}[\|G_i(w_k)\|^2|\mathcal{F}_k] \right) \end{aligned}$$

where we used the fact that batches $\mathcal{B}_k$ and $\mathcal{B}'_k$ are independently drawn. Now, using the expression

$$\mathbb{E}[G_i(w_k)|\mathcal{F}_k] = \nabla F(w_k) + r_k$$

in (89) along with (92) and (99) we can write that

$$\begin{aligned} \mathbb{E}[F(w_{k+1})|\mathcal{F}_k] &\le F(w_k) - \|\nabla F(w_k)\|^2 \left(\mathbb{E}[\beta_k|\mathcal{F}_k] - \frac{L_k}{2}\mathbb{E}[\beta_k^2|\mathcal{F}_k] \left(2 + \frac{40}{B}\right)\right) \\ &+ \mathbb{E}[\beta_k|\mathcal{F}_k]\|\nabla F(w_k)\|\|r_k\| + \frac{L_k}{2}\mathbb{E}[\beta_k^2|\mathcal{F}_k] \left(\frac{1}{B}\left(14\sigma^2 + \tilde{\sigma}^2\left(\frac{2}{D_o} + \frac{0.2}{D_{in}}\right)\right) + 0.08\frac{\tilde{\sigma}^2}{D_{in}}\right). \end{aligned} \tag{101}$$

Note that, using (91), we can show that

$$\|\nabla F(w_k)\|\|r_k\| \le \frac{\|\nabla F(w_k)\|^2}{10} + 10\|r_k\|^2 \le \frac{\|\nabla F(w_k)\|^2}{10} + 0.4\frac{\tilde{\sigma}^2}{D_{in}}.$$

Plugging this bound in (101) implies

$$\begin{aligned} \mathbb{E}[F(w_{k+1})|\mathcal{F}_k] &\le F(w_k) - \|\nabla F(w_k)\|^2 \left(\frac{9}{10}\mathbb{E}[\beta_k|\mathcal{F}_k] - \frac{L_k}{2}\mathbb{E}[\beta_k^2|\mathcal{F}_k] \left(2 + \frac{40}{B}\right)\right) \\ &+ \frac{L_k}{2}\mathbb{E}[\beta_k^2|\mathcal{F}_k] \left(\frac{1}{B}\left(14\sigma^2 + \tilde{\sigma}^2\left(\frac{2}{D_o} + \frac{0.2}{D_{in}}\right)\right) + 0.08\frac{\tilde{\sigma}^2}{D_{in}}\right) + 0.4\mathbb{E}[\beta_k|\mathcal{F}_k]\frac{\tilde{\sigma}^2}{D_{in}}. \end{aligned} \tag{102}$$

Note that $\beta_k = \tilde{\beta}(w_k)/12$, and hence, by using Lemma 4.11 along with $1/\tilde{\beta}(w_k), L_k \ge 4L$, we have

$$\frac{1}{48L} \ge \mathbb{E}[\beta_k|\mathcal{F}_k] \ge \frac{1}{15L_k}, \quad \frac{L_k}{2}\mathbb{E}[\beta_k^2|\mathcal{F}_k] \le \frac{1}{92L_k} \le \frac{1}{368L}.$$

Plugging these bounds in (102) and using the assumption $B \ge 20$ yields

$$\begin{aligned} &\mathbb{E}[F(w_{k+1})|\mathcal{F}_k] \\ &\quad\le F(w_k) - \frac{1}{100L_k}\|\nabla F(w_k)\|^2 + \frac{1}{368LB}\left(14\sigma^2 + \tilde{\sigma}^2\left(\frac{2}{D_o} + \frac{0.2}{D_{in}}\right)\right) + \frac{(0.4/48 + 0.08/368)\tilde{\sigma}^2}{LD_{in}} \\ &\quad\le F(w_k) - \frac{1}{100L_k}\|\nabla F(w_k)\|^2 + \frac{14\sigma^2 + 2\tilde{\sigma}^2/D_o}{368LB} + \frac{\tilde{\sigma}^2}{96LD_{in}}, \end{aligned} \tag{103}$$

where the last inequality is obtained by taking the $0.2\tilde{\sigma}^2/D_{in}$ from the second term and merging it with the third term. Next, note that

$$\frac{1}{L_k}\|\nabla F(w_k)\|^2 = \frac{\|\nabla F(w_k)\|^2}{4L + 2\rho\alpha\mathbb{E}_{i\sim p}\|\nabla f_i(w_k)\|} \ge \frac{\|\nabla F(w_k)\|^2}{4L + 2\rho\alpha\sigma + 2\rho\alpha\|\nabla f(w_k)\|} \tag{104}$$

where the last inequality follows from (51). Using Lemma A.2 along with the fact that $\alpha \le \frac{1}{6L}$, implies

$$\|\nabla f(w_k)\| \le 2\|\nabla F(w_k)\| + \sigma. \tag{105}$$

Plugging (105) in (104) yields

$$\begin{aligned}\frac{1}{L_k}\|\nabla F(w_k)\|^2 &\geq \frac{\|\nabla F(w_k)\|^2}{4L+4\rho\alpha\sigma+4\rho\alpha\|\nabla F(w_k)\|} \\ &\geq \frac{\|\nabla F(w_k)\|^2}{2\max\{4L+4\rho\alpha\sigma, 4\rho\alpha\|\nabla F(w_k)\|\}} = \min\left\{\frac{\|\nabla F(w_k)\|^2}{8L+8\rho\alpha\sigma}, \frac{\|\nabla F(w_k)\|}{8\rho\alpha}\right\}. \end{aligned} \tag{106}$$

Now, plugging (106) in (103) and taking expectation from both sides with respect to $\mathcal{F}_k$ along with using tower rule implies

$$\mathbb{E}[F(w_{k+1})] \leq \mathbb{E}[F(w_k)] - \frac{1}{800}\min\left\{\frac{\mathbb{E}[\|\nabla F(w_k)\|^2]}{L+\rho\alpha\sigma}, \frac{\mathbb{E}[\|\nabla F(w_k)\|]}{\rho\alpha}\right\} + \frac{14\sigma^2+2\tilde{\sigma}^2/D_o}{368LB} + \frac{\tilde{\sigma}^2}{96LD_{in}}. \tag{107}$$

Assume (86) does not hold at iteration k. Then, we have

$$\mathbb{E}[\|\nabla F(w_k)\|] \geq \max\{\sqrt{(1+\frac{\rho\alpha}{L}\sigma)\gamma_1}, \frac{\rho\alpha}{L}\gamma_1\}$$

with γ_1 given by

$$\gamma_1 = 61\left(\frac{\sigma^2}{B}+\frac{\tilde{\sigma}^2}{BD_o}+\frac{\tilde{\sigma}^2}{D_{in}}\right). \tag{108}$$

This implies

$$\frac{1}{1600}\min\left\{\frac{\mathbb{E}[\|\nabla F(w_k)\|^2]}{L+\rho\alpha\sigma}, \frac{\mathbb{E}[\|\nabla F(w_k)\|]}{\rho\alpha}\right\} \geq \frac{\gamma_1}{1600L} \geq \frac{14\sigma^2+2\tilde{\sigma}^2/D_o}{368LB} + \frac{\tilde{\sigma}^2}{96LD_{in}},$$

and hence, using (107), we obtain

$$\mathbb{E}[F(x_{w+1})] \leq \mathbb{E}[F(w_k)] - \frac{1}{1600}\min\left\{\frac{\mathbb{E}[\|\nabla F(w_k)\|^2]}{L+\rho\alpha\sigma}, \frac{\mathbb{E}[\|\nabla F(w_k)\|]}{\rho\alpha}\right\} \leq \mathbb{E}[F(w_k)] - \frac{\gamma_1}{1600L}.$$

Based on the assumption that (86) does not hold at iteration k we also know that $\mathbb{E}[\|\nabla F(w_k)\|] \geq \epsilon$ which implies that

$$\mathbb{E}[F(w_{k+1})] \leq \mathbb{E}[F(w_k)] - \frac{1}{1600}\min\left\{\frac{\epsilon^2}{L+\rho\alpha\sigma}, \frac{\epsilon}{\rho\alpha}\right\} \leq \mathbb{E}[F(w_k)] - \frac{1}{1600}\frac{\epsilon^2}{L+\rho\alpha(\sigma+\epsilon)}. \tag{109}$$

This result shows that if the condition in (86) is not satisfied the objective function value decreases by a constant value in expectation. If we assume that for all iterations $0, \ldots, T-1$ this condition does not hold then by summing both sides of (109) from 0 to $T-1$ we obtain that

$$\sum_{k=0}^{T-1}\mathbb{E}[F(w_{k+1})] \leq \sum_{k=0}^{T-1}\mathbb{E}[F(w_k)] - \sum_{k=0}^{T-1}\frac{1}{1600}\frac{\epsilon^2}{L+\rho\alpha(\sigma+\epsilon)}. \tag{110}$$

which implies that

$$\mathbb{E}[F(w_T)] \leq \mathbb{E}[F(w_0)] - \frac{T}{1600}\frac{\epsilon^2}{L+\rho\alpha(\sigma+\epsilon)}. \tag{111}$$

and hence

$$\begin{aligned} T &\leq (\mathbb{E}[F(w_0)] - \mathbb{E}[F(w_T)])1600\frac{L+\rho\alpha(\sigma+\epsilon)}{\epsilon^2} \\ &\leq (F(w_0) - F(w^*))1600\frac{L+\rho\alpha(\sigma+\epsilon)}{\epsilon^2} \end{aligned} \tag{112}$$

This argument shows that if the condition in (86) is not satisfied for all k form 0 to $T-1$, then the time T can not be larger than $(F(w_0)-F(w^*))1600\frac{L+\rho\alpha(\sigma+\epsilon)}{\epsilon^2}$. Hence, after $(F(w_0)-F(w^*))1600\frac{L+\rho\alpha(\sigma+\epsilon)}{\epsilon^2}$ iterations at least one of the iterates generated by MAML satisfies the condition in (86), and the proof is complete. □

G Proof of Theorem 4.18 (General Version)

Theorem G.1. *Consider the objective function F defined in (3) for the case that $\alpha \in (0, \frac{1}{10L}]$. Suppose that the conditions in Assumptions 4.3-4.6 are satisfied, and recall the definitions $L := \max L_i$ and $\rho := \max \rho_i$. Consider running FO-MAML with batch sizes satisfying the conditions $D_h \geq \lceil 2\alpha^2\sigma_H^2 \rceil$ and $B \geq 20$. Let $\beta_k = \tilde{\beta}(w_k)/18$ where $\tilde{\beta}(w)$ is given in defined in (26). Then, for any $\epsilon > 0$, first order MAML finds a solution w_ϵ such that*

$$\mathbb{E}[\|\nabla F(w_\epsilon)\|] \leq$$

$$\max\left\{\sqrt{14\left(1+\frac{\rho\alpha}{L}\sigma\right)\left(\sigma^2(1/B+20\alpha^2L^2)+\frac{\tilde{\sigma}^2}{BD_o}+\frac{\tilde{\sigma}^2}{D_{in}}\right)}, \frac{14\rho\alpha}{L}\left(\sigma^2(\frac{1}{B}+20\alpha^2L^2)+\frac{\tilde{\sigma}^2}{BD_o}+\frac{\tilde{\sigma}^2}{D_{in}}\right), \epsilon\right\} \tag{113}$$

after at most running for

$$\mathcal{O}(1)\Delta \min\left\{\frac{L+\rho\alpha(\sigma+\epsilon)}{\epsilon^2}, \frac{L}{\sigma^2(1/B+20\alpha^2L^2)} + \frac{L(BD_o+D_{in})}{\tilde{\sigma}^2}\right\} \tag{114}$$

iterations, where $\Delta := (F(w_0) - \min_{w\in\mathbb{R}^d} F(w))$.

Proof. First, note that the update of the first-order approximation of MAML can be written as $w_{k+1} = w_k - \frac{\beta_k}{B}\sum_{i\in\mathcal{B}_k} G_i(w_k)$, where

$$G_i(w_k) := \tilde{\nabla} f_i\left(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i\right).$$

To analyze this update, similar to the proof of Theorem F.1, we first characterize the first and second moment of the descent direction $G_i(w)$ conditioning on $\mathcal{F}_k$. Using the definition

$$e_{i,k} = \nabla f_i\left(w_k - \alpha\nabla f_i(w_k)\right) - \mathbb{E}_{\mathcal{D}_{in},\mathcal{D}_o}[\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i)]$$

we can write that

$$\mathbb{E}[G_i(w_k)] = \mathbb{E}_{i\sim p}\left[\nabla f_i\left(w_k - \alpha\nabla f_i(w_k)\right) + e_{i,k}\right]. \tag{115}$$

Further, based on the definition of F_i and the fact that its gradient is given by $\nabla F_i(w) = (I - \alpha\nabla^2 f_i(w))\nabla f_i\left(w - \alpha\nabla f_i(w)\right)$, we can rewrite the right hand side of (115) as

$$\mathbb{E}[G_i(w_k)] = \mathbb{E}_{i\sim p}\left[\left(I - \alpha\nabla^2 f_i(w_k)\right)^{-1}\nabla F_i(w_k) + e_{i,k}\right] \tag{116}$$

Now add and subtract $\nabla F_i(w_k)$ to the right hand side of (116) and use the fact that $\mathbb{E}_{i\sim p}[\nabla F_i(w_k)] = \nabla F(w_k)$ to obtain

$$\begin{aligned}\mathbb{E}[G_i(w_k)] &= \mathbb{E}_{i\sim p}\left[\left(I - \alpha\nabla^2 f_i(w_k)\right)^{-1}\nabla F_i(w_k) - \nabla F_i(w_k) + \nabla F_i(w_k) + e_{i,k}\right] \\ &= \nabla F(w_k) + \mathbb{E}_{i\sim p}\left[\left(I - \alpha\nabla^2 f_i(w_k)\right)^{-1}\nabla F_i(w_k) - \nabla F_i(w_k) + e_{i,k}\right] \\ &= \nabla F(w_k) + \mathbb{E}_{i\sim p}\left[\left(\left(I - \alpha\nabla^2 f_i(w_k)\right)^{-1} - I\right)\nabla F_i(w_k) + e_{i,k}\right]\end{aligned} \tag{117}$$

To simplify the expressions let us define r_k as

$$r_k = \mathbb{E}_{i\sim p}\left[\left((I - \alpha\nabla^2 f_i(w_k))^{-1} - I\right)\nabla F_i(w_k) + e_{i,k}\right] \tag{118}$$

Using the definition of r_k in (118) we can rewrite (117) as

$$\mathbb{E}[G_i(w_k)] = \nabla F(w_k) + r_k \tag{119}$$

Now we proceed to simplify the expression for r_k. Note that using the expansion

$$(I - \alpha\nabla^2 f_i(w_k))^{-1} = I + \sum_{j=1}^{\infty}\alpha^j(\nabla^2 f_i(w_k))^j. \tag{120}$$

we can rewrite r_k defined in (118) as

$$r_k = \sum_{j=1}^{\infty} \alpha^j \mathbb{E}_{i\sim p} \left[(\nabla^2 f_i(w_k))^j \nabla F_i(w_k) \right] + \mathbb{E}_{i\sim p}[e_{i,k}] \tag{121}$$

Next we derive an upper bound on the norm of r_k. The l_2 norm of the first term in (121) can be upper bounded by

$$\begin{aligned} \left\| \sum_{j=1}^{\infty} \alpha^j \mathbb{E}_{i\sim p} \left[(\nabla^2 f_i(w_k))^j \nabla F_i(w_k) \right] \right\| &\le \sum_{j=1}^{\infty} \alpha^j L^j \mathbb{E}_{i\sim p} \|\nabla F_i(w_k)\| \\ &\le \frac{\alpha L}{1-\alpha L} \mathbb{E}_{i\sim p} \|\nabla F_i(w_k)\| \\ &\le 0.22 \|\nabla F(w_k)\| + 2\alpha L \sigma \end{aligned} \tag{122}$$

where the last inequality follows from Lemma A.2 and the fact that $\alpha L \le 1/10$. Further, based on the result in Lemma 4.12 we know that $\|e_{i,k}\|$ for any i is bounded above by $\frac{\alpha L \tilde{\sigma}}{\sqrt{D_{in}}}$. Indeed, when norm of a random variable is bounded above by a constant, norm of its expectation is also upper bounded by that constant. Hence, we can write

$$\|\mathbb{E}_{i\sim p}[e_{i,k}]\| \le \frac{\alpha L \tilde{\sigma}}{\sqrt{D_{in}}} \tag{123}$$

Using the inequalities in (122) and (123) and the definition of r_k in (121) we can show that $\|r_k\|$ is upper bounded by

$$\|r_k\| \le 0.22 \|\nabla F(w_k)\| + 2\alpha L \sigma + 0.1 \frac{\tilde{\sigma}}{\sqrt{D_{in}}}. \tag{124}$$

Hence, by using the inequality $(a+b+c)^2 \le 3a^2 + 3b^2 + 3c^2$ we can show that

$$\|r_k\|^2 \le 0.15 \|\nabla F(w_k)\|^2 + 12\alpha^2 L^2 \sigma^2 + 0.03 \frac{\tilde{\sigma}^2}{D_{in}} \tag{125}$$

Considering this result and the expression in (119) we can write

$$\|\mathbb{E}[G_i(w_k)]\|^2 \le 2\|\nabla F(w_k)\|^2 + 2\|r_k\|^2 \le 2.3 \|\nabla F(w_k)\|^2 + 24\alpha^2 L^2 \sigma^2 + 0.06 \frac{\tilde{\sigma}^2}{D_{in}}. \tag{126}$$

Next, we can derive an upper bound on the second moment of $\|G_i(w_k)\|^2$ similar to the way that we derived (99) in the proof of Theorem F.1. More precisely, note that

$$\mathbb{E}[\|G_i(w_k)\|^2] = \mathbb{E}_{i\sim p} \left[\mathbb{E}_{\mathcal{D}_o^i, D_{in}^i} \left\| \tilde{\nabla} f_i \Big(w_k - \alpha \tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i \Big) \right\|^2 \right]. \tag{127}$$

Using Lemma 4.12 with $\phi = 1$, we have

$$\begin{aligned} \mathbb{E}_{\mathcal{D}_o^i, D_{in}^i} \left\| \tilde{\nabla} f_i \Big(w_k - \alpha \tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i \Big) \right\|^2 &\le 2\|\nabla f_i(w_k - \alpha \nabla f_i(w_k))\|^2 + 2\alpha^2 L^2 \frac{\tilde{\sigma}^2}{D_{in}} + \frac{\tilde{\sigma}^2}{D_o} \\ &\le 2 \frac{\|\nabla F_i(w_k)\|^2}{(1-\alpha L)^2} + 2\alpha^2 L^2 \frac{\tilde{\sigma}^2}{D_{in}} + \frac{\tilde{\sigma}^2}{D_o} \end{aligned} \tag{128}$$

where the last inequality follows from (45b) and the fact that $\|I - \alpha \nabla^2 f_i(w)\| \ge 1 - \alpha L$. Plugging (128) in (127) and using (44) in Lemma A.2 yields

$$\mathbb{E}[\|G_i(w_k)\|^2] \le 20 \|\nabla F(w_k)\|^2 + 7\sigma^2 + \tilde{\sigma}^2 (\frac{1}{D_o} + \frac{0.02}{D_{in}}). \tag{129}$$

Also, using the same argument in deriving (100), (101), and (102) in the proof of Theorem F.1, we obtain

$$\begin{aligned} &\mathbb{E}[F(w_{k+1})|\mathcal{F}_k] \\ &\le F(w_k) - \|\nabla F(w_k)\|^2 \left(\mathbb{E}[\beta_k|\mathcal{F}_k] - \frac{L_k}{2} \mathbb{E}[\beta_k^2|\mathcal{F}_k](2.3 + \frac{20}{B}) \right) \\ &+ \mathbb{E}[\beta_k|\mathcal{F}_k] \|\nabla F(w_k)\| \|r_k\| + \frac{L_k}{2} \mathbb{E}[\beta_k^2|\mathcal{F}_k] \left(\frac{1}{B} \left(7\sigma^2 + \tilde{\sigma}^2 (\frac{1}{D_o} + \frac{0.02}{D_{in}}) \right) + 24\alpha^2 L^2 \sigma^2 + 0.06 \frac{\tilde{\sigma}^2}{D_{in}} \right). \end{aligned} \tag{130}$$

Note that, using (125), we have

$$\|\nabla F(w_k)\|\|r_k\| \leq \frac{1}{2}\left(\frac{\|\nabla F(w_k)\|^2}{2} + 2\|r_k\|^2\right) \leq 0.4\|\nabla F(w_k)\|^2 + 0.03\frac{\tilde{\sigma}^2}{D_{in}} + 12\alpha^2L^2\sigma^2.$$

Plugging this bound in (130) implies

$$\begin{aligned}\mathbb{E}[F(w_{k+1})|\mathcal{F}_k] &\leq F(w_k) - \|\nabla F(w_k)\|^2\left(0.6\mathbb{E}[\beta_k|\mathcal{F}_k] - \frac{L_k}{2}\mathbb{E}[\beta_k^2|\mathcal{F}_k](2.3+\frac{20}{B})\right)\\ &+ \frac{L_k}{2}\mathbb{E}[\beta_k^2|\mathcal{F}_k]\left(\frac{1}{B}\left(7\sigma^2+\tilde{\sigma}^2(\frac{1}{D_o}+\frac{0.02}{D_{in}})\right)+24\alpha^2L^2\sigma^2+0.06\frac{\tilde{\sigma}^2}{D_{in}}\right)\\ &+ \mathbb{E}[\beta_k|\mathcal{F}_k](12\alpha^2L^2\sigma^2+0.03\frac{\tilde{\sigma}^2}{D_{in}}).\end{aligned}$$

Using $\beta_k = \tilde{\beta}(w_k)/18$, and with similar analysis as Theorem F.1, we obtain

$$\mathbb{E}[F(w_{k+1})|\mathcal{F}_k] \leq F(w_k) - \frac{1}{100L_k}\|\nabla F(w_k)\|^2 + \sigma^2(\frac{7}{828LB}+\frac{\alpha^2L}{6}) + \frac{\tilde{\sigma}^2/D_o}{828LB} + \frac{\tilde{\sigma}^2/D_{in}}{600L}$$

which is similar to (103), and the rest of proof follows same as the way that we derived (104)-(112) in the proof of Theorem F.1. □

H Proof of Theorem 4.21 (General Version)

Theorem H.1. *Consider the objective function F defined in (3) for the case that $\alpha \in (0, \frac{1}{6L}]$. Suppose that the conditions in Assumptions 4.3-4.6 are satisfied, and recall the definitions $L := \max L_i$ and $\rho := \max \rho_i$. Consider running HF-MAML with batch sizes satisfying the conditions $D_h \geq \lceil 36(\alpha\rho\sigma_H)^2\rceil$ and $B \geq 20$. Let $\beta_k = \tilde{\beta}(w_k)/25$ where $\tilde{\beta}(w)$ is defined in (26). Also, we choose the approximation parameter δ_k^i in HF-MAML as*

$$\delta_k^i = \frac{1}{6\rho\alpha\|\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i)\|}$$

Then, for any $\epsilon > 0$, HF-MAML finds a solution w_ϵ such that

$$\mathbb{E}[\|\nabla F(w_\epsilon)\|] \leq \max\left\{6\sqrt{(1+\frac{\rho\alpha}{L}\sigma)\left(\frac{\sigma^2}{B}+\frac{\tilde{\sigma}^2}{BD_o}+\frac{\tilde{\sigma}^2}{D_{in}}\right)}, 36\frac{\rho\alpha}{L}\left(\frac{\sigma^2}{B}+\frac{\tilde{\sigma}^2}{BD_o}+\frac{\tilde{\sigma}^2}{D_{in}}\right), \epsilon\right\} \tag{131}$$

after at most running for

$$\mathcal{O}(1)\Delta\min\left\{\frac{L+\rho\alpha(\sigma+\epsilon)}{\epsilon^2}, \frac{LB}{\sigma^2}+\frac{L(BD_o+D_{in})}{\tilde{\sigma}^2}\right\} \tag{132}$$

iterations, where $\Delta := (F(w_0) - \min_{w\in\mathbb{R}^d} F(w))$.

Proof. Note that the update of the first-order approximation of MAML can be written as $w_{k+1} = w_k - \frac{\beta_k}{B}\sum_{i\in\mathcal{B}_k} G_i(w_k)$, where

$$G_i(w) := \tilde{\nabla} f_i\Big(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i\Big) - \alpha d_k^i,$$

and d_k^i is given by (12). Similar to previous proofs, we first derive upper bounds on the first and second moment of $G_i(w_k)$. Using the definition

$$e_{i,k} = \nabla f_i\left(w_k - \alpha\nabla f_i(w_k)\right) - \mathbb{E}_{\mathcal{D}_{in},\mathcal{D}_o}[\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k, \mathcal{D}_{in}^i), \mathcal{D}_o^i)] \tag{133}$$

we can write that

$$\mathbb{E}[G_i(w)] = \mathbb{E}_{i\sim p}\left[\nabla f_i(w_k - \alpha\nabla f_i(w_k))\right] + \mathbb{E}_{i\sim p}[e_{i,k}] - \alpha\ \mathbb{E}_p[d_k^i] \tag{134}$$

Next, note that

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i,\mathcal{D}_h^i}[d_k^i] \\
&= \mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i}\left[\frac{\nabla f_i\Big(w_k + \delta_k^i \tilde{\nabla} f_i(w_k - \alpha \tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\Big) - \nabla f_i\Big(w_k - \delta_k^i \tilde{\nabla} f_i(w_k - \alpha \tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\Big)}{2\delta_k^i}\right] \\
&= \mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i}\left[\nabla^2 f_i(w_k)\tilde{\nabla} f_i(w_k - \alpha \tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i) + \tilde{e}_k^i\right] \qquad (135)
\end{aligned}
$$

with

$$
\tilde{e}_k^i = \nabla^2 f_i(w_k)v - \left[\frac{\nabla f_i(w_k + \delta_k^i v) - \nabla f_i(w_k - \delta_k^i v)}{2\delta_k^i}\right] \qquad (136)
$$

where $v = \tilde{\nabla} f_i(w_k - \alpha \tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)$. Next, by using the definition in (133) we can simplify (135) and write

$$
\mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i,\mathcal{D}_h^i}[d_k^i] = \nabla^2 f_i(w_k)\left(\nabla f_i(w_k - \alpha \nabla f_i(w_k)) + e_{i,k}\right) + \mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i}[\tilde{e}_k^i] \qquad (137)
$$

Plugging (137) in (134), we obtain

$$
\begin{aligned}
\mathbb{E}[G_i(w)] &= \mathbb{E}_{i\sim p}\left[\left(I - \alpha\nabla^2 f_i(w_k)\right)\left(\nabla f_i(w_k - \alpha\nabla f_i(w_k)) + e_{i,k}\right)\right] - \alpha\ \mathbb{E}[\tilde{e}_k^i] \\
&= \nabla F(w_k) + \mathbb{E}_{i\sim p}\left[\left(I - \alpha\nabla^2 f_i(w_k)\right)e_{i,k}\right] - \alpha\ \mathbb{E}[\tilde{e}_k^i]. \qquad (138)
\end{aligned}
$$

Now we proceed to bound the norm of each term in the right hand side of (138). First, note that according to Lemma 4.12 we know that $\|e_{i,k}\|$ is bounded above by

$$
\|e_{i,k}\| \leq \frac{\alpha L\tilde{\sigma}}{\sqrt{D_{in}}}. \qquad (139)
$$

Therefore, we can show that

$$
\mathbb{E}_{i\sim p}\left[\left(I - \alpha\nabla^2 f_i(w_k)\right)e_{i,k}\right] \leq \mathbb{E}_{i\sim p}\left[\|I - \alpha\nabla^2 f_i(w_k)\|\|e_{i,k}\|\right] \leq (1+\alpha L)\frac{\alpha L\tilde{\sigma}}{\sqrt{D_{in}}} \qquad (140)
$$

Next, we derive an upper bound on $\|\tilde{e}_k^i\|$. Note that for any vector v can show that

$$
\left\|\nabla^2 f_i(w_k)v - \left[\frac{\nabla f_i(w_k + \delta_k^i v) - \nabla f_i(w_k - \delta_k^i v)}{2\delta_k^i}\right]\right\| \leq \rho\delta_k^i\ \|v\|^2
$$

by using the fact the Hessians are ρ-Lipschitz continuous. Now if we set

$$
v = \tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i),
$$

then by the definition of $\tilde{e}_k^i$ in (136) we can show that

$$
\|\tilde{e}_k^i\| \leq \rho\ \delta_k^i\ \|\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\|^2 \qquad (141)
$$

If we replace δ_k^i by its definition

$$
\delta_k^i = \frac{\delta}{\|\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\|}, \qquad (142)
$$

where $\delta := \frac{1}{6\rho\alpha}$, then we can show that

$$
\|\tilde{e}_k^i\| \leq \rho\delta\ \|\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\| \qquad (143)
$$

Therefore, we have

$$
\begin{aligned}
\|\mathbb{E}[\tilde{e}_k^i]\| &= \left\|\mathbb{E}_{i\sim p}\left[\mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i}[\tilde{e}_k^i]\right]\right\| \leq \mathbb{E}\left[\mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i}[\|\tilde{e}_k^i\|]\right] \\
&\leq \rho\delta\ \mathbb{E}_{i\sim p}\left[\mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i}\left[\|\tilde{\nabla} f_i(w_k - \alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\|\right]\right] \\
&\leq \rho\delta\ \mathbb{E}_{i\sim p}\left[\nabla f_i(w_k - \alpha\nabla f_i(w_k))\right] + \rho\delta\ \frac{\alpha L\tilde{\sigma}}{\sqrt{D_{in}}} \qquad (144) \\
&\leq \rho\delta\ \mathbb{E}_{i\sim p}\left[\left(I - \alpha\nabla^2 f_i(w_k)\right)^{-1}\nabla F_i(w_k)\right] + \rho\delta\ \frac{\alpha L\tilde{\sigma}}{\sqrt{D_{in}}} \\
&\leq \frac{\rho\delta}{1-\alpha L}\ \nabla F(w_k) + \rho\delta\ \frac{\alpha L\tilde{\sigma}}{\sqrt{D_{in}}} \qquad (145)
\end{aligned}
$$

where (144) follows from Lemma 4.12. Considering the bounds (140) and (145) as well as the result in (138), we can write

$$\mathbb{E}[G_i(w)] = \nabla F(w_k) + s_k \tag{146}$$

with

$$\begin{aligned}\|s_k\| &\le \frac{\rho\delta\alpha}{1-\alpha L}\ \nabla F(w_k) + (1+\alpha L+\rho\delta\alpha)\ \frac{\alpha L\tilde{\sigma}}{\sqrt{D_{in}}}\\ &\le 0.2\nabla F(w_k) + 0.3\frac{\tilde{\sigma}}{\sqrt{D_{in}}}.\end{aligned} \tag{147}$$

where the last inequality is derived using $\alpha L, \rho\delta\alpha \le 1/6$. As a consequence, we also have

$$\|\mathbb{E}[G_i(w)]\|^2 \le 2\|\nabla F(w_k)\|^2 + 2\|s_k\|^2 \le 2.2\|\nabla F(w_k)\|^2 + 0.4\frac{\tilde{\sigma}^2}{D_{in}}. \tag{148}$$

Next, to bound the second moment of $G_i(w_k)$, note that

$$\begin{aligned}\mathbb{E}[\|G_i(w_k)\|^2] &\le 2\mathbb{E}\left[\|\tilde{\nabla} f_i\left(w_k - \alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i\right)\|^2\right] + 2\alpha^2\mathbb{E}[\|d_k^i\|^2]\\ &\le 2\mathbb{E}\left[\|\tilde{\nabla} f_i\left(w_k - \alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i\right)\|^2\right] + 2\alpha^2\mathbb{E}_{i\sim p}\left[\mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i,\mathcal{D}_h^i}[\|d_k^i\|^2]\right]\end{aligned} \tag{149}$$

where $\mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i,\mathcal{D}_h^i}[\|d_k^i\|^2]$ can be bounded as follows

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i,\mathcal{D}_h^i}[\|d_k^i\|^2]\\ &\le \mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i}\left[\left\|\frac{\nabla f_i\left(w_k+\delta_k^i\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\right)-\nabla f_i\left(w_k-\delta_k^i\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\right)}{2\delta_k^i}\right\|^2\right.\\ &\quad\left.+\frac{\sigma_H^2}{D_h(\delta_k^i)^2}\right]\end{aligned}$$

and the last inequality comes from the fact that $\mathrm{Var}(X+Y) \le 2(\mathrm{Var}(X)+\mathrm{Var}(Y))$. Moreover, according to the definition in (136) we can write that

$$\begin{aligned}&\frac{\nabla f_i\left(w_k+\delta_k^i\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\right)-\nabla f_i\left(w_k-\delta_k^i\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\right)}{2\delta_k^i}\\ &\quad= \nabla^2 f_i(w_k)\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i) + \tilde{e}_k^i\end{aligned}$$

which implies that

$$\mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i,\mathcal{D}_h^i}[\|d_k^i\|^2] \le \mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i}\left[\left\|\nabla^2 f_i(w_k)\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i) + \tilde{e}_k^i\right\|^2 + \frac{\sigma_H^2}{D_h(\delta_k^i)^2}\right] \tag{150}$$

Now, replace δ_k^i in the second term by its definition in (142) to obtain

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i,\mathcal{D}_h^i}[\|d_k^i\|^2]\\ &\le \mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i}\left[\left\|\nabla^2 f_i(w_k)\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i) + \tilde{e}_k^i\right\|^2 + \frac{\sigma_H^2}{D_h\delta^2}\|\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\|^2\right]\end{aligned}$$

Using this bound along with the inequalities $(a+b)^2 \le 2a^2+2b^2$ and

$$\left\|\nabla^2 f_i(w_k)\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\right\|^2 \le L^2\|\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\|^2$$

we can show that

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i,\mathcal{D}_h^i}[\|d_k^i\|^2]\\ &\le \mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i}\left[(\frac{\sigma_H^2}{D_h\delta^2}+2L^2)\|\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\|^2 + 2\|\tilde{e}_k^i\|^2\right]\\ &\le (\frac{\sigma_H^2}{D_h\delta^2}+2L^2+2\rho^2\delta^2)\mathbb{E}_{\mathcal{D}_o^i,\mathcal{D}_{in}^i}\left[\|\tilde{\nabla} f_i(w_k-\alpha\tilde{\nabla} f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i)\|^2\right]\end{aligned} \tag{151}$$

where the last inequality follows from (143). Plugging this bound in (149) leads to

$$\begin{aligned}\mathbb{E}[\|G_i(w_k)\|^2] &\leq (2+2\alpha^2(\frac{\sigma_H^2}{D_h\delta^2}+2L^2+2\rho^2\delta^2))\mathbb{E}\left[\|\tilde{\nabla}f_i\left(w_k-\alpha\tilde{\nabla}f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i\right)\|^2\right]\\ &\leq 4.3\ \mathbb{E}\left[\|\tilde{\nabla}f_i\left(w_k-\alpha\tilde{\nabla}f_i(w_k,\mathcal{D}_{in}^i),\mathcal{D}_o^i\right)\|^2\right] \qquad (152)\end{aligned}$$

where the last inequality is derived using $\alpha L \leq 1/6$ along with $\rho\delta\alpha = 1/6$ and $D_h \geq 36(\rho\alpha\sigma_H)^2$. Now, using Lemma 4.12 with $\phi = 10$, we can write

$$\begin{aligned}\mathbb{E}[\|G_i(w_k)\|^2] &\leq 4.3\left(1+\frac{1}{10}\right)\mathbb{E}[\|\nabla f_i(w_k-\alpha\nabla f_i(w_k)\|^2]+47.3\alpha^2L^2\frac{\tilde{\sigma}^2}{D_{in}}+4.3\frac{\tilde{\sigma}^2}{D_o} \qquad (153)\\ &\leq 5\mathbb{E}[\|\nabla f_i(w_k-\alpha\nabla f_i(w_k))\|^2]+5\tilde{\sigma}^2\left(\frac{1}{D_{in}}+\frac{1}{D_o}\right) \qquad (154)\\ &\leq \frac{5}{(1-\alpha L)^2}\mathbb{E}[\|\nabla F_i(w_k)\|^2]+5\tilde{\sigma}^2\left(\frac{1}{D_{in}}+\frac{1}{D_o}\right) \qquad (155)\end{aligned}$$

where (154) is a simplification of (153) using $\alpha L \leq 1/6$ and (155) comes from the fact that

$$\|\nabla f_i(w_k-\alpha\nabla f_i(w_k))\| = \|(I-\alpha\nabla^2 f_i(w_k))^{-1}\nabla F_i(w_k)\| \leq \frac{1}{1-\alpha L}\|\nabla F_i(w_k)\|.$$

Now, using (44) in Lemma (A.2), we can show that

$$\mathbb{E}[\|G_i(w_k)\|^2] \leq 50\|\nabla F(x)\|^2+18\sigma^2+5\tilde{\sigma}^2\left(\frac{1}{D_{in}}+\frac{1}{D_o}\right). \qquad (156)$$

Once again, the same argument as the proof of Theorem F.1, we obtain

$$\begin{aligned}\mathbb{E}[F(w_{k+1})|\mathcal{F}_k] &\leq F(w_k)-\|\nabla F(w_k)\|^2\left(\mathbb{E}[\beta_{i,k}|\mathcal{F}_k]-\frac{L_k}{2}\mathbb{E}[\beta_{i,k}^2|\mathcal{F}_k](2.2+\frac{50}{B})\right)\\ &+\mathbb{E}[\beta_{i,k}|\mathcal{F}_k]\|\nabla F(w_k)\|\|s_k\|+\frac{L_k}{2}\mathbb{E}[\beta_{i,k}^2|\mathcal{F}_k]\left(\frac{1}{B}\left(18\sigma^2+5\tilde{\sigma}^2\left(\frac{1}{D_o}+\frac{1}{D_{in}}\right)\right)+0.4\frac{\tilde{\sigma}^2}{D_{in}}\right). \qquad (157)\end{aligned}$$

Note that, using (147), we have

$$\|\nabla F(w_k)\|\|s_k\| \leq \frac{1}{2}\left(\frac{\|\nabla F(w_k)\|^2}{2}+2\|s_k\|^2\right) \leq 0.4\|\nabla F(w_k)\|^2+0.18\frac{\tilde{\sigma}^2}{D_{in}}.$$

Plugging this bound in (157) implies

$$\begin{aligned}\mathbb{E}[F(w_{k+1})|\mathcal{F}_k] &\leq F(w_k)-\|\nabla F(w_k)\|^2\left(0.6\mathbb{E}[\beta_{i,k}|\mathcal{F}_k]-\frac{L_k}{2}\mathbb{E}[\beta_{i,k}^2|\mathcal{F}_k](2.2+\frac{50}{B})\right)\\ &+\frac{L_k}{2}\mathbb{E}[\beta_{i,k}^2|\mathcal{F}_k]\left(\frac{1}{B}\left(18\sigma^2+5\tilde{\sigma}^2\left(\frac{1}{D_o}+\frac{1}{D_{in}}\right)\right)+0.4\frac{\tilde{\sigma}^2}{D_{in}}\right)+0.18\mathbb{E}[\beta_{i,k}|\mathcal{F}_k]\frac{\tilde{\sigma}^2}{D_{in}}.\end{aligned}$$

Using $\beta_k = \tilde{\beta}(w_k)/25$, and with similar analysis as Theorem F.1, we obtain

$$\mathbb{E}[F(w_{k+1})|\mathcal{F}_k] \leq F(w_k)-\frac{1}{200L_k}\|\nabla F(w_k)\|^2+\frac{18\sigma^2+5\tilde{\sigma}^2/D_o}{1600LB}+\frac{\tilde{\sigma}^2/D_{in}}{100L}$$

which is again similar to (103), and the rest of proof follows same as the way that we derived (104)-(112) in the proof of Theorem F.1. □