

---

# On the Stability and Generalization of Meta-Learning: the Impact of Inner-Levels

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1       Meta-learning has achieved significant advancements, with generalization emerging  
2       as a key metric for evaluating meta-learning algorithms. While recent studies have  
3       mainly focused on training strategies, data-split methods, and tightening general-  
4       ization bounds, they often ignore the impact of inner-levels on generalization. To  
5       bridge this gap, this paper focuses on several prominent meta-learning algorithms  
6       and establishes two generalization analytical frameworks for them based on their  
7       inner-processes: the Gradient Descent Framework (GDF) and the Proximal Descent  
8       Framework (PDF). Within these frameworks, we introduce two novel algorithmic  
9       stability definitions and derive the corresponding generalization bounds. Our find-  
10      ings reveal a trade-off of inner-levels under GDF, whereas PDF exhibits a beneficial  
11      relationship. Moreover, we highlight the critical role of the meta-objective function  
12      in minimizing generalization error. Inspired by this, we propose a new, simplified  
13      meta-objective function definition to enhance generalization performance. Many  
14      real-world experiments support our findings and show the improvement of the new  
15      meta-objective function.

## 16   1 Introduction

17   Meta-learning has been proven to be a powerful paradigm for extracting well-generalization from  
18   previous tasks and quickly learning new tasks [1]. It has received increasing attention in many  
19   machine learning applications such as few-shot learning [2], robust learning [3], and natural language  
20   processing [4]. The key idea of meta-learning is to improve the learning ability of agents through a  
21   learning-to-learn process. In recent years, optimization-based meta-learning algorithms have emerged  
22   as a popular approach [5–8]. These studies formulate the problem as a bi-level optimization problem  
23   and have demonstrated impressive performance across various domains, significant attention from  
24   the research community. In particular, at the outer-level, it trains a meta-learner to extract task-shared  
25   knowledge from meta-training tasks. At the inner-level, a basic model, which is initialized using the  
26   meta-parameters, adapts to each task by taking  $Q$  inner-level gradient updating, where the  $Q$  times  
27   inner-level update is commonly called an inner-process [5, 9–11].

28   Despite the remarkable success of meta-learning, its theoretical understanding of generalization  
29   remains largely unexplored. Recent studies have primarily focused on analyzing the effect of training  
30   strategy [12, 13], data splitting methods [14, 15] on generalization error or on advancing tighter  
31   generalization bounds [16, 17], while overlooking the impact of inner-levels on generalization, i.e.,  
32   *the relationship between the generalization error and the number of inner-levels  $Q$ .*

33   In particular, there are two main inner-processes frameworks in current meta-learning algorithms, as  
34   shown in Table 1. One is the Gradient Descent-based Framework (GDF) [5, 6, 18], where the key  
35   idea is to measure the closeness of the initial prior hypothesis to the target optimal hypothesis by the  
36   number of gradient descent steps. However, this approach incurs high computational costs due to the

Table 1: The summary of main meta-learning algorithms

Frame.	Algorithm	Inner-Level Process	Convex	Non-Convex
GDF	MAML[5]	$w_{T_i} = w - \alpha \sum_{q=0}^{Q-1} \nabla \widehat{\mathcal{L}}_i(w_{T_i}^q, \mathcal{D}_i)$	$\mathcal{O}\left(\frac{TQ}{mn^{\frac{1}{\tau}}} + \frac{\sqrt{F(w^0) - \min_W F + T}}{m}\right)$	$\mathcal{O}\left(\left(\frac{1+\frac{1}{m}}{m}\right)^{\frac{1}{c\gamma}} \left(1 + \frac{Q}{n^{\frac{1}{\tau}}}\right)^{\frac{1}{c\gamma}} (F(w^0)T)^{\frac{c\gamma}{1+c\gamma}}\right)$
	FO-MAML [5]	$= \arg \min_{w_{T_i}} \left\{ \left( \sum_{q=0}^{Q-1} \nabla \mathcal{L}_i(w_{T_i}^q, \mathcal{D}_i) \right)^2 \right\}$		
	Meta-SGD [6]	$w_{T_i} = w + \frac{1}{2\alpha} \ w_{T_i} - w\ _2^2$		
PDF	iMAML [20]	$w_{T_i} = \arg \min_{w_{T_i}} \widehat{\mathcal{L}}_i(w_{T_i}, \mathcal{D}_i)$	$\mathcal{O}\left(\frac{T}{mC^Q} + \frac{\sqrt{F(w^0) - \min_W F + T}}{m}\right)$	$\mathcal{O}\left(\left(\frac{1+\frac{1}{m}}{m}\right)^{\frac{1}{c\gamma}} \left(1 + \frac{1}{C^Q}\right)^{\frac{1}{c\gamma}} (F(w^0)T)^{\frac{c\gamma}{1+c\gamma}}\right)$
	Meta-MinibatchProx [7]	$+ \frac{\lambda}{2} \ w_{T_i} - w\ ^2$		
	FO-MuML [11]			

need for second-order derivatives and requires careful tuning of multiple hyper-parameters [19]. To address these limitations, another inner-process framework PDF, which is based on proximal descent, has been developed [20, 7, 11]. It only depends on the solution to the inner optimization and not the path taken by the inner optimization algorithm.

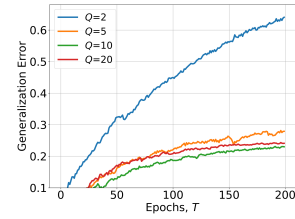
To validate the impact of  $Q$  on the two frameworks, we conducted two simple experiments on the Omniglot dataset [21] using MAML [5] and Meta-MinibatchProx [7] to be the examples. In Figure 1, the generalization error (Test Loss - Training Loss) first decreases but then increases with the number of inner-levels  $Q$  grows in MAML, while always decreases in Meta-MinibatchProx. This different behavior, attributed to the distinct inner processes of the two frameworks, motivates the need for a deeper analysis to help the design of  $Q$  for improved generalization performance.

To analyze the relationship between inner-levels  $Q$  and generalization error under these two frameworks: GDF and PDF, this paper leverages the algorithmic stability to characterize the generalization of algorithm [16, 22], which measures sensitivity to perturbations in the training dataset. To the best of our knowledge, this is the first study to investigate the influence of inner-levels on the generalization of meta-learning under two frameworks. Our findings offer valuable insights for developing efficient meta-learning algorithms. The main contributions can be summarized as follows:

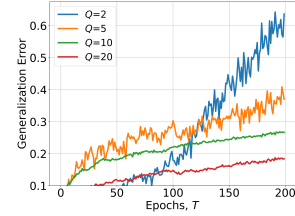
- (1) We summarize six mainstream meta-learning algorithms and extract their structural features. Based on their inner-processes, we classify these algorithms into two frameworks: GDF and PDF, and develop two definitions of on-average stability, respectively. Accordingly, we establish a quantitative relationship between inner-levels and the generalization error in convex and non-convex settings.
- (2) Our results reveal the influence of the inner-levels  $Q$  on generalization error. In particular, we identify a trade-off relationship in GDF, whereas PDF demonstrates a beneficial relationship in its generalization bound. The primary reason for this difference lies in the term introduced by the inner-process. For example, in convex setting, the term for GDF,  $\mathcal{O}\left(\frac{TQ}{mn^{\frac{1}{\tau}}}\right)$ , increases with  $Q$ , whereas the term for PDF,  $\mathcal{O}\left(\frac{T}{mC^Q}\right)$ , decreases with  $Q$ . These findings help to design a more efficient inner-process of meta-learning.
- (3) Based on the generalization results of GDF and PDF, we further derive the generalization bounds for six meta-learning algorithms and analyze their implications. In general, note that the meta-objective  $F(w)$  plays a crucial role in reducing the generalization bound. Motivated by this, we propose a new meta-objective  $F_{\text{new}}(w)$  and prove  $F_{\text{new}}(w) \leq F(w)$ , thereby enhancing generalization performance. Extensive experiments confirm the efficiency of the proposed objective.

## 2 Related Work

**Algorithm Stability.** Algorithmic stability is critical for learnability [23]. There are two main approaches to investigating the stability-based generalization bound of meta-learning: (1) The first approach, introduced by [24], focuses on deriving generalization bounds for the transfer error, based on the assumption of independent task environments. This line of research has been further



(a) MAML.



(b) Meta-MinibatchProx.

Figure 1: Effect of  $Q$  on Omniglot dataset.

developed in several influential works [12, 25, 13, 17]. In this approach, the generalization error is defined as the transfer error minus the training error. (2) The second approach, by quantifying the test error, successfully eliminates the assumption of independent task environments and provides tighter generalization bounds compared to the first approach [16]. Additionally, [22] re-examined the generalization performance of meta-learning from a bi-level perspective, i.e., the inner-level and outer-level, and concluded that inverted regularization at the inner level helps to reduce the generalization bound. Differently, we make a further development in the second group by re-examining the generalization ability from two main structures at inner-level and reveal the two distinct impacts of inner-levels. Notably, the algorithm stability of prior work is limited in one-step MAML, highlighting the need for a novel algorithm stability definition.

**Meta-Learning.** Recent advances in meta-learning have spurred significant progress. From the optimization perspective, [26] analyzed the convergence rate and computational complexity of ANIL, while [27] investigated the convergence properties with a single adaptation step. Extending this, [28] developed a theoretical framework for MAML with multiple inner-levels. [11] introduced a novel analysis of first-order meta-learning algorithms. To improve the in-context learning performance of pre-trained models, several meta-training-based approaches have been proposed. For example, MetaICT [29] uses BinaryCLFs and LAMA datasets to create tasks, while pre-pending human-generated instructions to each task. In contrast, MetaICL [30] leverages a wide variety of disjoint tasks to meta-train large language models. [31] demonstrates that meta-learning can aid in cross-task generalization for prompt tuning. Moreover, MAML-en-LLM [32] is capable of learning truly generalizable parameters that not only perform well across disjoint tasks but also adapt effectively to unseen tasks.

### 3 Preliminaries

In this paper, each data point  $z = (x, y) \in \mathcal{Z}$  consists of an input  $x \in \mathcal{X}$  and its corresponding label  $y \in \mathcal{Y}$ . We assume access to  $m$  tasks, denoted by  $\mathcal{T}_1, \dots, \mathcal{T}_m$  with the data for each task  $\mathcal{T}_i, i \in [m]$ , generated from an unknown distribution  $\mathcal{P}_i$ . We use the loss function  $\ell : \mathbb{R}^d \times \mathcal{Z} \rightarrow \mathbb{R}^+$  to evaluate the performance of a model parameterized by  $w \in \mathcal{W}$ , where  $\mathcal{W}$  is a closed subset of  $\mathbb{R}^d$ . The population loss corresponding for task  $\mathcal{T}_i$  is defined as  $\mathcal{L}_i(w) := \mathbb{E}_{z \sim \mathcal{P}_i} [\ell(w, z)]$ . In addition, we use the notation  $\hat{\mathcal{L}}_i(w)$  to denote the empirical loss for task dataset  $\mathcal{D}_i$ , i.e.,  $\hat{\mathcal{L}}_i(w, \mathcal{D}_i) := \frac{1}{|\mathcal{D}_i|} \sum_{z \in \mathcal{D}_i} \ell(w, z)$ , where  $|\mathcal{D}_i|$  is the size of  $\mathcal{D}_i$ . The goal of meta-learning is to learn good meta-parameters  $w$  that perform wells across different tasks, which can be written as follows:

$$\min_{w \in \mathcal{W}} F(w) := \frac{1}{m} \sum_{i=1}^m F_i(w), \quad (1)$$

where  $F_i(w)$  is used to denote the performance of  $w$  on task  $\mathcal{T}_i$ , after undergoing multiple (or one) gradient updates to adapt the parameters to the specific task. At a high level, a.k.a., the outer-level, the learner needs to figure out useful meta-information that can generalize across tasks. At the inner-level, the learner needs to find task-specific parameters  $w_{\mathcal{T}_i}$  that perform well on individual tasks after undergoing an inner-process.

#### 3.1 The GDF Framework

From the inner-level perspective, two distinct inner-processes have primarily been developed, giving rise to two dominant meta-learning frameworks, as illustrated in Table 1. One prominent meta-learning framework, called the Gradient Descent Framework (GDF), is exemplified by MAML [5], FO-MAML [5], and Meta-SGD [6]. In particular, the inner-process finds the task-specific optimal hypothesis by solving the following optimization problem:

$$\min_{w_{\mathcal{T}_i}} \left\langle \sum_{q=0}^{Q-1} \nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i}^q, \mathcal{D}_i), w_{\mathcal{T}_i} - w \right\rangle + \frac{1}{2\alpha} \|w_{\mathcal{T}_i} - w\|_2^2, \quad (2)$$

where  $Q$  is the number of inner-levels, i.e., gradient descent steps in GDF and  $\alpha$  is the inner step size. In fact, by taking the derivative on  $w_{\mathcal{T}_i}$  and making the gradient equal to 0, (2) can be written as  $w_{\mathcal{T}_i} = w - \alpha \sum_{q=0}^{Q-1} \nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i}^q, \mathcal{D}_i)$ , which reveals that task-specific parameters are iteratively learned by minimizing the empirical loss based using gradient descent. Specifically, in GDF, without loss of generality, we mainly follow MAML [5], and summarize the training procedure in Algorithm 1.

---

**Algorithm 1** GDF and PDF
 

---

```

1: The set of datasets  $\mathcal{S} = \{S_i\}_{i=1}^m$ , outer iterations  $T$ , inner-levels  $Q$ , regulation  $\lambda$ .
2: Choose arbitrary initial point  $w^0 \in W$ ;
3: for  $t = 0$  to  $T - 1$  do
4:   Randomly choose the task  $i$ .
5:   Inner-Level:  $w_{\mathcal{T}_i,0}^t = w_t$ 
6:   for  $q = 0, 1, \dots, Q - 1$  do
7:      $w_{\mathcal{T}_i,q+1}^t = w_{\mathcal{T}_i,q}^t - \alpha \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i,q}^t, S_i^{\text{tr}})$ ;
8:      $w_{\mathcal{T}_i,q+1}^t = w_{\mathcal{T}_i,q}^t - \alpha \nabla \widehat{\mathcal{K}}(w_{\mathcal{T}_i,q}^t, S_i)$ ;
9:   end for
10:   $w^{t+1} := w^t - \eta_t \nabla_w \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,Q}^t, S_i^{\text{ts}})$ 
11:   $w^{t+1} := w^t - \eta_t \lambda (w^t - w_{\mathcal{T}_i,Q}^t)$ 
12: end for
13:  $w^T$  and  $\bar{w}^T := \frac{1}{T+1} \sum_{t=0}^T w^t$ ;
  
```

---

For each task  $\mathcal{T}_i$ , we group the samples into two distinct sets among  $n$  training samples: a support set  $S_i^{\text{tr}}$  of size  $n^{\text{tr}}$  for meta-train at inner-level and a query set  $S_i^{\text{ts}}$  of size  $n^{\text{ts}}$  for meta-validation at outer-level. Then, for each task  $\mathcal{T}_i$ , we have one corresponding training set  $S_i := \{S_i^{\text{tr}}, S_i^{\text{ts}}\}$  in GDF. Referring back to (1), the formal definition of the meta-loss of task  $\mathcal{T}_i$  is:

$$F_i(w) := \mathbb{E}_{\mathcal{D}_i} \mathbb{E}_{z \in \mathcal{P}_i} [\ell(w_{\mathcal{T}_i}^Q(w, \mathcal{D}_i), z)]. \quad (3)$$

Because it is difficult to obtain the (3) in practical applications, we usually use the empirical loss  $\widehat{F}_i(w, S_i) := \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^Q(w, S_i^{\text{tr}}), S_i^{\text{ts}}) = \frac{1}{n^{\text{ts}}} \sum_{z \in S_i^{\text{ts}}} \ell(w_{\mathcal{T}_i}^Q(w, S_i^{\text{tr}}), z)$  to approximate it.

### 3.2 The PDF Framework

Another widely used framework of meta-learning, known as the Proximal Descent Framework (PDF), is based on a “proximal” descent, wherein task-specific parameters are iteratively learned by minimizing the empirical loss and an  $\ell_2$  regularizer. This framework is exemplified by iMAML [20], Meta-MinibatchProx [7], Fo-MuML [11]. In this framework, the inner-level optimization problem is formulated as follows:

$$\min_{w_{\mathcal{T}_i}} \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}, \mathcal{D}_i) + \frac{\lambda}{2} \|w_{\mathcal{T}_i} - w\|^2, \quad (4)$$

where  $\lambda \geq 0$  is a regularization constant. In PDF, without loss of generality, we mainly follow Meta-MinibatchProx [7]. In particular, the task-specific training set  $S_i$  of size  $n$  is used directly for meta-train at inner-level, without the data into two distinct sets. Then, we present the following formulation of the meta-loss for task  $\mathcal{T}_i$  corresponding to the PDF:

$$F_i(w) := \min_{w_{\mathcal{T}_i}} \left\{ \mathbb{E}_{z \sim \mathcal{P}_i} \ell(w_{\mathcal{T}_i}, z) + \frac{\lambda}{2} \|w_{\mathcal{T}_i} - w\|^2 \right\} = \mathbb{E}_{z \sim \mathcal{P}_i} \ell_\lambda(w, z). \quad (5)$$

and the empirical loss is  $\widehat{F}_i(w, S_i) := \min_{w_{\mathcal{T}_i}} \left\{ \frac{1}{n} \sum_{z \in S_i} \ell(w_{\mathcal{T}_i}, z) + \frac{\lambda}{2} \|w_{\mathcal{T}_i} - w\|^2 \right\} = \widehat{\mathcal{L}}_\lambda(w, S_i)$ . Note that we use  $\mathbb{E}_{z \sim \mathcal{P}_i} \ell_\lambda(w, z)$  and  $\widehat{\mathcal{L}}_\lambda(w, S_i)$  forlicity without impacting our analysis.

For a practical PDF-based algorithm, it’s difficult to get an exact solution of  $\widehat{\mathcal{L}}_\lambda(w, S_i)$ , thereby we usually turn to get the inexact solution,  $\widehat{\mathcal{K}}(w, S_i) = \frac{1}{n} \sum_{z \in S_i} \ell(w_{\mathcal{T}_i}, z) + \frac{\lambda}{2} \|w_{\mathcal{T}_i} - w\|^2$ . Instead of solving (1), we solve its sample average surrogate problem as  $\arg \min_{w \in \mathcal{W}} \widehat{F}(w, \mathcal{S}) := \frac{1}{m} \sum_{i=1}^m \widehat{F}_i(w, S_i)$ , in which each  $\widehat{F}_i$  is calculated by its corresponding empirical loss.

Specifically, by comparing (2) and (4), we observe that GDF only uses the first-order information of  $\widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}, \mathcal{D}_i)$ , leading to an inexact solution. In contrast, PDF focuses on optimizing  $\widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}, \mathcal{D}_i) = \left\langle \nabla \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}, \mathcal{D}_i), w_{\mathcal{T}_i} - w \right\rangle + \frac{1}{2} \left\langle \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}, \mathcal{D}_i), (w_{\mathcal{T}_i} - w)^{\otimes 2} \right\rangle +$

153  $\frac{1}{6} \left\langle \nabla^3 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i}, \mathcal{D}_i), (w_{\mathcal{T}_i} - w)^{\otimes 3} \right\rangle + \dots$ , and thus implicitly leveraging higher-order information.  
 154 This finding suggests that fewer inner-levels  $Q$  may be more suitable for GDF, while more inner-levels  
 155  $Q$  are better suited for PDF.

### 156 3.3 Stability and Generalization

157 Test error is generally considered the most critical metric for evaluating the performance of meta-  
 158 learning algorithms. To control it, most studies focus on two key perspectives: generalization error  
 159 and optimization error [33–35]. Specifically, let  $\mathcal{S} := \{S_i\}_{i=1}^m$  be the concatenation of all tasks data  
 160 sets. Given a randomized optimization algorithm  $\mathcal{A}$  that acts on the dataset  $\mathcal{S}$  and produces an output  
 161  $\mathcal{A}(\mathcal{S})$ , the generalization error is formally defined as  $\epsilon_{\text{gen}} := \mathbb{E}_{\mathcal{A}, \mathcal{S}}[F(\mathcal{A}(\mathcal{S})) - \hat{F}(\mathcal{A}(\mathcal{S}), \mathcal{S})]$  and  
 162 the optimization error is  $\epsilon_{\text{opt}} := \mathbb{E}_{\mathcal{A}, \mathcal{S}}[\hat{F}(\mathcal{A}(\mathcal{S}), \mathcal{S}) - \min_{\mathcal{W}} \hat{F}(\cdot, \mathcal{S})]$  [36, 37]. Then the test error of  
 163  $\mathcal{A}$  can be decomposed into three distinct terms:

$$\mathbb{E}_{\mathcal{A}, \mathcal{S}} \left[ F(\mathcal{A}(\mathcal{S})) - \min_{\mathcal{W}} F \right] = \epsilon_{\text{gen}} + \epsilon_{\text{opt}} + \underbrace{\mathbb{E}_{\mathcal{S}} \left[ \min_{\mathcal{W}} \hat{F}(\cdot, \mathcal{S}) \right] - \min_{\mathcal{W}} F}_{\leq 0}. \quad (6)$$

164 [28, 27] have shown that  $\epsilon_{\text{opt}}$  will converge to 0 as the number of outer iterations  $T$  increases, given  
 165 that the loss function  $\ell(w, z)$  satisfies certain assumptions, and the third term is non-positive. As  
 166 such, analyzing  $\epsilon_{\text{gen}}$  to improve the performance of the test error is more crucial. Note that [16, 22]  
 167 only provide  $\epsilon_{\text{gen}}$  of GDF in the strongly-convex setting. To fill the gap, in this paper, we establish  
 168 a comprehensive theoretical analysis of two meta-learning frameworks in convex and non-convex  
 169 settings. Before presenting our results, we state the following definition and assumptions widely used  
 170 in generalization analysis [37, 16, 28].

171 **Definition 1.** We say a function  $\ell(w)$  is  $\lambda$ -strongly convex if  $\forall w_1, w_2, \ell(w_1) \geq \ell(w_2) +$   
 172  $\langle \nabla \ell(w_2), w_1 - w_2 \rangle + \frac{\lambda}{2} \|w_1 - w_2\|^2$ . If  $\lambda = 0$ , then we say  $\ell(w)$  is convex.

173 **Assumption 1.** We assume  $\mathcal{Z}$  is a polish space (i.e., complete, separable, and metric) and  $\mathcal{F}_{\mathcal{Z}}$  is  
 174 the Borel  $\sigma$ -algebra over  $\mathcal{Z}$ . Moreover, for any  $i$ ,  $P_i$  is a non-atomic probability distribution over  
 175  $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$ , i.e.,  $P_i(z) = 0$  for every  $z \in \mathcal{Z}$ .

176 **Assumption 2.** For any  $z \in \mathcal{Z}$ , the function  $\ell(\cdot, z)$  is twice continuously differentiable. Furthermore,  
 177 we assume it satisfies the following properties for any  $w, u \in \mathbb{R}^d$ .

178 (i) The loss is  $G$ -Lipschitz over  $\mathbb{R}^d$ , i.e.,  $\|\ell(w, z) - \ell(u, z)\| \leq G\|w - u\|$ ;

179 (ii) The loss is  $L$ -smooth over  $\mathbb{R}^d$ , i.e.,  $\|\nabla \ell(w, z) - \nabla \ell(u, z)\| \leq L\|w - u\|$ ;

180 (iii) The second derivative is Lipschitz continuous with constant  $\rho$  over  $\mathbb{R}^d$ , i.e.,  $\|\nabla^2 \ell(w, z) -$   
 181  $\nabla^2 \ell(u, z)\| \leq \rho\|w - u\|$ ;

182 (iv) The third derivative is Lipschitz continuous with constant  $\kappa$  over  $\mathbb{R}^d$ , i.e.,  $\|\nabla^3 \ell(w, z) -$   
 183  $\nabla^3 \ell(u, z)\| \leq \kappa\|w - u\|$ .

184 **Assumption 3.** For any  $i \in [m]$  and any  $w \in \mathbb{R}^d$ , the stochastic gradients  $\nabla \hat{F}_i(w, S)$  have bounded  
 185 variance, i.e.,  $\mathbb{E}_S \|\nabla \hat{F}_i(w, S) - \nabla F(w)\|^2 \leq \sigma^2$ .

186 Stability-based generalization error analysis has been widely used to characterize the generalization  
 187 properties for optimization algorithms such as stochastic gradient [37], adversarial training [38],  
 188 and federated learning [39, 40]. Next, we will establish two different stability definitions and then  
 189 leverage them to prove the generalization bound for GDF and PDF, respectively.

## 190 4 Theoretical Analysis

191 Under stability analysis, [16] improves the previous generalization error bound in the homogeneous  
 192 case and provides a total variation-based analysis in the heterogeneous case. In addition, [22],  
 193 focusing on the inner-process, demonstrates that introducing inverted regularization at the inner-level  
 194 can enhance generalization performance. However, their analyses are limited to GDF with one  
 195 inner-level update ( $Q = 1$ ), which ignores the effect of multiple inner-levels ( $Q > 1$ ). To fully  
 196 characterize the effect of multiple inner-level updates, a different stability definition is required.

**Definition 2.** (on-average stability for GDF). A randomized algorithm  $\mathcal{A}$  with output  $w_S$  is called  $\epsilon$ -on-average stable if the following condition holds for any  $i \in [m]$ : Take the dataset  $\tilde{S}$  which is the same as  $S$ , except that  $\tilde{S}_{i,k}^{\text{tr}}$  and  $\tilde{S}_{i,j}^{\text{ts}}$  differ from  $S_i^{\text{tr}}$  and  $S_i^{\text{ts}}$  by replacing the  $k$ -th and  $j$ -th data points, respectively, where  $k \in [n^{\text{tr}}], j \in [n^{\text{ts}}]$ . Then, we have:

$$\max_{k \in [n^{\text{tr}}], j \in [n^{\text{ts}}]} \mathbb{E}_{S, \mathcal{A}, \tilde{S}_{i,k}^{\text{tr}}, \tilde{S}_{i,j}^{\text{ts}}} [\ell(w_{\mathcal{T}_i}(w_S, \tilde{S}_{i,k}^{\text{tr}}), \tilde{z}_{i,j}^{\text{ts}}) - \ell(w_{\mathcal{T}_i}(w_{\tilde{S}}, \tilde{S}_{i,k}^{\text{tr}}), \tilde{z}_{i,j}^{\text{ts}})] \leq \epsilon.$$

Different with the stability definition for stochastic gradient descent in [41, 42] which considers only a single-level structure, our definitions for meta-learning mean any perturbation of samples across levels cannot lead to a big change of the model trained by an algorithm in expectation. The main reason that we are interested in the stability of an algorithm is its connection with generalization error. Next, we formalize this connection for GDF and show that whether an Algorithm  $\mathcal{A}$  is  $\epsilon$ -on-average stable, then its generalization error is bounded above by  $\epsilon$ .

**Theorem 1.** Consider the population risk  $F(w)$  and empirical risk  $\hat{F}(w)$ . The corresponding  $F_i(w)$  and  $\hat{F}_i(w)$  are approached by GDF in Algorithm 1. Under Assumption 1 and Definition 2, if  $\mathcal{A}$  is a randomized GDF-based algorithm, then  $\epsilon_{\text{gen}} \leq \mathbb{E}_{\mathcal{A}, S} [F(w_S) - \hat{F}(w_S, S)] \leq \epsilon$ .

Building on the GDF analysis, we now extend this concept to the PDF setting, where the stability definition needs to account for differences in training procedures.

**Definition 3.** (on-average stability for PDF). A randomized algorithm  $\mathcal{A}$  with output  $w_S$  is called  $\epsilon$ -on-average stable if the following condition holds for any  $i \in [m]$ : Take the dataset  $\tilde{S}$  which is the same as  $S$ , except that  $\tilde{S}_{i,j}$  differ from  $S_i$  by replacing the  $j$ -th data point. Then, for any  $\tilde{z} \in \mathcal{Z}$ , we have:  $\max_{j \in [n]} \mathbb{E}_{S, \mathcal{A}, \tilde{S}_{i,j}} [\ell_\lambda(w_S, \tilde{z}_{i,j}) - \ell_\lambda(w_{\tilde{S}}, \tilde{z}_{i,j})] \leq \epsilon$ , where  $\ell_\lambda$  is defined in (5).

Under this definition, we can establish a connection between generalization and stability. Although similar to Theorem 1, i.e.,  $\epsilon_{\text{gen}} \leq \mathbb{E}_{\mathcal{A}, S} [F(w_S) - \hat{F}(w_S, S)] \leq \epsilon$ , the proof of this is significantly different from it. In addition,  $F(w)$  and  $\hat{F}(w)$  in PDF also differ from those in GDF, as composed of different  $F_i(w)$  and  $\hat{F}_i(w)$ , respectively.

Based on the above theorems, we will present the generalization bounds under convex and non-convex.

#### 4.1 Generalization bounds of GDF

**Theorem 2.** Let the outer-level step size and inner-level step size be chosen as  $\eta_t \leq \frac{1}{L_Q}$  and  $\alpha \leq \frac{1}{L}$ , respectively. Under Assumptions 1- 3, the generalization error  $\epsilon_{\text{gen}}$  of GDF can be bounded by:

$$\mathcal{O}\left(\sum_{t=0}^{T-1} \eta_t (1 + \alpha L)^{Q-1} \frac{(6QG + Q^2 \alpha^2 G^2 \rho)}{mn^{\text{tr}}} + \frac{1}{m} \sqrt{F(w^0) - \min_{\mathcal{W}} F + \frac{L_Q \sigma^2}{2} \sum_{t=0}^{T-1} \eta_t^2}\right),$$

where  $L_Q = \alpha \rho Q (1 + \alpha L)^{Q-1} + (1 + \alpha L)^Q L$ . If we further let  $\alpha \leq \frac{1}{L_Q}$  and  $\eta_t = \eta$  be fixed, we can obtain a concise result as follows:

$$\epsilon_{\text{gen}} \leq \mathcal{O}\left(\frac{TQ}{mn^{\text{tr}}} + \frac{1}{m} \sqrt{F(w^0) - \min_{\mathcal{W}} F + T}\right).$$

**Remark 1.** The first observation from the above results is that the generalization bound worsens as  $Q$  increases, primarily due to the first term,  $\mathcal{O}(\frac{TQ}{mn^{\text{tr}}})$ . Similarly, [15] provides a comparable bound,  $\mathcal{O}(\sqrt{T + TQ})$ , using the information-theoretic analysis. However, by revisiting (1) and (3), we observe that  $F(w)$  is influenced by  $w_{\mathcal{T}_i}^Q(w, \mathcal{D}_i)$ . Consequently, after  $Q$  steps of gradient descent, the divergence between  $F(w^0)$  and  $\min_{\mathcal{W}} F$  decreases, which in turn positively impacts the generalization bound. In summary, our findings indicate a trade-off in selecting the number of inner steps  $Q$ , which also explains the phenomenon in Figure 1a.

**Remark 2.** In previous work, [9] also provides a similar trade-off relationship,  $\mathcal{O}(\frac{(1+\alpha L)^Q}{n} + F(w^T) - \min_{\mathcal{W}} F)$ , with  $\alpha \leq \frac{1}{L}$ . However, under the choice  $\alpha \leq \frac{1}{QL}$ , the first term simplifies to  $\mathcal{O}(\frac{1}{n})$ , effectively mitigating the trade-off relationship. In contrast, our result demonstrates greater robustness to the choice of  $\alpha$ . Furthermore, unlike the generalization bound  $\mathcal{O}(\frac{1}{mn^{\text{tr}}})$  provided by [16], our generalization bound grows with  $T$  due to our consideration of a general convex setting. In contrast, their analysis assumes a strongly convex setting, which is consistent with [37].

**Theorem 3.** Let the outer-level step size be chosen as  $\eta_t = \frac{c}{t}$  satisfy  $c \leq \min\{\frac{1}{L_Q}, \frac{1}{4(2L_Q \ln(T))^2}\}$  and the inner-level step size be chosen as  $\alpha \leq \frac{1}{Q_L}$ . Under Assumptions 1-3, and further assume that  $\mathbb{E}_{S, \mathcal{A}}[F(w_S)] \leq F(w^0)$ , the generalization error  $\epsilon_{\text{gen}}$  of GDF can be bounded by:

$$\mathcal{O}\left(\left(\frac{1 + \frac{1}{c\gamma}}{m}\Phi_Q\right)^{\frac{1}{c\gamma}}(F(w^0)T)^{\frac{c\gamma}{1+c\gamma}}\right),$$

where  $\Phi_Q = 1 + \frac{Q}{n^{\text{tr}}}$ ,  $L_Q = \frac{3\rho(1+\alpha L)^{2(Q-1)}}{L} + (1 + \alpha L)^Q L$ ,  $\rho_Q = \frac{3\rho(1+\alpha)(1+\alpha L)^{2(Q-1)}}{L} + (1 + \alpha L)^{3Q}\rho + \alpha\kappa(1+\alpha L)^{2Q}$  and  $\gamma = \min\{L_Q, \mathbb{E}_S[\|\nabla^2 \hat{F}_i(w^0)\|] + \rho_Q(c\sigma + \sqrt{c(F(w^0) - \min_{\mathcal{W}} F)})\}$ .  
**Remark 3.** Note that we find that the trade-off relationship also exists in the non-convex setting, which is similar to Theorem 2. Specifically, on the one hand, we observe that  $\epsilon_{\text{gen}}$  increases with  $\Phi_Q = 1 + \frac{Q}{n^{\text{tr}}}$ . On the other hand,  $\epsilon_{\text{gen}}$  decreases with  $F(w^0)$ , where  $F(w^0)$  itself decrease with  $Q$ . In addition,  $\gamma$  also characterizes the effect of  $Q$  on  $\epsilon_{\text{gen}}$  through the term  $\rho_Q(c\sigma + \sqrt{c(F(w^0) - \min_{\mathcal{W}} F)})$ . Here,  $\rho_Q$  grows with  $Q$ , while  $F(w^0) - \min_{\mathcal{W}} F$  decreases with  $Q$ . Importantly, an inappropriate choice of the inner step size  $\alpha$  can exacerbate the generalization error. Even when the standard loss  $\ell(w, z)$  is  $L$ -smoothness and has  $\rho$ -Lipschitz hessian, the compositional loss  $\ell(w_{\mathcal{T}_i}^Q(w, S_i^{\text{tr}}), z)$  may become  $L_Q$ -smoothness and  $\rho_Q$ -Lipschitz, with  $L_Q$  and  $\rho_Q$  decreasing with  $Q$ . However, under our principled choice of  $\alpha \leq \frac{1}{Q_L}$ , we can ensure  $L_Q \leq \mathcal{O}(1)$  and  $\rho_Q \leq \mathcal{O}(1)$ , thereby mitigating these negative effects.

## 4.2 Generalization bounds of PDF

**Theorem 4.** Let the outer-level step size be chosen as  $\eta_t \leq \frac{1}{L_Q}$ , the inner-level step size be fixed and  $C > 1$  being a constant. Under Assumptions 2-3, the generalization error  $\epsilon_{\text{gen}}$  of PDF can be bounded by:

$$\mathcal{O}\left(\sum_{t=0}^{T-1} \frac{2\eta_t \lambda}{mC^Q} + \frac{1}{m}\sqrt{F(w^0) - \min_{\mathcal{W}} F} + \frac{L_Q \sigma^2}{2} \sum_{t=0}^{T-1} \eta_t^2\right),$$

where  $L_Q = \frac{\lambda L}{\lambda + L}$ . If we further let  $\eta_t = \eta \leq \frac{1}{L_Q}$  be a constant, we can obtain a more concise result as follows:

$$\epsilon_{\text{gen}} \leq \mathcal{O}\left(\frac{T}{mC^Q} + \frac{1}{m}\sqrt{F(w^0) - \min_{\mathcal{W}} F} + T\right).$$

**Remark 4.** An immediate conclusion from the first term,  $\frac{T}{mC^Q}$ , in the above result is that the generalization error in PDF decreases as the adaption steps  $Q$  increase. Additionally, the second term,  $\frac{1}{m}\sqrt{F(w^0) - \min_{\mathcal{W}} F} + T$ , also benefits from  $Q$  for similar reasons discussed in Remark 1. It is worth mentioning that, in a strong-convex setting, the generalization error does not grow with the number of iterations  $T$  [37]. However, the objective in (5) that we are minimizing is actually strongly convex w.r.t  $w_{\mathcal{T}_i}$  (due to the strongly-convex regularizer,  $\frac{\lambda}{2}\|w_{\mathcal{T}_i} - w\|^2$  and convex function  $\ell(\cdot, z)$ ), but only convex w.r.t  $w$  [43, 44, 35]. As a result, our generalization bound still grows with  $T$ . Furthermore, if we let  $Q = n^{\text{tr}}$ , the generalization error of GDF becomes  $\mathcal{O}(\frac{T}{m} + \frac{1}{m}\sqrt{F(w^0) - \min_{\mathcal{W}} F} + T)$ , which is greater than the generalization error of PDF,  $\mathcal{O}(\frac{T}{mC^{n^{\text{tr}}} + \frac{1}{m}\sqrt{F(w^0) - \min_{\mathcal{W}} F} + T)$ .

**Theorem 5.** Let the outer-level step size be chosen as  $\eta_t = \frac{c}{t}$  satisfy  $c \leq \min\{\frac{1}{L_Q}, \frac{1}{4(2L_Q \ln(T))^2}\}$ , inner-level step size be fixed and  $C > 1$  being a constant. Under Assumptions 2-3, and further assume that  $\mathbb{E}_{S, \mathcal{A}}[F(w_S)] \leq F(w^0)$ , the generalization error  $\epsilon_{\text{gen}}$  of PDF can be bounded by:

$$\mathcal{O}\left(\left(\frac{1 + \frac{1}{c\gamma}}{m}\Phi_Q\right)^{\frac{1}{c\gamma}}(F(w^0)T)^{\frac{c\gamma}{1+c\gamma}}\right),$$

where  $L_Q = \frac{\lambda L}{\lambda + L}$ ,  $\gamma = \min\{L_Q, \mathbb{E}_S[\|\nabla^2 \hat{F}_i(w^0)\|] + \rho_Q(c\sigma + \sqrt{c(F(w^0) - \min_{\mathcal{W}} F)})\}$ ,  $\Phi_Q = 1 + \frac{1}{C^Q}$  and  $\rho_Q = \rho$ .

**Remark 5.** Similarly, we get a lower generalization error since  $\Phi_Q = 1 + \frac{1}{C^Q}$  and  $F(w^0)$  gets smaller when  $Q$  increases. In addition, we observe that  $L_Q = \frac{\lambda L}{\lambda + L}$  and  $\rho_Q = \rho$  are independent of

Table 2: Summary of our results.

Frame.	Algorithm	Convex	Non-convex
GDF	MAML	$\mathcal{O}(\sum_{t=0}^{T-1} \eta_t (1 + \alpha L)^{Q-1} \frac{(6QG + Q^2 \alpha^2 G^2 \rho)}{mn^{\text{tr}}} + \frac{\mathcal{Q}(F(w^0))}{m})$	$\mathcal{O}(\frac{1+\frac{1}{\alpha}}{m} (1 + (1 + \alpha L)^{2(Q-1)} \frac{(6QG + Q^2 \alpha^2 G^2 \rho)}{n^{\text{tr}}}))^{\frac{1}{\alpha}} (F(w^0)T)^{\frac{\epsilon_{\text{gen}}}{1+\epsilon_{\text{gen}}}})$
	FOMAML	$\mathcal{O}(\sum_{t=0}^{T-1} \eta_t \frac{2Q\alpha LG}{mn^{\text{tr}}} + \frac{\mathcal{Q}(F(w^0))}{m})$	$\mathcal{O}(\frac{1+\frac{1}{\alpha}}{m} (1 + \frac{2Q(1+\alpha L)^2 \alpha LG}{n^{\text{tr}}}))^{\frac{1}{\alpha}} (F(w^0)T)^{\frac{\epsilon_{\text{gen}}}{1+\epsilon_{\text{gen}}}})$
	Meta-SGD	$\mathcal{O}(\sum_{t=0}^{T-1} \eta_t (1 + \hat{\alpha}_t L)^{Q-1} \frac{(6QG + Q^2 \hat{\alpha}_t^2 G^2 \rho)}{mn^{\text{tr}}} + \frac{\mathcal{Q}(F(w^0))}{m})$	$\mathcal{O}(\frac{1+\frac{1}{\alpha}}{m} (1 + (1 + \hat{\alpha} L)^{2(Q-1)} \frac{(6QG + Q^2 \hat{\alpha}^2 G^2 \rho)}{n^{\text{tr}}}))^{\frac{1}{\alpha}} (F(w^0)T)^{\frac{\epsilon_{\text{gen}}}{1+\epsilon_{\text{gen}}}})$
PDF	iMAML	$\mathcal{O}(\sum_{t=0}^{T-1} \frac{2L\eta_t(G^2+G)}{\lambda mn^{\text{tr}}} + \frac{\mathcal{Q}(F(w^0))}{m})$	$\mathcal{O}(\frac{1+\frac{1}{\alpha}}{m} (1 + \frac{2L(G^2+G)}{(\lambda-L)n^{\text{tr}}}))^{\frac{1}{\alpha}} (F(w^0)T)^{\frac{\epsilon_{\text{gen}}}{1+\epsilon_{\text{gen}}}})$
	Meta-MinibatchProx	$\mathcal{O}(\sum_{t=0}^{T-1} \frac{2\eta_t \lambda}{mC_Q} + \frac{\mathcal{Q}(F(w^0))}{m})$	$\mathcal{O}(\frac{1+\frac{1}{\alpha}}{m} (1 + \frac{2\lambda}{C_Q})^{\frac{1}{\alpha}} (F(w^0)T)^{\frac{\epsilon_{\text{gen}}}{1+\epsilon_{\text{gen}}}})$
	Fo-MuML	$\mathcal{O}(\sum_{t=0}^{T-1} \frac{2mG^2}{\lambda mn^{\text{tr}}} + \frac{\mathcal{Q}(F(w^0))}{m})$	$\mathcal{O}(\frac{1+\frac{1}{\alpha}}{m} (1 + \frac{2G^2}{(\lambda-L)n^{\text{tr}}})^{\frac{1}{\alpha}} (F(w^0)T)^{\frac{\epsilon_{\text{gen}}}{1+\epsilon_{\text{gen}}}})$

278  $Q$ , because PDF eliminates the compositional structure of  $\ell(w_{\mathcal{T}_i}(w, S_i^{\text{tr}}), z)$  in GDF. Furthermore, if  
 279 we set  $Q = n^{\text{tr}}$ , we can get  $\Phi_Q = 2$  in GDF, whereas  $\Phi_Q = 1 + \frac{1}{C_{n^{\text{tr}}}}$  in PDF, which indicate that  
 280 the generalization error of PDF is lower than GDF.

281 *Remark 6.* Based on Theorems 2-5, we present the generalization bounds for the corresponding algo-  
 282 rithms, summarized in Table 2. Here, the term  $\mathcal{Q}(F(w^0)) = \sqrt{F(w^0) - \min_{\mathcal{W}} F} + \frac{L_Q \sigma^2}{2} \sum_{t=0}^{T-1} \eta_t^2$   
 283 decreases as  $Q$  increases. Specifically, Meta-SGD redefines  $\alpha$  as a learnable parameter, denoted by  
 284  $\hat{\alpha}_t$ . As discussed in Remarks 2-3, the choice of the inner step size  $\alpha$  is critical, where  $\alpha \leq \frac{1}{QL}$  is  
 285 necessary to mitigate potential adverse effects caused by the inner process. In addition, our results  
 286 reveal that algorithms based on the GDF framework exhibit a trade-off between the inner process and  
 287 generalization error, as shown in Theorems 2-3. In contrast, the PDF framework avoids the adverse  
 288 effects of the adaptation process by implicitly leveraging higher-order information. It is important  
 289 to emphasize that our analysis focuses on the relationship between  $Q$  and  $\epsilon_{\text{gen}}$ . However, directly  
 290 comparing  $\epsilon_{\text{gen}}$  across algorithms remains challenging due to the influence of algorithm-specific  
 291 terms, such as  $L_Q$  and  $\rho_Q$ .

### 292 4.3 New Optimization Objective

293 Recalling from the Theorems 2-5 and the results in Table 2, we can observe that  $F(w^0)$  plays a  
 294 critical role in reducing the generalization bound, due to the terms,  $F(w^0) - \min_{\mathcal{W}} F$  and  $F(w^0)T$ .  
 295 To reduce the impact of  $F(w^0)$ , note that  $F(w^0) = \frac{1}{m} \sum_{i=1}^m F_i(w^0)$ , then let us consider a simple  
 296 relaxation technique: given a constant  $\beta \in [0, 1)$  and  $\hat{m} \leq \frac{m}{2}$  with the losses sorted by  $\{F_i(w^0)\}$   
 297 in *ascending* order, we can obtain that  $\sum_{i=1}^{\hat{m}} F_i(w^0)^{\psi(i)} \leq \sum_{i=\hat{m}+1}^m F_i(w^0)^{\psi(i)}$ , where  $\psi(i)$  is a  
 298 mapping function associating the index with the original loss  $F_i(w^0)$ . Based on this, we can derive  
 299  $F(w^0) = \frac{1}{m} \sum_{i=1}^m F_i(w^0)^{\psi(i)} \geq \frac{1+\beta}{m} \sum_{i=1}^{\hat{m}} F_i(w^0)^{\psi(i)} + \frac{1-\beta}{m} \sum_{i=\hat{m}+1}^m F_i(w^0)^{\psi(i)}$ . Motivated by  
 300 this result, we propose the following optimization objective:

$$F_{\text{new}}(w) = \frac{1+\beta}{\tilde{m}} \sum_{i=1}^m F_i(w)^{\psi(i)} + \frac{1-\beta}{\tilde{m}} \sum_{i=\hat{m}+1}^m F_i(w)^{\psi(i)}, \quad (7)$$

301 where  $\tilde{m} = (1+\beta)\hat{m} + (1-\beta)(m-\hat{m})$  is to ensure the lower bound derivation. A clear advantage  
 302 of our new optimization objective is its ability to achieve a smaller initial value,  $F_{\text{new}}(w^0)$ , thereby  
 303 reducing the generalization error. In addition, this new optimization objective is designed to easily  
 304 integrate with all types of algorithms in both the GDF and PDF frameworks without requiring  
 305 modification of their training procedures. During the training phase, we usually select a batch  
 306  $\mathcal{B}$  with size  $B$ . After performing the inner-process on each task, we evaluate their task losses  
 307  $\hat{\mathcal{L}}_i(w_{\mathcal{T}_i}^t), i \in [B]$ , at the outer-level and sort the losses in *ascending* order, yielding  $\{\hat{\mathcal{L}}_i(w_{\mathcal{T}_i}^t)\}^{\psi(i)}$ .  
 308 Then, based on the sorted losses, we then use a weighting factor  $\beta$  to adjust the weight of each  
 309 task when updating the meta-model  $w^t$ . Importantly, our strategy achieves a dynamical fairness for  
 310 each task without compromising convergence performance. The design most similar to our new  
 311 optimization goal is [45], however, their goal is to maximize the meta-loss as much as possible to  
 312 obtain task-robust meta-parameters, whereas our goal is to reduce the original meta-loss, and our  
 313 approach is easier to practice.

## 314 5 Experiment

315 **Few-shot regression.** This problem focuses on approximating a family of sine functions represented  
 316 as  $f(x) = a \sin(bx)$ . The task distribution, denoted as  $\mathcal{P}$ , corresponds to the joint distribution  $p(a, b)$ ,



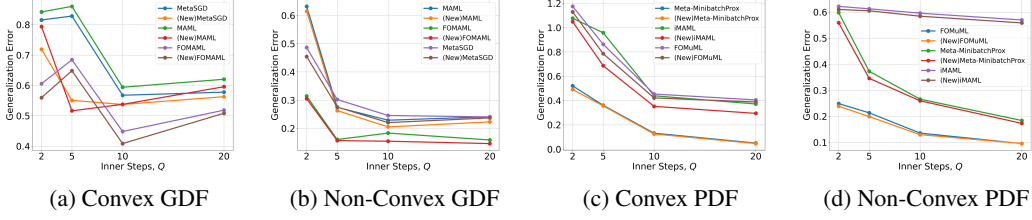


Figure 2: Relationship between generalization errors and  $Q$ .

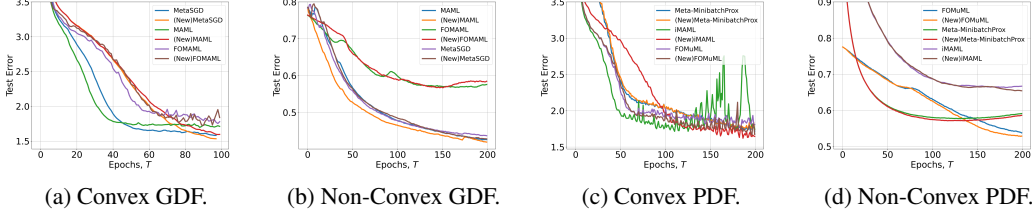


Figure 3: Convergence results with or without our new objective.

where  $a \sim U[0.1, 5]$  is the amplitude and  $b \sim U[0, \pi]$  is the phase. Both the training and test tasks are randomly generated from  $\mathcal{P} = p(a, b)$ . Following [7], we use an MLP network with Mean square error loss function. The generalization error is assessed as the difference between training and test errors. For each training task, we set the number of support samples and query samples to  $n^{\text{tr}} = 5$  and  $n^{\text{ts}} = 5$ , respectively, while for each test task, we use  $n^{\text{tr}} = 5, n^{\text{ts}} = 15$ .

**Few-shot classification.** We follow the standard experimental setup described in [21] using the real-world Omniglot dataset, which comprises 1,623 characters from 50 different alphabets, with each character having 20 instances drawn by different individuals. We employ a  $3 \times 3$  CNN to align with [5, 12], and use the Cross-Entropy Loss as the loss function. For each task, the number of support samples and query samples to  $n^{\text{tr}} = 1$  and  $n^{\text{ts}} = 5$ , respectively. For each test task, we use  $n^{\text{tr}} = 1, n^{\text{ts}} = 15$ . In addition, each task is formulated as a 5-way classification problem.

We fix the training task number  $m = 100$  and generate 10000 new tasks at test time by using the standard library [46]. To ensure a fair comparison, We report the average generalization error during the last 10 iterations of GDF and PDF under convex and non-convex settings. As shown in Figure 2, GDF’s generalization ability benefits considerably from smaller inner-levels  $Q$ , but deteriorates when  $Q$  becomes larger. In contrast, PDF’s generalization improves with increasing  $Q$ . These observations align with our findings in Table 2. Under our new optimization Objective, the generalization error of different algorithms can be effectively reduced. Furthermore, we also compare the convergence performance of GDF and PDF with or without our new objective within the same number of epochs. In particular, we set  $Q = 5$  for the convex setting and  $Q = 10$  for the non-convex setting. In Figure 3, the results indicate that our new optimization objective leads to lower test errors for most algorithms.

## 6 Conclusion

In this paper, we introduce two theoretical frameworks, GDF and PDF. They are summarized by several popular meta-learning algorithms based on inner processes. Within these frameworks, we derive generalization upper bounds for both convex and non-convex settings, which offer a detailed understanding of how the number of inner-levels influences the generalization error of GDF and PDF. Our analysis reveals a trade-off relationship induced by the inner-level in GDF, while PDF demonstrates a more favorable relationship that improves generalization. Building on these insights, we propose a novel meta-objective designed to significantly reduce generalization error. Extensive experiments validate the effectiveness of our findings and the proposed objective. We believe that the insights, the proof techniques, and the new meta-objective can inspire further research and open new directions in meta-learning and related areas.

## References

- [1] Bryony Hoskins and Ulf Fredriksson. *Learning to learn: What is it and can it be measured?* European Commission JRC, 2008.
- [2] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.
- [3] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR, 2018.
- [4] Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. Investigating meta-learning algorithms for low-resource natural language understanding tasks. *arXiv preprint arXiv:1908.10423*, 2019.
- [5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.
- [6] 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.
- [7] Pan Zhou, Xiaotong Yuan, Huan Xu, Shuicheng Yan, and Jiashi Feng. Efficient meta learning via minibatch proximal update. *NeurIPS*, 32, 2019.
- [8] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2017.
- [9] Pan Zhou, Yingtian Zou, Xiao-Tong Yuan, Jiashi Feng, Caiming Xiong, and Steven Hoi. Task similarity aware meta learning: Theory-inspired improvement on maml. In *Uncertainty in artificial intelligence*, pages 23–33. PMLR, 2021.
- [10] Thanh Nguyen, Tung Luu, Trung Pham, Sanzhar Rakhimkul, and Chang D Yoo. Robust maml: Prioritization task buffer with adaptive learning process for model-agnostic meta-learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3460–3464. IEEE, 2021.
- [11] Konstantin Mishchenko, Slavomir Hanzely, and Peter Richtárik. Convergence of first-order algorithms for meta-learning with moreau envelopes. *arXiv preprint arXiv:2301.06806*, 2023.
- [12] Jiaxin Chen, Xiao-Ming Wu, Yanke Li, Qimai Li, Li-Ming Zhan, and Fu-lai Chung. A closer look at the training strategy for modern meta-learning. *NeurIPS*, 33:396–406, 2020.
- [13] Jiechao Guan, Yong Liu, and Zhiwu Lu. Fine-grained analysis of stability and generalization for modern meta learning algorithms. *NeurIPS*, 35:18487–18500, 2022.
- [14] Yu Bai, Minshuo Chen, Pan Zhou, Tuo Zhao, Jason Lee, Sham Kakade, Huan Wang, and Caiming Xiong. How important is the train-validation split in meta-learning? In *International Conference on Machine Learning*, pages 543–553. PMLR, 2021.
- [15] Qi Chen, Changjian Shui, and Mario Marchand. Generalization bounds for meta-learning: An information-theoretic analysis. *NeurIPS*, 34:25878–25890, 2021.
- [16] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks. *NeurIPS*, 34:5469–5480, 2021.
- [17] Yunjuan Wang and Raman Arora. On the stability and generalization of meta-learning. In *NeurIPS*, 2024.
- [18] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.
- [19] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. In *International conference on learning representations*, 2018.

- [20] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *NeurIPS*, 32, 2019.
- [21] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.
- [22] Lianzhe Wang, Shiji Zhou, Shanghang Zhang, Xu Chu, Heng Chang, and Wenwu Zhu. Improving generalization of meta-learning with inverted regularization at inner-level. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7826–7835, 2023.
- [23] William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.
- [24] Andreas Maurer and Tommi Jaakkola. Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6(6), 2005.
- [25] Maruan Al-Shedivat, Liam Li, Eric Xing, and Ameet Talwalkar. On data efficiency of meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1369–1377. PMLR, 2021.
- [26] Kaiyi Ji, Jason D Lee, Yingbin Liang, and H Vincent Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. *NeurIPS*, 33:11490–11500, 2020.
- [27] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1082–1092. PMLR, 2020.
- [28] Kaiyi Ji, Junjie Yang, and Yingbin Liang. Theoretical convergence of multi-step model-agnostic meta-learning. *Journal of machine learning research*, 23(29):1–41, 2022.
- [29] Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning. *arXiv preprint arXiv:2110.07814*, 2021.
- [30] Budhaditya Deb, Guoqing Zheng, and Ahmed Hassan Awadallah. Boosting natural language generation from instructions with meta-learning. *arXiv preprint arXiv:2210.11617*, 2022.
- [31] Chengwei Qin, Qian Li, Ruochen Zhao, and Shafiq Joty. Learning to initialize: Can meta learning improve cross-task generalization in prompt tuning? *arXiv preprint arXiv:2302.08143*, 2023.
- [32] Sanchit Sinha, Yuguang Yue, Victor Soto, Mayank Kulkarni, Jianhua Lu, and Aidong Zhang. Maml-en-llm: Model agnostic meta-training of llms for improved in-context learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2711–2720, 2024.
- [33] Johannes O Royset. Stability and error analysis for optimization and generalized equations. *SIAM Journal on Optimization*, 30(1):752–780, 2020.
- [34] Yi Zhou, Yingbin Liang, and Huishuai Zhang. Understanding generalization error of sgd in nonconvex optimization. *Machine Learning*, pages 1–31, 2022.
- [35] Jiancong Xiao, Jiawei Zhang, Zhi-Quan Luo, and Asuman Ozdaglar. Uniformly stable algorithms for adversarial training and beyond. *arXiv preprint arXiv:2405.01817*, 2024.
- [36] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.
- [37] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.
- [38] Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. *NeurIPS*, 35:15446–15459, 2022.

- [39] Zhenyu Sun, Xiaochun Niu, and Ermin Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. In *International Conference on Artificial Intelligence and Statistics*, pages 676–684. PMLR, 2024.
- [40] Wenjun Ding, Ying An, Lixing Chen, Shichao Kan, Fan Wu, and Zhe Qu. How does the smoothness approximation method facilitate generalization for federated adversarial learning? *arXiv preprint arXiv:2412.08282*, 2024.
- [41] Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR, 2018.
- [42] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.
- [43] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.
- [44] Alexandros Georgogiannis. The generalization error of dictionary learning with moreau envelopes. In *International Conference on Machine Learning*, pages 1617–1625. PMLR, 2018.
- [45] Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Task-robust model-agnostic meta-learning. *Advances in Neural Information Processing Systems*, 33:18860–18871, 2020.
- [46] Tristan Deleu, Tobias Würfl, Mandana Samiei, Joseph Paul Cohen, and Yoshua Bengio. Torchmeta: A meta-learning library for pytorch. *arXiv preprint arXiv:1909.06576*, 2019.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: Our main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[No\]](#)

Justification: We didn’t discuss the limitations of our work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For each theorem and lemma in our paper, we have given the full set of assumptions and a complete and correct proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We fully disclose all the information needed to reproduce the main experimental results of this paper to the extent that it affects the main conclusions of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Due to some reasons, we don't provide open access to the code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: : We give the detail of experiment in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In the pictures of experiment result, we have depicted the range of variance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Our experiment has low requirements for configuration.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).



## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our paper conducted in the paper conform.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of our work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.



- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: We have listed the sources of our datasets in the reference list.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[NA\]](#)

Justification: Our paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

## 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: Our paper does not involve crowdsourcing nor research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Our paper does not use LLM.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## A Useful Lemmas

**Lemma 1.** Let  $\phi$  be a convex and  $L$ -smooth function. Then, for  $\eta \leq \frac{1}{L}$ , we have

$$\|(w - \eta \nabla \phi(w)) - (u - \eta \nabla \phi(u))\| \leq \|w - u\|,$$

for any  $w$  and  $u$ .

*Proof.* Since  $\phi$  is  $L$ -smooth and convex, we know that

$$\langle \nabla \phi(w) - \nabla \phi(u), w - u \rangle \geq \frac{1}{\eta} \|\nabla \phi(w) - \nabla \phi(u)\|^2.$$

Using this fact,

$$\begin{aligned} \|(w - \eta \nabla \phi(w)) - (u - \eta \nabla \phi(u))\|^2 &= \|w - u - \eta(\nabla \phi(w) - \nabla \phi(u))\|^2 \\ &= \|w - u\|^2 + \eta^2 \|\nabla \phi(w) - \nabla \phi(u)\|^2 - \eta \langle \nabla \phi(w) - \nabla \phi(u), w - u \rangle \\ &\leq \|w - u\|^2 + \eta(\eta - L^{-1}) \|\nabla \phi(w) - \nabla \phi(u)\|^2 \\ &\leq \|w - u\|^2 \end{aligned}$$

when  $\eta \leq \frac{1}{\gamma}$ . □

**Lemma 2.** [37] Let  $\phi$  be a  $\lambda$ -strongly convex and  $L$ -smooth function. Then, for any  $\delta \leq \frac{2}{\lambda+L}$ , we have

$$\|(u - \delta \nabla \phi(u)) - (v - \delta \nabla \phi(v))\| \leq (1 - \frac{L\delta\lambda}{\lambda+L})\|u - v\|$$

for any  $u$  and  $v$ .

**Lemma 3.** Let  $f_\lambda(w) = \min_u (f(u) + \frac{\lambda}{2}\|w - u\|^2)$ , if  $f$  is  $G$ -Lipschitz, then we have  $\|\nabla f_\lambda(w)\| \leq 2G$ .

*Proof.* Define  $\phi(u) \triangleq f(u) + \frac{\lambda}{2}\|w - u\|^2$  and  $u^* = \arg \min_u \phi(u)$ . Now, observe that

$$0 \leq \phi(w) - \phi(u^*) = f(w) - f(u^*) - \frac{\lambda}{2}\|w - u^*\|^2$$

Thus, we have

$$\frac{\lambda}{2}\|w - u^*\|^2 \leq f(w) - f(u^*) \leq L\|w - u^*\|$$

where the last inequality follows from the fact that  $f$  is  $G$ -Lipschitz. Thus, we get  $\|w - u^*\| \leq \frac{2G}{\lambda}$ . Since  $\|\nabla f_\lambda(w)\| = \lambda\|w - u^*\|$ . This together with the above bound gives the desired result. □

## B Stability and Generalization of MAML

### B.1 Proof of Theorem 1

To show the claim, it just suffices to show that for any  $i$ , we have

$$\mathbb{E}_{\mathcal{A}, \mathcal{S}} [F_i(w_{\mathcal{S}}) - \widehat{F}_i(w_{\mathcal{S}}, S_i)] \leq \epsilon.$$

Take the dataset  $\widetilde{\mathcal{S}}$  which is the same as  $\mathcal{S}$ , except that  $\widetilde{S}_{i,k}^{\text{tr}}$  and  $\widetilde{S}_{i,j}^{\text{ts}}$  differ from  $S_i^{\text{tr}}$  and  $S_i^{\text{ts}}$  in at most one data point, respectively. In particular,

$$\begin{aligned} S_i^{\text{tr}} &= \{z_{i,1}^{\text{tr}}, \dots, z_{i,n^{\text{tr}}}^{\text{tr}}\}, S_i^{\text{ts}} = \{z_{i,1}^{\text{ts}}, \dots, z_{i,n^{\text{ts}}}^{\text{ts}}\}, \\ \widetilde{S}_{i,k}^{\text{tr}} &= \{z_{i,1}^{\text{tr}}, \dots, \widetilde{z}_{i,k}^{\text{tr}}, \dots, z_{i,n^{\text{tr}}}^{\text{tr}}\}, \widetilde{S}_{i,j}^{\text{ts}} = \{z_{i,1}^{\text{ts}}, \dots, \widetilde{z}_{i,j}^{\text{ts}}, \dots, z_{i,n^{\text{ts}}}^{\text{ts}}\}. \end{aligned}$$

Under Assumption 1, we could assume  $\widetilde{z}_{i,j}^{\text{ts}}$  is different with  $\widetilde{z}_{i,k}^{\text{tr}}$ . Then, we relate empirical risk and population risk by

$$\begin{aligned} \mathbb{E}_{\mathcal{S}, \mathcal{A}} [\widehat{F}_i(w_{\mathcal{S}}, S_i)] &= \frac{1}{n^{\text{ts}}} \sum_{j=1}^{n^{\text{ts}}} \mathbb{E}_{\mathcal{S}, \mathcal{A}} [\ell(w_{\mathcal{T}_i}(w_{\mathcal{S}}, S_i^{\text{tr}}), z_{i,j}^{\text{ts}})]. \\ &= \frac{1}{n^{\text{ts}}} \sum_{j=1}^{n^{\text{ts}}} \mathbb{E}_{\mathcal{S}, \mathcal{A}, \widetilde{S}_{i,k}^{\text{tr}}, \widetilde{z}_{i,j}^{\text{ts}}} [\ell(w_{\mathcal{T}_i}(w_{\mathcal{S}}, \widetilde{S}_{i,k}^{\text{tr}}), \widetilde{z}_{i,j}^{\text{ts}})]. \end{aligned} \tag{8}$$

Moreover, we have

$$\mathbb{E}_{\mathcal{A}, \mathcal{S}} [F_i(w_{\mathcal{S}})] = \frac{1}{n^{\text{ts}}} \sum_{j=1}^{n^{\text{ts}}} \mathbb{E}_{\mathcal{S}, \mathcal{A}, \widetilde{S}_{i,k}^{\text{tr}}, \widetilde{z}_{i,j}^{\text{ts}}} [\ell(w_{\mathcal{T}_i}(w_{\mathcal{S}}, \widetilde{S}_{i,k}^{\text{tr}}), \widetilde{z}_{i,j}^{\text{ts}})]. \tag{9}$$

Putting (8) and (9) together, we have

$$\begin{aligned} \mathbb{E}_{\mathcal{A}, \mathcal{S}} [F(w_{\mathcal{S}}) - \widehat{F}(w_{\mathcal{S}}, S_i)] &\leq \frac{1}{m} \sum_{i=1}^m \frac{1}{n^{\text{ts}}} \sum_{j=1}^{n^{\text{ts}}} \mathbb{E}_{\mathcal{A}, \mathcal{S}, \widetilde{S}_{i,k}^{\text{tr}}, \widetilde{z}_{i,j}^{\text{ts}}} [\ell(w_{\mathcal{T}_i}(w_{\mathcal{S}}, \widetilde{S}_{i,k}^{\text{tr}}), \widetilde{z}_{i,j}^{\text{ts}}) - \ell(w_{\mathcal{T}_i}(w_{\mathcal{S}}, \widetilde{S}_{i,k}^{\text{tr}}), \widetilde{z}_{i,j}^{\text{ts}})] \\ &\leq \epsilon. \end{aligned}$$

Then we obtain the desired result.

## 806 B.2 Lemmas

807 In the following proofs, for simplicity, we use  $\widehat{\mathcal{L}}(w_{\mathcal{T}_i,q}) = \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q}, S_i^{\text{tr}})$ ,  $\ell(w_{\mathcal{T}_i,q}) = \ell(w_{\mathcal{T}_i,q}, z)$   
 808 without other explanation, where  $q \in [Q]$ .

809 **Lemma 4.** For any  $i \in [m]$ ,  $q = 0, \dots, Q-1$  and  $w, u \in \mathbb{R}^d$ , if  $\ell$  is a convex function, we have

$$\|w_{\mathcal{T}_i,q+1} - u_{\mathcal{T}_i,q+1}\| \leq \|w - u\|$$

810 *Proof.* Based on the updates that  $w_{\mathcal{T}_i,q+1} = w_{\mathcal{T}_i,q} - \alpha_q \nabla \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})$  and  $u_{\mathcal{T}_i,q+1} = u_{\mathcal{T}_i,q} -$   
 811  $\alpha_q \nabla \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})$ , we obtain

$$\begin{aligned} \|w_{\mathcal{T}_i,q+1} - u_{\mathcal{T}_i,q+1}\| &= \|(w_{\mathcal{T}_i,q} - \alpha_q \nabla \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) - (u_{\mathcal{T}_i,q} - \alpha_q \nabla \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q}))\| \\ &\leq \|w_{\mathcal{T}_i,q} - u_{\mathcal{T}_i,q}\| \end{aligned}$$

812 where we use the Lemma 1 in the first inequality. Unrolling it over  $q$  from 0 to  $q+1$ , we obtain

$$\|w_{\mathcal{T}_i,q+1} - u_{\mathcal{T}_i,q+1}\| \leq \|w - u\|$$

813 □

814 **Lemma 5.** Suppose the conditions in Assumption 2 are satisfied. Then, if  $\alpha \leq \frac{1}{L}$  and  $\ell$  is a convex  
 815 function, we have

$$\|\ell(w_{\mathcal{T}_i}(w, S_i^{\text{tr}}), z) - \ell(w_{\mathcal{T}_i}(u, S_i^{\text{tr}}), z)\| \leq G\|w - u\|$$

816 *Proof.* For any  $w, u \in \mathcal{W}$ , note that

$$\begin{aligned} \|\ell(w_{\mathcal{T}_i}(w, S_i^{\text{tr}}), z) - \ell(w_{\mathcal{T}_i}(u, S_i^{\text{tr}}), z)\| &\leq G\|w_{\mathcal{T}_i}(w, S_i^{\text{tr}}) - w_{\mathcal{T}_i}(u, S_i^{\text{tr}})\| \\ &= G\|(w_{\mathcal{T}_i,Q-1} - \alpha \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i,Q-1}, S_i^{\text{tr}})) - (u_{\mathcal{T}_i,Q-1} - \alpha \nabla \widehat{\mathcal{L}}(u_{\mathcal{T}_i,Q-1}, S_i^{\text{tr}}))\| \\ &\leq G\|w_{\mathcal{T}_i,Q-1} - u_{\mathcal{T}_i,Q-1}\| \\ &\leq G\|w - u\| \end{aligned}$$

817 where we use Lemma 1 in the second inequality and Lemma 4 in the last inequality. □

818 **Lemma 6.** For any  $i \in [m]$ ,  $q = 0, \dots, Q-1$  and  $w, u \in \mathbb{R}^d$ , we have

$$\|w_{\mathcal{T}_i,q+1} - u_{\mathcal{T}_i,q+1}\| \leq (1 + \alpha L)^{q+1} \|w - u\|$$

819 *Proof.* Based on the updates that  $w_{\mathcal{T}_i,q+1} = w_{\mathcal{T}_i,q} - \alpha_q \nabla \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})$  and  $u_{\mathcal{T}_i,q+1} = u_{\mathcal{T}_i,q} -$   
 820  $\alpha_q \nabla \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})$ , we obtain

$$\begin{aligned} \|w_{\mathcal{T}_i,q+1} - u_{\mathcal{T}_i,q+1}\| &= \|(w_{\mathcal{T}_i,q} - \alpha_q \nabla \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) - (u_{\mathcal{T}_i,q} - \alpha_q \nabla \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q}))\| \\ &\leq \|w_{\mathcal{T}_i,q} - u_{\mathcal{T}_i,q}\| + \alpha L \|w_{\mathcal{T}_i,q} - u_{\mathcal{T}_i,q}\| \\ &= (1 + \alpha L) \|w_{\mathcal{T}_i,q} - u_{\mathcal{T}_i,q}\| \end{aligned}$$

821 where we use the Assumption 2 in the first inequality. Unrolling it over  $q$  from 0 to  $q+1$ , we obtain

$$\|w_{\mathcal{T}_i,q+1} - u_{\mathcal{T}_i,q+1}\| \leq (1 + \alpha L)^{q+1} \|w - u\|$$

822 □

823 **Lemma 7.** Suppose the conditions in Assumption 2 are satisfied. Then, with  $\alpha \leq \frac{1}{QL}$  and  $\ell$  is a  
 824 non-convex function, we have

$$\|\ell(w_{\mathcal{T}_i}(w, S_i^{\text{tr}}), z) - \ell(w_{\mathcal{T}_i}(u, S_i^{\text{tr}}), z)\| \leq eG\|w - u\|$$

825 *Proof.* For any  $w, u \in \mathcal{W}$ , note that

$$\begin{aligned} \|\ell(w_{\mathcal{T}_i}(w, S_i^{\text{tr}}), z) - \ell(w_{\mathcal{T}_i}(u, S_i^{\text{tr}}), z)\| &\leq G\|w_{\mathcal{T}_i}(w, S_i^{\text{tr}}) - w_{\mathcal{T}_i}(u, S_i^{\text{tr}})\| \\ &= G\|[w_{\mathcal{T}_i,Q-1} - \alpha \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i,Q-1}, S_i^{\text{tr}})] - [u_{\mathcal{T}_i,Q-1} - \alpha \nabla \widehat{\mathcal{L}}(u_{\mathcal{T}_i,Q-1}, S_i^{\text{tr}})]\| \\ &= G\|w_{\mathcal{T}_i,Q-1} - u_{\mathcal{T}_i,Q-1}\| + \alpha \|\nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i,Q-1}, S_i^{\text{tr}}) - \nabla \widehat{\mathcal{L}}(u_{\mathcal{T}_i,Q-1}, S_i^{\text{tr}})\| \\ &\leq G(1 + \alpha L) \|w_{\mathcal{T}_i,Q-1} - u_{\mathcal{T}_i,Q-1}\| \\ &\leq G(1 + \alpha L)^Q \|w - u\| \\ &\leq eG\|w - u\| \end{aligned}$$

826 where we use Assumption 2 in the first and second inequality, Lemma 6 in the third inequality. If  
 827  $\alpha \leq \frac{1}{QL}$ , we have the last inequality.  $\square$

828 **Lemma 8.** For any  $i \in [m]$ ,  $q = 0, \dots, Q$  and  $w \in \mathbb{R}^d$ , we have

$$\|\nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q+1})\| \leq (1 + \alpha L)^q \|\nabla \hat{\mathcal{L}}_i(w)\|.$$

829 *Proof.* Recalling the update rule  $w_{\mathcal{T}_i, q+1} = w_{\mathcal{T}_i, q} - \alpha_q \nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})$ , Using Assumption 2, we have

$$\begin{aligned} \|\nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q+1})\| &= \|\nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q+1}) - \nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q}) + \nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})\| \\ &\leq \|\nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q+1}) - \nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})\| + \|\nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})\| \\ &\leq L \|w_{\mathcal{T}_i, q+1} - w_{\mathcal{T}_i, q}\| + \|\nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})\| \\ &= L \|w_{\mathcal{T}_i, q} - \alpha \nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q}) - w_{\mathcal{T}_i, q}\| + \|\nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})\| \leq (1 + \alpha L) \|\nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})\| \end{aligned}$$

830 where we use Assumption 2 in the second inequality. Then, unrolling the above inequality over  $q$   
 831 from 0 to  $q + 1$ , we get

$$\|\nabla \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q+1})\| \leq (1 + \alpha)^{q+1} \|\nabla \hat{\mathcal{L}}_i(w)\|$$

832  $\square$

833 **Lemma 9.** Suppose that Assumption 2 hold and the function  $F_i(w)$  defined in Eq (3). Then, for any  
 834  $w, u \in \mathbb{R}^d$ , if  $\ell$  is convex, we have

$$\|\nabla \hat{F}_i(w) - \nabla \hat{F}_i(u)\| \leq L_Q \|w - u\|$$

835 where  $L_Q = \alpha \rho Q(1 + \alpha L)^{Q-1} + (1 + \alpha L)^Q L$  with  $\alpha \leq \frac{1}{L}$ .

836 *Proof.* Similar to the proof of Lemma 9, we have

$$\begin{aligned} &\|\nabla \hat{F}_i(w) - \nabla \hat{F}_i(u)\| \\ &= \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})) \nabla \ell(w_{\mathcal{T}_i, Q}) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i, q})) \nabla \ell(u_{\mathcal{T}_i, Q}) \right\| \\ &\leq \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})) \nabla \ell(w_{\mathcal{T}_i, Q}) - \prod_{j=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i, q})) \nabla \ell(w_{\mathcal{T}_i, Q}) \right\| \\ &\quad + \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i, q})) \nabla \ell(w_{\mathcal{T}_i, Q}) - \prod_{j=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i, q})) \nabla \ell(u_{\mathcal{T}_i, Q}) \right\| \\ &\leq \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i, q})) \right\| \|\nabla \ell(w_{\mathcal{T}_i, Q})\| + \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i, q})) \right\| \|\nabla \ell(w_{\mathcal{T}_i, Q}) - \nabla \ell(u_{\mathcal{T}_i, Q})\| \\ &\leq \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i, q})) \right\| \|\nabla \ell(w_{\mathcal{T}_i, Q})\| + (1 + \alpha L)^Q \|\nabla \ell(w_{\mathcal{T}_i, Q}) - \nabla \ell(u_{\mathcal{T}_i, Q})\| \\ &\leq \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i, q})) \right\| \|\nabla \ell(w_{\mathcal{T}_i, Q})\| + (1 + \alpha L)^Q L \|w_{\mathcal{T}_i, Q} - u_{\mathcal{T}_i, Q}\| \\ &\leq \underbrace{\left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i, q})) \right\|}_{\Lambda(Q-1)} G + (1 + \alpha L)^Q L \|w - u\|, \end{aligned} \tag{10}$$

837 where in the last inequality, we use Assumption 2 and Lemma 4. We next upper-bound  $\Lambda$  in the above  
838 inequality. Specifically, we have

$$\begin{aligned}
\Lambda(Q-1) &\leq \left\| \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i,Q-1})) - \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i,Q-1})) \right\| \\
&\quad + \left\| \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i,Q-1})) - \prod_{q=0}^{Q-2} (I - \alpha^2 \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})) (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i,Q-1})) \right\| \\
&\leq \left\| \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) \right\| \left\| \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i,Q-1}) - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i,Q-1}) \right\| \\
&\quad + \left\| \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) - \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})) \right\| \left\| (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i,Q-1})) \right\| \\
&\leq (1 + \alpha L)^{Q-1} \left\| \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i,Q-1}) - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i,Q-1}) \right\| + (1 + \alpha L) \left\| \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) - \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})) \right\| \\
&\leq (1 + \alpha L)^{Q-1} \alpha \rho \|w_{\mathcal{T}_i,Q-1} - u_{\mathcal{T}_i,Q-1}\| + (1 + \alpha L) \Lambda(Q-2) \\
&\leq (1 + \alpha L)^{Q-1} \alpha \rho \|w - u\| + (1 + \alpha L) \Lambda(Q-2),
\end{aligned}$$

839 where we use Assumption 2 in the second inequality and Lemma 4 in the last inequality. Telescoping  
840 the above inequality over  $q$  from 1 to  $Q-1$  and noting  $\Lambda(0) \leq \alpha \rho \|w - u\|$ , we have

$$\begin{aligned}
\Lambda(Q-1) &\leq (1 + \alpha L)^{Q-1} \Lambda(0) + \alpha \rho (Q-1) (1 + \alpha L)^{Q-1} \|w - u\| \\
&\leq (1 + \alpha L)^{Q-1} \alpha \rho \|w - u\| + \alpha \rho (Q-1) (1 + \alpha L)^{Q-1} \|w - u\| \\
&\leq \alpha \rho Q (1 + \alpha L)^{Q-1} \|w - u\|
\end{aligned} \tag{11}$$

841 Substituting (11) into (10) and let  $L_Q = \alpha \rho Q (1 + \alpha L)^{Q-1} + (1 + \alpha L)^Q L$ , we get

$$\|\nabla F_i(w) - \nabla F_i(u)\| \leq L_Q \|w - u\|,$$

842 which completes the proof.  $\square$

843 **Lemma 10.** Suppose that Assumption 2 hold and the function  $F_i(w)$  defined in Eq (3). Then, for any  
844  $w, u \in \mathbb{R}^d$ , if  $\ell$  is non-convex, we have

$$\|\nabla \hat{F}_i(w) - \nabla \hat{F}_i(u)\| \leq L_Q \|w - u\|$$

845 where  $L_Q = \frac{3\rho(1+\alpha L)^{2(Q-1)}}{L} + (1 + \alpha L)^Q L$ .

846 *Proof.* Similar to the proof of Lemma 9, we have

$$\begin{aligned}
&\|\nabla \hat{F}_i(w) - \nabla \hat{F}_i(u)\| \\
&\leq \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})) \right\| \|\nabla \ell(w_{\mathcal{T}_i,Q})\| + (1 + \alpha L)^Q L \|w_{\mathcal{T}_i,Q} - u_{\mathcal{T}_i,Q}\| \\
&\leq \underbrace{\left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})) \right\|}_{\Lambda(Q-1)} G + (1 + \alpha L)^Q L \|w - u\|,
\end{aligned} \tag{12}$$

847 where in the last inequality, we use Assumption 2 and Lemma 4. We next upper-bound  $\Lambda$  in the above  
 848 inequality which is also similar to the proof of (11). Specifically, we have

$$\begin{aligned}
 \Lambda(Q-1) &\leq (1+\alpha L)^{Q-1} \Lambda(0) + \sum_{l=0}^{Q-2} \alpha \rho (1+\alpha L)^{2(Q-1-l)+l} \|w-u\| \\
 &\leq (1+\alpha L)^{Q-1} \alpha \rho \|w-u\| + \sum_{l=0}^{Q-2} \alpha \rho (1+\alpha L)^{2(Q-1-l)+l} \|w-u\| \\
 &\leq \left[ (1+\alpha L)^{Q-1} \alpha \rho + \alpha \rho (1+\alpha L)^Q \sum_{l=0}^{Q-2} (1+\alpha L)^l \right] \|w-u\| \\
 &\leq \left[ (1+\alpha L)^{Q-1} \alpha \rho + \frac{\rho}{L} (1+\alpha L)^Q ((1+\alpha L)^{Q-1} - 1) \right] \|w-u\| \\
 &\leq \frac{3\rho(1+\alpha L)^{2(Q-1)}}{L} \|w-u\|
 \end{aligned} \tag{13}$$

849 where we use  $\alpha \leq \frac{1}{L}$  in the last inequality. Substituting (13) into (12) and let  $L_Q = \frac{3\rho(1+\alpha L)^{2(Q-1)}}{L} +$   
 850  $(1+\alpha L)^Q L$ , we get

$$\|\nabla F_i(w) - \nabla F_i(u)\| \leq L_Q \|w-u\|,$$

851 which completes the proof.  $\square$

852 **Lemma 11.** Suppose that Assumption 2 hold and the function  $F_i(w)$  defined in Eq (3). Then, for any  
 853  $w, u \in \mathbb{R}^d$ , if  $\ell$  is convex, we have

$$\|\nabla^2 F_i(w) - \nabla^2 F_i(u)\| \leq \rho_Q \|w-u\|$$

854 where  $\rho_Q = (1+\alpha L)^{2Q-1} [\alpha \rho Q + 2\rho + Q\alpha \rho + G\alpha \kappa + G\alpha^2 \rho^2 Q]$

855 *Proof.* For simplicity, we denote  $J_q = I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i, q})$ .

$$\begin{aligned}
 &\|\nabla^2 \widehat{F}_i(w) - \nabla^2 \widehat{F}_i(u)\| \\
 &= \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i, q}))^2 \nabla^2 \ell(w_{\mathcal{T}_i, Q}) + \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial w} \left( \prod_{l=q+1}^{Q-1} J_l \right) \nabla \ell(w_{\mathcal{T}_i, Q}) \right. \\
 &\quad \left. - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i, q}))^2 \nabla^2 \ell(u_{\mathcal{T}_i, Q}) - \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial u} \left( \prod_{l=q+1}^{Q-1} J_l \right) \nabla \ell(u_{\mathcal{T}_i, Q}) \right\| \\
 &\leq \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i, q}))^2 \nabla^2 \ell(w_{\mathcal{T}_i, Q}) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i, q}))^2 \nabla^2 \ell(u_{\mathcal{T}_i, Q}) \right\| \\
 &\quad + \left\| \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial w} \left( \prod_{l=q+1}^{Q-1} J_l \right) \nabla \ell(w_{\mathcal{T}_i, Q}) - \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial u} \left( \prod_{l=q+1}^{Q-1} J_l \right) \nabla \ell(u_{\mathcal{T}_i, Q}) \right\|
 \end{aligned} \tag{14}$$

856 For the first term, we have

$$\begin{aligned}
& \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q}))^2 \nabla^2 \ell(w_{\mathcal{T}_i,Q}) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q}))^2 \nabla^2 \ell(u_{\mathcal{T}_i,Q}) \right\| \\
& \leq \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q}))^2 \nabla^2 \ell(w_{\mathcal{T}_i,Q}) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q}))^2 \nabla^2 \ell(w_{\mathcal{T}_i,Q}) \right\| \\
& \quad + \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q}))^2 \nabla^2 \ell(w_{\mathcal{T}_i,Q}) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q}))^2 \nabla^2 \ell(u_{\mathcal{T}_i,Q}) \right\| \\
& \leq \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q}))^2 - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q}))^2 \right\| \|\nabla^2 \ell(w_{\mathcal{T}_i,Q})\| + \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q}))^2 \right\| \|\nabla^2 \ell(w_{\mathcal{T}_i,Q}) - \nabla^2 \ell(u_{\mathcal{T}_i,Q})\| \\
& \leq \underbrace{\left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q}))^2 - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q}))^2 \right\|}_{\Lambda(Q-1)} L + (1 + \alpha L)^{2Q} \rho \|w - u\|
\end{aligned} \tag{15}$$

857 where we use Assumption 2 and Lemma 4 in the last inequality. We next upper-bound  $\Lambda(Q-1)$  in  
858 the above inequality. Specifically, we have

$$\begin{aligned}
\Lambda(Q-1) & \leq \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q}))^2 - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q}))(I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})) \right\| \\
& \quad + \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q}))(I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q}))^2 \right\| \\
& \leq \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) \right\| \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})) \right\| \\
& \quad + \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})) \right\| \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})) \right\| \\
& \leq (1 + \alpha L)^Q \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q})) \right\| \\
& \leq (1 + \alpha L)^Q \alpha \rho Q (1 + \alpha L)^{Q-1} \|w - u\|.
\end{aligned} \tag{16}$$

859 where we use (11) in the last inequality. Putting (16) into (15), we have

$$\begin{aligned}
& \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q}))^2 \nabla^2 \ell(w_{\mathcal{T}_i,Q}) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q}))^2 \nabla^2 \ell(u_{\mathcal{T}_i,Q}) \right\| \\
& \leq \left[ (1 + \alpha L)^Q \alpha \rho Q (1 + \alpha L)^{Q-1} + (1 + \alpha L)^{2Q} \rho \right] \|w - u\|
\end{aligned} \tag{17}$$



860 To bound the second term in (14), we have

$$\begin{aligned}
& \left\| \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial w} \left( \prod_{l=q+1}^{Q-1} J_l \right) \nabla \ell(w_{\mathcal{T}_i, Q}) - \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial u} \left( \prod_{l=q+1}^{Q-1} J_l \right) \nabla \ell(u_{\mathcal{T}_i, Q}) \right\| \\
& \leq \left\| \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial w} \left( \prod_{l=q+1}^{Q-1} J_l \right) \nabla \ell(w_{\mathcal{T}_i, Q}) - \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial w} \left( \prod_{l=q+1}^{Q-1} J_l \right) \nabla \ell(u_{\mathcal{T}_i, Q}) \right\| \\
& \quad + \left\| \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial w} \left( \prod_{l=q+1}^{Q-1} J_l \right) \nabla \ell(u_{\mathcal{T}_i, Q}) - \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial u} \left( \prod_{l=q+1}^{Q-1} J_l \right) \nabla \ell(u_{\mathcal{T}_i, Q}) \right\| \\
& \leq \sum_{q=0}^{Q-1} \left\| \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial w} \left( \prod_{l=q+1}^{Q-1} J_l \right) \right\| \|\nabla \ell(w_{\mathcal{T}_i, Q}) - \nabla \ell(u_{\mathcal{T}_i, Q})\| \\
& \quad + \left\| \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial w} \left( \prod_{l=q+1}^{Q-1} J_l \right) - \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial u} \left( \prod_{l=q+1}^{Q-1} J_l \right) \right\| \|\nabla \ell(u_{\mathcal{T}_i, Q})\| \\
& \leq \sum_{q=0}^{Q-1} \left\| \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial w} \left( \prod_{l=q+1}^{Q-1} J_l \right) \right\| \|w - u\| + \sum_{q=0}^{Q-1} \left\| \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial w} \left( \prod_{l=q+1}^{Q-1} J_l \right) - \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial u} \left( \prod_{l=q+1}^{Q-1} J_l \right) \right\| G \\
& \leq \sum_{q=0}^{Q-1} (1 + \alpha L)^{Q-1} \left\| \frac{\partial J_q}{\partial w} \right\| \|w - u\| + \sum_{q=0}^{Q-1} (1 + \alpha L)^{Q-1} G \left\| \frac{\partial J_q}{\partial w} - \frac{\partial J_q}{\partial u} \right\|.
\end{aligned} \tag{18}$$

861 Firstly, we have

$$\begin{aligned}
\left\| \frac{\partial J_q}{\partial w} \right\| &= \left\| \alpha \nabla^3 \widehat{\mathcal{L}}(w_{\mathcal{T}_i, q}) \prod_{l=0}^{q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i, q})) \right\| \\
&\leq (1 + \alpha L)^{q-1} \alpha \rho,
\end{aligned} \tag{19}$$

862 where we use Assumption 2. Secondly, we have

$$\begin{aligned}
\left\| \frac{\partial J_q}{\partial w} - \frac{\partial J_q}{\partial u} \right\| &= \left\| \alpha \nabla^3 \widehat{\mathcal{L}}(w_{\mathcal{T}_i, q}) \prod_{l=0}^{q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i, q})) - \alpha \nabla^3 \widehat{\mathcal{L}}(u_{\mathcal{T}_i, q}) \prod_{l=0}^{q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(u_{\mathcal{T}_i, q})) \right\| \\
&\leq \left\| \alpha \nabla^3 \widehat{\mathcal{L}}(w_{\mathcal{T}_i, q}) \prod_{l=0}^{q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i, q})) - \alpha \nabla^3 \widehat{\mathcal{L}}(u_{\mathcal{T}_i, q}) \prod_{l=0}^{q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i, q})) \right\| \\
&\quad + \left\| \alpha \nabla^3 \widehat{\mathcal{L}}(u_{\mathcal{T}_i, q}) \prod_{l=0}^{q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i, q})) - \alpha \nabla^3 \widehat{\mathcal{L}}(u_{\mathcal{T}_i, q}) \prod_{l=0}^{q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(u_{\mathcal{T}_i, q})) \right\| \\
&\leq \alpha (1 + \alpha L)^q \|\nabla^3 \widehat{\mathcal{L}}(w_{\mathcal{T}_i, q}) - \nabla^3 \widehat{\mathcal{L}}(u_{\mathcal{T}_i, q})\| + \alpha \rho \left\| \prod_{l=0}^{q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i, q})) - \prod_{l=0}^{q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(u_{\mathcal{T}_i, q})) \right\| \\
&\leq \alpha \kappa (1 + \alpha L)^q \|w - u\| + \alpha^2 \rho^2 q (1 + \alpha L)^{q-1} \|w - u\|
\end{aligned} \tag{20}$$

863 Putting (20) and (19) into (18), we have

$$\begin{aligned}
& \left\| \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial w} \left( \prod_{l=q+1}^{Q-1} J_l \right) \nabla \ell(w_{\mathcal{T}_i, Q}) - \sum_{q=0}^{Q-1} \left( \prod_{l=0}^{q-1} J_l \right) \frac{\partial J_q}{\partial u} \left( \prod_{l=q+1}^{Q-1} J_l \right) \nabla \ell(u_{\mathcal{T}_i, Q}) \right\| \\
& \leq Q \left[ \alpha \rho (1 + \alpha L)^{2(Q-2)} + G \alpha \kappa (1 + \alpha L)^{2Q-1} + G \alpha^2 \rho^2 Q (1 + \alpha L)^{2(Q-1)} \right] \|w - u\|.
\end{aligned} \tag{21}$$

864 Putting (17) and (21) into (14), we get

$$\begin{aligned}
\|\nabla^2 F_i(w) - \nabla^2 F_i(u)\| &\leq [\alpha\rho Q(1+\alpha L)^{2Q-1} + (1+\alpha L)^{2Q}\rho] \|w-u\| \\
&\quad + Q \left[ \alpha\rho(1+\alpha L)^{2(Q-2)} + G\alpha\kappa(1+\alpha L)^{2Q-1} + G\alpha^2\rho^2 Q(1+\alpha L)^{2(Q-1)} \right] \|w-u\| \\
&= (1+\alpha L)^{2Q-1} [\alpha\rho Q + 2\rho + Q\alpha\rho + G\alpha\kappa + G\alpha^2\rho^2 Q] \|w-u\| \\
&= \rho_Q \|w-u\|,
\end{aligned}$$

865 where we denote  $\rho_Q = (1+\alpha L)^{2Q-1} [\alpha\rho Q + 2\rho + Q\alpha\rho + G\alpha\kappa + G\alpha^2\rho^2 Q]$ .  $\square$

866 **Lemma 12.** Suppose that Assumption 2 hold and the function  $F_i(w)$  defined in Eq (3). Then, for any  
867  $w, u \in \mathbb{R}^d$ , if  $\ell$  is non-convex, we have

$$\|\nabla^2 F_i(w) - \nabla^2 F_i(u)\| \leq \rho_Q \|w-u\|$$

868 where  $\rho_Q = \left[ \frac{3\rho(1+\alpha L)^{2(Q-1)}}{L} + (1+\alpha L)^{3Q}\rho + \alpha\kappa(1+\alpha L)^{2Q} + \frac{3\alpha\rho^2(1+\alpha L)^{2(Q-1)}}{L} \right]$

869 *Proof.* Similar to the proof of (14), we have

$$\begin{aligned}
\|\nabla^2 F_i(w) - \nabla^2 F_i(u)\| &\leq \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \mathcal{L}_i(w_{i,q}))^2 \nabla^2 \ell(w_{i,Q}) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \mathcal{L}_i(u_{i,q}))^2 \nabla^2 \ell(u_{i,Q}) \right\| \\
&\quad + \left\| \sum_{q=0}^{Q-1} \left( \prod_{k=0}^{q-1} J_k \right) \frac{\partial J_q}{\partial w} \left( \prod_{k=q+1}^{Q-1} J_k \right) \nabla \ell(w_{i,Q}) - \sum_{q=0}^{Q-1} \left( \prod_{k=0}^{q-1} J_k \right) \frac{\partial J_q}{\partial u} \left( \prod_{k=q+1}^{Q-1} J_k \right) \nabla \ell(u_{i,Q}) \right\|
\end{aligned} \tag{22}$$

870 For the first term, we have

$$\begin{aligned}
&\left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \mathcal{L}_i(w_{i,q}))^2 \nabla^2 \ell(w_{i,Q}) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \mathcal{L}_i(u_{i,q}))^2 \nabla^2 \ell(u_{i,Q}) \right\| \\
&\leq \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \mathcal{L}_i(w_{i,q}))^2 - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \mathcal{L}_i(u_{i,q}))^2 \right\| \|\nabla^2 \ell(w_{i,Q})\| + \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \mathcal{L}_i(u_{i,q}))^2 \right\| \|\nabla^2 \ell(w_{i,Q}) - \nabla^2 \ell(u_{i,Q})\| \\
&\leq \underbrace{\left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \mathcal{L}_i(w_{i,q}))^2 - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \mathcal{L}_i(u_{i,q}))^2 \right\|}_{\Lambda(Q-1)} L + (1+\alpha L)^{3Q}\rho \|w-u\|
\end{aligned} \tag{23}$$

871 where we use Assumption 2 and Lemma 6 in the last inequality. We next upper-bound  $\Lambda(Q-1)$  in  
872 the above inequality. Similar to the proof of (16), we have

$$\Lambda(Q-1) \leq \frac{3\rho(1+\alpha L)^{2(Q-1)}}{L} \|w-u\| \tag{24}$$

873 Putting (24) into (23), we have

$$\begin{aligned}
&\left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q}))^2 \nabla^2 \ell(w_{\mathcal{T}_i,Q}) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}_i(u_{\mathcal{T}_i,q}))^2 \nabla^2 \ell(u_{\mathcal{T}_i,Q}) \right\| \\
&\leq \left[ \frac{3\rho(1+\alpha L)^{2(Q-1)}}{L} + (1+\alpha L)^{3Q}\rho \right] \|w-u\|
\end{aligned} \tag{25}$$

875 To bound the second term, similar to the proof of (18), we have

$$\begin{aligned}
& \left\| \sum_{q=0}^{Q-1} \left( \prod_{k=0}^{q-1} J_k \right) \frac{\partial J_q}{\partial w} \left( \prod_{k=q+1}^{Q-1} J_k \right) \nabla \ell(w_{i,Q}) - \sum_{q=0}^{Q-1} \left( \prod_{k=0}^{q-1} J_k \right) \frac{\partial J_q}{\partial u} \left( \prod_{k=q+1}^{Q-1} J_k \right) \nabla \ell(u_{i,Q}) \right\| \\
& \leq \left\| \sum_{q=0}^{Q-1} \left( \prod_{k=0}^{q-1} J_k \right) \frac{\partial J_q}{\partial w} \left( \prod_{k=q+1}^{Q-1} J_k \right) \right\| \|\nabla \ell(w_{i,Q}) - \nabla \ell(u_{i,Q})\| \\
& \quad + \left\| \sum_{q=0}^{Q-1} \left( \prod_{k=0}^{q-1} J_k \right) \frac{\partial J_q}{\partial w} \left( \prod_{k=q+1}^{Q-1} J_k \right) - \sum_{q=0}^{Q-1} \left( \prod_{k=0}^{q-1} J_k \right) \frac{\partial J_q}{\partial u} \left( \prod_{k=q+1}^{Q-1} J_k \right) \right\| \|\nabla \ell(u_{i,Q})\| \\
& \leq \sum_{q=0}^{Q-1} \left\| \left( \prod_{k=0}^{q-1} J_k \right) \frac{\partial J_q}{\partial w} \left( \prod_{k=q+1}^{Q-1} J_k \right) \right\| (1 + \alpha L)^Q \|w - u\| + \sum_{q=0}^{Q-1} \left\| \left( \prod_{k=0}^{q-1} J_k \right) \frac{\partial J_q}{\partial w} \left( \prod_{k=q+1}^{Q-1} J_k \right) - \left( \prod_{k=0}^{q-1} J_k \right) \frac{\partial J_q}{\partial u} \left( \prod_{k=q+1}^{Q-1} J_k \right) \right\| \\
& \leq \sum_{q=0}^{Q-1} (1 + \alpha L)^{2Q-1} L \left\| \frac{\partial J_q}{\partial w} \right\| \|w - u\| + \sum_{q=0}^{Q-1} (1 + \alpha L)^{Q-1} G \left\| \frac{\partial J_q}{\partial w} - \frac{\partial J_q}{\partial u} \right\|,
\end{aligned} \tag{26}$$

876 where we use Assumption 2 and Lemma 6 in the second inequality. First, we have

$$\begin{aligned}
\left\| \frac{\partial J_q}{\partial w} \right\| &= \left\| \alpha \nabla^3 L(w_{i,q}) \prod_{k=0}^{q-1} (I - \alpha \nabla^2 L_i(w_{i,q})) \right\| \\
&\leq (1 + \alpha L)^{q-1} \alpha \rho
\end{aligned} \tag{27}$$

877 Secondly, we have

$$\begin{aligned}
\left\| \frac{\partial J_q}{\partial w} - \frac{\partial J_q}{\partial u} \right\| &= \left\| \alpha \nabla^3 L(w_{i,q}) \prod_{k=0}^{q-1} (I - \alpha \nabla^2 L_i(w_{i,q})) - \alpha \nabla^3 L(u_{i,q}) \prod_{k=0}^{q-1} (I - \alpha \nabla^2 L_i(u_{i,q})) \right\| \\
&\leq \left\| \alpha \nabla^3 L(w_{i,q}) \prod_{k=0}^{q-1} (I - \alpha \nabla^2 L_i(w_{i,q})) - \alpha \nabla^3 L(u_{i,q}) \prod_{k=0}^{q-1} (I - \alpha \nabla^2 L_i(w_{i,q})) \right\| \\
&\quad + \left\| \alpha \nabla^3 L(u_{i,q}) \prod_{k=0}^{q-1} (I - \alpha \nabla^2 L_i(w_{i,q})) - \alpha \nabla^3 L(u_{i,q}) \prod_{k=0}^{q-1} (I - \alpha \nabla^2 L_i(u_{i,q})) \right\| \\
&\leq \alpha (1 + \alpha L)^q \|\nabla^3 L(w_{i,q}) - \nabla^3 L(u_{i,q})\| + \alpha \rho \left\| \prod_{k=0}^{q-1} (I - \alpha \nabla^2 L_i(w_{i,q})) - \prod_{k=0}^{q-1} (I - \alpha \nabla^2 L_i(u_{i,q})) \right\| \\
&\leq \alpha \kappa (1 + \alpha L)^{2q} \|w - u\| + \frac{3\alpha \rho^2 (1 + \alpha L)^{2(Q-1)}}{L} \|w - u\|.
\end{aligned} \tag{28}$$

878 Putting (20) and (19) into (18), we have

$$\begin{aligned}
& \left\| \sum_{q=0}^{Q-1} \left( \prod_{k=0}^{q-1} J_k \right) \frac{\partial J_q}{\partial w} \left( \prod_{k=q+1}^{Q-1} J_k \right) \nabla \ell(w_{i,Q}) - \sum_{q=0}^{Q-1} \left( \prod_{k=0}^{q-1} J_k \right) \frac{\partial J_q}{\partial u} \left( \prod_{k=q+1}^{Q-1} J_k \right) \nabla \ell(u_{i,Q}) \right\| \\
& \leq Q \left[ \alpha \rho (1 + \alpha L)^{3Q-2} + \frac{3G\alpha \rho^2 (1 + \alpha L)^{3(Q-1)}}{L} \right] \|w - u\|.
\end{aligned} \tag{29}$$

879 Putting (25) and (29) into (22), we get

$$\begin{aligned}
\|\nabla^2 F_i(w) - \nabla^2 F_i(u)\| &\leq \left[ \frac{3\rho(1 + \alpha L)^{2(Q-1)}}{L} + (1 + \alpha L)^{3Q} \rho \right] \|w - u\| + \alpha \kappa (1 + \alpha L)^{2Q} \|w - u\| + \frac{3\alpha \rho^2 (1 + \alpha L)^{2(Q-1)}}{L} \\
&= \rho_Q \|w - u\|,
\end{aligned}$$

880 where we denote  $\rho_Q = \left[ \frac{3\rho(1 + \alpha L)^{2(Q-1)}}{L} + (1 + \alpha L)^{3Q} \rho + \alpha \kappa (1 + \alpha L)^{2Q} \right]$ .  $\square$

As shown in Algorithm 1, in the outer-level, We are also performing SGD by considering  $S$  as a meta-sample (which is equivalent to  $z$ ), then we can obtain the following Lemmas.

**Lemma 13.** (Lemma 4 in [41]) Suppose that Assumption 2 and Assumption 3 hold and  $F$  is  $L_Q$ -smoothness. Then, for any  $w \in \mathcal{W}$ , we have

$$\mathbb{E}_S \left[ \sum_{t=0}^{T-1} \eta_t \|\nabla \widehat{F}(w^t, S_t)\| \right] \leq 2 \sqrt{\sum_{t=0}^{T-1} \eta_t} \sqrt{F(w^{t+1}) - \min_{\mathcal{W}} F + \frac{L_Q \sigma^2}{2} \sum_{t=0}^{T-1} \eta_t^2 + \sum_{t=0}^{T-1} \sigma \eta_t}.$$

**Lemma 14.** (Lemma 6 in [41]) Let  $G_t(w) := w - \eta_t \nabla F(w, S_t)$  and assume that the loss function  $F(\cdot, S_t)$  is  $L_Q$ -smooth and that its Hessian is  $\rho_Q$ -Lipschitz. Then,

$$\|G_t(w_{S,t}) - G_t(w_{\widetilde{S}^{(i)},t})\| \leq (1 + \eta_t \xi_t) \|w_{S,t} - w_{\widetilde{S}^{(i)},t}\|,$$

where  $\xi_t := \|\nabla^2 F(w_0, S_t)\| + \frac{\rho_Q}{2} \left\| \sum_{l=1}^{t-1} \beta_l \nabla F(w_S^l, S_l) \right\| + \frac{\rho_Q}{2} \left\| \sum_{l=1}^{t-1} \beta_l \nabla F(w_{\widetilde{S}}^l, S_l) \right\|$ . Furthermore, more, we have  $\mathbb{E}_{S,S}[\xi_t] = \mathbb{E}_{S,S}[\|\nabla^2 F(w_0, S_t)\|] + \rho_Q \mathbb{E}_{S,S} \left\| \sum_{k=1}^{t-1} \beta_k \nabla F(w_S^k, S_k) \right\|$ . Furthermore, for any  $t \in [T]$ ,

$$\begin{aligned} \mathbb{E}_{S,S}[\xi_t(S, S)] &\leq \mathbb{E}_{S,S}[\|\nabla^2 F(w_0, S_t)\|] \\ &\quad + 2\rho_Q \sqrt{(F(w_0) - \min_{\mathcal{W}} F)c(1 + \ln(T))} \\ &\quad + \sigma \rho_Q \left( \sqrt{2cL_Q} + c(1 + \ln(T)) \right). \end{aligned}$$

**Lemma 15.** (Lemma 5 in [41]) Suppose that Assumption 2 hold, then for every  $t_0 \in \{0, 1, 2, \dots, T\}$  we have that

$$\mathbb{E}_{S,\widetilde{z},\mathcal{A}}[\ell(w_{\mathcal{T}_i}(w_S^T, \widetilde{S}_i^{\text{tr}}), \widetilde{z}) - \ell(w_{\mathcal{T}_i}(\widetilde{w}_S^T, \widetilde{S}_i^{\text{tr}}), \widetilde{z})] \leq eG\mathbb{E}_{S,\widetilde{z},\mathcal{A}}[w_S^T - w_{\widetilde{S}}^T | w_S^{t_0} - w_{\widetilde{S}}^{t_0} = 0] + \mathbb{E}_{S,\mathcal{A}}[F(w_S)] \frac{t_0}{m}$$

### B.3 Proof of Theorem 2(convex)

Let's consider two parallel processes of generating iterates  $\{w^t\}$  and  $\{\widetilde{w}^t\}$  by using datasets  $S$  and  $\widetilde{S}$ , respectively. We use the tilde superscript to refer to the second process throughout the proof. By Lemma 9, we know  $\ell$  is  $L_Q$ -smooth and convex function. Hence, for a given time index  $t$ , with probability  $1 - \frac{1}{m}$ , the task  $\mathcal{T}_j$  is selected, where  $j \neq i$ . By using Lemma 1, we have

$$\begin{aligned} \|w^{t+1} - \widetilde{w}^{t+1}\| &= \|(w^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_j}(w^t, S_j^{\text{tr}}), S_j^{\text{ts}})) - (\widetilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_j}(\widetilde{w}^t, S_j^{\text{tr}}), S_j^{\text{ts}}))\| \\ &\leq \frac{1}{n^{\text{ts}}} \sum_{z^{\text{ts}} \in S_j^{\text{ts}}} \|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_j}(w^t, S_j^{\text{tr}}), z^{\text{ts}})) - (\widetilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_j}(\widetilde{w}^t, S_j^{\text{tr}}), z^{\text{ts}}))\| \\ &\leq \|w^t - \widetilde{w}^t\| \end{aligned} \tag{30}$$

Next, for a given time index  $t$ , with probability  $\frac{1}{m}$ , the task  $\mathcal{T}_i$  is selected. In this case, we have

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \widetilde{w}^{t+1}\| &= \mathbb{E}\|(w^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), S_i^{\text{ts}})) - (\widetilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}(\widetilde{w}^t, \widetilde{S}_i^{\text{tr}}), \widetilde{S}_i^{\text{ts}}))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), S_i^{\text{ts}})) - (\widetilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}(\widetilde{w}^t, \widetilde{S}_i^{\text{tr}}), \widetilde{S}_i^{\text{ts}}))\| \\ &\quad + \mathbb{E}\|\eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}(\widetilde{w}^t, \widetilde{S}_i^{\text{tr}}), \widetilde{S}_i^{\text{ts}}) - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}(\widetilde{w}^t, \widetilde{S}_i^{\text{tr}}), S_i^{\text{ts}})\| \\ &\leq \frac{1}{n^{\text{ts}}} \sum_{z^{\text{ts}} \in S_i^{\text{ts}}} \mathbb{E}\|(w^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\widetilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}(\widetilde{w}^t, \widetilde{S}_i^{\text{tr}}), z^{\text{ts}}))\| \\ &\quad + \eta_t \mathbb{E}\|\nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}(\widetilde{w}^t, \widetilde{S}_i^{\text{tr}}), \widetilde{S}_i^{\text{ts}}) - \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}(\widetilde{w}^t, \widetilde{S}_i^{\text{tr}}), S_i^{\text{ts}})\| \\ &\leq \frac{1}{n^{\text{ts}}} \sum_{z \in S_i^{\text{ts}}} \mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\widetilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\widetilde{w}^t, \widetilde{S}_i^{\text{tr}}), z^{\text{ts}}))\| \\ &\quad + 2\eta_t \mathbb{E}\|\nabla \widehat{F}_i(w)\|, \end{aligned} \tag{31}$$

896 where the last inequality follows that  $\tilde{S}_i^{\text{ts}}$  and  $S_i^{\text{ts}}$  are sampled from the same distribution, then  
 897  $\mathbb{E}\|\nabla\hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}}))\| = \mathbb{E}\|\nabla\hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, S_i^{\text{ts}}))\| = \mathbb{E}\|\nabla\hat{F}_i(w)\|$ . Note that

$$\begin{aligned} & \mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})))\| \\ & \leq \mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}, z^{\text{ts}})))\| \\ & \quad + \eta_t \mathbb{E}\|\nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}, z^{\text{ts}})) - \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}}))\|. \end{aligned} \quad (32)$$

898 Let us bound the two terms on the RHS of (32) separately. First, similar to how we derived (30), we  
 899 could bound the first term by

$$\mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}, z^{\text{ts}})))\| \leq \mathbb{E}\|w^t - \tilde{w}^t\|.$$

900 To bound the second term on the RHS of (32), we consider two parallel processes of generating  
 901 iterates  $\{\tilde{w}_{\mathcal{T}_i,q}^t\}$  and  $\{\tilde{w}_{\mathcal{T}_i,q}^{t,\prime}\}$  by using datasets  $S_i^{\text{tr}}$  and  $\tilde{S}_i^{\text{tr}}$ , respectively. Note that

$$\begin{aligned} & \mathbb{E}\|\nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}, z^{\text{ts}})) - \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}}))\| \\ & = \mathbb{E}\|\prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i,q}^t, S_i^{\text{tr}})) \nabla \ell(\tilde{w}_{\mathcal{T}_i,Q}^t, z^{\text{ts}}) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{i,q}^{t,\prime}, \tilde{S}_i^{\text{tr}})) \nabla \ell(\tilde{w}_{i,Q}^{t,\prime}, z^{\text{ts}})\| \\ & \leq \mathbb{E}\|\prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i,q}^t, S_i^{\text{tr}})) \nabla \ell(\tilde{w}_{\mathcal{T}_i,q}^t, z^{\text{ts}}) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i,q}^t, S_i^{\text{tr}})) \nabla \ell(\tilde{w}_{i,Q}^{t,\prime}, z^{\text{ts}})\| \\ & \quad + \mathbb{E}\|\prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i,q}^t, S_i^{\text{tr}})) \nabla \ell(\tilde{w}_{i,Q}^{t,\prime}, z^{\text{ts}}) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{i,q}^{t,\prime}, \tilde{S}_i^{\text{tr}})) \nabla \ell(\tilde{w}_{i,Q}^{t,\prime}, z^{\text{ts}})\| \\ & \leq \mathbb{E}\|\prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i,q}^t, S_i^{\text{tr}}))\| \|\nabla \ell(\tilde{w}^t, z^{\text{ts}}) - \nabla \ell(\tilde{w}_{i,Q}^{t,\prime}, z^{\text{ts}})\| \\ & \quad + \mathbb{E}\|\prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i,q}^t, S_i^{\text{tr}})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{i,q}^{t,\prime}, \tilde{S}_i^{\text{tr}}))\| \|\nabla \ell(\tilde{w}_{i,Q}^{t,\prime}, z^{\text{ts}})\| \\ & \leq (1 + \alpha L)^Q \mathbb{E}\|\nabla \ell(\tilde{w}_{\mathcal{T}_i,Q}^t, z^{\text{ts}}) - \nabla \ell(\tilde{w}_{i,Q}^{t,\prime}, z^{\text{ts}})\| + \underbrace{G \mathbb{E}\|\prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i,q}^t, S_i^{\text{tr}})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{i,q}^{t,\prime}, \tilde{S}_i^{\text{tr}}))\|}_{V(Q)} \end{aligned} \quad (33)$$

902 where we use Assumption 2 in the last inequality. Hence, what remains is to bound the two terms in  
 903 (33). To do so, notice that

$$\begin{aligned} & \mathbb{E}\|\nabla \ell(\tilde{w}_{\mathcal{T}_i,Q}^t, z^{\text{ts}}) - \nabla \ell(\tilde{w}_{i,Q}^{t,\prime}, z^{\text{ts}})\| \leq L \mathbb{E}\|\tilde{w}_{\mathcal{T}_i,Q}^t - \tilde{w}_{i,Q}^{t,\prime}\| \\ & = L \mathbb{E}\|[\tilde{w}_{i,Q-1}^t - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i,Q-1}^t, S_i^{\text{tr}})] - [\tilde{w}_{i,Q-1}^{t,\prime} - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i,Q-1}^{t,\prime}, \tilde{S}_i^{\text{tr}})]\| \\ & \leq L \mathbb{E}\|[\tilde{w}_{i,Q-1}^t - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i,Q-1}^t, S_i^{\text{tr}})] - [\tilde{w}_{i,Q-1}^{t,\prime} - \alpha \nabla \ell(\tilde{w}_{i,Q-1}^{t,\prime}, S_i^{\text{tr}})]\| \\ & \quad + L \mathbb{E}\|\alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i,Q-1}^{t,\prime}, S_i^{\text{tr}}) - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i,Q-1}^{t,\prime}, \tilde{S}_i^{\text{tr}})\| \\ & \leq L \mathbb{E}\|\tilde{w}_{i,Q-1}^t - \tilde{w}_{i,Q-1}^{t,\prime}\| + \frac{2\alpha LG}{n^{\text{tr}}} \\ & \leq L \mathbb{E}\|[\tilde{w}^t - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}^t, S_i^{\text{tr}})] - [\tilde{w}^t - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}^t, \tilde{S}_i^{\text{tr}})]\| + \frac{2(Q-1)\alpha LG}{n^{\text{tr}}} \\ & = \alpha L \mathbb{E}\|\nabla \hat{\mathcal{L}}(\tilde{w}^t, S_i^{\text{tr}}) - \nabla \hat{\mathcal{L}}(\tilde{w}^t, \tilde{S}_i^{\text{tr}})\| + \frac{2(Q-1)\alpha LG}{n^{\text{tr}}} \\ & \leq \frac{2Q\alpha LG}{n^{\text{tr}}}, \end{aligned} \quad (34)$$

904 where we use Lemma 1 and Assumption 2 in the third inequality. Next,

$$\begin{aligned}
V(Q) &= \left\| \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i, q}^t, S_i^{\text{tr}})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, q}^{t, \prime}, \tilde{S}_i^{\text{tr}})) \right\| \\
&= \left\| \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i, q}^t, S_i^{\text{tr}})) (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}})) - \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, q}^{t, \prime}, \tilde{S}_i^{\text{tr}})) (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}})) \right\| \\
&\leq \left\| \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i, q}^t, S_i^{\text{tr}})) (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}})) - \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i, q}^t, S_i^{\text{tr}})) (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}})) \right\| \\
&\quad + \left\| \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i, q}^t, S_i^{\text{tr}})) (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}})) - \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, q}^{t, \prime}, \tilde{S}_i^{\text{tr}})) (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}})) \right\| \\
&= \left\| \prod_{q=0}^{Q-2} (I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i, q}^t, S_i^{\text{tr}})) \right\| \left\| \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}}) - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}}) \right\| + V(Q-1) \left\| I - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}}) \right\| \\
&\leq (1 + \alpha L)^{Q-1} \left\| \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}}) - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}}) \right\| + (1 + \alpha L) V(Q-1)
\end{aligned}$$

905 where we use Assumption 2 in the last inequality. To bound the above inequality, we first consider  
906 the first term.

$$\begin{aligned}
&(1 + \alpha L)^{Q-1} \left\| \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}}) - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}}) \right\| \\
&\leq \alpha (1 + \alpha L)^{Q-1} \left[ \left\| \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}}) - \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, S_i^{\text{tr}}) \right\| + \left\| \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, S_i^{\text{tr}}) - \nabla^2 \widehat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}}) \right\| \right] \\
&\leq \alpha (1 + \alpha L)^{Q-1} \left[ \rho \left\| \tilde{w}_{i, Q-1}^t - \tilde{w}_{i, Q-1}^{t, \prime} \right\| + \frac{2L}{n^{\text{tr}}} \right] \\
&= \alpha (1 + \alpha L)^{Q-1} \left[ \rho \left\| (\tilde{w}_{i, Q-2}^t - \alpha \nabla \widehat{\mathcal{L}}(\tilde{w}_{i, Q-2}^t, S_i^{\text{tr}})) - (\tilde{w}_{i, Q-2}^{t, \prime} - \alpha \nabla \widehat{\mathcal{L}}(\tilde{w}_{i, Q-2}^{t, \prime}, \tilde{S}_i^{\text{tr}})) \right\| + \frac{2L}{n^{\text{tr}}} \right] \\
&\leq \alpha (1 + \alpha L)^{Q-1} \left[ \rho \left\| \tilde{w}_{i, Q-2}^t - \tilde{w}_{i, Q-2}^{t, \prime} \right\| + \frac{2(L + G\alpha\rho)}{n^{\text{tr}}} \right] \\
&\leq \alpha (1 + \alpha L)^{Q-1} \left[ \rho \left\| \left[ \tilde{w}^t - \alpha \nabla \widehat{\mathcal{L}}(\tilde{w}^t, S_i^{\text{tr}}) \right] - \left[ \tilde{w}^t - \alpha \nabla \widehat{\mathcal{L}}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}) \right] \right\| + \frac{2(L + G\alpha\rho(Q-2))}{n^{\text{tr}}} \right] \\
&\leq \alpha (1 + \alpha L)^{Q-1} \left[ \alpha \rho \left\| \nabla \widehat{\mathcal{L}}(\tilde{w}^t, S_i^{\text{tr}}) - \nabla \widehat{\mathcal{L}}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}) \right\| + \frac{2(L + G\alpha\rho(Q-2))}{n^{\text{tr}}} \right] \\
&\leq \alpha (1 + \alpha L)^{Q-1} \frac{2(L + G\alpha\rho(Q-1))}{n^{\text{tr}}},
\end{aligned}$$

907 where we use Assumption 2 to derive the inequalities from the second to the last step. Putting it in  
908 to  $V(Q)$  and Unrolling it, noting that  $V(1) = \left\| \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}^t, S_i^{\text{tr}}) - \alpha \nabla^2 \widehat{\mathcal{L}}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}) \right\| \leq \frac{2\alpha L}{n^{\text{tr}}}$ , then we  
909 have

$$\begin{aligned}
V(Q) &\leq (1 + \alpha L)^{Q-1} V(1) + \sum_{k=1}^{Q-1} (1 + \alpha L)^{Q-1} \frac{2\alpha(L + G\alpha\rho(Q-k))}{n^{\text{tr}}} \\
&\leq \frac{2\alpha L(1 + \alpha L)^{Q-1}}{n^{\text{tr}}} + \sum_{k=1}^{Q-1} (1 + \alpha L)^{Q-1} \frac{2\alpha(L + G\alpha\rho k)}{n^{\text{tr}}} \tag{35} \\
&= \sum_{k=0}^{Q-1} (1 + \alpha L)^{Q-1} \frac{2\alpha(L + G\alpha\rho k)}{n^{\text{tr}}}.
\end{aligned}$$

By plugging (34) and (35) into (33), then we have

$$\begin{aligned} \mathbb{E}\|\nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}) - \nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}})\| &\leq (1 + \alpha L)^{Q-1} \left[ \frac{2(1 + \alpha L)Q\alpha LG}{n^{\text{tr}}} + \sum_{k=0}^{Q-1} \frac{2\alpha G(L + G\alpha\rho k)}{n^{\text{tr}}} \right] \\ &\leq (1 + \alpha L)^{Q-1} \frac{(6QG + Q^2\alpha^2 G^2\rho)}{n^{\text{tr}}}. \end{aligned} \quad (36)$$

where we using  $\alpha L \leq 1$  in the second inequality. Substituting (36) and (32) into (31), we have

$$\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \leq \mathbb{E}\|w^t - \tilde{w}^t\| + \eta_t(1 + \alpha L)^{Q-1} \frac{(6QG + Q^2\alpha^2 G^2\rho)}{n^{\text{tr}}} + 2\eta_t \mathbb{E}\|\nabla\hat{F}_i(w^t)\|.$$

Combing the above two cases, we obtain

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &\leq (1 - \frac{1}{m})\mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m}\mathbb{E}\|w^t - \tilde{w}^t\| \\ &\quad + \frac{1}{m}\eta_t(1 + \alpha L)^{Q-1} \frac{(6QG + Q^2\alpha^2 G^2\rho)}{n^{\text{tr}}} + \frac{2}{m}\eta_t \mathbb{E}\|\nabla\hat{F}_i(w^t)\| \\ &= \mathbb{E}\|w^t - \tilde{w}^t\| + \eta_t(1 + \alpha L)^{Q-1} \frac{(6QG + Q^2\alpha^2 G^2\rho)}{mn^{\text{tr}}} + \frac{2\eta_t}{m} \mathbb{E}\|\nabla\hat{F}_i(w^t)\|. \end{aligned}$$

Unrolling it and noting that  $\|w^0 - \tilde{w}^0\| = 0$ , we have

$$\begin{aligned} \mathbb{E}[\|w^T - \tilde{w}^T\|] &\leq \sum_{t=0}^{T-1} \eta_t(1 + \alpha L)^{Q-1} \frac{(6QG + Q^2\alpha^2 G^2\rho)}{mn^{\text{tr}}} + \sum_{t=0}^{T-1} \frac{2\eta_t}{m} \mathbb{E}\|\nabla\hat{F}(w^t, S_i)\| \\ &\leq \sum_{t=0}^{T-1} \eta_t(1 + \alpha L)^{Q-1} \frac{(6QG + Q^2\alpha^2 G^2\rho)}{mn^{\text{tr}}} + \frac{1}{m} \sqrt{F(w^0) - \min_{\mathcal{W}} F + \frac{LQ\sigma^2}{2} \sum_{t=0}^{T-1} \eta_t^2}, \end{aligned}$$

where we use Lemma 13 in the last inequality. Now we are ready to conclude. For any  $i \in [m]$ , we have

$$\begin{aligned} \epsilon_{gen} &\leq \mathbb{E}\|\ell(w_{\mathcal{T}_i}(w^T, \tilde{S}_i^{\text{tr}}), \tilde{z}) - \ell(w_{\mathcal{T}_i}(\tilde{w}^T, \tilde{S}_i^{\text{tr}}), \tilde{z})\| \\ &\leq G\mathbb{E}\|w^T - \tilde{w}^T\| \end{aligned}$$

which completes the proof.

#### B.4 Proof of Theorem 3

In this section, we establish stability results that do not rely on convexity, and we consider two cases. For the first case, using Lemma 10 and Lemma 14, we have

$$\begin{aligned} \|w^{t+1} - \tilde{w}^{t+1}\| &= \|(w^t - \eta_t \nabla\hat{\mathcal{L}}(w_{\mathcal{T}_j}(w^t, S_j^{\text{tr}}), S_j^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla\hat{\mathcal{L}}(w_{\mathcal{T}_j}(\tilde{w}^t, S_j^{\text{tr}}), S_j^{\text{ts}}))\| \\ &\leq (1 + \eta_t \phi_t) \|w^t - \tilde{w}^t\|, \end{aligned} \quad (37)$$

where  $\phi_t = \min\{L_Q, \xi_t\}$  with  $\xi_t = \|\nabla^2 F(w_0, S_t)\| + \frac{\rho_Q}{2} \left\| \sum_{l=1}^{t-1} \beta_l \nabla\hat{F}(w_{\mathcal{S}}^l) \right\| + \frac{\rho_Q}{2} \left\| \sum_{l=1}^{t-1} \beta_l \nabla\hat{F}(w_{\tilde{\mathcal{S}}}^l) \right\|$ ,  $\rho_Q = \frac{3\rho(1+\alpha L)^{2(Q-1)}}{L} + (1 + \alpha L)^{3Q}\rho + \alpha\kappa(1 + \alpha L)^{2Q} + \frac{3\alpha\rho^2(1+\alpha L)^{2(Q-1)}}{L}$ ,  $L_Q = \frac{3\rho(1+\alpha L)^{2(Q-1)}}{L} + (1 + \alpha L)^Q L$ .

Next, for the second case, similar to the proof of (31), we have

$$\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \leq \frac{1}{n^{\text{ts}}} \sum_{z \in S_i^{\text{ts}}} \mathbb{E}\|(w^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}}))\| + 2\eta_t \mathbb{E}\|\nabla\hat{F}_i(w)\|. \quad (38)$$

Note that

$$\begin{aligned} &\mathbb{E}\|(w^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}}))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}))\| + \eta_t \mathbb{E}\|\nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}) - \nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}})\| \end{aligned} \quad (39)$$

Let us bound the two terms on the RHS of (39), separately. First, similar to how we bound (37), we could bound the first term by

$$\mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}))\| \leq (1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\|.$$

To bound the second term on the RHS of (39), similar to the proof of (33), we have

$$\begin{aligned} & \mathbb{E}\|\nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}) - \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}})\| \\ & \leq (1 + \alpha L)^Q \mathbb{E}\|\nabla \ell(\tilde{w}_{\mathcal{T}_i, Q}^t, z^{\text{ts}}) - \nabla \ell(\tilde{w}_{i, Q}^{t, \prime}, z^{\text{ts}})\| + G \underbrace{\mathbb{E}\left\|\prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{\mathcal{T}_i, Q}^t, S_i^{\text{tr}})) - \prod_{q=0}^{Q-1} (I - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{i, q}^{t, \prime}, \tilde{S}_i^{\text{tr}}))\right\|}_{V(Q)} \end{aligned} \quad (40)$$

where we use Assumption 2 in the last inequality. Hence, what remains is to bound the two terms in (40). To do so, notice that

$$\begin{aligned} \mathbb{E}\|\nabla \ell(\tilde{w}_{\mathcal{T}_i, Q}^t, z^{\text{ts}}) - \nabla \ell(\tilde{w}_{i, Q}^{t, \prime}, z^{\text{ts}})\| & \leq L \mathbb{E}\|\tilde{w}_{\mathcal{T}_i, Q}^t - \tilde{w}_{i, Q}^{t, \prime}\| \\ & = L \mathbb{E}\left\|\left[\tilde{w}_{i, Q-1}^t - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}})\right] - \left[\tilde{w}_{i, Q-1}^{t, \prime} - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}})\right]\right\| \\ & \leq L \mathbb{E}\left\|\left[\tilde{w}_{i, Q-1}^t - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}})\right] - \left[\tilde{w}_{i, Q-1}^{t, \prime} - \alpha \nabla \ell(\tilde{w}_{i, Q-1}^{t, \prime}, S_i^{\text{tr}})\right]\right\| \\ & \quad + L \mathbb{E}\|\alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, S_i^{\text{tr}}) - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}})\| \\ & \leq L(1 + \alpha L) \mathbb{E}\|\tilde{w}_{i, Q-1}^t - \tilde{w}_{i, Q-1}^{t, \prime}\| + \frac{2\alpha LG}{n^{\text{tr}}} \\ & \leq L(1 + \alpha L)^Q \mathbb{E}\left\|\left[\tilde{w}^t - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}^t, S_i^{\text{tr}})\right] - \left[\tilde{w}^t - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}^t, \tilde{S}_i^{\text{tr}})\right]\right\| + \frac{2(Q-1)(1 + \alpha L)}{n^{\text{tr}}} \\ & = \alpha L(1 + \alpha L)^Q \mathbb{E}\|\nabla \hat{\mathcal{L}}(\tilde{w}^t, S_i^{\text{tr}}) - \nabla \hat{\mathcal{L}}(\tilde{w}^t, \tilde{S}_i^{\text{tr}})\| + \frac{2(Q-1)(1 + \alpha L)^{Q-1} \alpha LG}{n^{\text{tr}}} \\ & \leq \frac{2Q(1 + \alpha L)^Q \alpha LG}{n^{\text{tr}}}, \end{aligned} \quad (41)$$

Next, we have

$$V(Q) \leq (1 + \alpha L)^{Q-1} \|\alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}}) - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}})\| + (1 + \alpha L)V(Q-1), \quad (42)$$

To bound the above inequality, we first consider the first term,

$$(1 + \alpha L)^{Q-1} \|\alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}}) - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}})\| \leq \alpha(1 + \alpha L)^{Q-1} \frac{2(L + G\alpha\rho(Q-1))}{n^{\text{tr}}},$$

Putting it in to  $V(Q)$  and Unrolling it, noting that  $V(1) = \|\alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}^t, S_i^{\text{tr}}) - \alpha \nabla^2 \hat{\mathcal{L}}(\tilde{w}^t, \tilde{S}_i^{\text{tr}})\| \leq \frac{2\alpha L}{n^{\text{tr}}}$ , then we have

$$\begin{aligned} V(Q) & \leq (1 + \alpha L)^{Q-1} V(1) + \sum_{k=1}^{Q-1} (1 + \alpha L)^{Q-1} \frac{2\alpha(L + G\alpha\rho(Q-k))}{n^{\text{tr}}} \\ & \leq \frac{2\alpha L(1 + \alpha L)^{Q-1}}{n^{\text{tr}}} + \sum_{k=1}^{Q-1} (1 + \alpha L)^{Q-1} \frac{2\alpha(L + G\alpha\rho k)}{n^{\text{tr}}} \\ & = \sum_{k=0}^{Q-1} (1 + \alpha L)^{Q-1} \frac{2\alpha(L + G\alpha\rho k)}{n^{\text{tr}}}. \end{aligned} \quad (43)$$

By plugging (41) and (42) into (40), then we have

$$\begin{aligned} \mathbb{E}\|\nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}) - \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}})\| & \leq (1 + \alpha L)^{2(Q-1)} \left[ \frac{2(1 + \alpha L)Q\alpha LG}{n^{\text{tr}}} + \sum_{k=0}^{Q-1} \frac{2\alpha G(L + G\alpha\rho k)}{n^{\text{tr}}} \right] \\ & \leq (1 + \alpha L)^{2(Q-1)} \frac{(6QG + Q^2\alpha^2 G^2\rho)}{n^{\text{tr}}}. \end{aligned} \quad (44)$$



936 where we using  $\alpha L \leq 1$  in the second inequality. Substituting (44) and (39) into (38), we obtain

$$\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \leq (1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \eta_t (1 + \alpha L)^{2(Q-1)} \frac{(6QG + Q^2 \alpha^2 G^2 \rho)}{n^{\text{tr}}} + 2\eta_t \mathbb{E}\|\nabla \hat{F}_i(w^t)\|.$$

937 From Lemma 7, we can know  $\mathbb{E}\|\nabla \hat{F}_i(w^t)\| \leq eG$ . Then combing the above two cases, we obtain

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &\leq (1 - \frac{1}{m})(1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m}(1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| \\ &\quad + \frac{1}{m} \eta_t (1 + \alpha L)^{2(Q-1)} \frac{(6QG + Q^2 \alpha^2 G^2 \rho)}{n^{\text{tr}}} + \frac{2\eta_t eG}{m} \\ &\leq \exp(\eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{\eta_t \Phi}{m}. \end{aligned}$$

938 where we use  $1 + x \leq \exp(x)$  and  $\alpha \leq \frac{1}{QL}$  in the second inequality, and we denote

939  $\Phi = 2eG + (1 + \alpha L)^{2(Q-1)} \frac{(6QG + Q^2 \alpha^2 G^2 \rho)}{n^{\text{tr}}}$ . Following the same proof technique in [41](Eq (23) in  
940 Theorem 4), we can easily get

941

$$\begin{aligned} \mathbb{E}\|w^T - \tilde{w}^T\| &\leq \sum_{t=t_0+1}^T \exp(2c\gamma \sum_{l=t+1}^T \frac{1}{k}) \frac{2c\Phi}{mt} \\ &\leq \sum_{t=t_0+1}^T \exp(2c\gamma \ln(\frac{T}{t})) \frac{2c\Phi}{mt} \\ &= \frac{2c\Phi}{m} (T^{2c\gamma}) \sum_{t=t_0+1}^T t^{-2c\gamma-1} \\ &\leq \frac{1}{2c\gamma} \frac{2c\Phi}{m} (\frac{T}{t_0})^{2c\gamma} \end{aligned}$$

942 and

$$\mathbb{E}\|\ell(w_{\mathcal{T}_i}(w^T, \tilde{\mathcal{S}}_i^{\text{tr}}, \tilde{z}) - \ell(w_{\mathcal{T}_i}(\tilde{w}^T, \tilde{\mathcal{S}}_i^{\text{tr}}, \tilde{z}))\| \leq \frac{eG\Phi}{\gamma m} (\frac{T}{t_0})^{2c\gamma} + r \frac{t_0}{m}, \quad (45)$$

943 where  $r = \mathbb{E}_{\mathcal{S}, \mathcal{A}}[F(w_{\mathcal{S}})], \gamma = \mathcal{O}(\min\{L_Q, \mathbb{E}_{\mathcal{S}}[\|\nabla^2 F_i(w^0, S)\|]) + \rho_Q(c\sigma +$   
944  $\sqrt{c(F(w^0) - \min_{\mathcal{W}} F)})\}$ . Next, let  $b = 2c\gamma$ . Then, setting

$$t_0 = (\frac{2ceG\Phi}{r})^{\frac{1}{1+b}} T^{\frac{b}{1+b}}$$

945 minimizes (45). Plugging  $t_0$  back we get that (44) equals to

$$\frac{1 + \frac{1}{b}}{m} (2ceG\Phi)^{\frac{1}{1+b}} (rT)^{\frac{b}{1+b}}$$

946 This completes the proof.

## 947 C Stability and Generalization of PDF

### 948 C.1 Proof of stability of PDF

949 To show the claim, it just suffices to show that for any  $i$ , we have

$$\mathbb{E}_{\mathcal{A}, \mathcal{S}} [F_i(w_{\mathcal{S}}) - \hat{F}_i(w_{\mathcal{S}}, S_i)] \leq \epsilon.$$

950 Take the dataset  $\tilde{\mathcal{S}}^{(i)}$  which is the same as  $\mathcal{S}$ , except that  $\tilde{S}_i$  differ from  $S_i$  in at most one data point.  
951 In particular,

$$S_i^{\text{tr}} = \{z_{i,1}, \dots, z_{i,n}\}, \tilde{S}_i = \{z_{i,1}, \dots, \tilde{z}_{i,j}, \dots, z_{i,n}\}.$$

952 Then, we relate empirical risk and population risk by

$$\begin{aligned}\mathbb{E}_{\mathcal{S},\mathcal{A}}[\widehat{F}_i(w_{\mathcal{S}}, z_{i,j})] &= \frac{1}{n} \sum_{j=1}^n \mathbb{E}_{\mathcal{S},\mathcal{A}} \ell_{\lambda}(w_{\mathcal{S}}, z_{i,j}) \\ &= \frac{1}{n} \sum_{j=1}^n \mathbb{E}_{\mathcal{S},\mathcal{A},\widetilde{z}_{i,j}} \ell_{\lambda}(w_{\mathcal{S}}, \widetilde{z}_{i,j}).\end{aligned}\tag{46}$$

953 Moreover, we have

$$\mathbb{E}_{\mathcal{A},\mathcal{S}}[F_i(w_{\mathcal{S}})] = \frac{1}{n} \sum_{j=1}^n \mathbb{E}_{\mathcal{S},\mathcal{A},\widetilde{z}_{i,j}} \ell_{\lambda}(w_{\mathcal{S}}, \widetilde{z}_{i,j}).\tag{47}$$

954 Putting (46) and (47) together, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{A},\mathcal{S}} \left[ F(w_{\mathcal{S}}) - \widehat{F}(w_{\mathcal{S}}, \mathcal{S}_i) \right] &\leq \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{j=1}^n \mathbb{E}_{\mathcal{S},\mathcal{A},\widetilde{z}_{i,j}} \ell_{\lambda}(w_{\mathcal{S}}, \widetilde{z}_{i,j}) - \ell_{\lambda}(w_{\mathcal{S}}, \widetilde{z}_{i,j}) \\ &\leq \epsilon.\end{aligned}$$

955 Then we obtain the desired result.

## 956 C.2 Lemmas

957 **Lemma 16.** Assume that  $\widehat{\mathcal{L}}$  is differentiable and  $w_{\mathcal{T}}^*$  is the unique minimizer of  $\widehat{\mathcal{L}}(w_{\mathcal{T}}) + \frac{\lambda}{2} \|w_{\mathcal{T}} - w\|^2$ .  
958 Then the gradient of  $\widehat{\mathcal{L}}_{\lambda}(w) = \widehat{\mathcal{L}}(w_{\mathcal{T}}^*) + \frac{\lambda}{2} \|w_{\mathcal{T}}^* - w\|^2$  is given by  $\nabla \widehat{\mathcal{L}}_{\lambda} = \lambda(w - w_{\mathcal{T}}^*)$

959 *Proof.* Since  $\widehat{\mathcal{L}}$  is differentiable, from the first-order optimality condition we know that

$$\nabla \widehat{\mathcal{L}}(w_{\mathcal{T}}^*) + \lambda(w_{\mathcal{T}}^* - w) = 0$$

960 From the chain rule we have

$$\begin{aligned}\nabla \widehat{\mathcal{L}}_{\lambda}(w) &= \left( \frac{\partial w_{\mathcal{T}}^*}{\partial w} \right)^{\top} \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}}^*) + \lambda \left( I - \left( \frac{\partial w_{\mathcal{T}}^*}{\partial w} \right)^{\top} \right) (w - w_{\mathcal{T}}^*) \\ &= \lambda(w - w_{\mathcal{T}}^*) + \left( \frac{\partial w_{\mathcal{T}}^*}{\partial w} \right)^{\top} (\nabla \widehat{\mathcal{L}}(w_{\mathcal{T}}^*) + \lambda(w - w_{\mathcal{T}}^*)) \\ &= \lambda(w - w_{\mathcal{T}}^*)\end{aligned}$$

961 □

962 **Lemma 17.** Suppose Assumption 2 hold. Then if  $\lambda \geq L$ ,  $\ell_{\lambda}(w, z) = \ell(w_{\mathcal{T}}^*, z) + \frac{\lambda}{2} \|w_{\mathcal{T}}^* - w\|^2$   
963 is  $L_Q$ -smoothness with  $L_Q = \frac{\lambda L}{\lambda + L}$  and  $\rho$ -Lipschitz Hessian with respect to  $w$ , where  $w_{\mathcal{T}}^* =$   
964  $\operatorname{argmin}_{w_{\mathcal{T}_i}} \ell(w_{\mathcal{T}_i}, z) + \frac{\lambda}{2} \|w_{\mathcal{T}_i} - w\|^2$ .

965 *Proof.* From the first-order optimality condition we know that

$$\nabla \ell(w_{\mathcal{T}_i}^*) + \lambda(w_{\mathcal{T}_i}^* - w) = 0.$$

966 Therefore, we can further obtain

$$\nabla^2 \ell(w_{\mathcal{T}_i}^*) \frac{\partial w_{\mathcal{T}_i}^*}{\partial w} + \lambda \left( \frac{\partial w_{\mathcal{T}_i}^*}{\partial w} - I \right) = 0.$$

967 and

$$\nabla^3 \ell(w_{\mathcal{T}_i}^*) \left( \frac{\partial w_{\mathcal{T}_i}^*}{\partial w} \right)^2 + \nabla^2 \ell(w_{\mathcal{T}_i}^*) \frac{\partial^2 w_{\mathcal{T}_i}^*}{\partial w^2} + \lambda \frac{\partial^2 w_{\mathcal{T}_i}^*}{\partial w^2} = 0.$$

968 This implies

$$\frac{\partial w_{\mathcal{T}_i}^*}{\partial w} = \lambda(\nabla^2 \ell(w_{\mathcal{T}_i}^*) + \lambda I)^{-1}, \quad \frac{\partial^2 w_{\mathcal{T}_i}^*}{\partial w^2} = -(\nabla^2 \ell(w_{\mathcal{T}_i}^*) + \lambda I)^{-1} \nabla^3 \ell(w_{\mathcal{T}_i}^*) \left( \frac{\partial w_{\mathcal{T}_i}^*}{\partial w} \right)^2.$$

969 From Lemma , we have  $\nabla \ell_\lambda(w) = \lambda(w - w_{\mathcal{T}_i}^*)$ . Therefore, we can further have

$$\nabla^2 \ell_\lambda(w) = \lambda(I - \frac{\partial w_{\mathcal{T}_i}^*}{\partial w}) = \lambda(I - \lambda(\nabla^2 \ell(w_{\mathcal{T}_i}^*) + \lambda I)^{-1})$$

970 and

$$\nabla^3 \ell_\lambda(w) = -\lambda \frac{\partial^2 w_{\mathcal{T}_i}^*}{\partial w^2} = \lambda^3(\nabla^2 \ell(w_{\mathcal{T}_i}^*) + \lambda I)^{-3} \nabla^3 \ell(w_{\mathcal{T}_i}^*).$$

971 Note that  $\ell(w_{\mathcal{T}_i})$  is  $L$ -smooth and  $\rho$ -Hessian Lipschitz with respect to  $w_{\mathcal{T}_i}$ . Then it yields

$$\|\nabla^2 \ell_\lambda(w)\| \leq \frac{\lambda L}{\lambda + L}, \quad \|\nabla^3 \ell_\lambda(w)\| \leq \rho$$

972

□

### 973 C.3 Proof of Theorem 4

974 To facilitate the analysis of stability, we rewrite  $\widehat{\mathcal{K}}(w_{\mathcal{T}}, w, S) = \widehat{\mathcal{L}}(w_{\mathcal{T}}, S) + \frac{\lambda}{2} \|w_{\mathcal{T}} - w\|^2$ ,  $\widehat{\mathcal{L}}_\lambda(w, S) = \min_{w_{\mathcal{T}}} \{\widehat{\mathcal{L}}(w_{\mathcal{T}}, S) + \frac{\lambda}{2} \|w_{\mathcal{T}} - w\|^2\}$ . From Lemma 17, we can know  $\widehat{\mathcal{L}}_\lambda$  is  $L_Q$ -smooth. Hence, by using Lemma 1, for the first case that task  $\mathcal{T}_j$  is selected, we have

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &= \mathbb{E}\|(w^t - \eta_t \nabla \widehat{\mathcal{K}}(w_{\mathcal{T}_j, Q}^t, w^t; S_j)) - (\tilde{w}^t - \eta_t \nabla \widehat{\mathcal{K}}(w_{\mathcal{T}_j, Q}^t, \tilde{w}^t; S_j))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \widehat{\mathcal{L}}_\lambda(w^t, S_j)) - (\tilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}_\lambda(\tilde{w}^t, S_j))\| + 2\eta_t \mathbb{E}\|\nabla \widehat{\mathcal{K}}(w_{\mathcal{T}_j, Q}^t, w^t; S_j) - \nabla \widehat{\mathcal{L}}_\lambda(w^t, S_j)\| \\ &= \mathbb{E}\|(w^t - \eta_t \nabla \widehat{\mathcal{L}}_\lambda(w^t, S_j)) - (\tilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}_\lambda(\tilde{w}^t, S_j))\| + 2\eta_t \lambda \mathbb{E}\|w_{j, Q}^t - w_{\mathcal{T}_i}^*(w^t)\|. \\ &\leq \mathbb{E}\|w^t - \tilde{w}^t\| + 2\eta_t \lambda \|w_{j, Q}^t - w_{\mathcal{T}_i}^*(w^t)\| \end{aligned} \quad (48)$$

977 where  $w_{\mathcal{T}}^*(w) = \arg \min_{w_{\mathcal{T}}} \widehat{\mathcal{L}}(w_{\mathcal{T}}, S) + \frac{\lambda}{2} \|w_{\mathcal{T}} - w\|^2$ . where we use Lemma 1 in the second  
978 inequality. In the second case, we have

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &= \mathbb{E}\|(w^t - \eta_t \nabla \mathcal{K}(w_{\mathcal{T}_i, Q}^t, w^t; S_i)) - (\tilde{w}^t - \eta_t \nabla \mathcal{K}(w_{\mathcal{T}_i, Q}^t, \tilde{w}^t; \tilde{S}_i))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \widehat{\mathcal{L}}_\lambda(w^t, S_i)) - (\tilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}_\lambda(\tilde{w}^t, \tilde{S}_i))\| + 2\mathbb{E}\|\nabla \mathcal{K}(w_{\mathcal{T}_i, Q}^t, w^t; S_i) - \nabla \widehat{\mathcal{L}}_\lambda(w^t, S_i)\|. \end{aligned} \quad (49)$$

979 For the first term in (49), we have

$$\begin{aligned} &\mathbb{E}\|(w^t - \eta_t \nabla \widehat{\mathcal{L}}_\lambda(w^t, S_i)) - (\tilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}_\lambda(\tilde{w}^t, \tilde{S}_i))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \widehat{\mathcal{L}}_\lambda(w^t, S_i)) - (\tilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}_\lambda(\tilde{w}^t, S_i))\| + \eta_t \mathbb{E}\|\nabla \widehat{\mathcal{L}}_\lambda(\tilde{w}^t, S_i) - \nabla \widehat{\mathcal{L}}_\lambda(w^t, \tilde{S}_i)\| \\ &\leq \mathbb{E}\|w^t - \tilde{w}^t\| + 2\eta_t \mathbb{E}\|\nabla \widehat{F}_i(w^t, S_i)\| \end{aligned} \quad (50)$$

980 For the second term in (50), we have

$$\mathbb{E}\|\nabla \mathcal{K}(\Phi_Q(w_{i, j}^t), w^t; j) - \nabla \widehat{\mathcal{L}}_\lambda(w^t, i)\| \leq \lambda \mathbb{E}\|w_{i, Q}^t - w_{\mathcal{T}_i}^*(w^t)\| \quad (51)$$

981 Putting (50) and (51) into (49), then we have

$$\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \leq \mathbb{E}\|w^t - \tilde{w}^t\| + 2\eta_t \mathbb{E}\|\nabla \widehat{F}_i(w^t, S_i)\| + 2\eta_t \lambda \mathbb{E}\|w_{i, Q}^t - w_{\mathcal{T}_i}^*(w^t)\| \quad (52)$$

982 Combining (48) and (52), we obtain

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &\leq (1 - \frac{1}{m}) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m} \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m} \left[ 2\eta_t \mathbb{E}\|\nabla \widehat{F}(w^t, i)\| + 2\eta_t \lambda \mathbb{E}\|w_{i, Q}^t - w_{\mathcal{T}_i}^*(w^t)\| \right] \\ &= \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{2\eta_t}{m} \mathbb{E}\|\nabla \widehat{F}(w^t, S_i)\| + \frac{2\eta_t \lambda}{m} \mathbb{E}\|w_{i, Q}^t - w_{\mathcal{T}_i}^*(w^t)\|. \end{aligned}$$

983 Unrolling it and noting that  $\|w^0 - \tilde{w}^0\| = 0$ , we have

$$\begin{aligned} \mathbb{E}\|w^T - \tilde{w}^T\| &\leq \sum_{t=0}^{T-1} \frac{2\eta_t \lambda}{m} \mathbb{E}\|w_{i, Q}^t - w_{\mathcal{T}_i}^*(w^t)\| + \sum_{t=0}^{T-1} \frac{2\eta_t}{m} \mathbb{E}\|\nabla \widehat{F}(w^t, S_i)\| \\ &\leq \sum_{t=0}^{T-1} \frac{2\eta_t \lambda}{m} \mathbb{E}\|w_{i, Q}^t - w_{\mathcal{T}_i}^*(w^t)\| + \frac{1}{m} \sqrt{F(w^0) - \min_{\mathcal{W}} F + \frac{L_Q \sigma^2}{2} \sum_{t=0}^{T-1} \eta_t^2} \end{aligned}$$

984 where we use Lemma 14 and Lemma 17 in the second inequality, and we denote  $L_Q = \frac{\lambda L}{\lambda + L}$ . Now  
 985 we are ready to conclude. By using Lemma A, we have

$$\begin{aligned} \mathbb{E}\|\ell_\lambda(w^T, \tilde{z}) - \ell_\lambda(\tilde{w}^T, \tilde{z})\| &\leq 2G\mathbb{E}\|w^T - \tilde{w}^T\| \\ &\leq \sum_{t=0}^{T-1} \frac{2\eta_t \lambda}{m} \mathbb{E}\|w_{i,Q}^t - w_{\mathcal{T}_i}^*(w^t)\| + \frac{1}{m} \sqrt{F(w^0) - \min_{\mathcal{W}} F + \frac{L_Q \sigma^2}{2} \sum_{t=0}^{T-1} \eta_t^2} \end{aligned}$$

986 Then we completes the proof.

#### 987 C.4 Proof of Theorem 5

988 For the non-convex case, under Lemma 17, for the first case we have

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &= \mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{K}}(w_{\mathcal{T}_j,q}^t, w^t; S_j)) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{K}}(w_{\mathcal{T}_j,q}^t, \tilde{w}^t; S_j))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{L}}_\lambda(w^t, S_j)) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}_\lambda(\tilde{w}^t, S_j))\| + 2\eta_t \mathbb{E}\|\nabla \hat{\mathcal{K}}(w_{\mathcal{T}_j,q}^t, w^t; S_j) - \nabla \hat{\mathcal{L}}_\lambda(w^t, S_j)\|. \\ &= \mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{L}}_\lambda(w^t, S_j)) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}_\lambda(\tilde{w}^t, S_j))\| + 2\eta_t \lambda \|w_{j,Q}^t - w_{\mathcal{T}_i}^*(w^t)\|. \\ &\leq (1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + 2\eta_t \lambda \|w_{j,Q}^t - w_{\mathcal{T}_i}^*(w^t)\| \end{aligned} \quad (53)$$

989 where  $\phi_t = \min\{L_Q, \xi_t\}$  with  $\xi_t = \|\nabla^2 F(w_0, S_t)\| + \frac{\rho}{2} \left\| \sum_{l=1}^{t-1} \beta_l \nabla \hat{F}(w_{\mathcal{S}}^l) \right\| +$   
 990  $\frac{\rho}{2} \left\| \sum_{l=1}^{t-1} \beta_l \nabla \hat{F}(w_{\mathcal{S}}^l) \right\|$ ,  $w_{\mathcal{T}}^*(w) = \arg \min_{w_{\mathcal{T}}} \hat{\mathcal{L}}(w_{\mathcal{T}}, S) + \frac{\lambda}{2} \|w_{\mathcal{T}} - w\|$ . In the second case, we  
 991 have

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &= \mathbb{E}\|(w^t - \eta_t \nabla \mathcal{K}(w_{\mathcal{T}_i,q}^t, w^t; S_i)) - (\tilde{w}^t - \eta_t \nabla \mathcal{K}(w_{\mathcal{T}_i,q}^{t'}, \tilde{w}^t; \tilde{S}_i))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{L}}_\lambda(w^t, S_i)) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}_\lambda(\tilde{w}^t, \tilde{S}_i))\| + 2\mathbb{E}\|\nabla \mathcal{K}(w_{\mathcal{T}_i,q}^t, w^t; S_i) - \nabla \hat{\mathcal{L}}_\lambda(w^t, S_i)\|. \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{L}}_\lambda(w^t, S_i)) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}_\lambda(\tilde{w}^t, \tilde{S}_i))\| + 2\eta_t \lambda \|w_{j,Q}^t - w_{\mathcal{T}_i}^*(w^t)\| \end{aligned} \quad (54)$$

992 For the first term in (54),

$$\begin{aligned} &\mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{L}}_\lambda(w^t, S_i)) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}_\lambda(\tilde{w}^t, \tilde{S}_i))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{L}}_\lambda(w^t, S_i)) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}_\lambda(\tilde{w}^t, S_i))\| + \eta_t \mathbb{E}\|\nabla \hat{\mathcal{L}}_\lambda(\tilde{w}^t, S_i) - \nabla \hat{\mathcal{L}}_\lambda(w^t, \tilde{S}_i)\| \end{aligned}$$

993 Putting it into (54), then we have

$$\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \leq (1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + 2\eta_t \mathbb{E}\|\nabla \hat{F}_i(w^t, S_i)\| + 2\eta_t \lambda \|w_{j,Q}^t - w_{\mathcal{T}_i}^*(w^t)\| \quad (55)$$

994 we can know  $\mathbb{E}\|\nabla \hat{F}_i(w^t)\| \leq G$ . Then combing two cases, we obtain

$$\begin{aligned} &\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \\ &\leq (1 - \frac{1}{m})(1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m}(1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m} \left[ 2\eta_t \mathbb{E}\|\nabla \hat{F}(w^t, S_i)\| + 2\eta_t \lambda \mathbb{E}\|w_{i,Q}^t - w_{\mathcal{T}_i}^*(w^t)\| \right] \\ &\leq (1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m} \left[ 2\eta_t \mathbb{E}\|\nabla \hat{F}(w^t, S_i)\| + \frac{2\eta_t \lambda G}{Q} \right] \\ &\leq \exp(\eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{2\eta_t \Phi}{m} \end{aligned}$$

995 where we use ... in the second inequality,  $1 + x \leq \exp(x)$  in the third inequality. Additionally, we  
 996 denote  $\Phi = G + \frac{(\lambda+L)G}{Q}$ . Following the same proof technique in [41](Eq (23) in Theorem 4), we  
 997 can easily get

$$\begin{aligned}
\mathbb{E}\|w^T - \tilde{w}^T\| &\leq \sum_{t=t_0+1}^T \exp(2c\gamma \sum_{l=t+1}^T \frac{1}{k}) \frac{2c\Phi}{mt} \\
&\leq \sum_{t=t_0+1}^T \exp(2c\gamma \ln(\frac{T}{t})) \frac{2c\Phi}{mt} \\
&= \frac{2c\Phi}{m} (T^{2c\gamma}) \sum_{t=t_0+1}^T t^{-2c\gamma-1} \\
&\leq \frac{1}{2c\gamma} \frac{2c\Phi}{m} (\frac{T}{t_0})^{2c\gamma}
\end{aligned}$$

999 and

$$\mathbb{E}\|\ell(w_{\mathcal{T}_i}(w^T, \tilde{\mathcal{S}}_i^{\text{tr}}), \tilde{z}) - \ell(w_{\mathcal{T}_i}(\tilde{w}^T, \tilde{\mathcal{S}}_i^{\text{tr}}), \tilde{z})\| \leq \frac{eG\Phi}{\gamma m} (\frac{T}{t_0})^{2c\gamma} + r \frac{t_0}{m}, \quad (56)$$

1000 where  $r = \mathbb{E}_{\mathcal{S}, \mathcal{A}}[F(w_S)]$ ,  $\gamma = \mathcal{O}(\min\{L_Q, \mathbb{E}_S[\|\nabla^2 F_i(w^0, S)\|] + \rho(c\sigma +$   
 1001  $\sqrt{c(F(w^0) - \min_{\mathcal{W}} F)})\}$ ). Next, let  $b = 2c\gamma$ . Then, setting

$$t_0 = (\frac{2cG\Phi}{r})^{\frac{1}{1+b}} T^{\frac{b}{1+b}}$$

1002 minimizes (56). Plugging  $t_0$  back we get that (44) equals to

$$\frac{1 + \frac{1}{b}}{m} (2cG\Phi)^{\frac{1}{1+b}} (rT)^{\frac{b}{1+b}}$$

1003 This completes the proof.

## 1004 D Results of Table 2

1005 In the subsequent proofs of other algorithms, we provide only the proof of generalization bound  
 1006 under the assumptions of the  $L_Q$ -smoothness constant and  $\rho_Q$ -Hessian Lipschitz continuity, which  
 1007 can be established by referring our previous proof. We first present their corresponding algorithm.

### 1008 D.1 Algorithms

---

#### Algorithm 2 MAML

---

**Require:** The set of datasets  $\mathcal{S} = \{S_i\}_{i=1}^m$  with  $S_i = \{S_i^{\text{tr}}, S_i^{\text{ts}}\}$ , outer iterations  $T$ , adaptation steps  $Q$ .

**Require:** Choose arbitrary initial point  $w^0 \in W$ ;

- 1: **for**  $t = 0$  **to**  $T - 1$  **do**
  - 2:   Randomly choose the task  $i$ .
  - 3:   Inner-Level:  $w_{\mathcal{T}_i,0}^t = w_t$
  - 4:   **for**  $q = 0, 1, \dots, Q - 1$  **do**
  - 5:      $w_{\mathcal{T}_i,q+1}^t = w_{\mathcal{T}_i,q}^t - \alpha \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i,q}^t, S_i^{\text{tr}})$ ;
  - 6:   **end for**
  - 7:   Outer-level:  $w_{\mathcal{T}_i} = w_{\mathcal{T}_i,Q}^t$
  - 8:    $w^{t+1} := w^t - \eta_t \nabla_w \hat{\mathcal{L}}_i(w_{\mathcal{T}_i}, S_i^{\text{ts}})$
  - 9: **end for**
  - 10:  $w^T$  and  $\bar{w}^T := \frac{1}{T+1} \sum_{t=0}^T w^t$ ;
-

---

**Algorithm 3** FOMAML

---

**Require:** The set of datasets  $\mathcal{S} = \{S_i\}_{i=1}^m$  with  $S_i = \{S_i^{\text{tr}}, S_i^{\text{ts}}\}$ , outer iterations  $T$ , adaptation steps  $Q$ .

**Require:** Choose arbitrary initial point  $w^0 \in W$ ;

```
1: for  $t = 0$  to  $T - 1$  do
2:   Randomly choose the task  $i$ .
3:   Inner-Level:  $w_{\mathcal{T}_i,0}^t = w_t$ 
4:   for  $q = 0, 1, \dots, Q - 1$  do
5:      $w_{\mathcal{T}_i,q+1}^t = w_{\mathcal{T}_i,q}^t - \alpha \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i,q}^t, S_i^{\text{tr}})$ ;
6:   end for
7:   Outer-level:  $w_{\mathcal{T}_i} = w_{\mathcal{T}_i,Q}^t$ 
8:    $w^{t+1} := w^t - \eta_t \nabla_{w_{\mathcal{T}_i}} \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}, S_i^{\text{ts}})$ 
9: end for
10:  $w^T$  and  $\bar{w}^T := \frac{1}{T+1} \sum_{t=0}^T w^t$ ;
```

---

---

**Algorithm 4** MetaSGD

---

**Require:** The set of datasets  $\mathcal{S} = \{S_i\}_{i=1}^m$  with  $S_i = \{S_i^{\text{tr}}, S_i^{\text{ts}}\}$ , outer iterations  $T$ , adaptation steps  $Q$ .

**Require:** Choose arbitrary initial point  $w^0 \in W$ ;

```
1: for  $t = 0$  to  $T - 1$  do
2:   Randomly choose the task  $i$ .
3:   Inner-Level:  $w_{\mathcal{T}_i,0}^t = w_t$ 
4:   for  $q = 0, 1, \dots, Q - 1$  do
5:      $w_{\mathcal{T}_i,q+1}^t = w_{\mathcal{T}_i,q}^t - \alpha \circ \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i,q}^t, S_i^{\text{tr}})$ ;
6:   end for
7:   Outer-level:  $w_{\mathcal{T}_i} = w_{\mathcal{T}_i,Q}^t$ 
8:    $w^{t+1} := w^t - \eta_t \nabla_w \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}, S_i^{\text{ts}})$ 
9:    $\alpha^{t+1} := \alpha^t - \eta_t \nabla_\alpha \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}, S_i^{\text{ts}})$ 
10: end for
11:  $w^T$  and  $\bar{w}^T := \frac{1}{T+1} \sum_{t=0}^T w^t$ ;
```

---

---

**Algorithm 5** iMAML

---

**Require:** The set of datasets  $\mathcal{S} = \{S_i\}_{i=1}^m$  with  $S_i = \{S_i^{\text{tr}}, S_i^{\text{ts}}\}$ , outer iterations  $T$ , adaptation steps  $Q$ , regularization constant  $\lambda$

**Require:** Choose arbitrary initial point  $w^0 \in W$ ;

```
1: for  $t = 0$  to  $T - 1$  do
2:   Randomly choose the task  $i$ .
3:   Inner-Level:  $w_{\mathcal{T}_i,0}^t = w_t$ 
4:   for  $q = 0, 1, \dots, Q - 1$  do
5:      $w_{\mathcal{T}_i,q+1}^t = w_{\mathcal{T}_i,q}^t - \alpha \nabla \widehat{\mathcal{L}}_\lambda(w_{\mathcal{T}_i,q}^t, S_i^{\text{tr}})$ ;
6:   end for
7:   Outer-level:  $w_{\mathcal{T}_i} = w_{\mathcal{T}_i,Q}^t$ 
8:    $w^{t+1} := w^t - \eta_t (I + \frac{1}{\lambda} \nabla_{w_{\mathcal{T}_i}}^2 \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}, S_i^{\text{tr}}))^{-1} \nabla_{w_{\mathcal{T}_i}} \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}, S_i^{\text{ts}})$ 
9: end for
10:  $w^T$  and  $\bar{w}^T := \frac{1}{T+1} \sum_{t=0}^T w^t$ ;
```

---

---

**Algorithm 6** Meta-MinibatchProx

---

**Require:** The set of datasets  $\mathcal{S} = \{S_i\}_{i=1}^m$ , outer iterations  $T$ , adaption steps  $Q$ , regularization constant  $\lambda$ .

**Require:** Choose arbitrary initial point  $w^0 \in W$ ;

```

1: for  $t = 0$  to  $T - 1$  do
2:   Randomly choose the task  $i$ .
3:   Inner-Level:  $w_{\mathcal{T}_i,0}^t = w_t$ 
4:   for  $q = 0, 1, \dots, Q - 1$  do
5:      $w_{\mathcal{T}_i,q+1}^t = w_{\mathcal{T}_i,q}^t - \alpha \nabla \widehat{\mathcal{K}}(w_{\mathcal{T}_i,q}^t, S_i)$ ;
6:   end for
7:   Outer-level:  $w_{\mathcal{T}_i} = w_{\mathcal{T}_i,Q}^t$ 
8:    $w^{t+1} := w^t - \eta_t \lambda (w^t - w_{\mathcal{T}_i})$ 
9: end for
10:  $w^T$  and  $\bar{w}^T := \frac{1}{T+1} \sum_{t=0}^T w^t$ ;

```

---



---

**Algorithm 7** FoMuML

---

**Require:** The set of datasets  $\mathcal{S} = \{S_i\}_{i=1}^m$  with  $S_i = \{S_i^{\text{tr}}, S_i^{\text{ts}}\}$ , outer iterations  $T$ , adaptation steps  $Q$ , regularization constant  $\lambda$ .

**Require:** Choose arbitrary initial point  $w^0 \in W$ ;

```

1: for  $t = 0$  to  $T - 1$  do
2:   Randomly choose the task  $i$ .
3:   Inner-Level:  $w_{\mathcal{T}_i,0}^t = w_t$ 
4:   for  $q = 0, 1, \dots, Q - 1$  do
5:      $w_{\mathcal{T}_i,q+1}^t = w_{\mathcal{T}_i,q}^t - \alpha \nabla \widehat{\mathcal{L}}_\lambda(w_{\mathcal{T}_i,q}^t, S_i^{\text{tr}})$ ;
6:   end for
7:   Outer-level:  $w_{\mathcal{T}_i} = w_{\mathcal{T}_i,Q}^t$ 
8:    $w^{t+1} := w^t - \eta_t \nabla_{w_{\mathcal{T}_i}} \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}, S_i^{\text{ts}})$ 
9: end for
10:  $w^T$  and  $\bar{w}^T := \frac{1}{T+1} \sum_{t=0}^T w^t$ ;

```

---

1009 **D.2 FOMAML(convex)**

1010 This proof is similar to the proof of Theorem 2, we have

$$\begin{aligned}
\|w^{t+1} - \tilde{w}^{t+1}\| &= \|(w^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_j}(w^t, S_j^{\text{tr}}), S_j^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_j}(\tilde{w}^t, S_j^{\text{tr}}), S_j^{\text{ts}}))\| \\
&\leq \frac{1}{n^{\text{ts}}} \sum_{z^{\text{ts}} \in S_j^{\text{ts}}} \|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_j}(w^t, S_j^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_j}(\tilde{w}^t, S_j^{\text{tr}}), z^{\text{ts}}))\| \\
&\leq \|w^t - \tilde{w}^t\|
\end{aligned} \tag{57}$$

1011 Next, for a given time index  $t$ , with probability  $\frac{1}{m}$ , the task  $\mathcal{T}_i$  is selected. In this case, we have

$$\begin{aligned}
\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &= \mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}, S_i^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}})))\| \\
&\leq \mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}, S_i^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}})))\| \\
&\quad + \mathbb{E}\|\eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}})) - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, S_i^{\text{ts}}))\| \\
&\leq \frac{1}{n^{\text{ts}}} \sum_{z^{\text{ts}} \in S_i^{\text{ts}}} \mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})))\| \\
&\quad + \eta_t \mathbb{E}\|\nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}})) - \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, S_i^{\text{ts}}))\| \\
&\leq \frac{1}{n^{\text{ts}}} \sum_{z \in S_i^{\text{ts}}} \mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})))\| \\
&\quad + 2\eta_t \mathbb{E}\|\nabla \hat{F}_i(w)\|,
\end{aligned} \tag{58}$$

1012 where the last inequality follows that  $\tilde{S}_i^{\text{ts}}$  and  $S_i^{\text{ts}}$  are sampled from the same distribution, then

1013  $\mathbb{E}\|\nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}}))\| = \mathbb{E}\|\nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, S_i^{\text{ts}}))\| = \mathbb{E}\|\nabla \hat{F}_i(w)\|$ . Note that

$$\begin{aligned}
&\mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})))\| \\
&\leq \mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})))\| \\
&\quad + \eta_t \mathbb{E}\|\nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})) - \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}}))\|.
\end{aligned} \tag{59}$$

1014 Let us bound the two terms on the RHS of (59) separately. First, similar to how we derived (57), we  
1015 could bound the first term by

$$\mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})))\| \leq \mathbb{E}\|w^t - \tilde{w}^t\|.$$

1016 To bound the second term on the RHS of (59), we consider two parallel processes of generating

1017 iterates  $\{\tilde{w}_{\mathcal{T}_i, q}^t\}$  and  $\{\tilde{w}_{\mathcal{T}_i, q}^{t, \prime}\}$  by using datasets  $S_i^{\text{tr}}$  and  $\tilde{S}_i^{\text{tr}}$ , respectively. Note that

$$\begin{aligned}
\mathbb{E}\|\nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}, z^{\text{ts}})) - \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}}))\| &= \mathbb{E}\|\nabla \ell(\tilde{w}_{\mathcal{T}_i, Q}^t, z^{\text{ts}}) - \nabla \ell(\tilde{w}_{\mathcal{T}_i, Q}^{t, \prime}, z^{\text{ts}})\| \\
&\leq L \mathbb{E}\|\tilde{w}_{\mathcal{T}_i, Q}^t - \tilde{w}_{\mathcal{T}_i, Q}^{t, \prime}\| \\
&= L \mathbb{E}\|[\tilde{w}_{i, Q-1}^t - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}})] - [\tilde{w}_{i, Q-1}^{t, \prime} - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}})]\| \\
&\leq L \mathbb{E}\|[\tilde{w}_{i, Q-1}^t - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^t, S_i^{\text{tr}})] - [\tilde{w}_{i, Q-1}^{t, \prime} - \alpha \nabla \ell(\tilde{w}_{i, Q-1}^{t, \prime}, S_i^{\text{tr}})]\| \\
&\quad + L \mathbb{E}\|\alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, S_i^{\text{tr}}) - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}_{i, Q-1}^{t, \prime}, \tilde{S}_i^{\text{tr}})\| \\
&\leq L \mathbb{E}\|\tilde{w}_{i, Q-1}^t - \tilde{w}_{i, Q-1}^{t, \prime}\| + \frac{2\alpha LG}{n^{\text{tr}}} \\
&\leq L \mathbb{E}\|[\tilde{w}^t - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}^t, S_i^{\text{tr}})] - [\tilde{w}^t - \alpha \nabla \hat{\mathcal{L}}(\tilde{w}^t, \tilde{S}_i^{\text{tr}})]\| + \frac{2(Q-1)\alpha LG}{n^{\text{tr}}} \\
&= \alpha L \mathbb{E}\|\nabla \hat{\mathcal{L}}(\tilde{w}^t, S_i^{\text{tr}}) - \nabla \hat{\mathcal{L}}(\tilde{w}^t, \tilde{S}_i^{\text{tr}})\| + \frac{2(Q-1)\alpha LG}{n^{\text{tr}}} \\
&\leq \frac{2Q\alpha LG}{n^{\text{tr}}},
\end{aligned} \tag{60}$$

1018 Substituting (60) and (59) into (58), we have

$$\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \leq \mathbb{E}\|w^t - \tilde{w}^t\| + \eta_t \frac{2Q\alpha LG}{n^{\text{tr}}} + 2\eta_t \mathbb{E}\|\nabla \hat{F}_i(w^t)\|.$$

1019 Combing the above two cases, we obtain

$$\begin{aligned}
\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &\leq (1 - \frac{1}{m}) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m} \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m} \eta_t \frac{2Q\alpha LG}{n^{\text{tr}}} + \frac{2}{m} \eta_t \mathbb{E}\|\nabla \hat{F}_i(w^t)\| \\
&= \mathbb{E}\|w^t - \tilde{w}^t\| + \eta_t \frac{2Q\alpha LG}{mn^{\text{tr}}} + \frac{2\eta_t}{m} \mathbb{E}\|\nabla \hat{F}_i(w^t)\|.
\end{aligned}$$



1020 Unrolling it and noting that  $\|w^0 - \tilde{w}^0\| = 0$ , we have

$$\begin{aligned}\mathbb{E}[\|w^T - \tilde{w}^T\|] &\leq \sum_{t=0}^{T-1} \eta_t \frac{2Q\alpha LG}{mn^{\text{tr}}} + \sum_{t=0}^{T-1} \frac{2\eta_t}{m} \mathbb{E}\|\nabla \hat{F}(w^t, S_t)\| \\ &\leq \sum_{t=0}^{T-1} \eta_t \frac{2Q\alpha LG}{mn^{\text{tr}}} + \frac{1}{m} \sqrt{F(w^0) - \min_{\mathcal{W}} F + \frac{L_Q \sigma^2}{2} \sum_{t=0}^{T-1} \eta_t^2},\end{aligned}$$

1021 where we use Lemma 13 in the last inequality. which completes the proof.

### 1022 D.3 FOMAML(non-convex)

1023 This proof is similar to the proof of Theorem 3, we have

$$\begin{aligned}\|w^{t+1} - \tilde{w}^{t+1}\| &= \|(w^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_j}(w^t, S_j^{\text{tr}}), S_j^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_j}(\tilde{w}^t, S_j^{\text{tr}}), S_j^{\text{ts}}))\| \\ &\leq (1 + \eta_t \phi_t) \|w^t - \tilde{w}^t\|,\end{aligned}\tag{61}$$

1024 where  $\phi_t = \min\{L_Q, \xi_t\}$  with  $\xi_t = \|\nabla^2 F(w_0, S_t)\| + \frac{\rho_Q}{2} \left\| \sum_{l=1}^{t-1} \beta_l \nabla \hat{F}(w_S^l) \right\| +$   
 1025  $\frac{\rho_Q}{2} \left\| \sum_{l=1}^{t-1} \beta_l \nabla \hat{F}(w_{\tilde{S}}^l) \right\|$ .

1026 Next, for the second case, similar to the proof of (58), we have

$$\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \leq \frac{1}{n^{\text{ts}}} \sum_{z \in S_i^{\text{ts}}} \mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}}))\| + 2\eta_t \mathbb{E}\|\nabla \hat{F}_i(w)\|.\tag{62}$$

1027 Note that

$$\begin{aligned}&\mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}}))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}))\| + \eta_t \mathbb{E}\|\nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}) - \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}}))\|.\end{aligned}\tag{63}$$

1028 Let us bound the two terms on the RHS of (63), separately. First, similar to how we bound (61), we  
 1029 could bound the first term by

$$\mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}))\| \leq (1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\|.$$

1030 To bound the second term on the RHS of (63), similar to the proof of (60), we have

$$\mathbb{E}\|\nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}) - \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}})\| \leq \frac{2Q(1 + \alpha L)^Q \alpha LG}{n^{\text{tr}}}\tag{64}$$

1031 Substituting (64) and (63) into (62), we obtain

$$\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \leq (1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \eta_t \frac{2Q(1 + \alpha L)^Q \alpha LG}{n^{\text{tr}}} + 2\eta_t \mathbb{E}\|\nabla \hat{F}_i(w^t)\|.$$

1032 From Lemma 7, we can know  $\mathbb{E}\|\nabla \hat{F}_i(w^t)\| \leq eG$ . Then combining the above two cases, we obtain

$$\begin{aligned}\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &\leq (1 - \frac{1}{m})(1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m}(1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| \\ &\quad + \frac{2Q(1 + \alpha L)^Q \alpha LG}{mn^{\text{tr}}} + \frac{2\eta_t eG}{m} \\ &\leq \exp(\eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{\eta_t \Phi}{m}.\end{aligned}$$

1033 where we use  $1 + x \leq \exp(x)$  and  $\alpha \leq \frac{1}{Q_L}$  in the second inequality, and we denote  $\Phi = 2eG +$   
 1034  $\frac{2Q(1 + \alpha L)^Q \alpha LG}{n^{\text{tr}}}$ . Using the same technique in Theorem 3. Then we complete the proof.

#### 1035 D.4 iMAML(convex)

1036 Since our goal is to demonstrate the extensibility of our framework analysis, here we consider the exact  
 1037 version of the solved iMAML algorithm, and of course we believe that this can be equally generalized  
 1038 to the iMAML algorithm with an error term. Let  $w_{\mathcal{T}_i}^{t,*} = \operatorname{argmin}_{w_{\mathcal{T}_i}} \{\widehat{\mathcal{L}}(w_{\mathcal{T}_i}, S_j^{\text{tr}}) + \frac{\lambda}{2} \|w_{\mathcal{T}_i} - w^t\|^2\}$ ,  
 1039 we have

$$\begin{aligned} \|w^{t+1} - \tilde{w}^{t+1}\| &= \|(w^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_j}^*(w^t, S_j^{\text{tr}}, S_j^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_j}^*(\tilde{w}^t, S_j^{\text{tr}}, S_j^{\text{ts}})))\| \\ &\leq \frac{1}{n^{\text{ts}}} \sum_{z^{\text{ts}} \in S_j^{\text{ts}}} \|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_j}^*(w^t, S_j^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_j}^*(\tilde{w}^t, S_j^{\text{tr}}, z^{\text{ts}})))\| \\ &\leq \|w^t - \tilde{w}^t\| \end{aligned} \quad (65)$$

1040 Next, for a given time index  $t$ , with probability  $\frac{1}{m}$ , the task  $\mathcal{T}_i$  is selected. In this case, we have

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &= \mathbb{E}\|(w^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*(w^t, S_i^{\text{tr}}, S_i^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}})))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*(w^t, S_i^{\text{tr}}, S_i^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, S_i^{\text{ts}})))\| \\ &\quad + \mathbb{E}\|\eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}})) - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, S_i^{\text{ts}}))\| \\ &\leq \frac{1}{n^{\text{ts}}} \sum_{z^{\text{ts}} \in S_i^{\text{ts}}} \mathbb{E}\|(w^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})))\| \\ &\quad + \eta_t \mathbb{E}\|\nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}})) - \nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, S_i^{\text{ts}}))\| \\ &\leq \frac{1}{n^{\text{ts}}} \sum_{z \in S_i^{\text{ts}}} \mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}^*(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})))\| \\ &\quad + 2\eta_t \mathbb{E}\|\nabla \widehat{F}_i(w)\|, \end{aligned} \quad (66)$$

1041 where the last inequality follows that  $\tilde{S}_i^{\text{ts}}$  and  $S_i^{\text{ts}}$  are sampled from the same distribution, then

1042  $\mathbb{E}\|\nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}}))\| = \mathbb{E}\|\nabla \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, S_i^{\text{ts}}))\| = \mathbb{E}\|\nabla \widehat{F}_i(w)\|$ . Note that

$$\begin{aligned} &\mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}^*(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}^*(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}^*(\tilde{w}^t, S_i^{\text{tr}}, z^{\text{ts}})))\| \\ &\quad + \eta_t \mathbb{E}\|\nabla \ell(w_{\mathcal{T}_i}^*(\tilde{w}^t, S_i^{\text{tr}}, z^{\text{ts}})) - \nabla \ell(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}}))\|. \end{aligned} \quad (67)$$

1043 Let us bound the two terms on the RHS of (67) separately. First, similar to how we derived (65), we  
 1044 could bound the first term by

$$\mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}^*(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}^*(\tilde{w}^t, S_i^{\text{tr}}, z^{\text{ts}})))\| \leq \mathbb{E}\|w^t - \tilde{w}^t\|.$$

1045 To bound the second term on the RHS of (67), note that

$$\begin{aligned} &\mathbb{E}\|\nabla \ell(w_{\mathcal{T}_i}^*(\tilde{w}^t, S_i^{\text{tr}}, z^{\text{ts}})) - \nabla \ell(w_{\mathcal{T}_i}^*(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}}))\| \\ &= \mathbb{E}\|(I + \frac{1}{\lambda} \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*, S_i^{\text{tr}}))^{-1} \nabla \ell(w_{\mathcal{T}_i}^*, z^{\text{ts}}) - (I + \frac{1}{\lambda} \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*, \tilde{S}_i^{\text{tr}}))^{-1} \nabla \ell(\tilde{w}_{\mathcal{T}_i}^*, z^{\text{ts}})\| \\ &\leq \mathbb{E}\|(I + \frac{1}{\lambda} \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*, S_i^{\text{tr}}))^{-1} \nabla \ell(w_{\mathcal{T}_i}^*, z^{\text{ts}}) - (I + \frac{1}{\lambda} \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*, S_i^{\text{tr}}))^{-1} \nabla \ell(\tilde{w}_{\mathcal{T}_i}^*, z^{\text{ts}})\| \\ &\quad + \mathbb{E}\|(I + \frac{1}{\lambda} \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*, S_i^{\text{tr}}))^{-1} \nabla \ell(\tilde{w}_{\mathcal{T}_i}^*, z^{\text{ts}}) - (I + \frac{1}{\lambda} \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*, \tilde{S}_i^{\text{tr}}))^{-1} \nabla \ell(\tilde{w}_{\mathcal{T}_i}^*, z^{\text{ts}})\| \\ &\leq \mathbb{E}\|(I + \frac{1}{\lambda} \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*, S_i^{\text{tr}}))^{-1}\| \|\nabla \ell(w_{\mathcal{T}_i}^*, z^{\text{ts}}) - \nabla \ell(\tilde{w}_{\mathcal{T}_i}^*, z^{\text{ts}})\| + \mathbb{E}\|(I + \frac{1}{\lambda} \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*, S_i^{\text{tr}}))^{-1} - (I + \frac{1}{\lambda} \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*, \tilde{S}_i^{\text{tr}}))^{-1}\| \end{aligned} \quad (68)$$

1046 For the first term in (68), since the inner-optimization problem is in  $\lambda$  strongly-convex setting, then  
 1047 by using the standard stability result in [37], we have

$$\mathbb{E}\|(I + \frac{1}{\lambda} \nabla^2 \widehat{\mathcal{L}}(w_{\mathcal{T}_i}^*, S_i^{\text{tr}}))^{-1}\| \|\nabla \ell(w_{\mathcal{T}_i}^*, z^{\text{ts}}) - \nabla \ell(\tilde{w}_{\mathcal{T}_i}^*, z^{\text{ts}})\| \leq \frac{2LG^2}{\lambda n^{\text{tr}}} \quad (69)$$

For the second term in (68), since  $S_i^{\text{tr}}$  different with  $\tilde{S}_i^{\text{tr}}$  at most one point, then we have

$$\mathbb{E} \left\| \left( I + \frac{1}{\lambda} \nabla^2 \hat{\mathcal{L}}(w_{\mathcal{T}_i}^*, S_i^{\text{tr}}) \right)^{-1} - \left( I + \frac{1}{\lambda} \nabla^2 \hat{\mathcal{L}}(w_{\mathcal{T}_i}^*, \tilde{S}_i^{\text{tr}}) \right)^{-1} \right\| \|\nabla \ell(\tilde{w}_{\mathcal{T}_i}^*, z^{\text{ts}})\| \leq \frac{2LG}{\lambda n^{\text{tr}}} \quad (70)$$

By plugging (69) and (70) into (68), then we have

$$\mathbb{E} \|\nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}) - \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}})\| \leq \frac{2L(G^2 + G)}{\lambda n^{\text{tr}}} \quad (71)$$

Substituting (71) and (67) into (66), we have

$$\mathbb{E} \|w^{t+1} - \tilde{w}^{t+1}\| \leq \mathbb{E} \|w^t - \tilde{w}^t\| + \eta_t \frac{2L(G^2 + G)}{\lambda n^{\text{tr}}} + 2\eta_t \mathbb{E} \|\nabla \hat{F}_i(w^t)\|.$$

Combing the above two cases, we obtain

$$\begin{aligned} \mathbb{E} \|w^{t+1} - \tilde{w}^{t+1}\| &\leq \left(1 - \frac{1}{m}\right) \mathbb{E} \|w^t - \tilde{w}^t\| + \frac{1}{m} \mathbb{E} \|w^t - \tilde{w}^t\| \\ &\quad + \frac{1}{m} \frac{2\eta_t L(G^2 + G)}{\lambda n^{\text{tr}}} + \frac{2\eta_t}{m} \mathbb{E} \|\nabla \hat{F}_i(w^t)\| \\ &= \mathbb{E} \|w^t - \tilde{w}^t\| + \frac{2\eta_t L(G^2 + G)}{\lambda m n^{\text{tr}}} + \frac{2\eta_t}{m} \mathbb{E} \|\nabla \hat{F}_i(w^t)\|. \end{aligned}$$

Unrolling it and noting that  $\|w^0 - \tilde{w}^0\| = 0$ , we have

$$\begin{aligned} \mathbb{E} [\|w^T - \tilde{w}^T\|] &\leq \sum_{t=0}^{T-1} \frac{2\eta_t L(G^2 + G)}{\lambda m n^{\text{tr}}} + \sum_{t=0}^{T-1} \frac{2\eta_t}{m} \mathbb{E} \|\nabla \hat{F}(w^t, S_i)\| \\ &\leq \sum_{t=0}^{T-1} \frac{2\eta_t L(G^2 + G)}{\lambda m n^{\text{tr}}} + \frac{1}{m} \sqrt{F(w^0) - \min_{\mathcal{W}} F + \frac{L_Q \sigma^2}{2} \sum_{t=0}^{T-1} \eta_t^2}, \end{aligned}$$

where we use Lemma 13 in the last inequality. which completes the proof.

## D.5 iMAML(non-convex)

This proof is similar to the proof of Theorem 3, we have

$$\begin{aligned} \|w^{t+1} - \tilde{w}^{t+1}\| &= \|(w^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_j}(w^t, S_j^{\text{tr}}), S_j^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_j}(\tilde{w}^t, S_j^{\text{tr}}), S_j^{\text{ts}}))\| \\ &\leq (1 + \eta_t \phi_t) \|w^t - \tilde{w}^t\|, \end{aligned} \quad (72)$$

where  $\phi_t = \min\{L_Q, \xi_t\}$  with  $\xi_t = \|\nabla^2 F(w_0, S_t)\| + \frac{\rho_Q}{2} \left\| \sum_{l=1}^{t-1} \beta_l \nabla \hat{F}(w_{\mathcal{S}}^l) \right\| + \frac{\rho_Q}{2} \left\| \sum_{l=1}^{t-1} \beta_l \nabla \hat{F}(w_{\mathcal{S}}^l) \right\|$ .

Next, for the second case, similar to the proof of (66), we have

$$\mathbb{E} \|w^{t+1} - \tilde{w}^{t+1}\| \leq \frac{1}{n^{\text{ts}}} \sum_{z \in S_i^{\text{ts}}} \mathbb{E} \|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}}))\| + 2\eta_t \mathbb{E} \|\nabla \hat{F}_i(w)\|. \quad (73)$$

Note that

$$\begin{aligned} &\mathbb{E} \|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}}))\| \\ &\leq \mathbb{E} \|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}))\| + \eta_t \mathbb{E} \|\nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}) - \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}})\|. \end{aligned} \quad (74)$$

Let us bound the two terms on the RHS of (74), separately. First, similar to how we bound (72), we could bound the first term by

$$\mathbb{E} \|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}))\| \leq (1 + \eta_t \phi_t) \mathbb{E} \|w^t - \tilde{w}^t\|.$$

To bound the second term on the RHS of (74), similar to the proof of (68), under the  $\lambda - L$  strongly-convex setting, we have

$$\mathbb{E} \|\nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}) - \nabla \ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}})\| \leq \frac{2L(G^2 + G)}{(\lambda - L)n^{\text{tr}}} \quad (75)$$

1064 Substituting (75) and (74) into (73), we obtain

$$\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \leq (1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \eta_t \frac{2L(G^2 + G)}{(\lambda - L)n^{\text{tr}}} + 2\eta_t \mathbb{E}\|\nabla \hat{F}_i(w^t)\|.$$

1065 From Lemma 7, we can know  $\mathbb{E}\|\nabla \hat{F}_i(w^t)\| \leq eG$ . Then combining the above two cases, we obtain

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &\leq (1 - \frac{1}{m})(1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m}(1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| \\ &\quad + \frac{2L(G^2 + G)}{(\lambda - L)mn^{\text{tr}}} + \frac{2\eta_t eG}{m} \\ &\leq \exp(\eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{\eta_t \Phi}{m}. \end{aligned}$$

1066 where we use  $1 + x \leq \exp(x)$  and  $\alpha \leq \frac{1}{QL}$  in the second inequality, and we denote  $\Phi = 2eG +$   
1067  $\frac{2L(G^2 + G)}{(\lambda - L)n^{\text{tr}}}$ . Using the same technique in Theorem 3. Then we complete the proof.

## 1068 D.6 Fo-MuML(convex)

1069 Fo-MuML is more like an application of FoMAML in PDF. In particular, in the outer-level, we no  
1070 longer derive the derivative of  $w$ , we take the derivative of  $w_{\mathcal{T}_i, Q}$ . Then we have

$$\begin{aligned} \|w^{t+1} - \tilde{w}^{t+1}\| &= \|(w^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_j, Q}(w^t, S_j^{\text{tr}}, S_j^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_j, Q}(\tilde{w}^t, S_j^{\text{tr}}, S_j^{\text{ts}})))\| \\ &\leq \frac{1}{n^{\text{ts}}} \sum_{z^{\text{ts}} \in S_j^{\text{ts}}} \|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_j, Q}(w^t, S_j^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_j, Q}(\tilde{w}^t, S_j^{\text{tr}}, z^{\text{ts}})))\| \\ &\leq \|w^t - \tilde{w}^t\| \end{aligned} \tag{76}$$

1071 Next, for a given time index  $t$ , with probability  $\frac{1}{m}$ , the task  $\mathcal{T}_i$  is selected. In this case, we have

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &= \mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i, Q}(w^t, S_i^{\text{tr}}, S_i^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i, Q}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}})))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i, Q}(w^t, S_i^{\text{tr}}, S_i^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i, Q}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, S_i^{\text{ts}})))\| \\ &\quad + \mathbb{E}\|\eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i, Q}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}})) - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i, Q}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, S_i^{\text{ts}}))\| \\ &\leq \frac{1}{n^{\text{ts}}} \sum_{z^{\text{ts}} \in S_i^{\text{ts}}} \mathbb{E}\|(w^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i, Q}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i, Q}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})))\| \\ &\quad + \eta_t \mathbb{E}\|\nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i, Q}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}})) - \nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i, Q}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, S_i^{\text{ts}}))\| \\ &\leq \frac{1}{n^{\text{ts}}} \sum_{z \in S_i^{\text{ts}}} \mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i, Q}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i, Q}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})))\| \\ &\quad + 2\eta_t \mathbb{E}\|\nabla \hat{F}_i(w)\|, \end{aligned} \tag{77}$$

1072 where the last inequality follows that  $\tilde{S}_i^{\text{ts}}$  and  $S_i^{\text{ts}}$  are sampled from the same distribution, then

1073  $\mathbb{E}\|\nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, \tilde{S}_i^{\text{ts}}))\| = \mathbb{E}\|\nabla \hat{\mathcal{L}}(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, S_i^{\text{ts}}))\| = \mathbb{E}\|\nabla \hat{F}_i(w)\|$ . Note that

$$\begin{aligned} &\mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i, Q}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}})))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i, Q}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i, Q}(\tilde{w}^t, S_i^{\text{tr}}, z^{\text{ts}})))\| \\ &\quad + \eta_t \mathbb{E}\|\nabla \ell(w_{\mathcal{T}_i, Q}(\tilde{w}^t, S_i^{\text{tr}}, z^{\text{ts}})) - \nabla \ell(w_{\mathcal{T}_i, Q}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}, z^{\text{ts}}))\|. \end{aligned} \tag{78}$$

1074 Let us bound the two terms on the RHS of (78) separately. First, similar to how we derived (76), we  
1075 could bound the first term by

$$\mathbb{E}\|(w^t - \eta_t \nabla \ell(w_{\mathcal{T}_i, Q}(w^t, S_i^{\text{tr}}, z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla \ell(w_{\mathcal{T}_i, Q}(\tilde{w}^t, S_i^{\text{tr}}, z^{\text{ts}})))\| \leq \mathbb{E}\|w^t - \tilde{w}^t\|.$$

1076 To bound the second term on the RHS of (78), we consider two parallel processes of generating  
 1077 iterates  $\{\tilde{w}_{\mathcal{T}_i,q}^t\}$  and  $\{\tilde{w}_{\mathcal{T}_i,q}^{t'}\}$  by using datasets  $S_i^{\text{tr}}$  and  $\tilde{S}_i^{\text{tr}}$ , respectively. Note that

$$\begin{aligned} \mathbb{E}\|\nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}) - \nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}})\| &= \mathbb{E}\|\nabla\ell(\tilde{w}_{\mathcal{T}_i,Q}^t, z^{\text{ts}}) - \nabla\ell(\tilde{w}_{\mathcal{T}_i,Q}^{t'}, z^{\text{ts}})\| \\ &\leq L\mathbb{E}\|\tilde{w}_{\mathcal{T}_i,Q}^t - \tilde{w}_{\mathcal{T}_i,Q}^{t'}\| \\ &\leq \frac{2G^2}{\lambda n^{\text{tr}}}, \end{aligned} \quad (79)$$

1078 Substituting (79) and (78) into (77), we have

$$\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \leq \mathbb{E}\|w^t - \tilde{w}^t\| + \eta_t \frac{2G^2}{\lambda n^{\text{tr}}} + 2\eta_t \mathbb{E}\|\nabla\hat{F}_i(w^t)\|.$$

1079 Combing the above two cases, we obtain

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &\leq (1 - \frac{1}{m})\mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m}\mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m}\eta_t \frac{2G^2}{\lambda n^{\text{tr}}} + \frac{2}{m}\eta_t \mathbb{E}\|\nabla\hat{F}_i(w^t)\| \\ &= \mathbb{E}\|w^t - \tilde{w}^t\| + \eta_t \frac{2G^2}{\lambda m n^{\text{tr}}} + \frac{2\eta_t}{m}\mathbb{E}\|\nabla\hat{F}_i(w^t)\|. \end{aligned}$$

1080 Unrolling it and noting that  $\|w^0 - \tilde{w}^0\| = 0$ , we have

$$\begin{aligned} \mathbb{E}[\|w^T - \tilde{w}^T\|] &\leq \sum_{t=0}^{T-1} \eta_t \frac{2G^2}{\lambda m n^{\text{tr}}} + \sum_{t=0}^{T-1} \frac{2\eta_t}{m} \mathbb{E}\|\nabla\hat{F}(w^t, S_i)\| \\ &\leq \sum_{t=0}^{T-1} \eta_t \frac{2G^2}{\lambda m n^{\text{tr}}} + \frac{1}{m} \sqrt{F(w^0) - \min_{\mathcal{W}} F + \frac{L_Q \sigma^2}{2} \sum_{t=0}^{T-1} \eta_t^2}, \end{aligned}$$

1081 where we use Lemma 13 in the last inequality. which completes the proof.

## 1082 D.7 FoMuML(non-convex)

1083 This proof is similar to the proof of Theorem 3, we have

$$\begin{aligned} \|w^{t+1} - \tilde{w}^{t+1}\| &= \|(w^t - \eta_t \nabla\hat{\mathcal{L}}(w_{\mathcal{T}_j}(w^t, S_j^{\text{tr}}), S_j^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla\hat{\mathcal{L}}(w_{\mathcal{T}_j}(\tilde{w}^t, S_j^{\text{tr}}), S_j^{\text{ts}}))\| \\ &\leq (1 + \eta_t \phi_t) \|w^t - \tilde{w}^t\|, \end{aligned} \quad (80)$$

1084 where  $\phi_t = \min\{L_Q, \xi_t\}$  with  $\xi_t = \|\nabla^2 F(w_0, S_t)\| + \frac{\rho_Q}{2} \left\| \sum_{l=1}^{t-1} \beta_l \nabla\hat{F}(w_S^l) \right\| +$   
 1085  $\frac{\rho_Q}{2} \left\| \sum_{l=1}^{t-1} \beta_l \nabla\hat{F}(w_{\tilde{S}}^l) \right\|$ .

1086 Next, for the second case, similar to the proof of (77), we have

$$\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \leq \frac{1}{n^{\text{ts}}} \sum_{z \in S_i^{\text{ts}}} \mathbb{E}\|(w^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}}))\| + 2\eta_t \mathbb{E}\|\nabla\hat{F}_i(w)\|. \quad (81)$$

1087 Note that

$$\begin{aligned} &\mathbb{E}\|(w^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}}))\| \\ &\leq \mathbb{E}\|(w^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}))\| + \eta_t \mathbb{E}\|\nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}) - \nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}})\|. \end{aligned} \quad (82)$$

1088 Let us bound the two terms on the RHS of (82), separately. First, similar to how we bound (80), we  
 1089 could bound the first term by

$$\mathbb{E}\|(w^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(w^t, S_i^{\text{tr}}), z^{\text{ts}})) - (\tilde{w}^t - \eta_t \nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}))\| \leq (1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\|.$$

1090 To bound the second term on the RHS of (82), similar to the proof of (79), we have

$$\mathbb{E}\|\nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, S_i^{\text{tr}}), z^{\text{ts}}) - \nabla\ell(w_{\mathcal{T}_i}(\tilde{w}^t, \tilde{S}_i^{\text{tr}}), z^{\text{ts}})\| \leq \frac{2G^2}{(\lambda - L)n^{\text{tr}}} \quad (83)$$

1091 Substituting (83) and (82) into (81), we obtain

$$\mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| \leq (1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \eta_t \frac{2G^2}{(\lambda - L)n^{\text{tr}}} + 2\eta_t \mathbb{E}\|\nabla \hat{F}_i(w^t)\|.$$

1092 From Lemma 7, we can know  $\mathbb{E}\|\nabla \hat{F}_i(w^t)\| \leq eG$ . Then combining the above two cases, we obtain

$$\begin{aligned} \mathbb{E}\|w^{t+1} - \tilde{w}^{t+1}\| &\leq (1 - \frac{1}{m})(1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{1}{m}(1 + \eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| \\ &\quad + \frac{2G^2}{(\lambda - L)mn^{\text{tr}}} + \frac{2\eta_t eG}{m} \\ &\leq \exp(\eta_t \phi_t) \mathbb{E}\|w^t - \tilde{w}^t\| + \frac{\eta_t \Phi}{m}. \end{aligned}$$

1093 where we use  $1 + x \leq \exp(x)$  and  $\alpha \leq \frac{1}{QL}$  in the second inequality, and we denote  $\Phi = 2eG +$

1094  $\frac{2G^2}{(\lambda - L)n^{\text{tr}}}$ . Using the same technique in Theorem 3. Then we complete the proof.