IEEE Open Journal of Signal Processing

Received 7 June 2021; revised 9 December 2021; accepted 10 December 2021. Date of publication 4 January 2022; date of current version 31 January 2022. The review of this article was arranged by Associate Editor D. Orlando.

Digital Object Identifier 10.1109/OJSP.2021.3140000

Dif-MAML: Decentralized Multi-Agent Meta-Learning

MERT KAYAALP [1] (Graduate Student Member, IEEE), STEFAN VLASKI [2] (Member, IEEE), AND ALI H. SAYED [1] (Fellow, IEEE)

[1]Adaptive Systems Laboratory, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland
[2]Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K.

CORRESPONDING AUTHOR: MERT KAYAALP (e-mail: mert.kayaalp@epfl.ch)

An earlier version of this paper was presented at EUSIPCO 2021 [DOI: 10.23919/EUSIPCO54536.2021.9616256].

ABSTRACT The objective of meta-learning is to exploit knowledge obtained from observed tasks to improve adaptation to unseen tasks. Meta-learners are able to generalize better when they are trained with a larger number of observed tasks and with a larger amount of data per task. Given the amount of resources that are needed, it is generally difficult to expect the tasks, their respective data, and the necessary computational capacity to be available at a single central location. It is more natural to encounter situations where these resources are spread across several agents connected by some graph topology. The formalism of meta-learning is actually well-suited for this decentralized setting, where the learner benefits from information and computational power spread across the agents. Motivated by this observation, we propose a cooperative fully-decentralized multi-agent meta-learning algorithm, referred to as Diffusion-based MAML or Dif-MAML. Decentralized optimization algorithms are superior to centralized implementations in terms of scalability, robustness, avoidance of communication bottlenecks, and privacy guarantees. The work provides a detailed theoretical analysis to show that the proposed strategy allows a collection of agents to attain agreement at a linear rate and to converge to a stationary point of the *aggregate* MAML objective even in non-convex environments. Simulation results illustrate the theoretical findings and the superior performance relative to the traditional non-cooperative setting.

INDEX TERMS Decentralized optimization, diffusion algorithm, distributed learning, learning to learn, meta-learning, multi-agent systems, networked agents.

I. INTRODUCTION

Training of highly expressive learning architectures, such as deep neural networks, requires large amounts of data in order to ensure high generalization performance. However, the generalization guarantees apply only to test data following the same distribution as the training data. Human intelligence, on the other hand, is characterized by a remarkable ability to leverage prior knowledge to accelerate adaptation to new tasks. This evident gap has motivated a growing number of works on learning architectures that *learn to learn* (see [2] for a recent survey).

The work [3] proposed a model-agnostic meta-learning (MAML) approach, which is an initial parameter-transfer methodology where the goal is to learn a good "*launch model*". Several works have extended and/or analyzed this approach to great effect such as [4]–[11]. Furthermore, some works used MAML for signal processing applications such as image segmentation [12], speech recognition [13], and demodulation [14]. However, there does not appear to exist works that consider model agnostic meta-learning in a decentralized multi-agent setting. This setting is very natural to consider for meta-learning, where different agents can be assumed to have local meta-learners based on their own experiences. Interactions with neighbors can help infuse their models with new information and speed up adaptation to new tasks.

Decentralized multi-agent systems consist of a collection of agents with access to data and computational capabilities, and a graph topology that imposes constraints on peer-to-peer communications. In contrast to centralized architectures,

which require some central aggregation of data, decentralized solutions rely solely on the diffusion of information over connected graphs through successive local aggregations over neighborhoods. While decentralized methods have been shown to be capable of matching the performance of centralized solutions [15], [16], the absence of a fusion center is advantageous in the presence of communication bottlenecks, and concerns over robustness or privacy. Applications that can benefit from decentralized meta-learning algorithms include but are not limited to the following:

- A robotic swarm might be assigned to do environmental monitoring [17]. The individual robots can share spatially and temporally dispersed data such as images or temperatures in order to learn better meta-models to adapt to new scenes. This teamwork is vital for circumstances where data collection is hard, such as natural disasters.
- Different hospitals or research groups can work on clinical risk prediction with limited patient health records [18] or drug discovery with small amount of data [19]. The individual agents in this context will benefit from cooperation, while avoiding the need for a central hub in order to preserve the privacy of medical data.
- In some situations, it is advantageous to distribute a single agent problem over multiple agents. For example, training a MAML can be computationally demanding since it requires Hessian calculations [3]. In order to speed up the process, tasks can be divided into different workers or machines.

The contributions in this paper are three-fold:

- By combining MAML with the diffusion strategy for decentralized stochastic optimization [16], we propose Diffusion-based Model-Agnostic Meta-Learning (Dif-MAML). The result is a decentralized algorithm for meta-learning over a collection of distributed agents, where each agent is provided with tasks stemming from potentially different task distributions.
- We establish that, despite the decentralized nature of the algorithm, all agents agree quickly on a common launch model, which subsequently converges to a stationary point of the *aggregate* MAML objective over the task distribution across the network. This implies that Dif-MAML matches the performance of a centralized solution, which would have required central aggregation of data stemming from *all tasks across the network*. In this way, agents will not only learn from locally observed tasks to accelerate future adaptation, but will also *learn from each other*, and from tasks seen by the other agents.
- We illustrate through numerical experiments across a number of benchmark datasets that Dif-MAML outperforms the traditional non-cooperative solution and matches the performance of the centralized solution.

Notation: We denote random variables in bold. Single data points are denoted by small letters like x and batches of data are denoted by big calligraphic letters like $\mathcal{X}$. $\mathbb{1}_K$ denotes a $K \times 1$ vector with all entries equal to one. The Kronecker product is denoted by $\otimes$. col{} stacks its arguments on top of each other. To refer to a loss function evaluated at a batch $\mathcal{X}$ with elements $\{x_n\}_{n=1}^N$, we use the notation $Q(w;\mathcal{X}) \triangleq \frac{1}{N}\sum_{n=1}^N Q(w;x_n)$, where w denotes the model parametrization (such as the parameters of a neural network). To denote expectation with respect to task-specific data, we use $\mathbb{E}_{x^{(t)}}$, where t corresponds to the task. This is an expectation over the distribution of $x^{(t)}$, and it is conditioned on every other random variable.

A. PROBLEM FORMULATION

We consider a collection of K agents (e.g., robots, workers, machines, processors) where each agent k is provided with data stemming from tasks in a set $\mathcal{T}_k$. We denote the probability distribution over $\mathcal{T}_k$ by π_k, i.e., the probability of drawing task t from $\mathcal{T}_k$ is $\pi_k(t)$. In principle, for any particular task $t \in \mathcal{T}_k$, each agent could learn a separate model $w_k^{o(t)}$ by solving:

$$w_k^{o(t)} \triangleq \arg\min_{w\in\mathbb{R}^M} J_k^{(t)}(w) \triangleq \arg\min_{w\in\mathbb{R}^M} \mathbb{E}_{x_k^{(t)}} Q_k^{(t)}\left(w;\boldsymbol{x}_k^{(t)}\right) \quad (1)$$

where $\boldsymbol{x}_k^{(t)}$ denotes the random data corresponding to task t observed at agent k. The loss $Q_k^{(t)}(w;\boldsymbol{x}_k^{(t)})$ denotes the penalization encountered by w under the random data $\boldsymbol{x}_k^{(t)}$, while $J_k^{(t)}(w)$ represents the *stochastic* risk. Note that the expectation in (1) is with respect to random data of a particular task t, i.e., within-task uncertainty.

Instead of training separately in this manner, meta-learning presumes an *a priori* relation between the tasks in $\mathcal{T}_k$ and exploits this fact. In particular, MAML seeks a "*launch model*" such that when faced with data arising from a previously unseen task, the agent would be able to update the "launch model" with a small number of task-specific gradient updates. It is common to allow for multiple gradient steps for task adaptation. For the analytical part of this work, we will restrict ourselves to a single gradient step for simplicity. Nevertheless, our experimental results suggest that the theoretical conclusions hold more broadly even when allowing for multiple gradient updates to the launch model. With a single gradient step, agent k can seek a launch model by minimizing the average risk over all tasks evaluated at an adjusted argument:

$$\min_{w\in\mathbb{R}^M} \overline{J}_k(w) \triangleq \mathbb{E}_{t\sim\pi_k} J_k^{(\boldsymbol{t})}\left(w - \alpha\nabla J_k^{(\boldsymbol{t})}(w)\right) \quad (2)$$

where $\alpha > 0$ is a step size parameter. In effect, this step amounts to minimizing the expected risk at a look-ahead step across all tasks. Observe that the expectation in (2) is with respect to distribution π_k over the agent-specific collection of tasks $\mathcal{T}_k$. The resulting gradient vector is given by (assuming the possibility of exchanging expectations and gradient operations, which is valid under mild technical conditions):

$$\nabla\overline{J}_k(w) \triangleq$$
$$\mathbb{E}_{t\sim\pi_k}\left[\left(I - \alpha\nabla^2 J_k^{(\boldsymbol{t})}(w)\right)\nabla J_k^{(\boldsymbol{t})}\left(w - \alpha\nabla J_k^{(\boldsymbol{t})}(w)\right)\right] \quad (3)$$

In practice, due to the lack of information about π_k and the distribution of $\boldsymbol{x}_k^{(t)}$, evaluation of (2) and (3) is not feasible. It is common to collect data realizations and replace (3) by a stochastic gradient approximation:

$$\nabla\overline{Q_k}(w) \triangleq \frac{1}{|\mathcal{S}_k|}\sum_{\boldsymbol{t}\in\boldsymbol{\mathcal{S}}_k}\Big[\Big(I-\alpha\nabla^2 Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\Big) \times \nabla Q_k^{(\boldsymbol{t})}\Big(w-\alpha\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)});\ \boldsymbol{\mathcal{X}}_o^{(t)}\Big)\Big] \tag{4}$$

where $\boldsymbol{\mathcal{X}}_{in}^{(t)}$, $\boldsymbol{\mathcal{X}}_o^{(t)}$ are two independently-selected random batches of data, $\boldsymbol{\mathcal{S}}_k \subset \mathcal{T}_k$ is a random batch of tasks, and $|\mathcal{S}_k|$ is the number of selected tasks. Recall the notation:

$$\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}) = \frac{1}{|\boldsymbol{\mathcal{X}}_{in}^{(t)}|}\sum_{n=1}^{|\boldsymbol{\mathcal{X}}_{in}^{(t)}|}\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{x}_n^{(t)}) \tag{5}$$

where the batch $\boldsymbol{\mathcal{X}}_{in}^{(t)}$ consists of $|\boldsymbol{\mathcal{X}}_{in}^{(t)}|$ number of elements $\{\boldsymbol{x}_n^{(t)}\}_{n=1}^{|\boldsymbol{\mathcal{X}}_{in}^{(t)}|}$. A similar definition holds for the Hessian. We assume that all elements of $\boldsymbol{\mathcal{X}}_{in}^{(t)}$, $\boldsymbol{\mathcal{X}}_o^{(t)}$ are independently sampled from the distribution of $\boldsymbol{x}_k^{(t)}$ and all tasks $\boldsymbol{t}\in\boldsymbol{\mathcal{S}}_k$ are independently sampled from $\mathcal{T}_k$.

In a *non-cooperative* MAML setting, each agent k would optimize (2) in an effort to obtain a launch model that is likely to adapt quickly to tasks similar to those encountered in $\mathcal{T}_k$. In a *cooperative* multi-agent setting, however, one would expect transfer learning to occur between agents. This motivates us to seek a decentralized scheme where the launch model obtained by agent k is likely to generalize well to tasks similar to those observed by agent ℓ during training, for any pair of agents k,ℓ. This can be achieved by pursuing a launch model that optimizes instead the aggregate risk:

$$\min_{w\in\mathbb{R}^M}\overline{J}(w) \triangleq \frac{1}{K}\sum_{k=1}^{K}\overline{J}_k(w) \tag{6}$$

By pursuing this network objective in place of the individual objectives, the effective number of tasks and data each agent is trained on is increased and hence a better generalization performance is expected. Even though both the centralized and decentralized strategies seek a solution to (6), in the decentralized strategy, the agents rely only on their immediate neighbors and there is no central processor.

B. RELATED WORK

Early works on meta-learning or learning to learn date back to [20]–[23]. Recently, there has been increased interest in meta-learning with various approaches such as learning an optimization rule [24], [25] or learning a metric that compares support and query samples for few-shot classification [26], [27].

In this paper, we consider a parameter-initialization-based meta-learning algorithm. This kind of approach was introduced by MAML [3], which aims to find a good initialization (launch model) that can be adapted to new tasks rapidly. It is model-agnostic, which means it can be applied to any model that is trained with gradient descent. MAML has shown competitive performance on benchmark few-shot learning tasks. Many algorithmic extensions have also been proposed by [4]–[7] and several works have focused on the theoretical analysis and convergence of MAML [8]–[11] in single-agent settings.

A different line of work [28]–[31] studies meta-learning in a federated setting where the agents communicate with a central processor in a manner that keeps the privacy of their data. In particular, [30] and [31] propose algorithms that learn a global shared launch model, which can be updated by a few agent-specific gradients for personalized learning. In contrast, we consider a decentralized scheme where there is no central node and only *localized* communications with neighbors occur. This leads to a more scalable and flexible system and avoids communication bottleneck at the central processor.

Our extension of MAML is based on the diffusion algorithm for decentralized optimization [16], [32]. While there exist other useful decentralized optimization strategies that are based on primal-dual methods [33], alternating direction method of multipliers [34], or consensus [35]–[37], diffusion strategies have been shown to be particularly suitable for adaptive scenarios where the solutions need to adapt to drifts in the data and models. Diffusion strategies have also been shown to lead to wider stability ranges and lower mean-square-error performance than other techniques in the context of adaptation and learning due to an inherent symmetry in their structure. Several works analyzed the performance of diffusion algorithms such as [32], [38]–[40]. The works [41], [42] examined diffusion under non-convex losses and stochastic gradient conditions, which are applicable to our work but only after proper adjustment in order to account for the fact that the risk function for MAML includes a gradient term as part of the argument for the risk function.

II. DIF-MAML

Our algorithm is based on the Adapt-then-Combine variant of the diffusion strategy [16].

A. DIFFUSION (ADAPT-THEN-COMBINE)

The diffusion strategy is applicable to scenarios where K agents, connected via a graph topology $A=[a_{\ell k}]$ (see Fig. 1), collectively try to minimize an aggregate risk $\overline{J}(w)\triangleq\frac{1}{K}\sum_{k=1}^{K}\overline{J}_k(w)$, which includes the setting (6) considered in this work. To solve this objective, at every iteration i, each agent k simultaneously performs the following steps:

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\nabla\overline{Q_k}(\boldsymbol{w}_{k,i-1}) \tag{7a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell=1}^{K} a_{\ell k}\boldsymbol{\phi}_{\ell,i} \tag{7b}$$

The coefficients $\{a_{\ell k}\}$ are non-negative and add up to one:

$$\sum_{\ell=1}^{K} a_{\ell k} = 1,\ \ a_{\ell k} > 0 \ \text{ if agents } \ell \text{ and } k \text{ are connected}$$

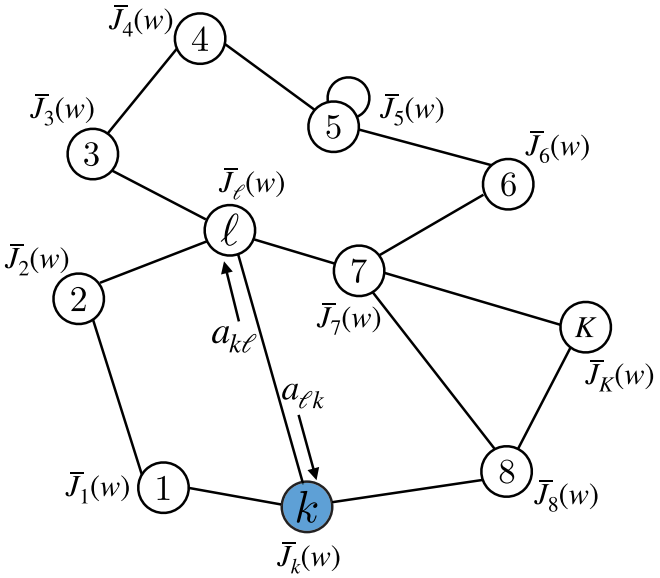

FIGURE 1. An example of a graph topology and interactions between the agents.

For example, the coefficients can be set using the Metropolis rule [43].

Expression (7a) is an *adaptation* step where all agents simultaneously obtain intermediate states $\boldsymbol{\phi}_{k,i}$ by a stochastic gradient update. Recall that $\nabla\overline{Q_k}(\boldsymbol{w}_{k,i-1})$ from (4) is the stochastic approximation of the exact gradient $\nabla\overline{J_k}(\boldsymbol{w}_{k,i-1})$ from (3). Expression (7b) is a *combination* step where the agents combine their neighbors' intermediate steps to obtain updated iterates $\boldsymbol{w}_{k,i}$.

B. DIFFUSION-BASED MAML (DIF-MAML)

We present the proposed algorithm for decentralized meta-learning in Algorithm 1. A visual representation of it is provided in Fig. 2. Each agent is assigned an initial launch model. At every iteration, the agents sample a batch of i.i.d. tasks from their agent-specific distribution of tasks. Then, in the inner loop, task-specific models are found by applying task-specific stochastic gradients to the launch models. Subsequently, in the outer loop, each agent computes an intermediate state for its launch model based on an update consisting of the sampled batch of tasks. A standard MAML algorithm would assign the intermediate states as the revised launch models and stop there, without any cooperation among the agents. However, in Dif-MAML, the agents cooperate and update their launch models by combining their intermediate states with the intermediate states of their neighbors. This helps in the transfer of knowledge among agents.

III. THEORETICAL RESULTS

In this section, we provide convergence analysis for Dif-MAML in non-convex environments. We start by listing conditions that are commonplace in the analysis of learning algorithms under such scenarios.

Algorithm 1: Dif-MAML.

Initialize the *launch* models $\{w_{k,0}\}_{k=1}^K$
while not done **do**
  for all agents **do**
    Agent k samples a batch of $i.i.d.$ tasks $\boldsymbol{\mathcal{S}}_{k,i}$ from $\mathcal{T}_k$
    for all tasks $\boldsymbol{t} \in \boldsymbol{\mathcal{S}}_{k,i}$ **do**
      Evaluate $\nabla Q_k^{(\boldsymbol{t})}(\boldsymbol{w}_{k,i-1}; \boldsymbol{\mathcal{X}}_{in,i}^{(t)})$ using a batch of $i.i.d.$ data $\boldsymbol{\mathcal{X}}_{in,i}^{(t)}$
      Set *task-specific* models $\boldsymbol{w}_{k,i}^{(t)} = \boldsymbol{w}_{k,i-1} - \alpha\nabla Q_k^{(\boldsymbol{t})}(\boldsymbol{w}_{k,i-1}; \boldsymbol{\mathcal{X}}_{in,i}^{(t)})$
    end for
    Compute *intermediate states* $\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - (\mu/|\mathcal{S}_{k,i}|)\sum_{\boldsymbol{t}\in\boldsymbol{\mathcal{S}}_{k,i}} \nabla Q_k^{(\boldsymbol{t})}(\boldsymbol{w}_{k,i}^{(t)}; \boldsymbol{\mathcal{X}}_{o,i}^{(t)})$ using a batch of $i.i.d.$ data $\boldsymbol{\mathcal{X}}_{o,i}^{(t)}$ for each task (The gradient here is with respect to the task-specific model — see (4) for the gradient expression with respect to the launch model explicitly.)
  end for
  for all agents **do**
    Update the launch models by combining the intermediate states $\boldsymbol{w}_{k,i} = \sum_{\ell=1}^K a_{\ell k}\boldsymbol{\phi}_{\ell,i}$
  end for
  $i \leftarrow i+1$
end while

A. ASSUMPTIONS

Assumption 1 (Lipschitz gradients): For each agent k and task $t \in \mathcal{T}_k$, the gradient $\nabla Q_k^{(t)}(\cdot;\cdot)$ is Lipschitz, namely, for any $w, u \in \mathbb{R}^M$ and $\boldsymbol{x}_k^{(t)}$ denoting a data point:

$$\left\|\nabla Q_k^{(t)}\left(w;\boldsymbol{x}_k^{(t)}\right) - \nabla Q_k^{(t)}\left(u;\boldsymbol{x}_k^{(t)}\right)\right\| \le L^{\boldsymbol{x}_k^{(t)}}\|w-u\| \quad (8)$$

We assume the second-order moment of the Lipschitz constant is bounded by a data-independent constant:

$$\mathbb{E}_{x_k^{(t)}}\left(L^{\boldsymbol{x}_k^{(t)}}\right)^2 \le \left(L_k^{(t)}\right)^2 \quad (9)$$

□

We establish in Appendix A that under (8) and (9), a similar property will also hold for gradients involving a batch of data. In this paper, for simplicity, we will mostly work with $L \triangleq \max_k \max_t L_k^{(t)}$.

Assumption 2 (Lipschitz Hessians): For each agent k and task $t \in \mathcal{T}_k$, the Hessian $\nabla^2 Q_k^{(t)}(\cdot;\cdot)$ is Lipschitz in expectation, namely, for any $w, u \in \mathbb{R}^M$ and $\boldsymbol{x}_k^{(t)}$ denoting a data point:

$$\mathbb{E}_{x_k^{(t)}}\left\|\nabla^2 Q_k^{(t)}\left(w;\boldsymbol{x}_k^{(t)}\right) - \nabla^2 Q_k^{(t)}\left(u;\boldsymbol{x}_k^{(t)}\right)\right\| \le \rho_k^{(t)}\|w-u\| \quad (10)$$

□

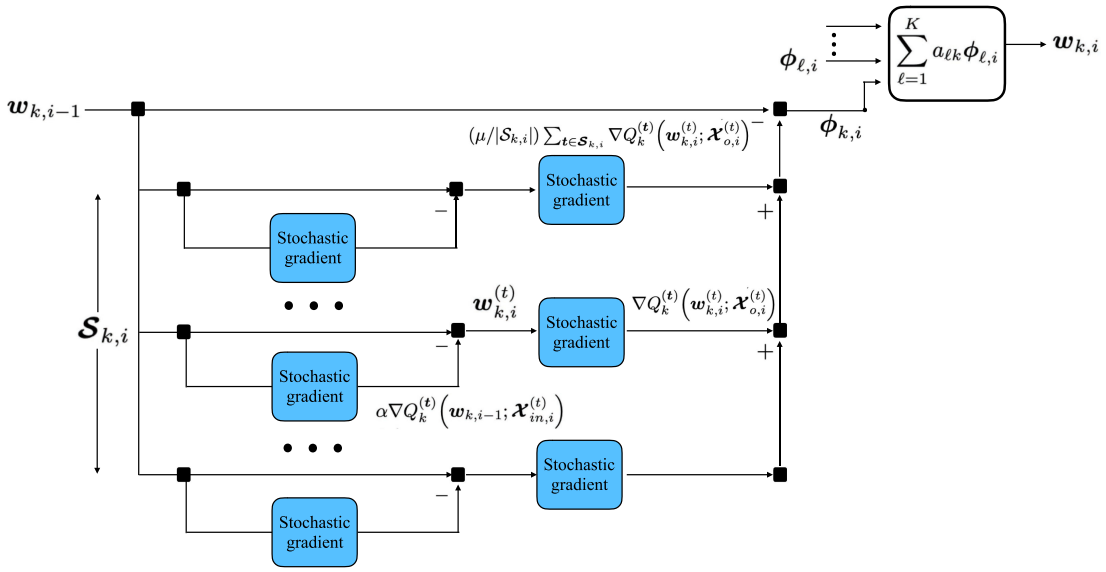

FIGURE 2. Pictorial representation of Algorithm 1.

We establish in Appendix B that under (10), a similar relation holds for Hessians involving a batch of data. In this paper, for simplicity, we will mostly work with $\rho \triangleq \max_k \max_t \rho_k^{(t)}$.

Assumption 3 (Bounded gradients): For each agent k and task $t \in \mathcal{T}_k$, the gradient $\nabla Q_k^{(t)}(\cdot;\cdot)$ is bounded in expectation, namely, for any $w \in \mathbb{R}^M$ and $\boldsymbol{x}_k^{(t)}$ denoting a data point:

$$\mathbb{E}_{x_k^{(t)}} \left\| \nabla Q_k^{(t)}\left(w;\boldsymbol{x}_k^{(t)}\right)\right\| \leq B_k^{(t)} \tag{11}$$

□

We establish in Appendix C that under (11), a similar relation holds for gradients involving a batch of data. In this paper, for simplicity, we will mostly work with $B \triangleq \max_k \max_t B_k^{(t)}$.

Assumption 4 (Bounded noise moments): For each agent k and task $t \in \mathcal{T}_k$, the gradient $\nabla Q_k^{(t)}(\cdot;\cdot)$ and the Hessian $\nabla^2 Q_k^{(t)}(\cdot;\cdot)$ have bounded fourth-order central moments, namely, for any $w \in \mathbb{R}^M$:

$$\mathbb{E}_{x_k^{(t)}} \left\| \nabla Q_k^{(t)}\left(w;\boldsymbol{x}_k^{(t)}\right) - \nabla J_k^{(t)}(w)\right\|^4 \leq \sigma_G^4 \tag{12}$$

$$\mathbb{E}_{x_k^{(t)}} \left\| \nabla^2 Q_k^{(t)}\left(w;\boldsymbol{x}_k^{(t)}\right) - \nabla^2 J_k^{(t)}(w)\right\|^4 \leq \sigma_H^4 \tag{13}$$

□

We establish in Appendix D that under (12) and (13), similar relations hold for gradients and Hessians involving a batch of data.

Denoting the mean of the risk functions of the tasks in $\mathcal{T}_k$ by $J_k(w) \triangleq \mathbb{E}_{t\sim\pi_k} J_k^{(t)}(w)$, we introduce the following assumption on the relations between the tasks of a particular agent.

Assumption 5 (Bounded task variability): For each agent k, the gradient $\nabla J_k^{(t)}(\cdot)$ and the Hessian $\nabla^2 J_k^{(t)}(\cdot)$ have bounded fourth-order central moments, namely, for any $w \in \mathbb{R}^M$:

$$\mathbb{E}_{t\sim\pi_k} \left\| \nabla J_k^{(t)}(w) - \nabla J_k(w)\right\|^4 \leq \gamma_G^4 \tag{14}$$

$$\mathbb{E}_{t\sim\pi_k} \left\| \nabla^2 J_k^{(t)}(w) - \nabla^2 J_k(w)\right\|^4 \leq \gamma_H^4 \tag{15}$$

□

Note that we do not assume any constraint on the relations between tasks of different agents.

Assumption 6 (Doubly-stochastic combination matrix): The weighted combination matrix $A = [a_{\ell k}]$ representing the graph is doubly-stochastic and symmetric. This means that the matrix has non-negative elements and satisfies:

$$A\mathbb{1}_K = \mathbb{1}_K, \quad A = A^{\mathsf{T}} \tag{16}$$

We further assume that the matrix A is primitive, which means that a path with positive weights can be found between any arbitrary nodes (k, ℓ), and moreover at least one $a_{kk} > 0$ for some k. □

B. ALTERNATIVE MAML OBJECTIVE

The stochastic MAML gradient (4), because of the gradient within a gradient form, is not an *unbiased* estimator of (3). We consider the following alternative objective in place of (2):

$$\widehat{J}_k(w) \triangleq \mathbb{E}_{t\sim\pi_k} \mathbb{E}_{\mathcal{X}_{in}^{(t)}} J_k^{(t)}\left(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right) \tag{17}$$

The gradient corresponding to this objective is the expectation of the stochastic MAML gradient (4):

$$\nabla\widehat{J}_k(w) = \mathbb{E}\nabla\widehat{Q_k}(w) \tag{18}$$

For ease of reference, Table 1 lists the notation used in this paper. We establish (18) in Appendix E. This means that the stochastic MAML gradient (4) is an unbiased estimator for the gradient of the alternative objective (17).

While the MAML objective (2) captures the goal of coming up with a launch model that performs well after a *gradient step*, the *adjusted objective* (17) searches for a launch model that performs well after a *stochastic gradient step*. Using the

TABLE 1 Summary of Some Notation Used in the Paper

	Single-task	Meta-objective	Meta-gradient
Risk function	$J_k^{(t)}$	$\overline{J_k}$	$\nabla\overline{J_k}$
Adjusted Risk	–	$\widehat{J_k}$	$\nabla\widehat{J_k}$
Stochastic Approx.	$Q_k^{(t)}$	–	$\nabla\overline{Q_k}$

adjusted objective allows us to analyze the convergence of Dif-MAML by exploiting the fact that it results in an unbiased stochastic gradient approximation. This allows the use of standard non-convex decentralized optimization techniques.

In the following two lemmas, we will perform perturbation analyses on the MAML objective $\overline{J_k}(w)$ and the adjusted objective $\widehat{J}_k(w)$. We will work with $\widehat{J}_k(w)$ afterwards. At the end of our theoretical analysis, we will use the perturbation results to establish convergence to stationary points for both objectives.

Lemma 1 (Objective perturbation bound): Under assumptions 1,3,4, for each agent k, the disagreement between $\overline{J_k}(\cdot)$ and $\widehat{J}_k(\cdot)$ is bounded, namely, for any $w \in \mathbb{R}^M$:[1]

$$\left|\overline{J_k}(w) - \widehat{J}_k(w)\right| \leq \frac{\alpha^2 L \sigma_G^2}{2|\mathcal{X}_{in}|} + \frac{B\alpha\sigma_G}{\sqrt{|\mathcal{X}_{in}|}} \tag{19}$$

Proof: See Appendix F. ■

Next, we perform a perturbation analysis at the gradient level.

Lemma 2 (Gradient perturbation bound): Under assumptions 1,3,4, for each agent k, the disagreement between $\nabla\overline{J_k}(\cdot)$ and $\nabla\widehat{J}_k(\cdot)$ is bounded, namely, for any $w \in \mathbb{R}^M$:

$$\left\|\nabla\overline{J_k}(w) - \nabla\widehat{J}_k(w)\right\| \leq (1+\alpha L)\frac{\alpha L\sigma_G}{\sqrt{|\mathcal{X}_{in}|}} + \frac{B\alpha\sigma_H}{\sqrt{|\mathcal{X}_{in}|}} \tag{20}$$

Proof: See Appendix G. ■

Lemmas 1 and 2 suggest that the standard MAML objective and the adjusted objective get closer to each other with decreasing inner learning rate α and increasing inner batch size $|\mathcal{X}_{in}|$. This is intuitive since the adjusted objective is a look-ahead risk under stochastic gradient update whereas the standard MAML objective is under gradient update. Stochastic gradient under a batch of data gets closer to gradient with increasing batch size. Next, we establish some properties of the adjusted objective, which will be called upon in the analysis and will let us use the standard techniques for non-convex optimization.

Lemma 3 (Bounded gradient of adjusted objective): Under assumptions 1,3, for each agent k, the gradient $\nabla\widehat{J}_k(\cdot)$ of the adjusted objective is bounded, namely, for any $w \in \mathbb{R}^M$:

$$\left\|\nabla\widehat{J}_k(w)\right\| \leq \widehat{B} \tag{21}$$

where $\widehat{B} \triangleq (1+\alpha L)B$ is a non-negative constant.

[1]In this paper, for simplicity, we assume that for each agent k and task $t \in \mathcal{T}_k$, $|\boldsymbol{\mathcal{X}}_{in}^{(t)}| = |\mathcal{X}_{in}|$ and $|\boldsymbol{\mathcal{X}}_{o}^{(t)}| = |\mathcal{X}_{o}|$.

Proof: See Appendix H. ■

Lemma 4 (Lipschitz gradient of adjusted objective): Under assumptions 1-3, for each agent k, the gradient $\nabla\widehat{J}_k(\cdot)$ of adjusted objective is Lipschitz, namely, for any $w, u \in \mathbb{R}^M$:

$$\left\|\nabla\widehat{J}_k(w) - \nabla\widehat{J}_k(u)\right\| \leq \widehat{L}\|w-u\| \tag{22}$$

where $\widehat{L} \triangleq (L(1+\alpha L)^2 + \alpha\rho B)$ is a non-negative constant.

Proof: See Appendix I. ■

Lemma 5 (Gradient noise for adjusted objective): Under assumptions 1-5, the gradient noise defined as $\nabla\overline{Q_k}(w) - \nabla\widehat{J}_k(w)$ is bounded for any $w \in \mathbb{R}^M$, namely:

$$\mathbb{E}\left\|\nabla\overline{Q_k}(w) - \nabla\widehat{J}_k(w)\right\|^2 \leq C^2 \tag{23}$$

for a non-negative constant C^2, whose expression is given in (130) in Appendix J.

Proof: See Appendix J. ■

The upper bound on the gradient noise, C, increases with parameters $\alpha, L, \sigma_G, \sigma_H, B, \gamma_G, \gamma_H$, and decreases with batch-sizes $|\mathcal{X}_{in}|, |\mathcal{X}_o|$.

C. EVOLUTION ANALYSIS

In this section, we analyze the Dif-MAML algorithm over the network. The analysis is similar to [41], [42]. We first prove that agents cluster around the network centroid in $O(\log\mu) = o(1/\mu)$ iterations, then show that this centroid reaches an $O(\mu)$-mean-square-stationary point in at most $O(1/\mu^2)$ iterations. Fig. 3 summarizes the analysis.

The network centroid is defined as $\boldsymbol{w}_{c,i} \triangleq \frac{1}{K}\sum_{k=1}^K \boldsymbol{w}_{k,i}$. It is an average of the agents' parameters. In the following theorem, we study the difference between the centroid launch model and the launch model for each agent k.

Theorem 1 (Network disagreement): Under assumptions 1-6, network disagreement between the centroid launch model and the launch models of each agent k is bounded after $O(\log\mu) = o(1/\mu)$ iterations, namely:

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left\|\boldsymbol{w}_{k,i} - \boldsymbol{w}_{c,i}\right\|^2 \leq \mu^2 \frac{\lambda_2^2}{(1-\lambda_2)^2}\left(\widehat{B}^2 + C^2\right) + O(\mu^3) \tag{24}$$

for

$$i \geq \frac{3\log\mu}{\log\lambda_2} + O(1) = o(1/\mu) \tag{25}$$

where λ_2 is the mixing rate of the combination matrix A, i.e., it is the spectral radius of $A^{\mathsf{T}} - \frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}}$.

Proof: See Appendix K. ■

In Theorem 1, we proved that the disagreement between the centroid launch model and agent-specific launch models is bounded after sufficient number of iterations. Therefore, we can use the centroid model as a deputy for all models and examine its convergence properties.

Theorem 2 (Stationary points of adjusted objective): In addition to assumptions 1-6, assume that $\widehat{J}(w)$ is bounded from below, i.e., $\widehat{J}(w) \geq \widehat{J}^o$. Then, the centroid launch model $\boldsymbol{w}_{c,i}$

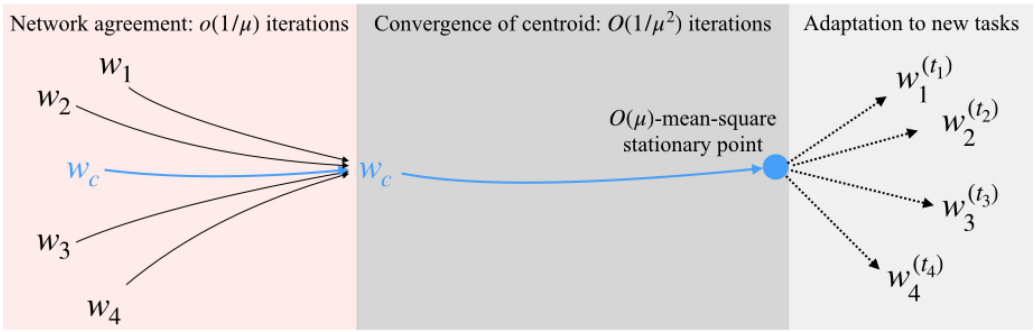

FIGURE 3. Diagram of the analysis. Agents cluster around a common network centroid, and this centroid reaches a stationary point of the MAML objective during meta-training. Subsequently, agents can use this launch model in order to adapt to new tasks.

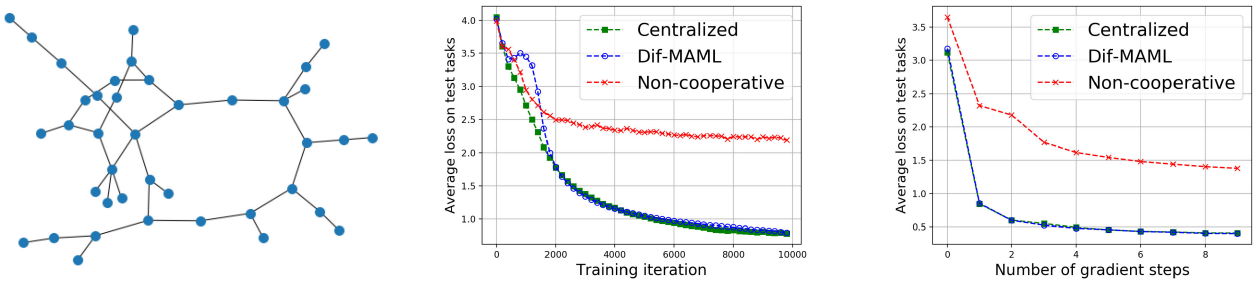

FIGURE 4. Regression. (a) 40-agent sparse network. (b) Test losses during training- Metropolis - Adam. (c) Test losses with respect to number of gradient steps after training.

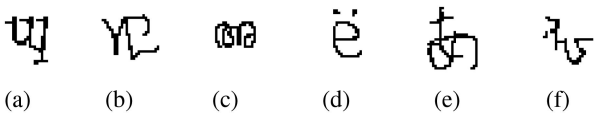

FIGURE 5. Omniglot dataset: Samples from six different characters.

FIGURE 6. MiniImagenet dataset: Samples from three different classes.

will reach an $O(\mu)$-mean-square-stationary point in at most $O(1/\mu^2)$ iterations. In particular, there exists a time instant $i^\star$ such that:

$$\mathbb{E}\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i^\star})\right\|^2 \leq 2\mu\widehat{L}C^2 + O(\mu^2) \tag{26}$$

and

$$i^\star \leq \left(\frac{2(\widehat{J}(w_0)-\widehat{J}^o)}{\widehat{L}C^2}\right)1/\mu^2 + O(1/\mu) \tag{27}$$

Proof: See Appendix L. ■

Next, we prove that the same analysis holds for the standard MAML objective, using the gradient perturbation bound for the adjusted objective (Lemma 2).

Corollary 1 (Stationary points of MAML objective): Assume that the same conditions of Theorem 2 hold. Then, the centroid launch model $\boldsymbol{w}_{c,i}$ will reach an $O(\mu)$-mean-square-stationary point, up to a constant, in at most $O(1/\mu^2)$ iterations. Namely, for time instant $i^\star$ defined in (27):

$$\mathbb{E}\left\|\nabla\overline{J}(\boldsymbol{w}_{c,i^\star})\right\|^2 \leq 4\mu\widehat{L}C^2 + O(\mu^2) + 2\left((1+\alpha L)\frac{\alpha L\sigma_G}{\sqrt{|\mathcal{X}_{in}|}} + \frac{B\alpha\sigma_H}{\sqrt{|\mathcal{X}_{in}|}}\right)^2 \tag{28}$$

Proof: See Appendix M. ■

Corollary 1 states that the centroid launch model can reach an $O(\mu)$-mean-square-stationary point for sufficiently small inner learning rate α and for sufficiently large inner batch size $|\mathcal{X}_{in}|$, in at most $O(1/\mu^2)$ iterations. Note that as $\mu \to 0$, the number of iterations required for network agreement ($O(\log\mu) = o(1/\mu)$) becomes negligible compared to the number of iterations necessary for convergence ($O(1/\mu^2)$). This follows from:

$$\lim_{\mu\to 0}\frac{1/\mu}{1/\mu^2} = \lim_{\mu\to 0}\mu = 0 \tag{29}$$

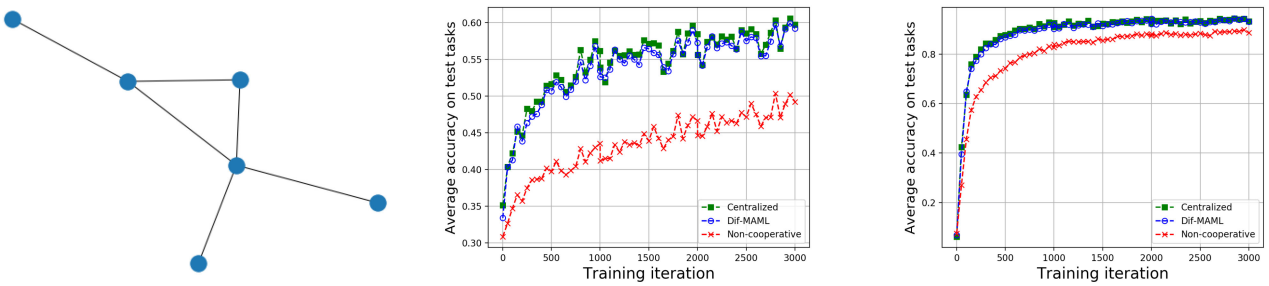

FIGURE 7. Classification. (a) The network. (b) Test losses during training - MiniImagenet 5-way 5-shot -Averaging Rule- SGD. (c) Test losses during training - Omniglot 20-way 1-shot - Metropolis- Adam.

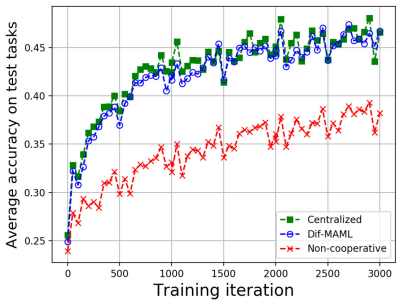

FIGURE 8. MiniImagenet 5-way 1-shot test accuracies during training process: Metropolis - Adam.

IV. EXPERIMENTS

In this section, we provide experimental evaluations. In particular, we present comparisons between the centralized, diffusion-based decentralized, and non-cooperative strategies. Our demonstrations cover both regression and classification tasks. Even though our theoretical analysis is general with respect to various learning models, for the experiments, our focus is on neural networks.

The centralized strategy corresponds to a central processor that has access to all data and tasks. Note that this is equivalent to having a network with a fully-connected graph in terms of loss/accuracy performance. The non-cooperative strategy represents a solution where agents do not communicate with each other. In other words, they all try to learn separate launch models.

A. REGRESSION

For regression, we consider the benchmark from [3]. In this setting, each task requires predicting the output of a sine wave from its input. Different tasks have different amplitudes and phases. Specifically, the phases are varied between $[0, \pi]$ for each agent. However, the agents have access to different task distributions since the amplitude interval $[0.1, 5.0]$ is evenly partitioned into $K = 40$ different intervals and each agent is equipped with one of them. The outer-loop optimization is based on Adam [44] and combination weights are set with Metropolis rule [43].

The same model architecture (a neural network with 2 hidden layers of 40 neurons with ReLu activations) is used for each agent. The loss function is the mean-squared error. As in [3], while training, 10 random points (10-shot) are chosen from each sinusoid and used with 1 stochastic gradient update ($\alpha = 0.01$). Adam optimizer is used with $\mu = 0.001$. Each agent is trained on 1000 tasks over 10 epochs (total number of iterations $=$ 10000). As in training, 10 data points from each sinusoid with 1 gradient update is used for adaptation.

Every 200th iteration, the agents are tested over 1000 tasks. All agents are evaluated with the same tasks, which stem from the intervals $[0.1, 5.0]$ for amplitude and $[0, \pi]$ for phase. The results are shown in Fig. 4(b). It can be seen that Dif-MAML converges to the centralized solution and clearly outperforms the non-cooperative solution. This suggests that cooperation helps even when agents have access to different task distributions. Furthermore, even though our analysis was based on stochastic gradient descent (SGD), Fig. 4(b) suggests that our results can extend to other optimization methods. Moreover, we also test the performance after training with respect to number of gradient updates for adaptation in Fig. 4(c). It is visible that the match between the centralized and decentralized solutions does not change and the performance of the non-cooperative solution is still inferior. Note that this plot is also showing the average performance over all agents on 1000 tasks.

B. CLASSIFICATION

For classification, we consider widely used few-shot image recognition tasks on the Omniglot [45] and MiniImagenet [25] datasets. The Omniglot dataset comprises 1623 characters from 50 different alphabets. Each character has 20 samples, which were hand drawn by 20 different people —see Fig. 5 for sample characters. Therefore, it is suitable for few-shot learning scenarios as there is small number of data per class. The MiniImagenet dataset consists of 100 classes from ImageNet [46] with 600 samples from each class —see Fig. 6 for samples. It captures the complexity of ImageNet samples while not working on the full dataset which is huge.

Following [47] and [3], Omniglot is augmented with multiples of 90 degree rotations of the images. All agents are

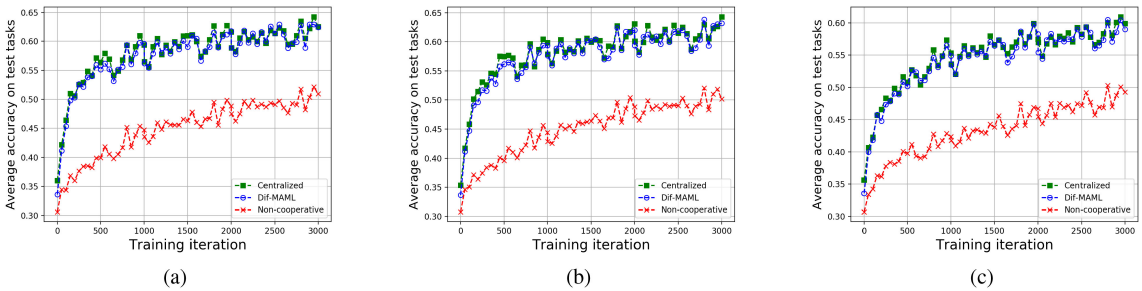

FIGURE 9. MiniImagenet test accuracies during training process 5-way 5-shot. (a) Averaging-Adam. (b) Metropolis-Adam. (c) Metropolis-SGD.

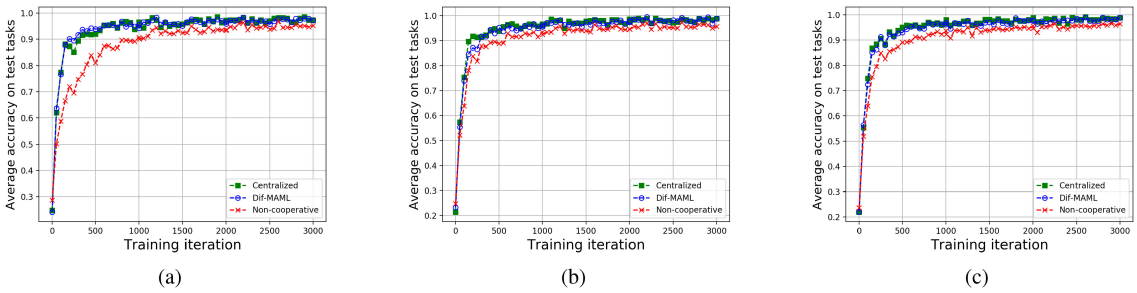

FIGURE 10. Omniglot test accuracies during training process 5-way 1-shot. (a) Averaging - SGD. (b) Averaging - Adam. (c) Metropolis - Adam.

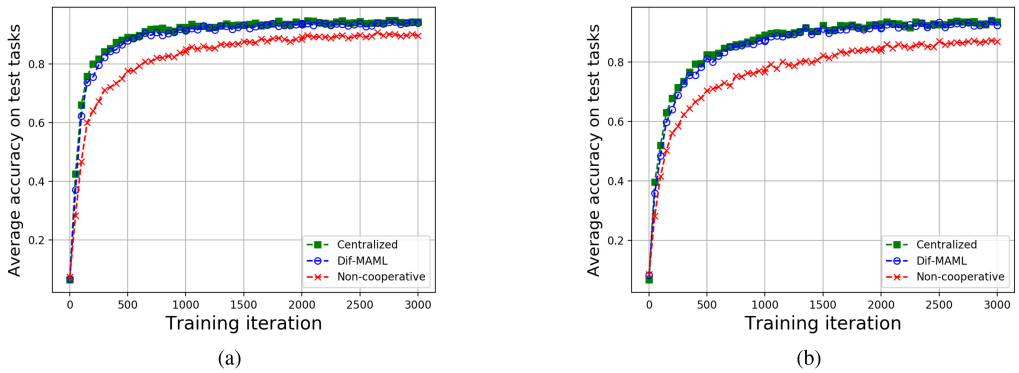

FIGURE 11. Omniglot test accuracies during training process 20-way 1-shot Averaging rule. (a) Adam. (b) SGD.

equipped with the same convolutional neural network architecture. Convolutional neural network architectures are based on the architectures in [3] which are based on [27].

In all simulations, each agent runs over 1000 batches of tasks over 3 epochs. In Omniglot experiments, for the Adam experiments $\mu = 0.001$ and for the SGD experiments $\mu = 0.1$. A single gradient step is used for adaptation in both training and testing and $\alpha = 0.4$. Training meta-batch size is equal to 16 for 5-way 1-shot and 8 for 5-way 5-shot. The plots are showing an average result of 100 tasks as testing meta-batch consists of 100 tasks. For MiniImagenet experiments, 10-query examples are used, testing meta-batch consists of 25 tasks and $\alpha = 0.01$. For the Adam experiments $\mu = 0.001$ and for the SGD experiments $\mu = 0.01$. The number of gradient updates is equal to 5 for training, 10 for testing. For 5-way 1-shot, training meta-batch has 4 tasks whereas 5-way 5-shot training meta-batch has 2 tasks. Note that the first testings occur after the first training step. In other words, the first data of all classification plots are at 1st iteration, not at 0th iteration.

In contrast to the regression experiment, in these simulations, all agents have access to the same tasks and data. However, in the centralized and decentralized strategies, the effective number of samples is larger as we limit the number of data and tasks processed in one agent.

Average accuracy on test tasks at every 50th training iteration is shown in Fig. 7(b) for MiniImageNet 5-way 5-shot setting trained with SGD and in Fig. 7(c) for Omniglot 20-way 1-shot setting trained with Adam. Note that the combination weights are set with averaging rule [16] in Fig. 7(b) and with Metropolis rule [43] in Fig. 7(c). Similar to the regression experiment, the decentralized solutions quickly match the centralized solutions and are substantially better than the

non-cooperative solutions. Additionally, the experiments suggest that our analysis can extend to left-stochastic combination matrices as well as multiple gradient updates to find the task-specific models in the inner-loop of the algorithm.

In Figs. 8–11 additional plots for MiniImagenet 5-way 1-shot, MiniImagenet 5-way 5-shot, Omniglot 5-way 1-shot and Omniglot 20-way 1-shot can be found, respectively. The results illustrate our conclusions are valid for the specified settings as well.

During experimentation, we observe that batch normalization [48] is necessary for applying Dif-MAML, and diffusion in general, on neural networks since the combination step (7b) reduces the variance of the weights due to averaging.

V. CONCLUSION

In this paper, we proposed a decentralized algorithm for meta-learning. Our theoretical analysis establishes that the agents' launch models cluster quickly in a small region around the centroid model and this centroid model reaches a stationary point after sufficient iterations. We illustrated by means of experiments on regression and classification problems that the performance of Dif-MAML consistently coincides with the centralized strategy and surpasses the non-cooperative strategy significantly. For future work, decentralized learning under imperfections [49], multiple updates during the adaptation step before the combination step [50], or active task sampling strategies [51] can be considered.

APPENDIX A
THE IMPLICATION OF ASSUMPTION 1

In Appendices A-D we denote a batch of data by $\boldsymbol{\mathcal{X}}_k^{(t)}$, its size by $N_k^{(t)}$, and its elements by $\{\boldsymbol{x}_{k,n}^{(t)}\}_{n=1}^{N_k^{(t)}}$.

Assumption 1 implies for the stochastic gradient constructed using a batch:

$$\left\|\nabla Q_k^{(t)}\left(w;\boldsymbol{\mathcal{X}}_k^{(t)}\right)-\nabla Q_k^{(t)}\left(u;\boldsymbol{\mathcal{X}}_k^{(t)}\right)\right\|\leq L^{\boldsymbol{\mathcal{X}}_k^{(t)}}\|w-u\| \tag{30}$$

where we have:

$$L^{\boldsymbol{\mathcal{X}}_k^{(t)}}\triangleq\frac{1}{N_k^{(t)}}\sum_{n=1}^{N_k^{(t)}}L^{\boldsymbol{x}_{k,n}^{(t)}} \tag{31}$$

Moreover,

$$\mathbb{E}_{\mathcal{X}_k^{(t)}}\left(L^{\boldsymbol{\mathcal{X}}_k^{(t)}}\right)^2\leq\left(L_k^{(t)}\right)^2 \tag{32}$$

Proof: For the stochastic gradients under a batch of data:

$$\begin{aligned}&\left\|\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_k^{(t)})-\nabla Q_k^{(t)}(u;\boldsymbol{\mathcal{X}}_k^{(t)})\right\|\\&=\left\|\frac{1}{N_k^{(t)}}\sum_{n=1}^{N_k^{(t)}}\left(\nabla Q_k^{(t)}(w;\boldsymbol{x}_{k,n}^{(t)})-\nabla Q_k^{(t)}(u;\boldsymbol{x}_{k,n}^{(t)})\right)\right\|\\&\overset{(a)}{\leq}\frac{1}{N_k^{(t)}}\sum_{n=1}^{N_k^{(t)}}\left\|\nabla Q_k^{(t)}(w;\boldsymbol{x}_{k,n}^{(t)})-\nabla Q_k^{(t)}(u;\boldsymbol{x}_{k,n}^{(t)})\right\|\\&\overset{(b)}{\leq}\frac{1}{N_k^{(t)}}\sum_{n=1}^{N_k^{(t)}}L^{\boldsymbol{x}_{k,n}^{(t)}}\|w-u\|\\&=L^{\boldsymbol{\mathcal{X}}_k^{(t)}}\|w-u\|\end{aligned} \tag{33}$$

where (a) follows from Jensen's inequality, and (b) follows from (8). Likewise,

$$\begin{aligned}\mathbb{E}_{\mathcal{X}_k^{(t)}}(L^{\boldsymbol{\mathcal{X}}_k^{(t)}})^2&=\mathbb{E}_{\mathcal{X}_k^{(t)}}\left(\frac{1}{N_k^{(t)}}\sum_{n=1}^{N_k^{(t)}}L^{\boldsymbol{x}_{k,n}^{(t)}}\right)^2\\&\overset{(a)}{\leq}\frac{1}{N_k^{(t)}}\mathbb{E}_{\mathcal{X}_k^{(t)}}\sum_{n=1}^{N_k^{(t)}}(L^{\boldsymbol{x}_{k,n}^{(t)}})^2\\&\overset{(b)}{\leq}(L_k^{(t)})^2\end{aligned} \tag{34}$$

where (a) follows from Jensen's inequality, and (b) follows from (9).

APPENDIX B
THE IMPLICATION OF ASSUMPTION 2

Assumption 2 implies for the loss Hessian under a batch of data:

$$\mathbb{E}_{\mathcal{X}_k^{(t)}}\left\|\nabla^2 Q_k^{(t)}\left(w;\boldsymbol{\mathcal{X}}_k^{(t)}\right)-\nabla^2 Q_k^{(t)}\left(u;\boldsymbol{\mathcal{X}}_k^{(t)}\right)\right\|\leq\rho_k^{(t)}\|w-u\| \tag{35}$$

Proof: For the loss Hessians under a batch of data:

$$\begin{aligned}&\mathbb{E}_{\mathcal{X}_k^{(t)}}\left\|\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_k^{(t)})-\nabla^2 Q_k^{(t)}(u;\boldsymbol{\mathcal{X}}_k^{(t)})\right\|\\&=\mathbb{E}_{\mathcal{X}_k^{(t)}}\left\|\frac{1}{N_k^{(t)}}\sum_{n=1}^{N_k^{(t)}}\left(\nabla^2 Q_k^{(t)}(w;\boldsymbol{x}_{k,n}^{(t)})-\nabla^2 Q_k^{(t)}(u;\boldsymbol{x}_{k,n}^{(t)})\right)\right\|\\&\overset{(a)}{\leq}\frac{1}{N_k^{(t)}}\sum_{n=1}^{N_k^{(t)}}\mathbb{E}_{x_{k,n}^{(t)}}\left\|\nabla^2 Q_k^{(t)}(w;\boldsymbol{x}_{k,n}^{(t)})-\nabla^2 Q_k^{(t)}(u;\boldsymbol{x}_{k,n}^{(t)})\right\|\\&\overset{(b)}{\leq}\frac{1}{N_k^{(t)}}\sum_{n=1}^{N_k^{(t)}}\rho_k^{(t)}\|w-u\|\\&=\rho_k^{(t)}\|w-u\|\end{aligned} \tag{36}$$

where (a) follows from Jensen's inequality, and (b) follows from (10).

APPENDIX C
THE IMPLICATION OF ASSUMPTION 3

Assumption 3 implies for the loss gradient under a batch of data:

$$\mathbb{E}_{\mathcal{X}_k^{(t)}} \left\| \nabla Q_k^{(t)} \left(w; \boldsymbol{\mathcal{X}}_k^{(t)} \right) \right\| \leq B_k^{(t)} \tag{37}$$

Proof: The bound for the norm of the stochastic gradients constructed using a batch is derived as follows:

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}_k^{(t)}} \left\| \nabla Q_k^{(t)}(w; \boldsymbol{\mathcal{X}}_k^{(t)}) \right\| &= \mathbb{E}_{\mathcal{X}_k^{(t)}} \left\| \frac{1}{N_k^{(t)}} \sum_{n=1}^{N_k^{(t)}} \nabla Q_k^{(t)}(w; \boldsymbol{x}_{k,n}^{(t)}) \right\| \\
&\overset{(a)}{\leq} \frac{1}{N_k^{(t)}} \sum_{n=1}^{N_k^{(t)}} \mathbb{E}_{x_{k,n}^{(t)}} \left\| \nabla Q_k^{(t)}(w; \boldsymbol{x}_{k,n}^{(t)}) \right\| \\
&\overset{(b)}{\leq} \frac{1}{N_k^{(t)}} \sum_{n=1}^{N_k^{(t)}} B_k^{(t)} \\
&= B_k^{(t)}
\end{aligned} \tag{38}$$

where (a) follows from Jensen's inequality, and (b) follows from (11).

APPENDIX D
THE IMPLICATION OF ASSUMPTION 4

Assumption 4 implies for the gradient and the Hessian under a batch of data:

$$\mathbb{E}_{\mathcal{X}_k^{(t)}} \left\| \nabla Q_k^{(t)} \left(w; \boldsymbol{\mathcal{X}}_k^{(t)} \right) - \nabla J_k^{(t)}(w) \right\|^4 \leq \frac{3\sigma_G^4}{(N_k^{(t)})^2} \tag{39}$$

$$\mathbb{E}_{\mathcal{X}_k^{(t)}} \left\| \nabla^2 Q_k^{(t)} \left(w; \boldsymbol{\mathcal{X}}_k^{(t)} \right) - \nabla^2 J_k^{(t)}(w) \right\|^4 \leq \frac{3\sigma_H^4}{(N_k^{(t)})^2} \tag{40}$$

Proof: We apply induction on $N_k^{(t)}$ [52]. For $N_k^{(t)} = 1$, expression (39) trivially holds since (12) is a tighter bound than (39). Now assume that (39) holds for $N_k^{(t)} - 1$. Define:

$$\begin{aligned}
\boldsymbol{r}(w) &\triangleq \nabla Q_k^{(t)}(w; \boldsymbol{x}_k^{(t)}) - \nabla J_k^{(t)}(w) \\
\boldsymbol{r}_{N_k^{(t)}}(w) &\triangleq \nabla Q_k^{(t)}(w; \boldsymbol{\mathcal{X}}_k^{(t)}) - \nabla J_k^{(t)}(w) \\
&= \frac{1}{N_k^{(t)}} \sum_{n=1}^{N_k^{(t)}} \nabla Q_k^{(t)}(w; \boldsymbol{x}_{k,n}^{(t)}) - \nabla J_k^{(t)}(w)
\end{aligned} \tag{41}$$

Then, we get:

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{r}_{N_k^{(t)}}(w)\|^4 \\
&= \mathbb{E} \left\| \frac{N_k^{(t)} - 1}{N_k^{(t)}} \boldsymbol{r}_{N_k^{(t)}-1}(w) + \frac{1}{N_k^{(t)}} \boldsymbol{r}(w) \right\|^4 \\
&= \mathbb{E} \left(\left\| \frac{N_k^{(t)} - 1}{N_k^{(t)}} \boldsymbol{r}_{N_k^{(t)}-1}(w) + \frac{1}{N_k^{(t)}} \boldsymbol{r}(w) \right\|^2 \right)^2 \\
&= \mathbb{E} \left(\frac{(N_k^{(t)} - 1)^2}{(N_k^{(t)})^2} \|\boldsymbol{r}_{N_k^{(t)}-1}(w)\|^2 + \frac{1}{(N_k^{(t)})^2} \|\boldsymbol{r}(w)\|^2 \right. \\
&\quad \left. + 2 \frac{N_k^{(t)} - 1}{(N_k^{(t)})^2} \boldsymbol{r}_{N_k^{(t)}-1}(w)^\mathsf{T} \boldsymbol{r}(w) \right)^2 \\
&\overset{(a)}{=} \frac{(N_k^{(t)} - 1)^4}{(N_k^{(t)})^4} \mathbb{E} \left[\|\boldsymbol{r}_{N_k^{(t)}-1}(w)\|^4 \right] + \frac{1}{(N_k^{(t)})^4} \mathbb{E} \left[\|\boldsymbol{r}(w)\|^4 \right] \\
&\quad + \frac{4(N_k^{(t)} - 1)^2}{(N_k^{(t)})^4} \mathbb{E} \left[(\boldsymbol{r}_{N_k^{(t)}-1}(w)^\mathsf{T} \boldsymbol{r}(w))^2 \right] \\
&\quad + \frac{2(N_k^{(t)} - 1)^2}{(N_k^{(t)})^4} \mathbb{E} \left[\|\boldsymbol{r}_{N_k^{(t)}-1}(w)\|^2 \|\boldsymbol{r}(w)\|^2 \right] \\
&\overset{(b)}{\leq} \frac{(N_k^{(t)} - 1)^4}{(N_k^{(t)})^4} \mathbb{E} \left[\|\boldsymbol{r}_{N_k^{(t)}-1}(w)\|^4 \right] + \frac{1}{(N_k^{(t)})^4} \mathbb{E} \left[\|\boldsymbol{r}(w)\|^4 \right] \\
&\quad + \frac{6(N_k^{(t)} - 1)^2}{(N_k^{(t)})^4} \mathbb{E} \left[\|\boldsymbol{r}_{N_k^{(t)}-1}(w)\|^2 \|\boldsymbol{r}(w)\|^2 \right] \\
&\overset{(c)}{=} \frac{(N_k^{(t)} - 1)^4}{(N_k^{(t)})^4} \mathbb{E} \left[\|\boldsymbol{r}_{N_k^{(t)}-1}(w)\|^4 \right] + \frac{1}{(N_k^{(t)})^4} \mathbb{E} \left[\|\boldsymbol{r}(w)\|^4 \right] \\
&\quad + \frac{6(N_k^{(t)} - 1)^2}{(N_k^{(t)})^4} \mathbb{E} \left[\|\boldsymbol{r}_{N_k^{(t)}-1}(w)\|^2 \right] \mathbb{E} \left[\|\boldsymbol{r}(w)\|^2 \right] \\
&\overset{(d)}{\leq} \frac{(N_k^{(t)} - 1)^4}{(N_k^{(t)})^4} \frac{3\sigma_G^4}{(N_k^{(t)} - 1)^2} + \frac{1}{(N_k^{(t)})^4} \sigma_G^4 \\
&\quad + \frac{6(N_k^{(t)} - 1)^2}{(N_k^{(t)})^4} \frac{\sigma_G^2}{(N_k^{(t)} - 1)} \sigma_G^2 \\
&= \frac{\sigma_G^4}{(N_k^{(t)})^2} \left(\frac{3(N_k^{(t)} - 1)^2}{(N_k^{(t)})^2} + \frac{1}{(N_k^{(t)})^2} + \frac{6(N_k^{(t)} - 1)}{(N_k^{(t)})^2} \right) \\
&= \frac{\sigma_G^4}{(N_k^{(t)})^2} \left(\frac{3(N_k^{(t)})^2 - 2}{(N_k^{(t)})^2} \right) \\
&\leq \frac{3\sigma_G^4}{(N_k^{(t)})^2}
\end{aligned} \tag{42}$$

where (a) follows from expansion of the square and dropping the cross-terms that are zero due to the independence assumption on the data,(b) follows from Cauchy-Schwarz, (c) follows from independence assumption on the data, and (d) follows from the induction hypothesis, (12), and the following variance reduction formula:

$$\mathbb{E}\|\boldsymbol{r}_{N_k^{(t)}-1}(w)\|^2 = \frac{1}{N_k^{(t)} - 1} \mathbb{E}\|\boldsymbol{r}(w)\|^2 \tag{43}$$

For proving (40), just replacing the gradients with the Hessians in (41) is enough.

APPENDIX E
PROOF OF (18)

Recall the definition of the adjusted objective:

$$\widehat{J}_k(w) = \mathbb{E}_{t\sim\pi_k}\mathbb{E}_{\mathcal{X}_{in}^{(t)}} J_k^{(t)}\left(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right) \tag{44}$$

The gradient corresponding to this objective is:

$$\begin{aligned}\nabla\widehat{J}_k(w) = \mathbb{E}_{t\sim\pi_k}\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\Big[&\left(I - \alpha\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right)\\ &\times \nabla J_k^{(t)}\left(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right)\Big]\end{aligned} \tag{45}$$

Expectation of the stochastic MAML gradient is given by:

$$\begin{aligned}\mathbb{E}\nabla\overline{Q_k}(w) &= \mathbb{E}\Bigg[\frac{1}{|\mathcal{S}_k|}\sum_{t\in\boldsymbol{\mathcal{S}}_k}\left(I - \alpha\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right)\\ &\quad\times \nabla Q_k^{(t)}\left(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)});\ \boldsymbol{\mathcal{X}}_{o}^{(t)}\right)\Bigg]\\ &\overset{(a)}{=} \mathbb{E}\Big[\left(I - \alpha\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right)\\ &\quad\times \nabla Q_k^{(t)}\left(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)});\ \boldsymbol{\mathcal{X}}_{o}^{(t)}\right)\Big]\\ &\overset{(b)}{=} \mathbb{E}_{t\sim\pi_k}\Big[\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\Big[\mathbb{E}_{\mathcal{X}_{o}^{(t)}}\Big[\left(I - \alpha\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right)\\ &\quad\times \nabla Q_k^{(t)}\left(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)});\ \boldsymbol{\mathcal{X}}_{o}^{(t)}\right)\Big]\Big]\Big]\\ &= \mathbb{E}_{t\sim\pi_k}\Big[\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\Big[\left(I - \alpha\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right)\\ &\quad\times \mathbb{E}_{\mathcal{X}_{o}^{(t)}}\Big[\nabla Q_k^{(t)}\left(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)});\ \boldsymbol{\mathcal{X}}_{o}^{(t)}\right)\Big]\Big]\Big]\\ &\overset{(c)}{=} \mathbb{E}_{t\sim\pi_k}\Big[\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\Big[\left(I - \alpha\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right)\\ &\quad\times \nabla J_k^{(t)}\left(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right)\Big]\Big]\end{aligned} \tag{46}$$

where (a) follows from the *i.i.d.* assumption on the batch of tasks, (b) follows from conditioning, and (c) follows from the relation between loss functions and stochastic risks.

APPENDIX F
PROOF OF LEMMA 1

The disagreement between (2) and (17) is:

$$\begin{aligned}\left|\overline{J_k}(w) - \widehat{J}_k(w)\right| &= \left|\mathbb{E}\left[J_k^{(t)}(\widetilde{\boldsymbol{w}}_1) - J_k^{(t)}(\widetilde{\boldsymbol{w}}_2)\right]\right|\\ &\overset{(a)}{\leq} \mathbb{E}\left|J_k^{(t)}(\widetilde{\boldsymbol{w}}_1) - J_k^{(t)}(\widetilde{\boldsymbol{w}}_2)\right|\end{aligned} \tag{47}$$

where $\widetilde{\boldsymbol{w}}_1 \triangleq w - \alpha\nabla J_k^{(t)}(w)$, $\widetilde{\boldsymbol{w}}_2 \triangleq w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})$ and (a) follows from Jensen's inequality. Lipschitz property of the gradient (Assumption 1) implies:

$$\begin{aligned}J_k^{(t)}(\widetilde{\boldsymbol{w}}_1) - J_k^{(t)}(\widetilde{\boldsymbol{w}}_2) &\leq (\nabla J_k^{(t)}(\widetilde{\boldsymbol{w}}_2))^{\mathsf{T}}(\widetilde{\boldsymbol{w}}_1 - \widetilde{\boldsymbol{w}}_2)\\ &\quad + \frac{L}{2}\|\widetilde{\boldsymbol{w}}_1 - \widetilde{\boldsymbol{w}}_2\|^2\end{aligned} \tag{48}$$

$$\begin{aligned}J_k^{(t)}(\widetilde{\boldsymbol{w}}_2) - J_k^{(t)}(\widetilde{\boldsymbol{w}}_1) &\leq (\nabla J_k^{(t)}(\widetilde{\boldsymbol{w}}_1))^{\mathsf{T}}(\widetilde{\boldsymbol{w}}_2 - \widetilde{\boldsymbol{w}}_1)\\ &\quad + \frac{L}{2}\|\widetilde{\boldsymbol{w}}_1 - \widetilde{\boldsymbol{w}}_2\|^2\end{aligned} \tag{49}$$

Combining the inequalities yields:

$$\begin{aligned}\mathbb{E}\left|J_k^{(t)}(\widetilde{\boldsymbol{w}}_1) - J_k^{(t)}(\widetilde{\boldsymbol{w}}_2)\right| &\leq \mathbb{E}\Bigg[\frac{L}{2}\|\widetilde{\boldsymbol{w}}_1 - \widetilde{\boldsymbol{w}}_2\|^2\\ &\quad + \max\left\{\|\nabla J_k^{(t)}(\widetilde{\boldsymbol{w}}_1)\|, \|\nabla J_k^{(t)}(\widetilde{\boldsymbol{w}}_2)\|\right\}\|\widetilde{\boldsymbol{w}}_1 - \widetilde{\boldsymbol{w}}_2\|\Bigg]\\ &\overset{(a)}{\leq} \mathbb{E}\left[\frac{L}{2}\|\widetilde{\boldsymbol{w}}_1 - \widetilde{\boldsymbol{w}}_2\|^2 + B\|\widetilde{\boldsymbol{w}}_1 - \widetilde{\boldsymbol{w}}_2\|\right]\\ &\overset{(b)}{\leq} \frac{\alpha^2 L}{2}\mathbb{E}\left[\left\|\nabla J_k^{(t)}(w) - \nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right\|^2\right]\\ &\quad + B\alpha\mathbb{E}\left[\left\|\nabla J_k^{(t)}(w) - \nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right\|\right]\\ &\overset{(c)}{\leq} \frac{\alpha^2 L\sigma_G^2}{2|\mathcal{X}_{in}|} + \frac{B\alpha\sigma_G}{\sqrt{|\mathcal{X}_{in}|}}\end{aligned} \tag{50}$$

where (a) follows from Assumption 3, (b) follows from inserting $\widetilde{\boldsymbol{w}}_1$ and $\widetilde{\boldsymbol{w}}_2$ expressions, and (c) follows from Assumption 4.

APPENDIX G
PROOF OF LEMMA 2

Using (3) and (18) we have:

$$\begin{aligned}\nabla\overline{J_k}(w) &= \mathbb{E}_{t\sim\pi_k}\left[(I - \alpha\nabla^2 J_k^{(t)}(w))\nabla J_k^{(t)}(w - \alpha\nabla J_k^{(t)}(w))\right]\\ &= \mathbb{E}_{t\sim\pi_k}\Big[\nabla J_k^{(t)}(w - \alpha\nabla J_k^{(t)}(w))\\ &\quad - \alpha\nabla^2 J_k^{(t)}(w)\nabla J_k^{(t)}(w - \alpha\nabla J_k^{(t)}(w))\Big]\end{aligned} \tag{51}$$

and

$$\begin{aligned}\nabla\widehat{J}_k(w) &= \mathbb{E}_{t\sim\pi_k}\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\Big[(I - \alpha\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}))\\ &\quad\times \nabla J_k^{(t)}(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}))\Big]\\ &= \mathbb{E}_{t\sim\pi_k}\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\Big[\nabla J_k^{(t)}(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}))\\ &\quad - \alpha\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla J_k^{(t)}(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}))\Big]\end{aligned} \tag{52}$$

The norm of the disagreement then follows:

$$\begin{aligned}&\left\|\nabla\overline{J_k}(w) - \nabla\widehat{J}_k(w)\right\|\\ &= \Big\|\mathbb{E}\Big[\nabla J_k^{(t)}(w - \alpha\nabla J_k^{(t)}(w))\\ &\quad - \alpha\nabla^2 J_k^{(t)}(w)\nabla J_k^{(t)}(w - \alpha\nabla J_k^{(t)}(w))\\ &\quad - \nabla J_k^{(t)}(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}))\\ &\quad + \alpha\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla J_k^{(t)}(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}))\Big]\Big\|\\ &\overset{(a)}{\leq} \mathbb{E}\Big[\Big\|\nabla J_k^{(t)}(w - \alpha\nabla J_k^{(t)}(w)) - \nabla J_k^{(t)}(w - \alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}))\end{aligned}$$

$$+\alpha\nabla^2 Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)})\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))$$
$$-\alpha\nabla^2 J_k^{(t)}(w)\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))\Big\|\Big]$$
$$\overset{(b)}{\leq}\mathbb{E}\Big[\Big\|\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))-\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))\Big\|$$
$$+\alpha\Big\|\nabla^2 Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)})\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))$$
$$-\nabla^2 J_k^{(t)}(w)\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))\Big\|\Big] \tag{53}$$

where (a) follows from applying Jensen's inequality and rearranging terms, and (b) follows from applying the triangle inequality. We bound the terms in (53) separately. For the first term we have:

$$\mathbb{E}\Big\|\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))-\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))\Big\|$$
$$\overset{(a)}{\leq} L\alpha\mathbb{E}\Big[\Big\|\nabla J_k^{(t)}(w)-\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)})\Big\|\Big]$$
$$\overset{(b)}{\leq} L\alpha\frac{\sigma_G}{\sqrt{|\mathcal{X}_{in}|}} \tag{54}$$

where (a) follows from Assumption 1, and (b) follows from Assumption 4.

Rewriting the second term in (53):

$$\mathbb{E}\Big[\Big\|\nabla^2 Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)})\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))$$
$$-\nabla^2 J_k^{(t)}(w)\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))\Big\|\Big]$$
$$\overset{(a)}{\leq}\mathbb{E}\Big[\Big\|\nabla^2 Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)})\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))$$
$$-\nabla^2 J_k^{(t)}(w)\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))\Big\|$$
$$+\Big\|\nabla^2 J_k^{(t)}(w)\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))$$
$$-\nabla^2 J_k^{(t)}(w)\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))\Big\|\Big] \tag{55}$$

where (a) follows from adding and subtracting the term $\nabla^2 J_k^{(t)}(w)\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))$ and applying the triangle inequality. We bound the terms in (55) separately. For the first term:

$$\mathbb{E}\Big[\Big\|\nabla^2 Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)})\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))$$
$$-\nabla^2 J_k^{(t)}(w)\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))\Big\|\Big]$$
$$\overset{(a)}{\leq}\mathbb{E}\Big[\Big\|\nabla^2 Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)})-\nabla^2 J_k^{(t)}(w)\Big\|$$
$$\times\Big\|\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))\Big\|\Big]$$
$$\overset{(b)}{\leq}\mathbb{E}\Big[\Big\|\nabla^2 Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)})-\nabla^2 J_k^{(t)}(w)\Big\|\Big]B$$
$$\overset{(c)}{\leq} B\frac{\sigma_H}{\sqrt{|\mathcal{X}_{in}|}} \tag{56}$$

where (a) follows from sub-multiplicity of the norm, (b) follows from Assumption 3, and (c) follows from Assumption 4. For the second term in (55):

$$\mathbb{E}\Big[\Big\|\nabla^2 J_k^{(t)}(w)\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))$$
$$-\nabla^2 J_k^{(t)}(w)\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))\Big\|\Big]$$
$$\overset{(a)}{\leq}\mathbb{E}\Big[\Big\|\nabla^2 J_k^{(t)}(w)\Big\|\Big\|\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))$$
$$-\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))\Big\|\Big]$$
$$\overset{(b)}{\leq} L\mathbb{E}\Big[\Big\|\nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))$$
$$-\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))\Big\|\Big]$$
$$\overset{(c)}{\leq}\alpha L^2\mathbb{E}\Big[\Big\|\nabla J_k^{(t)}(w)-\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)})\Big\|\Big]$$
$$\overset{(d)}{\leq}\alpha L^2\frac{\sigma_G}{\sqrt{|\mathcal{X}_{in}|}} \tag{57}$$

where (a) follows from sub-multiplicity of the norm, (b) and (c) follow from Assumption 1, and (d) follows from Assumption 4. Combining the results completes the proof.

APPENDIX H
PROOF OF LEMMA 3

Recall the formula for the gradient of the adjusted objective (18):

$$\|\nabla\widehat{J}_k(w)\|=\Big\|\mathbb{E}\Big[(I-\alpha\nabla^2 Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))$$
$$\times\nabla Q_k^{(t)}\Big(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)});\mathcal{X}_o^{(t)}\Big)\Big]\Big\|$$
$$\overset{(a)}{\leq}\mathbb{E}\Big\|(I-\alpha\nabla^2 Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))$$
$$\times\nabla Q_k^{(t)}\Big(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)});\mathcal{X}_o^{(t)}\Big)\Big\|$$
$$\overset{(b)}{\leq}\mathbb{E}\Big[\Big\|(I-\alpha\nabla^2 Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))\Big\|$$
$$\times\Big\|\nabla Q_k^{(t)}\Big(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)});\mathcal{X}_o^{(t)}\Big)\Big\|\Big]$$
$$\overset{(c)}{\leq}\mathbb{E}_{t\sim\pi_k}\Big[\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\Big[\mathbb{E}_{\mathcal{X}_o^{(t)}}\Big[\Big\|(I-\alpha\nabla^2 Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))\Big\|$$
$$\times\Big\|\nabla Q_k^{(t)}\Big(w-\alpha\nabla Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)});\mathcal{X}_o^{(t)}\Big)\Big\|\Big]\Big]\Big]$$
$$\overset{(d)}{\leq}\mathbb{E}_{t\sim\pi_k}\Big[\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\Big[\Big\|(I-\alpha\nabla^2 Q_k^{(t)}(w;\mathcal{X}_{in}^{(t)}))\Big\|B\Big]\Big]$$
$$\overset{(e)}{\leq}(1+\alpha L)B \tag{58}$$

where (a) follows from Jensen's inequality, (b) follows from sub-multiplicity of the norm, (c) follows from conditioning,

(d) follows from Assumption 3, and (e) follows from Assumption 1.

APPENDIX I
PROOF OF LEMMA 4

Define the following variables:

$$\widetilde{\boldsymbol{w}}_2 \triangleq w - \alpha \nabla Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)}) \tag{59}$$

$$\widetilde{\boldsymbol{u}}_2 \triangleq u - \alpha \nabla Q_k^{(\boldsymbol{t})}(u; \boldsymbol{\mathcal{X}}_{in}^{(t)}) \tag{60}$$

Recall the formula for the gradient of the adjusted objective (18):

$$\begin{aligned}\nabla \widehat{J}_k(w) &= \mathbb{E}\Big[(I - \alpha \nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)}))\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big] \\ &= \mathbb{E}\Big[\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) - \alpha \nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big]\end{aligned} \tag{61}$$

and

$$\begin{aligned}&\nabla \widehat{J}_k(u) \\ &= \mathbb{E}\Big[(I - \alpha \nabla^2 Q_k^{(\boldsymbol{t})}(u; \boldsymbol{\mathcal{X}}_{in}^{(t)}))\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big] \\ &= \mathbb{E}\Big[\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) - \alpha \nabla^2 Q_k^{(\boldsymbol{t})}(u; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big]\end{aligned} \tag{62}$$

Bounding the disagreement:

$$\begin{aligned}&\Big\|\nabla \widehat{J}_k(w) - \nabla \widehat{J}_k(u)\Big\| \\ &\quad= \Big\|\mathbb{E}\Big[\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) - \nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) \\ &\qquad - \alpha\Big(\nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) \\ &\qquad - \nabla^2 Q_k^{(\boldsymbol{t})}(u; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big)\Big]\Big\| \\ &\quad\overset{(a)}{\leq} \mathbb{E}\Big[\Big\|\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) - \nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) \\ &\qquad - \alpha\Big(\nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) \\ &\qquad - \nabla^2 Q_k^{(\boldsymbol{t})}(u; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big)\Big\|\Big] \\ &\quad\overset{(b)}{\leq} \mathbb{E}\Big[\Big\|\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) - \nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big\|\Big] \\ &\qquad + \alpha\mathbb{E}\Big[\Big\|\nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) \\ &\qquad - \nabla^2 Q_k^{(\boldsymbol{t})}(u; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big\|\Big]\end{aligned} \tag{63}$$

where (a) follows from Jensen's inequality, and (b) follows from the triangle inequality. We bound the terms in (63) separately. For the first term,

$$\begin{aligned}&\mathbb{E}\Big[\Big\|\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) - \nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big\|\Big] \\ &\overset{(a)}{\leq} \mathbb{E}\Big[L^{\boldsymbol{\mathcal{X}}_o^{(t)}}\|\widetilde{\boldsymbol{w}}_2 - \widetilde{\boldsymbol{u}}_2\|\Big] \\ &\overset{(b)}{\leq} \mathbb{E}\Big[L^{\boldsymbol{\mathcal{X}}_o^{(t)}}\Big(\|w - u\| \\ &\qquad + \alpha\Big\|\nabla Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)}) - \nabla Q_k^{(\boldsymbol{t})}(u; \boldsymbol{\mathcal{X}}_{in}^{(t)})\Big\|\Big)\Big] \\ &\overset{(c)}{\leq} \mathbb{E}\Big[L^{\boldsymbol{\mathcal{X}}_o^{(t)}}\Big(\|w - u\| + \alpha L^{\boldsymbol{\mathcal{X}}_{in}^{(t)}}\|w - u\|\Big)\Big]\end{aligned} \tag{64}$$

$$\overset{(d)}{\leq} L(1 + \alpha L)\|w - u\| \tag{65}$$

where (a) follows from Assumption 1, (b) follows from replacing $\widetilde{\boldsymbol{w}}_2, \widetilde{\boldsymbol{u}}_2$ and applying triangle inequality, (c) follows from Assumption 1, (d) follows from the independence assumption on $\boldsymbol{\mathcal{X}}_{in}^{(t)}, \boldsymbol{\mathcal{X}}_o^{(t)}$ and taking the expectation. For the second term we have:

$$\begin{aligned}&\mathbb{E}\Big\|\nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) \\ &\qquad - \nabla^2 Q_k^{(\boldsymbol{t})}(u; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big\| \\ &\overset{(a)}{\leq} \mathbb{E}\Big[\Big\|\nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) \\ &\qquad - \nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big\|\Big] \\ &\qquad + \mathbb{E}\Big[\Big\|\nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) \\ &\qquad - \nabla^2 Q_k^{(\boldsymbol{t})}(u; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big\|\Big]\end{aligned} \tag{66}$$

where (a) follows from adding and subtracting the same term and triangle inequality. For the first term in (66), we have:

$$\begin{aligned}&\mathbb{E}\Big\|\nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) \\ &\qquad - \nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big\| \\ &\overset{(a)}{\leq} \mathbb{E}\Big[\Big\|\nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)})\Big\| \\ &\qquad \times \Big\|\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) - \nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big\|\Big] \\ &\overset{(b)}{\leq} \mathbb{E}\Big[(L^{\boldsymbol{\mathcal{X}}_{in}^{(t)}})\Big(L^{\boldsymbol{\mathcal{X}}_o^{(t)}}(1 + \alpha L^{\boldsymbol{\mathcal{X}}_{in}^{(t)}})\|w - u\|\Big)\Big] \\ &\overset{(c)}{\leq} L^2(1 + \alpha L)\|w - u\|\end{aligned} \tag{67}$$

where (a) follows from sub-multiplicity of the norm, (b) follows from Assumption 1 and (64), (c) follows from the independence assumption on $\boldsymbol{\mathcal{X}}_{in}^{(t)}, \boldsymbol{\mathcal{X}}_o^{(t)}$ and taking the expectation.

For the second term in (66), we have:

$$\begin{aligned}&\mathbb{E}\Big\|\nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)}) \\ &\qquad - \nabla^2 Q_k^{(\boldsymbol{t})}(u; \boldsymbol{\mathcal{X}}_{in}^{(t)})\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big\| \\ &\overset{(a)}{\leq} \mathbb{E}\Big[\Big\|\nabla^2 Q_k^{(\boldsymbol{t})}(w; \boldsymbol{\mathcal{X}}_{in}^{(t)}) - \nabla^2 Q_k^{(\boldsymbol{t})}(u; \boldsymbol{\mathcal{X}}_{in}^{(t)})\Big\| \\ &\qquad \times \Big\|\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{u}}_2; \boldsymbol{\mathcal{X}}_o^{(t)})\Big\|\Big]\end{aligned}$$

$$\overset{(b)}{\leq} \mathbb{E}_{t\sim\pi_k}\Big[\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\Big[\mathbb{E}_{\mathcal{X}_o^{(t)}}\Big[\Big\|\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})-\nabla^2 Q_k^{(t)}(u;\boldsymbol{\mathcal{X}}_{in}^{(t)})\Big\| \times \Big\|\nabla Q_k^{(t)}(\widetilde{\boldsymbol{u}}_2;\boldsymbol{\mathcal{X}}_o^{(t)})\Big\|\Big]\Big]\Big]$$

$$\overset{(c)}{\leq} \mathbb{E}_{t\sim\pi_k}\Big[\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\Big[\Big\|\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}) - \nabla^2 Q_k^{(t)}(u;\boldsymbol{\mathcal{X}}_{in}^{(t)})\Big\|B\Big]\Big]$$

$$\overset{(d)}{\leq} \rho B\|w-u\| \tag{68}$$

where (a) follows from sub-multiplicity of the norm, (b) follows from conditioning, (c) follows from Assumption 3 and (d) follows from Assumption 2. Combining the results completes the proof.

APPENDIX J
PROOF OF LEMMA 5

We will first prove three intermediate lemmas, then conclude the proof.

First, we define the task-specific meta-gradient and task-specific meta-stochastic gradient:

$$\nabla\overline{Q_k}^{(t)}(w) \triangleq (I-\alpha\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})) \times \nabla Q_k^{(t)}\Big(w-\alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)});\boldsymbol{\mathcal{X}}_o^{(t)}\Big) \tag{69}$$

$$\nabla\overline{J_k}^{(t)}(w) \triangleq (I-\alpha\nabla^2 J_k^{(t)}(w))\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w)) \tag{70}$$

Lemma 6: Under assumptions 1,3,4, for each agent k, the disagreement between $\nabla\overline{Q_k}^{(t)}(\cdot)$ and $\nabla\overline{J_k}^{(t)}(\cdot)$ is bounded in expectation, namely, for any $w\in\mathbb{R}^M$:

$$\mathbb{E}\|\nabla\overline{Q_k}^{(t)}(w)-\nabla\overline{J_k}^{(t)}(w)\|^2 \leq C_1^2 \tag{71}$$

where C_1^2 is a non-negative constant, given by:

$$C_1^2 \triangleq 6(1+\alpha L)^2\sigma_G^2\Big(\frac{1}{|\mathcal{X}_o|}+\frac{L^2\alpha^2}{|\mathcal{X}_{in}|}\Big) + \frac{6\alpha^2\sigma_H^2}{|\mathcal{X}_{in}|}\Big(B^2+\frac{\sigma_G^2}{|\mathcal{X}_o|}\Big)+\frac{9\alpha^4}{|\mathcal{X}_{in}|^2}(\sigma_H^4+L^4\sigma_G^4) \tag{72}$$

Proof: We introduce the error terms:

$$\boldsymbol{e}_{h,x}^{(t)} \triangleq \alpha\nabla^2 J_k^{(t)}(w)-\alpha\nabla^2 Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}) \tag{73}$$

$$\boldsymbol{e}_{g,o}^{(t)} \triangleq \nabla Q_k^{(t)}\Big(w-\alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)});\boldsymbol{\mathcal{X}}_o^{(t)}\Big) - \nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})) \tag{74}$$

$$\boldsymbol{e}_{g,x}^{(t)} \triangleq \nabla J_k^{(t)}(w-\alpha\nabla Q_k^{(t)}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})) - \nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w)) \tag{75}$$

Rewriting (69):

$$\nabla\overline{Q_k}^{(t)}(w) = \Big(I-\alpha\nabla^2 J_k^{(t)}(w)+\boldsymbol{e}_{h,x}^{(t)}\Big) \times \Big(\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))+\boldsymbol{e}_{g,o}^{(t)}+\boldsymbol{e}_{g,x}^{(t)}\Big) \tag{76}$$

It then follows:

$$\begin{aligned}&\nabla\overline{Q_k}^{(t)}(w)-\nabla\overline{J_k}^{(t)}(w)\\ &= (I-\alpha\nabla^2 J_k^{(t)}(w))\boldsymbol{e}_{g,o}^{(t)}+(I-\alpha\nabla^2 J_k^{(t)}(w))\boldsymbol{e}_{g,x}^{(t)}\\ &\quad+\boldsymbol{e}_{h,x}^{(t)}\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))+\boldsymbol{e}_{h,x}^{(t)}\boldsymbol{e}_{g,o}^{(t)}+\boldsymbol{e}_{h,x}^{(t)}\boldsymbol{e}_{g,x}^{(t)}\end{aligned} \tag{77}$$

Bounding the disagreement:

$$\begin{aligned}&\Big\|\nabla\overline{Q_k}^{(t)}(w)-\nabla\overline{J_k}^{(t)}(w)\Big\|\\ &\overset{(a)}{\leq}\Big\|(I-\alpha\nabla^2 J_k^{(t)}(w))\Big\|\|\boldsymbol{e}_{g,o}^{(t)}\|+\Big\|(I-\alpha\nabla^2 J_k^{(t)}(w))\Big\|\|\boldsymbol{e}_{g,x}^{(t)}\|\\ &\quad+\|\boldsymbol{e}_{h,x}^{(t)}\|\Big\|\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))\Big\|+\|\boldsymbol{e}_{h,x}^{(t)}\|\|\boldsymbol{e}_{g,o}^{(t)}\|\\ &\quad+\|\boldsymbol{e}_{h,x}^{(t)}\|\|\boldsymbol{e}_{g,x}^{(t)}\|\\ &\overset{(b)}{\leq}\Big\|(I-\alpha\nabla^2 J_k^{(t)}(w))\Big\|\|\boldsymbol{e}_{g,o}^{(t)}\|+\Big\|(I-\alpha\nabla^2 J_k^{(t)}(w))\Big\|\|\boldsymbol{e}_{g,x}^{(t)}\|\\ &\quad+\|\boldsymbol{e}_{h,x}^{(t)}\|\Big\|\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))\Big\|+\|\boldsymbol{e}_{h,x}^{(t)}\|\|\boldsymbol{e}_{g,o}^{(t)}\|\\ &\quad+\frac{\|\boldsymbol{e}_{h,x}^{(t)}\|^2}{2}+\frac{\|\boldsymbol{e}_{g,x}^{(t)}\|^2}{2}\end{aligned} \tag{78}$$

where (a) follows from sub-multiplicative property of the norm and the triangle inequality, while (b) follows from $ab\leq\frac{a^2+b^2}{2}$.

Taking the square of the norm and using $(\sum_{i=1}^6 x_i)^2\leq 6(\sum_{i=1}^6 x_i^2)$ yields:

$$\begin{aligned}&\Big\|\nabla\overline{Q_k}^{(t)}(w)-\nabla\overline{J_k}^{(t)}(w)\Big\|^2\\ &\leq 6\Big\|(I-\alpha\nabla^2 J_k^{(t)}(w))\Big\|^2\|\boldsymbol{e}_{g,o}^{(t)}\|^2\\ &\quad+6\Big\|(I-\alpha\nabla^2 J_k^{(t)}(w))\Big\|^2\|\boldsymbol{e}_{g,x}^{(t)}\|^2\\ &\quad+6\|\boldsymbol{e}_{h,x}^{(t)}\|^2\Big\|\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))\Big\|^2+6\|\boldsymbol{e}_{h,x}^{(t)}\|^2\|\boldsymbol{e}_{g,o}^{(t)}\|^2\\ &\quad+3\|\boldsymbol{e}_{h,x}^{(t)}\|^4+3\|\boldsymbol{e}_{g,x}^{(t)}\|^4\end{aligned} \tag{79}$$

Taking the expectation with respect to the inner and outer batches of data yields:

$$\begin{aligned}&\mathbb{E}_{\mathcal{X}_{in}^{(t)},\mathcal{X}_o^{(t)}}\Big\|\nabla\overline{Q_k}^{(t)}(w)-\nabla\overline{J_k}^{(t)}(w)\Big\|^2\\ &\leq 6\Big\|(I-\alpha\nabla^2 J_k^{(t)}(w))\Big\|^2\mathbb{E}_{\mathcal{X}_{in}^{(t)},\mathcal{X}_o^{(t)}}\Big[\|\boldsymbol{e}_{g,o}^{(t)}\|^2\Big]\\ &\quad+6\Big\|(I-\alpha\nabla^2 J_k^{(t)}(w))\Big\|^2\mathbb{E}_{\mathcal{X}_{in}^{(t)},\mathcal{X}_o^{(t)}}\Big[\|\boldsymbol{e}_{g,x}^{(t)}\|^2\Big]\\ &\quad+6\mathbb{E}_{\mathcal{X}_{in}^{(t)},\mathcal{X}_o^{(t)}}\Big[\|\boldsymbol{e}_{h,x}^{(t)}\|^2\Big]\Big\|\nabla J_k^{(t)}(w-\alpha\nabla J_k^{(t)}(w))\Big\|^2\\ &\quad+6\mathbb{E}_{\mathcal{X}_{in}^{(t)},\mathcal{X}_o^{(t)}}\Big[\|\boldsymbol{e}_{h,x}^{(t)}\|^2\|\boldsymbol{e}_{g,o}^{(t)}\|^2\Big]+3\mathbb{E}_{\mathcal{X}_{in}^{(t)},\mathcal{X}_o^{(t)}}\Big[\|\boldsymbol{e}_{h,x}^{(t)}\|^4\Big]\\ &\quad+3\mathbb{E}_{\mathcal{X}_{in}^{(t)},\mathcal{X}_o^{(t)}}\Big[\|\boldsymbol{e}_{g,x}^{(t)}\|^4\Big]\end{aligned} \tag{80}$$

We bound the terms of (80) one by one:

$$\left\|(I-\alpha\nabla^2 J_k^{(\boldsymbol{t})}(w))\right\| \overset{(a)}{\leq} (1+\alpha L) \tag{81}$$

$$\left\|(I-\alpha\nabla^2 J_k^{(\boldsymbol{t})}(w))\right\|^2 \leq (1+\alpha L)^2 \tag{82}$$

where (a) follows from Assumption 1. Moreover,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{X}_{in}^{(t)},\mathcal{X}_o^{(t)}}\|\boldsymbol{e}_{g,o}^{(t)}\|^2 \\
&\quad= \mathbb{E}_{\mathcal{X}_{in}^{(t)},\mathcal{X}_o^{(t)}}\Big\|\nabla Q_k^{(\boldsymbol{t})}\left(w-\alpha\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)});\boldsymbol{\mathcal{X}}_o^{(t)}\right) \\
&\quad- \nabla J_k^{(\boldsymbol{t})}(w-\alpha\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}))\Big\|^2 \\
&\quad\overset{(a)}{=} \mathbb{E}_{\mathcal{X}_{in}^{(t)},\mathcal{X}_o^{(t)}}\left\|\nabla Q_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2;\boldsymbol{\mathcal{X}}_o^{(t)})-\nabla J_k^{(\boldsymbol{t})}(\widetilde{\boldsymbol{w}}_2)\right\|^2 \\
&\quad\overset{(b)}{\leq} \frac{\sigma_G^2}{|\mathcal{X}_o|}
\end{aligned} \tag{83}$$

where (a) follows from defining $\widetilde{\boldsymbol{w}}_2 \triangleq w-\alpha\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})$, and (b) follows from conditioning on $\boldsymbol{\mathcal{X}}_{in}^{(t)}$ and Assumption 4. Likewise,

$$\begin{aligned}
&\|\boldsymbol{e}_{g,x}^{(t)}\| \\
&= \left\|\nabla J_k^{(\boldsymbol{t})}(w-\alpha\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}))-\nabla J_k^{(\boldsymbol{t})}(w-\alpha\nabla J_k^{(\boldsymbol{t})}(w))\right\| \\
&\overset{(a)}{\leq} \alpha L\left\|\nabla J_k^{(\boldsymbol{t})}(w)-\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right\|
\end{aligned} \tag{84}$$

where (a) follows from Assumption 1. It then follows:

$$\|\boldsymbol{e}_{g,x}^{(t)}\|^4 \leq \alpha^4 L^4\left\|\nabla J_k^{(\boldsymbol{t})}(w)-\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right\|^4 \tag{85}$$

Taking expectations:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\left[\|\boldsymbol{e}_{g,x}^{(t)}\|^4\right] \\
&\leq \alpha^4 L^4\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\left[\left\|\nabla J_k^{(\boldsymbol{t})}(w)-\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right\|^4\right] \\
&\overset{(a)}{\leq} \alpha^4 L^4\frac{3\sigma_G^4}{|\mathcal{X}_{in}|^2}
\end{aligned} \tag{86}$$

where (a) follows from Assumption 4. Similarly:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\left[\|\boldsymbol{e}_{g,x}^{(t)}\|^2\right] \\
&\overset{(a)}{\leq} \alpha^2 L^2\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\left[\left\|\nabla J_k^{(\boldsymbol{t})}(w)-\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right\|^2\right] \\
&\overset{(b)}{\leq} \alpha^2 L^2\frac{\sigma_G^2}{|\mathcal{X}_{in}|}
\end{aligned} \tag{87}$$

where (a) follows from taking square and expectation of (84), and (b) follows from Assumption 4. Furthermore,

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\|\boldsymbol{e}_{h,x}^{(t)}\|^4 &= \alpha^4\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\left\|\nabla^2 J_k^{(\boldsymbol{t})}(w)-\alpha\nabla^2 Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right\|^4 \\
&\overset{(a)}{\leq} \alpha^4\frac{3\sigma_H^4}{|\mathcal{X}_{in}|^2}
\end{aligned} \tag{88}$$

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\|\boldsymbol{e}_{h,x}^{(t)}\|^2 &= \alpha^2\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\left\|\nabla^2 J_k^{(\boldsymbol{t})}(w)-\alpha\nabla^2 Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})\right\|^2 \\
&\overset{(b)}{\leq} \alpha^2\frac{\sigma_H^2}{|\mathcal{X}_{in}|}
\end{aligned} \tag{89}$$

where (a) and (b) follow from Assumption 4. Moreover,

$$\left\|\nabla J_k^{(\boldsymbol{t})}(w-\alpha\nabla J_k^{(\boldsymbol{t})}(w))\right\|^2 \leq B^2 \tag{90}$$

because of Assumption 3, and

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}_{in}^{(t)},\mathcal{X}_o^{(t)}}\left[\|\boldsymbol{e}_{h,x}^{(t)}\|^2\|\boldsymbol{e}_{g,o}^{(t)}\|^2\right] &= \mathbb{E}_{\mathcal{X}_{in}^{(t)}}\left[\mathbb{E}_{\mathcal{X}_o^{(t)}}\left[\|\boldsymbol{e}_{h,x}^{(t)}\|^2\|\boldsymbol{e}_{g,o}^{(t)}\|^2\right]\right] \\
&\overset{(a)}{\leq} \mathbb{E}_{\mathcal{X}_{in}^{(t)}}\left[\|\boldsymbol{e}_{h,x}^{(t)}\|^2\frac{\sigma_G^2}{|\mathcal{X}_o|}\right] \\
&\overset{(b)}{\leq} \alpha^2\frac{\sigma_H^2}{|\mathcal{X}_{in}|}\frac{\sigma_G^2}{|\mathcal{X}_o|}
\end{aligned} \tag{91}$$

where (a) follows from (83), and (b) follows from (89). Substituting the results into (80) completes the proof.

Defining $\overline{J_k^*}(w) := J_k(w-\alpha\nabla J_k(w))$ where $J_k(w) = \mathbb{E}_{t\sim\pi_k}[J_k^{(\boldsymbol{t})}(w)]$, we have the following two lemmas.

Lemma 7: Under assumptions 1,3,5, for each agent k, the disagreement between $\nabla\overline{J_k}^{(\boldsymbol{t})}(\cdot)$ and $\nabla\overline{J_k^*}(\cdot)$ is bounded in expectation, namely, for any $w\in\mathbb{R}^M$:

$$\mathbb{E}\left\|\nabla\overline{J_k}^{(\boldsymbol{t})}(w)-\nabla\overline{J_k^*}(w)\right\|^2 \leq C_2^2 \tag{92}$$

where C_2^2 is a non-negative constant, given by:

$$\begin{aligned}
C_2^2 \triangleq\ &8(1+\alpha L)^2(1+\alpha^2L^2)\gamma_G^2+4B^2\alpha^2\gamma_H^2 \\
&+2\alpha^4\gamma_H^4+16(1+\alpha^4L^4)\gamma_G^4
\end{aligned} \tag{93}$$

Proof: Recall the definitions:

$$\nabla\overline{J_k}^{(\boldsymbol{t})}(w) = (I-\alpha\nabla^2 J_k^{(\boldsymbol{t})}(w))\nabla J_k^{(\boldsymbol{t})}(w-\alpha\nabla J_k^{(\boldsymbol{t})}(w)) \tag{94}$$

$$\nabla\overline{J_k^*}(w) = (I-\alpha\nabla^2 J_k(w))\nabla J_k(w-\alpha\nabla J_k(w)) \tag{95}$$

Defining the error terms:

$$\boldsymbol{e}_{h,t}^{(t)} \triangleq \alpha\nabla^2 J_k(w)-\alpha\nabla^2 J_k^{(\boldsymbol{t})}(w) \tag{96}$$

$$\boldsymbol{e}_{g,t}^{(t)} \triangleq \nabla J_k^{(\boldsymbol{t})}(w-\alpha\nabla J_k^{(\boldsymbol{t})}(w))-\nabla J_k(w-\alpha\nabla J_k(w)) \tag{97}$$

We have:

$$\begin{aligned}
&\nabla\overline{J_k}^{(\boldsymbol{t})}(w) \\
&= (I-\alpha\nabla^2 J_k(w)+\boldsymbol{e}_{h,t}^{(t)})(\nabla J_k(w-\alpha\nabla J_k(w))+\boldsymbol{e}_{g,t}^{(t)})
\end{aligned} \tag{98}$$

It then follows:

$$\begin{aligned}
&\nabla\overline{J_k}^{(\boldsymbol{t})}(w)-\nabla\overline{J_k^*}(w) \\
&= (I-\alpha\nabla^2 J_k(w))\boldsymbol{e}_{g,t}^{(t)}+\boldsymbol{e}_{h,t}^{(t)}\nabla J_k(w-\alpha\nabla J_k(w))+\boldsymbol{e}_{h,t}^{(t)}\boldsymbol{e}_{g,t}^{(t)}
\end{aligned} \tag{99}$$

Taking the norms:

$$\left\|\nabla\overline{J_k}^{(\boldsymbol{t})}(w)-\nabla\overline{J_k^*}(w)\right\|$$

$$\overset{(a)}{\leq} \left\|(I-\alpha\nabla^2 J_k(w))\right\| \|\boldsymbol{e}_{g,t}^{(t)}\| + \|\boldsymbol{e}_{h,t}^{(t)}\| \left\|\nabla J_k(w-\alpha\nabla J_k(w))\right\| + \|\boldsymbol{e}_{h,t}^{(t)}\|\|\boldsymbol{e}_{g,t}^{(t)}\|$$

$$\overset{(b)}{\leq} \left\|(I-\alpha\nabla^2 J_k(w))\right\| \|\boldsymbol{e}_{g,t}^{(t)}\| + \|\boldsymbol{e}_{h,t}^{(t)}\| \left\|\nabla J_k(w-\alpha\nabla J_k(w))\right\| + \frac{\|\boldsymbol{e}_{h,t}^{(t)}\|^2}{2} + \frac{\|\boldsymbol{e}_{g,t}^{(t)}\|^2}{2} \tag{100}$$

where (a) follows from the sub-multiplicative property of norms and the triangle inequality, (b) follows from $ab \leq \frac{a^2+b^2}{2}$. Using $(\sum_{i=1}^4 x_i)^2 \leq 4(\sum_{i=1}^4 x_i^2)$ and taking expectation yield:

$$\begin{aligned}\mathbb{E}\left\|\nabla\overline{J_k}^{(\boldsymbol{t})}(w) - \nabla\overline{J_k^*}(w)\right\|^2 &\leq 4\left\|(I-\alpha\nabla^2 J_k(w))\right\|^2 \mathbb{E}\left[\|\boldsymbol{e}_{g,t}^{(t)}\|^2\right] \\ &\quad + 4\left\|\nabla J_k(w-\alpha\nabla J_k(w))\right\|^2 \mathbb{E}\left[\|\boldsymbol{e}_{h,t}^{(t)}\|^2\right] \\ &\quad + 2\mathbb{E}\left[\|\boldsymbol{e}_{h,t}^{(t)}\|^4\right] + 2\mathbb{E}\left[\|\boldsymbol{e}_{g,t}^{(t)}\|^4\right] \end{aligned} \tag{101}$$

We bound terms in (101) one by one. Note that

$$\left\|(I-\alpha\nabla^2 J_k(w))\right\|^2 \leq (1+\alpha L)^2 \tag{102}$$

by Assumption 1, while

$$\begin{aligned}&\|\boldsymbol{e}_{g,t}^{(t)}\| \\ &= \left\|\nabla J_k^{(\boldsymbol{t})}(w-\alpha\nabla J_k^{(\boldsymbol{t})}(w)) - \nabla J_k(w-\alpha\nabla J_k(w))\right\| \\ &\overset{(a)}{=} \left\|\nabla J_k^{(\boldsymbol{t})}(w-\alpha\nabla J_k^{(\boldsymbol{t})}(w)) - \nabla J_k^{(\boldsymbol{t})}(w-\alpha\nabla J_k(w)) + \nabla J_k^{(\boldsymbol{t})}(\widetilde{w}_3) - \nabla J_k(\widetilde{w}_3)\right\| \\ &\overset{(b)}{\leq} \left\|\nabla J_k^{(\boldsymbol{t})}(w-\alpha\nabla J_k^{(\boldsymbol{t})}(w)) - \nabla J_k^{(\boldsymbol{t})}(w-\alpha\nabla J_k(w))\right\| + \left\|\nabla J_k^{(\boldsymbol{t})}(\widetilde{w}_3) - \nabla J_k(\widetilde{w}_3)\right\| \\ &\overset{(c)}{\leq} \alpha L\left\|\nabla J_k^{(\boldsymbol{t})}(w) - \nabla J_k(w)\right\| + \left\|\nabla J_k^{(\boldsymbol{t})}(\widetilde{w}_3) - \nabla J_k(\widetilde{w}_3)\right\| \end{aligned} \tag{103}$$

where (a) follows from the definition $\widetilde{w}_3 \triangleq w - \alpha\nabla J_k(w)$, (b) follows from triangle inequality, and (c) follows from Assumption 1.

For the second-order moment of $\boldsymbol{e}_{g,t}^{(t)}$, using $(a+b)^2 \leq 2(a^2+b^2)$ and taking expectation result in:

$$\begin{aligned}\mathbb{E}_{t\sim\pi_k}\|\boldsymbol{e}_{g,t}^{(t)}\|^2 &\leq 2\alpha^2 L^2 \mathbb{E}_{t\sim\pi_k}\left[\left\|\nabla J_k^{(\boldsymbol{t})}(w) - \nabla J_k(w)\right\|^2\right] + 2\mathbb{E}_{t\sim\pi_k}\left[\left\|\nabla J_k^{(\boldsymbol{t})}(\widetilde{w}_3) - \nabla J_k(\widetilde{w}_3)\right\|^2\right] \\ &\overset{(a)}{\leq} 2\alpha^2 L^2\gamma_G^2 + 2\gamma_G^2 \\ &= 2\gamma_G^2(1+\alpha^2 L^2) \end{aligned} \tag{104}$$

where (a) follows from Assumption 5.

For the fourth-order moment of $\boldsymbol{e}_{g,t}^{(t)}$, using $(a+b)^4 \leq 8(a^4+b^4)$ and taking expectation result in:

$$\begin{aligned}\mathbb{E}_{t\sim\pi_k}\left[\|\boldsymbol{e}_{g,t}^{(t)}\|^4\right] &\leq 8\alpha^4 L^4 \mathbb{E}_{t\sim\pi_k}\left[\left\|\nabla J_k^{(\boldsymbol{t})}(w) - \nabla J_k(w)\right\|^4\right] + 8\mathbb{E}_{t\sim\pi_k}\left[\left\|\nabla J_k^{(\boldsymbol{t})}(\widetilde{w}_3) - \nabla J_k(\widetilde{w}_3)\right\|^4\right] \\ &\overset{(a)}{\leq} 8\alpha^4 L^4\gamma_G^4 + 8\gamma_G^4 \\ &\leq 8\gamma_G^4(1+\alpha^4 L^4) \end{aligned} \tag{105}$$

where (a) follows from Assumption 5. Also,

$$\left\|\nabla J_k(w-\alpha\nabla J_k(w))\right\|^2 \leq B^2 \tag{106}$$

by Assumption 3, and

$$\begin{aligned}\mathbb{E}_{t\sim\pi_k}\|\boldsymbol{e}_{h,t}^{(t)}\|^4 &= \mathbb{E}_{t\sim\pi_k}\left\|\alpha\nabla^2 J_k(w) - \alpha\nabla^2 J_k^{(\boldsymbol{t})}(w)\right\|^4 \\ &= \alpha^4\mathbb{E}_{t\sim\pi_k}\left\|\nabla^2 J_k(w) - \nabla^2 J_k^{(\boldsymbol{t})}(w)\right\|^4 \\ &\overset{(a)}{\leq} \alpha^4\gamma_H^4 \end{aligned} \tag{107}$$

$$\mathbb{E}_{t\sim\pi_k}\|\boldsymbol{e}_{h,t}^{(t)}\|^2 \overset{(b)}{\leq} \sqrt{\mathbb{E}_{t\sim\pi_k}\|\boldsymbol{e}_{h,t}^{(t)}\|^4} \overset{(c)}{\leq} \alpha^2\gamma_H^2 \tag{108}$$

where (a) follows from Assumption 5, (b) follows from Jensen's inequality, and (c) follows from taking square root of (a). Inserting all the results into (101) completes the proof.

Next, we prove the last intermediate lemma.

Lemma 8: Under assumptions 1,3,4,5, for each agent k, the disagreement between $\nabla\overline{J_k^*}(\cdot)$ and $\nabla\widehat{J}_k(\cdot)$ is bounded, namely, for any $w \in \mathbb{R}^M$:

$$\left\|\nabla\overline{J_k^*}(w) - \nabla\widehat{J}_k(w)\right\| \leq C_3 \tag{109}$$

where C_3 is a non-negative constant, given by:

$$\begin{aligned}C_3 \triangleq\ & (1+\alpha L)\alpha L\frac{\sigma_G}{\sqrt{|\mathcal{X}_{in}|}} + (1+\alpha L)^2\gamma_G + B\alpha\frac{\sigma_H}{\sqrt{|\mathcal{X}_{in}|}} \\ &+ B\alpha\gamma_H + \alpha^2\frac{\sigma_H^2}{|\mathcal{X}_{in}|} + \alpha^2\gamma_H^2 + \alpha^2 L^2\frac{\sigma_G^2}{|\mathcal{X}_{in}|} \\ &+ 2(1+\alpha^2 L^2)\gamma_G^2 \end{aligned} \tag{110}$$

Proof: Recall the definitions:

$$\nabla\overline{J_k^*}(w) = (I-\alpha\nabla^2 J_k(w))\nabla J_k(w-\alpha\nabla J_k(w)) \tag{111}$$

$$\begin{aligned}\nabla\widehat{J}_k(w) &= \mathbb{E}\Big[(I-\alpha\nabla^2 Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})) \times \nabla Q_k^{(\boldsymbol{t})}\left(w-\alpha\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)});\boldsymbol{\mathcal{X}}_o^{(t)}\right)\Big] \\ &= \mathbb{E}_{t\sim\pi_k}\Big[\mathbb{E}_{\mathcal{X}_{in}^{(t)}}\Big[(I-\alpha\nabla^2 Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})) \times \nabla J_k^{(\boldsymbol{t})}(w-\alpha\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}))\Big]\Big] \end{aligned} \tag{112}$$

Recall the error terms:

$$\boldsymbol{e}_{h,x}^{(t)} \triangleq \alpha\nabla^2 J_k^{(\boldsymbol{t})}(w) - \alpha\nabla^2 Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}) \tag{113}$$

$$\boldsymbol{e}_{h,t}^{(t)} \triangleq \alpha\nabla^2 J_k(w) - \alpha\nabla^2 J_k^{(\boldsymbol{t})}(w) \tag{114}$$

$$\begin{aligned}\boldsymbol{e}_{g,x}^{(t)} \triangleq{}& \nabla J_k^{(\boldsymbol{t})}(w - \alpha\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})) \\ &- \nabla J_k^{(\boldsymbol{t})}(w - \alpha\nabla J_k^{(\boldsymbol{t})}(w))\end{aligned} \tag{115}$$

$$\boldsymbol{e}_{g,t}^{(t)} \triangleq \nabla J_k^{(\boldsymbol{t})}(w - \alpha\nabla J_k^{(\boldsymbol{t})}(w)) - \nabla J_k(w - \alpha\nabla J_k(w)) \tag{116}$$

We can rewrite the components of the adjusted objective gradient as:

$$I - \alpha\nabla^2 Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)}) = I - \alpha\nabla^2 J_k(w) + \boldsymbol{e}_{h,x}^{(t)} + \boldsymbol{e}_{h,t}^{(t)} \tag{117}$$

$$\begin{aligned}&\nabla J_k^{(\boldsymbol{t})}(w - \alpha\nabla Q_k^{(\boldsymbol{t})}(w;\boldsymbol{\mathcal{X}}_{in}^{(t)})) \\ &\qquad = \nabla J_k(w - \alpha\nabla J_k(w)) + \boldsymbol{e}_{g,x}^{(t)} + \boldsymbol{e}_{g,t}^{(t)}\end{aligned} \tag{118}$$

and the distance as:

$$\begin{aligned}&\left\|\nabla\widehat{J}_k(w) - \nabla\overline{J_k^*}(w)\right\| \\ &= \Big\|\mathbb{E}\Big[(I - \alpha\nabla^2 J_k(w))(\boldsymbol{e}_{g,x}^{(t)} + \boldsymbol{e}_{g,t}^{(t)}) \\ &\quad + \nabla J_k(w - \alpha\nabla J_k(w))(\boldsymbol{e}_{h,x}^{(t)} + \boldsymbol{e}_{h,t}^{(t)}) \\ &\quad + (\boldsymbol{e}_{g,x}^{(t)} + \boldsymbol{e}_{g,t}^{(t)})(\boldsymbol{e}_{h,x}^{(t)} + \boldsymbol{e}_{h,t}^{(t)})\Big]\Big\| \\ &\overset{(a)}{\leq} \mathbb{E}\Big[\Big\|(I - \alpha\nabla^2 J_k(w))(\boldsymbol{e}_{g,x}^{(t)} + \boldsymbol{e}_{g,t}^{(t)}) \\ &\quad + \nabla J_k(w - \alpha\nabla J_k(w))(\boldsymbol{e}_{h,x}^{(t)} + \boldsymbol{e}_{h,t}^{(t)}) \\ &\quad + (\boldsymbol{e}_{g,x}^{(t)} + \boldsymbol{e}_{g,t}^{(t)})(\boldsymbol{e}_{h,x}^{(t)} + \boldsymbol{e}_{h,t}^{(t)})\Big\|\Big] \\ &\overset{(b)}{\leq} \left\|(I - \alpha\nabla^2 J_k(w))\right\|\mathbb{E}\left[\|\boldsymbol{e}_{g,x}^{(t)}\|\right] \\ &\quad + \left\|(I - \alpha\nabla^2 J_k(w))\right\|\mathbb{E}\left[\|\boldsymbol{e}_{g,t}^{(t)}\|\right] \\ &\quad + \left\|\nabla J_k(w - \alpha\nabla J_k(w))\right\|\mathbb{E}\left[\|\boldsymbol{e}_{h,x}^{(t)}\|\right] \\ &\quad + \left\|\nabla J_k(w - \alpha\nabla J_k(w))\right\|\mathbb{E}\left[\|\boldsymbol{e}_{h,t}^{(t)}\|\right] \\ &\quad + \mathbb{E}\left[\|\boldsymbol{e}_{h,x}^{(t)}\|^2\right] + \mathbb{E}\left[\|\boldsymbol{e}_{h,t}^{(t)}\|^2\right] + \mathbb{E}\left[\|\boldsymbol{e}_{g,x}^{(t)}\|^2\right] \\ &\quad + \mathbb{E}\left[\|\boldsymbol{e}_{g,t}^{(t)}\|^2\right]\end{aligned} \tag{119}$$

where (a) follows from Jensen's inequality, (b) follows from triangle inequality and $\|(a+b)(c+d)\| \leq \|a\|^2 + \|b\|^2 + \|c\|^2 + \|d\|^2$.

We bound the terms in (119) one by one. Note that

$$\left\|(I - \alpha\nabla^2 J_k(w))\right\| \leq (1 + \alpha L) \tag{120}$$

by Assumption 1. Also,

$$\mathbb{E}\|\boldsymbol{e}_{g,x}^{(t)}\| \overset{(a)}{\leq} \sqrt{\mathbb{E}\|\boldsymbol{e}_{g,x}^{(t)}\|^2} \overset{(b)}{\leq} \alpha L\frac{\sigma_G}{\sqrt{|\mathcal{X}_{in}|}} \tag{121}$$

where (a) follows from Jensen's inequality, and (b) follows from (87). Likewise,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{e}_{g,t}^{(t)}\| &\overset{(a)}{\leq} \alpha L\mathbb{E}_{t\sim\pi_k}\left[\left\|\nabla J_k^{(\boldsymbol{t})}(w) - \nabla J_k(w)\right\|\right] \\ &\quad + \mathbb{E}_{t\sim\pi_k}\left[\left\|\nabla J_k^{(\boldsymbol{t})}(\widetilde{w}_3) - \nabla J_k(\widetilde{w}_3)\right\|\right] \\ &\overset{(b)}{\leq} \alpha L\gamma_G + \gamma_G \\ &\leq (1 + \alpha L)\gamma_G\end{aligned} \tag{122}$$

where (a) follows from (103) and taking the expectation, and (b) follows from Assumption 5. Moreover,

$$\left\|\nabla J_k(w - \alpha\nabla J_k(w))\right\| \leq B \tag{123}$$

by Assumption 3, and

$$\mathbb{E}\|\boldsymbol{e}_{h,x}^{(t)}\| \overset{(a)}{\leq} \sqrt{\mathbb{E}\|\boldsymbol{e}_{h,x}^{(t)}\|^2} \overset{(b)}{\leq} \alpha\frac{\sigma_H}{\sqrt{|\mathcal{X}_{in}|}} \tag{124}$$

where (a) follows from Jensen's inequality, and (b) follows from (89). Also,

$$\mathbb{E}\|\boldsymbol{e}_{h,t}^{(t)}\| \overset{(a)}{\leq} \sqrt{\mathbb{E}\|\boldsymbol{e}_{h,t}^{(t)}\|^2} \overset{(b)}{\leq} \alpha\gamma_H \tag{125}$$

where (a) follows from Jensen's inequality, and (b) follows from (108). Moreover,

$$\mathbb{E}\|\boldsymbol{e}_{h,x}^{(t)}\|^2 \leq \alpha^2\frac{\sigma_H^2}{|\mathcal{X}_{in}|} \tag{126}$$

by (89),

$$\mathbb{E}\|\boldsymbol{e}_{h,t}^{(t)}\|^2 \leq \alpha^2\gamma_H^2 \tag{127}$$

by (108),

$$\mathbb{E}\|\boldsymbol{e}_{g,x}^{(t)}\|^2 \leq \alpha^2 L^2\frac{\sigma_G^2}{|\mathcal{X}_{in}|} \tag{128}$$

by (87), and

$$\mathbb{E}\|\boldsymbol{e}_{g,t}^{(t)}\|^2 \leq 2(1 + \alpha^2 L^2)\gamma_G^2 \tag{129}$$

by (104). Inserting all the bounds into (119) completes the proof.

Now, combining the results of the previous three intermediate lemmas, we will prove that $C^2 = \frac{3}{|\mathcal{S}_k|}(C_1^2 + C_2^2 + C_3^2)$, i.e.,

$$\mathbb{E}\left\|\nabla\overline{Q_k}(w) - \nabla\widehat{J}_k(w)\right\|^2 \leq \frac{3}{|\mathcal{S}_k|}(C_1^2 + C_2^2 + C_3^2) \tag{130}$$

where C_1, C_2 and C_3 expressions are given in Lemma 6, Lemma 7 and Lemma 8, respectively.

Proof:

$$\begin{aligned}&\mathbb{E}\left\|\nabla\overline{Q_k}(w) - \nabla\widehat{J}_k(w)\right\|^2 \\ &= \mathbb{E}\left\|\frac{1}{|\mathcal{S}_k|}\sum_{\boldsymbol{t}\in\boldsymbol{\mathcal{S}}_k}\nabla\overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla\widehat{J}_k(w)\right\|^2\end{aligned}$$

$$= \mathbb{E}\Big\| \frac{1}{|\mathcal{S}_k|} \sum_{\boldsymbol{t}\in\boldsymbol{\mathcal{S}}_k} (\nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \widehat{J}_k(w)) \Big\|^2$$

$$= \frac{1}{|\mathcal{S}_k|^2} \sum_{\boldsymbol{t}\in\boldsymbol{\mathcal{S}}_k} \mathbb{E}\left[\left\| \nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \widehat{J}_k(w) \right\|^2 \right]$$

$$+ \frac{1}{|\mathcal{S}_k|^2} \sum_{\boldsymbol{t}_1\neq\boldsymbol{t}_2} \mathbb{E}\Big[(\nabla Q_k^{(\boldsymbol{t}_1)}(w) - \nabla \widehat{J}_k(w))^{\mathsf{T}}$$

$$\times (\nabla Q_k^{(\boldsymbol{t}_2)}(w) - \nabla \widehat{J}_k(w)) \Big]$$

$$\overset{(a)}{=} \frac{1}{|\mathcal{S}_k|^2} \sum_{\boldsymbol{t}\in\boldsymbol{\mathcal{S}}_k} \mathbb{E}\left[\left\| \nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \widehat{J}_k(w) \right\|^2 \right]$$

$$+ \frac{1}{|\mathcal{S}_k|^2} \sum_{\boldsymbol{t}_1\neq\boldsymbol{t}_2} \left[\mathbb{E}\left[(\nabla Q_k^{(\boldsymbol{t}_1)}(w) - \nabla \widehat{J}_k(w))^{\mathsf{T}} \right] \right.$$

$$\left. \times\, \mathbb{E}\left[\nabla Q_k^{(\boldsymbol{t}_2)}(w) - \nabla \widehat{J}_k(w) \right] \right]$$

$$\overset{(b)}{=} \frac{1}{|\mathcal{S}_k|^2} \sum_{\boldsymbol{t}\in\boldsymbol{\mathcal{S}}_k} \mathbb{E}\left\| \nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \widehat{J}_k(w) \right\|^2$$

$$= \frac{1}{|\mathcal{S}_k|} \mathbb{E}\left\| \nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \widehat{J}_k(w) \right\|^2 \tag{131}$$

where (a) follows from independence assumption on batch of tasks, and (b) follows from the definition of the adjusted objective. Now, bounding the term in (131):

$$\nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \widehat{J}_k(w)$$

$$= (\nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \overline{J_k}^{(\boldsymbol{t})}(w)) + (\nabla \overline{J_k}^{(\boldsymbol{t})}(w) - \nabla \overline{J_k^*}(w))$$

$$+ (\nabla \overline{J_k^*}(w) - \nabla \widehat{J}_k(w)) \tag{132}$$

By triangle inequality:

$$\left\| \nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \widehat{J}_k(w) \right\| \tag{133}$$

$$\leq \left\| \nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \overline{J_k}^{(\boldsymbol{t})}(w) \right\| + \left\| \nabla \overline{J_k}^{(\boldsymbol{t})}(w) - \nabla \overline{J_k^*}(w) \right\|$$

$$+ \left\| \nabla \overline{J_k^*}(w) - \nabla \widehat{J}_k(w) \right\| \tag{134}$$

Using $(\sum_{i=1}^3 x_i)^2 \leq 3(\sum_{i=1}^3 x_i^2)$:

$$\left\| \nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \widehat{J}_k(w) \right\|^2$$

$$\leq 3\left\| \nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \overline{J_k}^{(\boldsymbol{t})}(w) \right\|^2 + 3\left\| \nabla \overline{J_k}^{(\boldsymbol{t})}(w) - \nabla \overline{J_k^*}(w) \right\|^2$$

$$+ 3\left\| \nabla \overline{J_k^*}(w) - \nabla \widehat{J}_k(w) \right\|^2 \tag{135}$$

Taking expectations:

$$\mathbb{E}\left\| \nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \widehat{J}_k(w) \right\|^2$$

$$\leq 3\mathbb{E}\left[\left\| \nabla \overline{Q_k}^{(\boldsymbol{t})}(w) - \nabla \overline{J_k}^{(\boldsymbol{t})}(w) \right\|^2 \right]$$

$$+ 3\mathbb{E}\left[\left\| \nabla \overline{J_k}^{(\boldsymbol{t})}(w) - \nabla \overline{J_k^*}(w) \right\|^2 \right] + 3\left\| \nabla \overline{J_k^*}(w) - \nabla \widehat{J}_k(w) \right\|^2 \tag{136}$$

$$\overset{(a)}{\leq} 3(C_1^2 + C_2^2 + C_3^2) \tag{137}$$

and (a) follows from definitions of C_1, C_2, C_3. Inserting (137) into (131) completes the proof.

APPENDIX K
PROOF OF THEOREM 1

For analyzing the centroid model recursion it is useful to define the following variables, which collect all variables from across the network:

$$\mathcal{W}_i \triangleq \operatorname{col}\left\{ \boldsymbol{w}_{1,i}, \ldots, \boldsymbol{w}_{K,i} \right\} \tag{138}$$

$$\mathcal{A} \triangleq A \otimes I_M \tag{139}$$

$$\widehat{\boldsymbol{g}}(\mathcal{W}_i) \triangleq \operatorname{col}\left\{ \nabla \overline{Q_1}(\boldsymbol{w}_{1,i}), \ldots, \nabla \overline{Q_K}(\boldsymbol{w}_{K,i}) \right\} \tag{140}$$

Then, we rewrite the diffusion equations (7a)–(7b) in a more compact form as:

$$\mathcal{W}_i = \mathcal{A}^{\mathsf{T}} \left(\mathcal{W}_{i-1} - \mu \widehat{\boldsymbol{g}}(\mathcal{W}_{i-1}) \right) \tag{141}$$

Multiplying this equation by $\frac{1}{K}\mathbb{1}_K^{\mathsf{T}} \otimes I$ from the left and using (16) we get the recursion:

$$\left(\frac{1}{K}\mathbb{1}_K^{\mathsf{T}} \otimes I \right) \mathcal{W}_i$$

$$= \left(\frac{1}{K}\mathbb{1}_K^{\mathsf{T}} \otimes I \right) \mathcal{W}_{i-1} - \mu \left(\frac{1}{K}\mathbb{1}_K^{\mathsf{T}} \otimes I \right) \widehat{\boldsymbol{g}}(\mathcal{W}_{i-1}) \tag{142}$$

Rewriting the centroid launch model as:

$$\boldsymbol{w}_{c,i} = \sum_{k=1}^{K} \frac{1}{K} \boldsymbol{w}_{k,i} = \left(\frac{1}{K}\mathbb{1}_K^{\mathsf{T}} \otimes I \right) \mathcal{W}_i \tag{143}$$

and defining the extended centroid matrix,

$$\mathcal{W}_{c,i} \triangleq \mathbb{1}_K \otimes \boldsymbol{w}_{c,i} = \left(\frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}} \otimes I \right) \left(\mathcal{W}_{i-1} - \mu \widehat{\boldsymbol{g}}(\mathcal{W}_{i-1}) \right) \tag{144}$$

it follows that:

$$\mathcal{W}_i - \mathcal{W}_{c,i}$$

$$= \left(\mathcal{A}^{\mathsf{T}} - \frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}} \otimes I \right) \left(\mathcal{W}_{i-1} - \mu \widehat{\boldsymbol{g}}(\mathcal{W}_{i-1}) \right)$$

$$\overset{(a)}{=} \left(\mathcal{A}^{\mathsf{T}} - \frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}} \otimes I \right) \left(I - \frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}} \otimes I \right)$$

$$\times \left(\mathcal{W}_{i-1} - \mu \widehat{\boldsymbol{g}}(\mathcal{W}_{i-1}) \right)$$

$$= \left(\mathcal{A}^{\mathsf{T}} - \frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}} \otimes I \right)$$

$$\times \left(\mathcal{W}_{i-1} - \mathcal{W}_{c,i-1} - \left(I - \frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}} \otimes I\right)\mu\widehat{\boldsymbol{g}}(\mathcal{W}_{i-1})\right)$$

$$\overset{(b)}{=} \left(\mathcal{A}^{\mathsf{T}} - \frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}} \otimes I\right)\left(\mathcal{W}_{i-1} - \mathcal{W}_{c,i-1} - \mu\widehat{\boldsymbol{g}}(\mathcal{W}_{i-1})\right) \tag{145}$$

where (a) and (b) follows from the equality:

$$\left(\mathcal{A}^{\mathsf{T}} - \frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}} \otimes I\right)\left(I - \frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}} \otimes I\right) = \mathcal{A}^{\mathsf{T}} - \frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}} \otimes I \tag{146}$$

which follows from doubly-stochastic combination matrix assumption. Taking the squared norms:

$$\begin{aligned}
&\left\|\mathcal{W}_i - \mathcal{W}_{c,i}\right\|^2 \\
&= \left\|\left(\mathcal{A}^{\mathsf{T}} - \frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}} \otimes I\right)\left(\mathcal{W}_{i-1} - \mathcal{W}_{c,i-1} - \mu\widehat{\boldsymbol{g}}(\mathcal{W}_{i-1})\right)\right\|^2 \\
&\overset{(a)}{\leq} \lambda_2^2\left\|\mathcal{W}_{i-1} - \mathcal{W}_{c,i-1} - \mu\widehat{\boldsymbol{g}}(\mathcal{W}_{i-1})\right\|^2 \\
&\overset{(b)}{\leq} \lambda_2\left\|\mathcal{W}_{i-1} - \mathcal{W}_{c,i-1}\right\|^2 + \mu^2\frac{\lambda_2^2}{1-\lambda_2}\|\widehat{\boldsymbol{g}}(\mathcal{W}_{i-1})\|^2
\end{aligned} \tag{147}$$

where (a) follows from sub-multiplicative property of the norms and (b) follows from $\|a+b\|^2 \leq \frac{1}{\beta}\|a\|^2 + \frac{1}{1-\beta}\|b\|^2$ for $0 < \beta < 1$ with the choice of β:

$$\beta = \lambda_2 = \left\|\mathcal{A}^{\mathsf{T}} - \frac{1}{K}\mathbb{1}_K\mathbb{1}_K^{\mathsf{T}} \otimes I\right\| < 1 \tag{148}$$

Taking expectation conditioned on $\mathcal{W}_{i-1}$:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathcal{W}_i - \mathcal{W}_{c,i}\right\|^2 | \mathcal{W}_{i-1}\right] \\
&\leq \lambda_2\mathbb{E}\left[\left\|\mathcal{W}_{i-1} - \mathcal{W}_{c,i-1}\right\|^2 | \mathcal{W}_{i-1}\right] \\
&\quad + \mu^2\frac{\lambda_2^2}{1-\lambda_2}\mathbb{E}\left[\|\widehat{\boldsymbol{g}}(\mathcal{W}_{i-1})\|^2 | \mathcal{W}_{i-1}\right] \\
&\leq \lambda_2\mathbb{E}\left[\left\|\mathcal{W}_{i-1} - \mathcal{W}_{c,i-1}\right\|^2 | \mathcal{W}_{i-1}\right] \\
&\quad + \mu^2\frac{\lambda_2^2}{1-\lambda_2}\sum_{k=1}^K\mathbb{E}\left[\left\|\nabla\overline{Q_k}(\boldsymbol{w}_{k,i-1})\right\|^2 | \mathcal{W}_{i-1}\right] \\
&\overset{(a)}{=} \lambda_2\mathbb{E}\left[\left\|\mathcal{W}_{i-1} - \mathcal{W}_{c,i-1}\right\|^2 | \mathcal{W}_{i-1}\right] \\
&\quad + \mu^2\frac{\lambda_2^2}{1-\lambda_2}\sum_{k=1}^K\left\|\nabla\widehat{J}_k(\boldsymbol{w}_{k,i-1})\right\|^2 \\
&\quad + \mu^2\frac{\lambda_2^2}{1-\lambda_2}\sum_{k=1}^K\mathbb{E}\left[\left\|\nabla\overline{Q_k}(\boldsymbol{w}_{k,i-1}) - \nabla\widehat{J}_k(\boldsymbol{w}_{k,i-1})\right\|^2 | \mathcal{W}_{i-1}\right] \\
&\overset{(b)}{\leq} \lambda_2\mathbb{E}\left[\left\|\mathcal{W}_{i-1} - \mathcal{W}_{c,i-1}\right\|^2 | \mathcal{W}_{i-1}\right] + \mu^2\frac{\lambda_2^2}{1-\lambda_2}K\widehat{B}^2 \\
&\quad + \mu^2\frac{\lambda_2^2}{1-\lambda_2}KC^2 \\
&= \lambda_2\mathbb{E}\left[\left\|\mathcal{W}_{i-1} - \mathcal{W}_{c,i-1}\right\|^2 | \mathcal{W}_{i-1}\right] + \mu^2\frac{\lambda_2^2}{1-\lambda_2}K\left(\widehat{B}^2 + C^2\right)
\end{aligned} \tag{149}$$

where (a) follows from dropping the cross-terms due to unbiasedness of the stochastic gradient update, and (b) follows from Lemmas 3 and 5. Taking expectation again to remove the conditioning:

$$\begin{aligned}
\mathbb{E}\left\|\mathcal{W}_i - \mathcal{W}_{c,i}\right\|^2 &\leq \lambda_2\mathbb{E}\left\|\mathcal{W}_{i-1} - \mathcal{W}_{c,i-1}\right\|^2 \\
&\quad + \mu^2\frac{\lambda_2^2}{1-\lambda_2}K\left(\widehat{B}^2 + C^2\right)
\end{aligned} \tag{150}$$

We can iterate, starting from $i = 0$, to obtain:

$$\begin{aligned}
&\mathbb{E}\left\|\mathcal{W}_i - \mathcal{W}_{c,i}\right\|^2 \\
&\leq \lambda_2^i\|\mathcal{W}_0 - \mathcal{W}_{c,0}\|^2 + \mu^2\frac{\lambda_2^2}{1-\lambda_2}K\left(\widehat{B}^2 + C^2\right)\sum_{k=0}^i\lambda_2^{i-k} \\
&\leq \lambda_2^i\|\mathcal{W}_0 - \mathcal{W}_{c,0}\|^2 + \mu^2\frac{\lambda_2^2}{(1-\lambda_2)^2}K\left(\widehat{B}^2 + C^2\right) \\
&\overset{(a)}{\leq} \mu^2\frac{\lambda_2^2}{(1-\lambda_2)^2}K\left(\widehat{B}^2 + C^2\right) + O(\mu^3)
\end{aligned} \tag{151}$$

where (a) holds whenever:

$$\begin{aligned}
&\lambda_2^i\|\mathcal{W}_0 - \mathcal{W}_{c,0}\|^2 \leq c\mu^3 \\
&\iff i\log\lambda_2 \leq 3\log\mu + \log c - 2\log\|\mathcal{W}_0 - \mathcal{W}_{c,0}\| \\
&\iff i \geq \frac{3\log\mu}{\log\lambda_2} + O(1) = o(1/\mu)
\end{aligned} \tag{152}$$

where c is an arbitrary constant.

APPENDIX L
PROOF OF THEOREM 2

We first prove two intermediate lemmas, then conclude the proof.

Recall the centroid launch model:

$$\boldsymbol{w}_{c,i} = \sum_{k=1}^K\frac{1}{K}\boldsymbol{w}_{k,i} = \left(\frac{1}{K}\mathbb{1}_K^{\mathsf{T}} \otimes I\right)\mathcal{W}_i \tag{153}$$

which leads to the recursion:

$$\boldsymbol{w}_{c,i} = \boldsymbol{w}_{c,i-1} - \frac{\mu}{K}\sum_{k=1}^K\nabla\overline{Q_k}(\boldsymbol{w}_{k,i-1}) \tag{154}$$

This is almost an exact gradient descent on the aggregate cost (6) except for the perturbation terms. Decoupling them gives us:

$$\boldsymbol{w}_{c,i} = \boldsymbol{w}_{c,i-1} - \frac{\mu}{K}\sum_{k=1}^K\nabla\widehat{J}_k(\boldsymbol{w}_{c,i-1}) - \mu\boldsymbol{d}_{i-1} - \mu\boldsymbol{s}_i \tag{155}$$

where the perturbation terms are:

$$\boldsymbol{d}_{i-1} \triangleq \frac{1}{K}\sum_{k=1}^{K}\left(\nabla\widehat{J}_k(\boldsymbol{w}_{k,i-1}) - \nabla\widehat{J}_k(\boldsymbol{w}_{c,i-1})\right) \tag{156}$$

$$\boldsymbol{s}_i \triangleq \frac{1}{K}\sum_{k=1}^{K}\left(\nabla\overline{Q_k}(\boldsymbol{w}_{k,i-1}) - \nabla\widehat{J}_k(\boldsymbol{w}_{k,i-1})\right) \tag{157}$$

Here, $\boldsymbol{d}_{i-1}$ measures the disagreement with the average launch model whereas $\boldsymbol{s}_i$ represents the average stochastic gradient noise in the process. Based on the network disagreement result Theorem 1, we can bound the perturbation terms in (155).

Lemma 9 (Perturbation bounds): Under assumptions 1-6, perturbation terms are bounded for sufficently small outer-step sizes μ after sufficient number of iterations, namely:

$$\mathbb{E}\|\boldsymbol{d}_{i-1}\|^2 \le \mu^2\widehat{L}^2\frac{\lambda_2^2}{(1-\lambda_2)^2}\left(\widehat{B}^2 + C^2\right) + O(\mu^3) \tag{158}$$

$$\mathbb{E}\|\boldsymbol{s}_i\|^2 \le C^2 \tag{159}$$

Proof: We begin by studying the perturbation term $\boldsymbol{s}_i$ arising from the gradient approximations. We have:

$$\begin{aligned}\mathbb{E}\|\boldsymbol{s}_i\|^2 &= \mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\left(\nabla\overline{Q_k}(\boldsymbol{w}_{k,i-1}) - \nabla\widehat{J}_k(\boldsymbol{w}_{k,i-1})\right)\right\|^2\\ &\overset{(a)}{\le} \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla\overline{Q_k}(\boldsymbol{w}_{k,i-1}) - \nabla\widehat{J}_k(\boldsymbol{w}_{k,i-1})\right\|^2\\ &\overset{(b)}{\le} \frac{1}{K}\sum_{k=1}^{K}\left(C^2\right)\\ &= C^2\end{aligned} \tag{160}$$

where (a) follows from Jensen's inequality, and (b) follows from Lemma 5. For the second perturbation term arising from the disagreement within the network, we can bound:

$$\begin{aligned}\|\boldsymbol{d}_{i-1}\|^2 &= \left\|\sum_{k=1}^{K}\frac{1}{K}\left(\nabla\widehat{J}_k(\boldsymbol{w}_{k,i-1}) - \nabla\widehat{J}_k(\boldsymbol{w}_{c,i-1})\right)\right\|^2\\ &\overset{(a)}{\le} \sum_{k=1}^{N}\frac{1}{K}\left\|\nabla\widehat{J}_k(\boldsymbol{w}_{k,i-1}) - \nabla\widehat{J}_k(\boldsymbol{w}_{c,i-1})\right\|^2\\ &\overset{(b)}{\le} \frac{\widehat{L}^2}{K}\sum_{k=1}^{K}\left\|\boldsymbol{w}_{k,i-1} - \boldsymbol{w}_{c,i-1}\right\|^2\\ &= \frac{\widehat{L}^2}{K}\left\|\mathcal{W}_{i-1} - \mathcal{W}_{c,i-1}\right\|^2\end{aligned} \tag{161}$$

where (a) follows from Jensen's inequality, and (b) follows from Lemma 4. Taking the expectation and using Theorem 1 we complete the proof:

$$\mathbb{E}\|\boldsymbol{d}_{i-1}\|^2 \le \mu^2\widehat{L}^2\frac{\lambda_2^2}{(1-\lambda_2)^2}\left(\widehat{B}^2 + C^2\right) + O(\mu^3) \tag{162}$$

Next, we present the second lemma.

Lemma 10 (Descent relation): Under asssumptions 1-6 we have the descent relation:

$$\begin{aligned}&\mathbb{E}\left[\widehat{J}(\boldsymbol{w}_{c,i})|\boldsymbol{w}_{c,i-1}\right]\\ &\le \widehat{J}(\boldsymbol{w}_{c,i-1}) - \frac{\mu}{2}(1-2\mu\widehat{L})\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i-1})\right\|^2 + \frac{1}{2}\mu^2\widehat{L}C^2\\ &\quad + O(\mu^3)\end{aligned} \tag{163}$$

Proof: First, observe that since each individual $\widehat{J}_k(\cdot)$ has Lipschitz gradients by Lemma 4, the same holds for the average:

$$\begin{aligned}&\left\|\nabla\widehat{J}(w) - \nabla\widehat{J}(u)\right\|\\ &\quad= \left\|\nabla\left(\sum_{k=1}^{K}\frac{1}{K}\widehat{J}_k(w)\right) - \nabla\left(\sum_{k=1}^{K}\frac{1}{K}\widehat{J}_k(u)\right)\right\|\\ &\quad= \left\|\sum_{k=1}^{K}\frac{1}{K}\left(\nabla\widehat{J}_k(w) - \nabla\widehat{J}_k(u)\right)\right\|\\ &\quad\overset{(a)}{\le} \sum_{k=1}^{N}\frac{1}{K}\left\|\nabla\widehat{J}_k(w) - \nabla\widehat{J}_k(u)\right\|\\ &\quad\overset{(b)}{\le} \sum_{k=1}^{N}\frac{1}{K}\widehat{L}\left\|w-u\right\|\\ &\quad= \widehat{L}\left\|w-u\right\|\end{aligned} \tag{164}$$

where (a) follows from Jensen's inequality, and (b) follows from Lemma 4. This property then implies the following bound:

$$\begin{aligned}&\widehat{J}(\boldsymbol{w}_{c,i})\\ &\le \widehat{J}(\boldsymbol{w}_{c,i-1}) + \nabla\widehat{J}(\boldsymbol{w}_{c,i-1})^{\mathsf{T}}\left(\boldsymbol{w}_{c,i} - \boldsymbol{w}_{c,i-1}\right)\\ &\quad+ \frac{\widehat{L}}{2}\left\|\boldsymbol{w}_{c,i} - \boldsymbol{w}_{c,i-1}\right\|^2\\ &\overset{(a)}{\le} \widehat{J}(\boldsymbol{w}_{c,i-1}) - \mu\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i-1})\right\|^2 - \mu\nabla\widehat{J}(\boldsymbol{w}_{c,i-1})^{\mathsf{T}}\\ &\quad\times(\boldsymbol{d}_{i-1} + \boldsymbol{s}_i) + \mu^2\frac{\widehat{L}}{2}\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i-1}) + \boldsymbol{d}_{i-1} + \boldsymbol{s}_i\right\|^2\end{aligned} \tag{165}$$

where (a) follows from (155). Taking expectations, conditioned on $\mathcal{W}_{i-1}$ yields:

$$\begin{aligned}&\mathbb{E}\left[\widehat{J}(\boldsymbol{w}_{c,i})|\,\mathcal{W}_{i-1}\right]\\ &\overset{(a)}{\le} \widehat{J}(\boldsymbol{w}_{c,i-1}) - \mu\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i-1})\right\|^2 - \mu\nabla\widehat{J}(\boldsymbol{w}_{c,i-1})^{\mathsf{T}}\boldsymbol{d}_{i-1}\\ &\quad+ \mu^2\frac{\widehat{L}}{2}\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i-1}) + \boldsymbol{d}_{i-1}\right\|^2 + \mu^2\frac{\widehat{L}}{2}\mathbb{E}\left[\|\boldsymbol{s}_i\|^2|\,\mathcal{W}_{i-1}\right]\\ &\overset{(b)}{\le} \widehat{J}(\boldsymbol{w}_{c,i-1}) - \mu\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i-1})\right\|^2 + \frac{\mu}{2}\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i-1})\right\|^2\\ &\quad+ \frac{\mu}{2}\|\boldsymbol{d}_{i-1}\|^2 + \mu^2\widehat{L}\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i-1})\right\|^2 + \mu^2\widehat{L}\|\boldsymbol{d}_{i-1}\|^2\end{aligned}$$

$$+\mu^2\frac{\widehat{L}}{2}\mathbb{E}\left[\|\boldsymbol{s}_i\|^2|\,\mathcal{W}_{i-1}\right]$$
$$\leq \widehat{J}(\boldsymbol{w}_{c,i-1})-\frac{\mu}{2}\left(1-2\mu\widehat{L}\right)\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i-1})\right\|^2$$
$$+\frac{\mu}{2}\left(1+2\mu\widehat{L}\right)\|\boldsymbol{d}_{i-1}\|^2+\mu^2\frac{\widehat{L}}{2}\mathbb{E}\left[\|\boldsymbol{s}_i\|^2|\,\mathcal{W}_{i-1}\right] \quad (166)$$

where (a) follows from $\mathbb{E}\boldsymbol{s}_i=0$, and (b) follows from Cauchy-Schwarz and $ab\leq\frac{a^2+b^2}{2}$.

Taking expectations to remove the conditioning and the bounds from Lemma 9 yields:

$$\mathbb{E}\widehat{J}(\boldsymbol{w}_{c,i})$$
$$\leq \mathbb{E}\widehat{J}(\boldsymbol{w}_{c,i-1})-\frac{\mu}{2}\left(1-2\mu\widehat{L}\right)\mathbb{E}\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i-1})\right\|^2$$
$$+\frac{\mu^3}{2}\left(1+2\mu\widehat{L}\right)\widehat{L}^2\frac{\lambda_2^2}{(1-\lambda_2)^2}\left(\widehat{B}^2+C^2\right)+\mu^2\frac{\widehat{L}}{2}C^2$$
$$+O(\mu^4)$$
$$=\mathbb{E}\widehat{J}(\boldsymbol{w}_{c,i-1})-\frac{\mu}{2}\left(1-2\mu\widehat{L}\right)\mathbb{E}\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i-1})\right\|^2+\mu^2\frac{\widehat{L}}{2}C^2$$
$$+O(\mu^3) \quad (167)$$

The proof of the theorem is based on contradiction. First define:

$$c_1\triangleq\frac{1-2\mu\widehat{L}}{2} \quad (168)$$
$$c_2\triangleq\frac{\widehat{L}C^2}{2}+O(\mu) \quad (169)$$

We will prove that

$$\mathbb{E}\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i^\star})\right\|^2\leq 2\mu\frac{c_2}{c_1} \quad (170)$$
$$i^\star\leq\left(\frac{\widehat{J}(w_0)-\widehat{J}^o}{c_2}\right)1/\mu^2 \quad (171)$$

which correspond to (26) and (27), respectively. Descent relation (163) can be rewritten as:

$$\mathbb{E}\left[\widehat{J}(\boldsymbol{w}_{c,i})|\boldsymbol{w}_{c,i-1}\right]\leq\widehat{J}(\boldsymbol{w}_{c,i-1})-\mu c_1\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i-1})\right\|^2+\mu^2c_2 \quad (172)$$

Suppose there is no time instant $i^\star$ satisfying $\|\nabla\widehat{J}(\boldsymbol{w}_{c,i^\star})\|^2\leq 2\mu\frac{c_2}{c_1}$. Then, for any time i we obtain:

$$\mathbb{E}\widehat{J}(\boldsymbol{w}_{c,i})\overset{(a)}{\leq}\widehat{J}(w_0)-\mu c_1\sum_{k=1}^{i}\left(\mathbb{E}\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,k-1})\right\|^2-\mu\frac{c_2}{c_1}\right)$$
$$\leq\widehat{J}(w_0)-\mu^2c_2i \quad (173)$$

where (a) follows from starting from the first iteration and iterating over (172). But when the limit is taken we get $\lim_{i\to\infty}\mathbb{E}\widehat{J}(\boldsymbol{w}_{c,i})\leq-\infty$, which contradicts the boundedness from below assumption $\widehat{J}(w)\geq\widehat{J}^o$ for all w. This proves (26). In order to prove (27), we iterate over (172) up to time $i^\star$, then, the first time instant where $\mathbb{E}\|\nabla\widehat{J}(\boldsymbol{w}_{c,i^\star})\|^2\leq 2\mu\frac{c_2}{c_1}$ holds:

$$\widehat{J}^o\leq\mathbb{E}\widehat{J}(\boldsymbol{w}_{c,i^\star})$$
$$\leq\widehat{J}(w_0)-\mu c_1\sum_{k=1}^{i^\star}\left(\mathbb{E}\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,k-1})\right\|^2-\mu\frac{c_2}{c_1}\right)$$
$$\leq\widehat{J}(w_0)-\mu^2c_2i^\star \quad (174)$$

Rearranging terms completes the proof.

APPENDIX M
PROOF OF COROLLARY 1

We begin by adding and subtracting $\|\nabla\widehat{J}(\boldsymbol{w}_{c,i^\star})\|^2$:

$$\mathbb{E}\left\|\nabla\overline{J}(\boldsymbol{w}_{c,i^\star})\right\|^2$$
$$=\mathbb{E}\left\|\nabla\overline{J}(\boldsymbol{w}_{c,i^\star})-\nabla\widehat{J}(\boldsymbol{w}_{c,i^\star})+\nabla\widehat{J}(\boldsymbol{w}_{c,i^\star})\right\|^2$$
$$\overset{(a)}{\leq}2\mathbb{E}\left\|\nabla\overline{J}(\boldsymbol{w}_{c,i^\star})-\nabla\widehat{J}(\boldsymbol{w}_{c,i^\star})\right\|^2+2\mathbb{E}\left\|\nabla\widehat{J}(\boldsymbol{w}_{c,i^\star})\right\|^2 \quad (175)$$

where (a) follows from the inequality $\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2$. Inserting (20) and (26) completes the proof.

ACKNOWLEDGMENT

The authors would like to thank Y. Efe Erginbas for helpful discussions on the experiments.

REFERENCES

[1] M. Kayaalp, S. Vlaski, and A. H. Sayed, "Distributed meta-learning with networked agents," in *Proc. Eur. Signal Process. Conf.*, Dublin, Ireland, 2021, pp. 1361–1365, doi: 10.23919/EUSIPCO54536.2021.9616256.

[2] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2021.3079209.

[3] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, Sydney, Australia, Aug. 2017, pp. 1126–1135.

[4] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," Mar. 2018, *arXiv:1803.02999*.

[5] C. Finn, K. Xu, and S. Levine, "Probabilistic model-agnostic meta-learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montréal, Canada, Dec. 2018, pp. 9537–9548.

[6] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals, "Rapid learning or feature reuse? Towards understanding the effectiveness of MAML," in *Proc. Int. Conf. Learn. Representations*, 2020.

[7] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-SGD: Learning to learn quickly for few shot learning," Jul. 2017, *arXiv:1707.09835*.

[8] A. Fallah, A. Mokhtari, and A. Ozdaglar, "On the convergence theory of gradient-based model-agnostic meta-learning algorithms," in *Proc. Int. Conf. Artif. Intell. Statist.*, Aug. 2020, pp. 1082–1092.

[9] K. Ji, J. Yang, and Y. Liang, "Theoretical convergence of multi-step model-agnostic meta-learning," Feb. 2020, *arXiv:2002.07836*.

[10] Z. Zhuang, Y. Wang, K. Yu, and S. Lu, "No-regret non-convex online meta-learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Barcelona, Spain, 2020, pp. 3942–3946.

[11] M.-F. Balcan, M. Khodak, and A. Talwalkar, "Provable guarantees for gradient-based meta-learning," in *Proc. Int. Conf. Mach. Learn.*, Long Beach, CA, USA, Jun. 2019, pp. 424–433.

[12] S. M. Hendryx, A. B. Leach, P. D. Hein, and C. T. Morrison, "Meta-learning initializations for image segmentation," 2019, *arXiv:1912.06290*.

[13] J.-Y. Hsu, Y.-J. Chen, and H. Y. Lee, "Meta learning for end-to-end low-resource speech recognition," in *Proc. IEEE IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7844–7848.

[14] S. Park, H. Jang, O. Simeone, and J. Kang, "Learning to demodulate from few pilots via offline and online meta-learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 226–239, 2021.

[15] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5330–5340.

[16] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Found. Trends Mach. Learn.*, vol. 7, no. 4/5, pp. 311–801, Jul. 2014.

[17] M. Dunbabin and L. Marques, "Robots for environmental monitoring: Significant advancements and applications," *IEEE Robot. Automat. Mag.*, vol. 19, no. 1, pp. 24–39, Mar. 2012.

[18] X. S. Zhang, F. Tang, H. H. Dodge, J. Zhou, and F. Wang, "MetaPred: Meta-learning for clinical risk prediction with limited patient electronic health records," in *Proc. Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 2487–2495.

[19] H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande, "Low data drug discovery with one-shot learning," *ACS Central Sci.*, vol. 3, no. 4, pp. 283–293, 2017.

[20] J. Schmidhuber, "Evolutionary principles in self-referential learning. on learning how to learn: The meta-meta-meta...-hook," Diploma thesis, Institut f. Informatik, Technische Universitat Munchen, Germany, May 1987.

[21] J. Schmidhuber, "Learning to control fast-weight memories: An alternative to dynamic recurrent networks," *Neural Comput.*, vol. 4, no. 1, pp. 131–139, Jan. 1992.

[22] Y. Bengio, S. Bengio, and J. Cloutier, "Learning a synaptic learning rule," in *Proc. Int. Joint Conf. Neural Netw.*, 1991.

[23] S. Bengio, Y. Bengio, J. Cloutier, and J. Gecsei, "On the optimization of a synaptic learning rule," in *Proc. Conf. Optimality Artif. Biol. Neural Netw.*, 1992.

[24] M. Andrychowicz *et al.*, "Learning to learn by gradient descent by gradient descent," in *Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 3981–3989.

[25] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. Int. Conf. Learn. Representations*, Toulon, France, Apr. 2017.

[26] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. Int. Conf. Mach. Learn. Deep Learn. Workshop*, Lille, France, Jul. 2015.

[27] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 3630–3638.

[28] M. Khodak, M.-F. Balcan, and A. Talwalkar, "Adaptive gradient-based meta-learning methods," in *Adv. Neural Inf. Process. Syst.*, Vancouver, Canada, Dec. 2019, pp. 5917–5928.

[29] Y. Jiang, J. Konecný, K. Rush, and S. Kannan, "Improving federated learning personalization via model agnostic meta learning," Sep. 2019, *arXiv:1909.12488*.

[30] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 3557–3568.

[31] F. Chen, M. Luo, Z. Dong, Z. Li, and X. He, "Federated meta-learning with fast convergence and efficient communication," Feb. 2018, *arXiv:1802.07876*.

[32] A. H. Sayed, "Adaptive networks," *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.

[33] A. Koppel, F. Y. Jakubiec, and A. Ribeiro, "A saddle point algorithm for networked online convex optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5149–5164, Oct. 2015.

[34] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[35] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[36] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," in *Proc. IEEE Int. Conf. Decis. Control*, 2003, pp. 4997–5002.

[37] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *Soc. Ind. Appl. Math. J. Optim.*, vol. 26, no. 3, pp. 1835–1854, 2016.

[38] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Proximal multitask learning over networks with sparsity-inducing coregularization," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6329–6344, Dec. 2016.

[39] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks – Part I: Transient analysis," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3487–3517, Jun. 2015.

[40] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks – Part II: Performance analysis," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3518–3548, Jun. 2015.

[41] S. Vlaski and A. H. Sayed, "Diffusion learning in non-convex environments," in *Proc. IEEE IEEE Int. Conf. Acoust., Speech Signal Process.*, Brighton, U.K., May 2019, pp. 5262–5266.

[42] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments – Part I: Agreement at a linear rate," *IEEE Trans. Signal Process.*, vol. 69, pp. 1242–1256, 2021.

[43] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, Jun. 1953.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, San Diego, CA, USA, May 2015, 15 pages.

[45] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[46] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[47] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, New York, NY, USA, Jun. 2016, pp. 1842–1850.

[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[49] X. Zhao and A. H. Sayed, "Asynchronous adaptation and learning over networks – Part II: Performance analysis," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 827–842, Feb. 2015.

[50] J. Wang and G. Joshi, "Adaptive communication strategies to achieve the best error-runtime trade-off in local-update SGD," in *Proc. Mach. Learn. Syst.*, 2019, pp. 212–229.

[51] B. Wang, A. Koppel, and V. Krishnamurthy, "A Markov decision process approach to active meta learning," 2020, *arXiv:2009.04950*.

[52] S. Vlaski and A. H. Sayed, "Second-order guarantees in centralized, federated and decentralized nonconvex optimization," *Commun. Inf. Syst.*, vol. 20, pp. 353–388, 2020.