PARS-Push: Personalized, Asynchronous and Robust Decentralized Optimization

Anonymous Authors[1]

Abstract

We study the decentralized multi-step Model-Agnostic Meta-Learning (MAML) framework where a group of n agents seeks to find a common point that enables "few-shot" learning (personalization) via local stochastic gradient steps on their local functions. We formulate the personalized optimization problem under the MAML framework and propose PARS-Push, a decentralized asynchronous algorithm robust to message failures, communication delays, and directed message sharing. We characterize the convergence rate of PARS-Push for smooth and strongly convex and smooth and non-convex functions under arbitrary multi-step personalization. Moreover, we provide numerical experiments showing its performance under heterogeneous data setups.

1. Introduction

Distributed and decentralized optimization problems considers a set of n agents where the objective is to jointly minimize the sum of a set of functions where each function is accessible to one agent only. Conventionally, we consider the following stochastic optimization problem:

$$\mathbf{z}^* = \arg\min_{\mathbf{z}\in\mathbb{R}^d} f(\mathbf{z}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{z}), \tag{1}$$

$$\text{where} \quad f_i(\mathbf{z}) := \mathbb{E}_{\xi\sim p_i}\left[\ell(\mathbf{z},\xi)\right], \tag{2}$$

the function $\ell : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$ generally denotes the loss function for a model that we seek to learn in the parameter space $\mathbb{R}^d$, and $\mathcal{S}$ denotes the sample space of random variable ξ which represents the data. In (2), $f_i(\mathbf{z})$ is the population loss of model $\mathbf{z}$ over samples selected uniformly with respect to a distribution p_i, i.e., agent i data distribution. In general, the data distribution is unknown and agents have access to empirical realizations of ξ only.

In a homogeneous data setting, all the agents have the same data distributions. Thus, the quality of a minimizer of the empirical loss of $f_i(\cdot)$ on a finite data set of realizations of ξ will improve as the number of data samples increases. Therefore, minimizing an empirical estimator of $f(\cdot)$ is guaranteed to achieve lower costs with a high probability as it will include more realizations from the data space. On the other hand, if p_i's are different, there is no guarantee that a minimizer of the empirical estimator of $f(\cdot)$ will have a lower cost when evaluated on each $f_i(\cdot)$. A minimizer of $f(\cdot)$ will only perform well on $f_i(\cdot)$ on average.

Finn et al. (2017) proposed Model-Agnostic Meta-Learning (MAML) as an optimization-based technique to address data heterogeneity for any optimization problem that can be solved by (stochastic) gradient descent. Fallah et al. (2020a); Ji et al. (2022); Rajeswaran et al. (2019); Finn et al. (2019) studied the convergence of MAML for different function classes. Moreover, this problem has been extensively studied for various learning setups (Finn et al., 2019; Fallah et al., 2021b; Collins et al., 2020; Fallah et al., 2021a; Charles & Konečnỳ, 2021; Kayaalp et al., 2022). Most of prior works analyze MAML with one gradient descent step as the personalization budget, while Fallah et al. (2021a); Ji et al. (2022) study this problem with multi-step fine-tuning budget. Moreover, several recent works have studied personalization in collaborative learning as a form of multi-task learning (Fallah et al., 2020b; Dinh et al., 2020; Gasanov et al., 2021; Collins et al., 2021).

Most of the recent works on the personalized collaborative learning problem, often assume the existence of reliable synchronized communications to a single server (Fallah et al., 2020b; Dinh et al., 2020; Gasanov et al., 2021; Collins et al., 2021), or synchronous local communications over an undirected network (Kayaalp et al., 2022). These strong communication assumptions restrict the application of personalization for modern machine learning problems. Xie et al. (2019); Chen et al. (2021); Nguyen et al. (2021) propose updates with staleness to deal with asynchronous communications with the server in distributed scenarios. Here, we leverage recent works on unreliable communications for distributed/decentralized inference (Mojica-Nava et al., 2021), consensus (Hadjicostis et al., 2015; 2018; Spiridonoff et al., 2020b; Assran et al., 2019), and optimization problems (Spiridonoff et al., 2020b; Assran et al., 2019; Xie et al., 2019; Chen et al., 2021; Nguyen et al.,

2021). The main objective of this work is to study the convergence of the personalized multi-step MAML problem under robust asynchronous communications over directed networks. We summarize our contributions as follows:

- We study the multi-step MAML cost for smooth and strongly convex and non-convex functions. We show that the multi-step model preserves the function class, and we explicitly characterize its condition number as a function of the personalization budget.
- We propose an algorithm named PARS-Push, to minimize the personalized optimization objective function on a directed network with asynchronous communications under message delays and losses.
- We establish convergence guarantees for PARS-Push and provide numerical examples to show the advantages of our method for collaborative learning with heterogeneous data samples over the agents.

The remainder of this paper is arranged as follows. In Section 2, we introduce the personalization setup and discuss the communication framework between the agents. In Section 3, we describe the PARS-Push algorithm for personalized decentralized optimization and present the assumptions and convergence results. We discuss the sketch of proofs in Section 4, and provide numerical experiments in Section 5. We end by concluding the remarks in Section 6.

2. Problem Setup

This section formalizes the personalized decentralized optimization problem under asynchronous communications over a directed network. Moreover, we discuss the problem setup and challenges for the personalization and communication network model.

Personalization via Multi-Step MAML: We denote the $\tilde{f}_i(\mathbf{z}, \mathcal{D}_i)$ as the empirical cost of agent i,

$$\tilde{f}_i(\mathbf{z}, \mathcal{D}_i) \coloneqq \frac{1}{|\mathcal{D}_i|} \sum_{\xi \in \mathcal{D}_i} \ell(\mathbf{z}, \xi), \tag{3}$$

where $\mathcal{D}_i$ is a data batch ($|\mathcal{D}_i| = b$) with samples drawn from a probability distribution p_i, hence $\mathbb{E}_{p_i} \tilde{f}_i(\mathbf{z}, \mathcal{D}_i) = f_i(\mathbf{z})$.

The goal in personalized distributed learning under the MAML approach is to collaboratively find $\mathbf{z}$ that performs well on average after applying $u \geq 1$ (integer) local stochastic gradient descent steps for each agent. Formally, instead of solving Problem (1), we seek to solve the following problem:

$$\mathbf{z}^{*(u)} = \arg\min_{\mathbf{z} \in \mathbb{R}^d} F^{(u)}(\mathbf{z}) \coloneqq \frac{1}{n} \sum_{i=1}^{n} F_i^{(u)}(\mathbf{z}), \tag{4}$$

$$F_i^{(u)}(\mathbf{z}) \coloneqq \mathbb{E}_{p_i} \left[f_i\big(\Psi_i \left(\ldots \big(\Psi_i\big(\mathbf{z}, \mathcal{D}_{i,0}^{\text{test}}\big) \ldots\big), \mathcal{D}_{i,u-1}^{\text{test}}\big)\big)\right],$$

where $\Psi_i(\mathbf{z}, \mathcal{D}_i) \coloneqq \mathbf{z} - \alpha \nabla \tilde{f}_i(\mathbf{z}, \mathcal{D}_i)$ with personalized learning rate $\alpha \geq 0$, and independent data batches $\mathcal{D}_{i,r}^{\text{test}}$ selected uniformly at random with respect to distribution p_i, for all $i \in [n]$, and $r \in \{0\} \cup [u]$. For simplicity of exposition, we assume that all batches have the same size of b, i.e., $|\mathcal{D}_{i,r}^{\text{test}}| = b$. Note that u represents the available budget for few-shot learning. For example, Problem (4) becomes Problem (1) when $u = 0$, i.,e., no budget for personalization. On the contrary, $u \to \infty$ implies there is no need for cooperation, as each agent can find a minimizer of its local function.

Communication Network Model: We consider a static, directed, and strongly connected network $\mathcal{G} = \{[n], \mathcal{E}\}$ with no self-loops, and $\mathcal{E} \subseteq [n] \times [n]$, where $(i, j) \in \mathcal{E}$ if there is an edge from node i to j. For each agent $i \in [n]$, we define $\mathcal{N}_i^- = \{j \text{ s.t. } (j, i) \in \mathcal{E}\}$ as the set of in-neighbors to i, as well as $\mathcal{N}_i^+ = \{j \text{ s.t. } (i, j) \in \mathcal{E}\}$ as the set of out-neighbors from i, where the communications take place over the edges of $\mathcal{G}$. We also denote $d_i^- = |\mathcal{N}_i^-|$, $d_i^+ = |\mathcal{N}_i^+|$, and $m = \sum_{i=1}^n d_i^-$. In this work, we assume that the agents perform updates and communicate asynchronously. We moreover consider the possibility of message losses and delays. Hence, inspired by Spiridonoff et al. (2020b), we assume the following communication model.

Assumption 1 (Spiridonoff et al. (2020b))**.** *The communication network has the following properties:*

(a) graph $\mathcal{G}$ is static, directed, and strongly connected
(b) the delays on each communication link $(i, j) \in \mathcal{E}$ are bounded by some $\Gamma_d \geq 1$,
(c) each communication link $(i, j) \in \mathcal{E}$ fails at most $\Gamma_f \geq 0$ consecutive times,
(d) every agent $i \in [n]$ wakes up and performs updates at least once every $\Gamma_w \geq 1$ iterations.

Assumption 1 is a rather weak model that allows robust asynchronous communications under the possibility of idle agents, as well as a finite number of consecutive failures and bounded delays in the communication. Moreover, it implies an effective maximum delay $\Gamma_e \coloneqq \Gamma_w + \Gamma_d - 1$, where for each receives a message from its in-neighbors at least once every $\Gamma_s \coloneqq \Gamma_w(\Gamma_f + 1) + \Gamma_e$ iterations. Based on this assumption, we will discuss an augmented communication model in Section 3.

3. PARS-Push Algorithm

We first introduce empirical cost

$$\tilde{F}_i^{(u)}(\mathbf{z}, \vartheta_i) \coloneqq \tilde{f}_i(\Psi_i \left(\ldots (\Psi_i(\mathbf{z}, \mathcal{D}_{i,0}) \ldots), \mathcal{D}_{i,u-1}\right), \mathcal{D}_{i,u}), \tag{5}$$

where $\vartheta_i = \{\mathcal{D}_{i,r}\}_{r=0}^{u}$ denotes a set of independent data batches uniformly drawn from distribution p_i, and $F_i^{(u)}(\mathbf{z}) = \mathbb{E}_{p_i}\left[\tilde{F}_i^{(u)}(\mathbf{z}, \vartheta_i)\right]$. Several prior works (Finn et al., 2017; Fallah et al., 2020b; Ji et al., 2022; Kayaalp et al., 2021) consider surrogate cost functions with biased gradient estimators which lead to a non-vanishing bias in the convergence. Following Fallah et al. (2021a;b), we minimize the cost in (4) which has an unbiased stochastic estimators of its deterministic gradients for each $i \in [n]$.

Algorithm 1 PARS-Push: Personalized, Asynchronous, Robust Stochastic Gradient-Push

1: **Initialize:** $y_i = 1$, $\kappa_i = -1$, $\boldsymbol{\phi}_i^{\mathbf{x}} = \mathbf{0}$, $\phi_i^y = 0$, $\forall i \in [n]$, and $\kappa_{ij} = -1$, $\boldsymbol{\rho}_{ij}^{\mathbf{x}} = \mathbf{0}$, $\rho_{ij}^y = 0$, $\forall (j,i) \in \mathcal{E}$.
2: **for** $t = 0, 1, 2, \ldots,$ in parallel for all $i \in [n]$ **do**
3: **if** node i wakes up **then**
4: $\eta_i(t) := \sum_{r=\kappa_i+1}^{t} \theta(r)$
5: $\mathbf{w}_i^{(0)} := \mathbf{z}_i$
6: **for** $r = 0, 1, 2, \ldots, u-1$ **do**
7: Sample a batch $\mathcal{D}_{i,r}^t$ with size b from p_i
8: $\mathbf{w}_i^{(r+1)} := \mathbf{w}_i^{(r)} - \alpha \nabla \tilde{f}_i\left(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r}^t\right)$
9: **end for**
10: Sample a batch $\mathcal{D}_{i,u}^t$ with size b from p_i
11: $\mathbf{x}_i := \mathbf{x}_i - \eta_i(t)\left[\prod_{r=0}^{u-1}\left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i\left(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r}^t\right)\right)\right] \times \nabla \tilde{f}_i(\mathbf{w}_i^{(u)}, \mathcal{D}_{i,u}^t)$
12: $\kappa_i := t$
13: $\mathbf{x}_i := \frac{\mathbf{x}_i}{d_i^+ + 1}$, $y_i := \frac{y_i}{d_i^+ + 1}$
14: $\boldsymbol{\phi}_i^{\mathbf{x}} := \boldsymbol{\phi}_i^{\mathbf{x}} + \mathbf{x}_i$, $\phi_i^y := \phi_i^y + y_i$
15: Node i sends $(\boldsymbol{\phi}_i^{\mathbf{x}}, \phi_i^y, \kappa_i)$ to $\mathcal{N}_i^+$
16: $\mathcal{R}_i :=$ messages received from $\mathcal{N}_i^-$
17: **for** $(\boldsymbol{\phi}_j^{\mathbf{x}}, \phi_j^y, \kappa_j)$ in $\mathcal{R}_i$ **do**
18: **if** $\kappa_j > \kappa_{ij}$ **then**
19: $\boldsymbol{\rho}_{ij}^{*\mathbf{x}} := \boldsymbol{\phi}_j^{\mathbf{x}}$, $\rho_{ij}^{*y} := \phi_j^y$, $\kappa_{ij} := \kappa_j$
20: **end if**
21: **end for**
22: $\mathbf{x}_i := \mathbf{x}_i + \sum_{j \in \mathcal{N}_i^-} (\boldsymbol{\rho}_{ij}^{*\mathbf{x}} - \boldsymbol{\rho}_{ij}^{\mathbf{x}})$
23: $y_i := y_i + \sum_{j \in \mathcal{N}_i^-} (\rho_{ij}^{*y} - \rho_{ij}^{y})$
24: $\boldsymbol{\rho}_{ij}^{\mathbf{x}} := \boldsymbol{\rho}_{ij}^{*\mathbf{x}}$, $\boldsymbol{\rho}_{ij}^{y} := \boldsymbol{\rho}_{ij}^{*y}$, $\mathbf{z}_i := \frac{\mathbf{x}_i}{y_i}$
25: **end if**
26: **end for**

We consider vector $\mathbf{x}_i(t) \in \mathbb{R}^d$, and slack scalar $y_i(t) \in \mathbb{R}^+$ as the parameters of node $i \in [n]$, where $\mathbf{z}_i(t) \in \mathbb{R}^d$ represents the ratio of $\mathbf{x}_i(t)/y_i(t)$. Moreover, each agent $i \in [n]$ allocates $\boldsymbol{\phi}_i^{\mathbf{x}}(t) \in \mathbb{R}^d$ and $\boldsymbol{\rho}_{ij}^{\mathbf{x}}(t) \in \mathbb{R}^d$, for all $j \in \mathcal{N}^-$, to keep the sum of $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$, respectively over the time. Similarly, node i has the same parameters $\phi_i^y(t) \in \mathbb{R}^+$ and $\rho_{ij}^y(t) \in \mathbb{R}^+$ for variables $y_i(t)$ and $y_j(t)$. For the sake of simplicity, we drop t from the description of the algorithm.

At any round t that node $i \in [n]$ is idle, its parameters will not be updated. Once node i wakes up, it performs three set of operations. First, agent i selects $u+1$ independent data batches $\vartheta_i^t = \{\mathcal{D}_{i,r}^t\}_{r=0}^u$ with respect to distribution p_i, and after performing u steps of stochastic gradient descent starting from $\mathbf{z}_i$ (Lines 5-10), agent i computes an unbiased stochastic gradient of (4), and then updates $\mathbf{x}_i$ in Line 11. Second, node i updates its parameters $\mathbf{x}_i$, y_i, $\boldsymbol{\phi}_i^{\mathbf{x}}$, and ϕ_i^y according to Line 13-14, and sends the running sum parameters to its out-neighbors $\mathcal{N}_i^+$ (Line 15). Finally, agent i processes the received messages from its in-neighbors $\mathcal{N}_i^-$ which leads to selecting the most recent updates (Lines 16-21). Consequently, agent i updates $\mathbf{x}_i$, y_i, and $\mathbf{z}_i$ in Lines 22-24 by combining newly received messages. In a nutshell, PARS-Push contains two key operations, (i) robust asynchronous aggregation over a virtual, augmented graph, and (ii) stochastic gradient descent with respect to unbiased stochastic gradients of (4).

Dynamics of the Update Rule: Next, we introduce a linear formulation for Algorithm 1. The variables $\mathbf{x}_i(t)$ and $y_i(t)$ denote node i's parameters at round t. Next, we will discuss the update model for vectors $\mathbf{x}_i(t)$ only, the same discussion would also hold for $\mathbf{y}_i(t)$. According to Spiridonoff et al. (2020b, Section 2), we can model message losses and delays between every two nodes $(j,i) \in \mathcal{E}$, using Γ_e+1 additional virtual nodes, where one of them represents the information that has not successfully been transferred to node i, and the other Γ_e nodes indicate the information that will be delivered with a delay. We formally define variables $\hat{\mathbf{x}}_{ji}^l(t)$ to denote the message sent from node j to i and arriving with an effective delay of l, for all $l \in [\Gamma_e]$. We also consider variable $\tilde{\mathbf{x}}_{ji}(t)$ as the information that has failed to be sent. We moreover define indicator variables $\tau_i(t)$ and $\tau_{ji}^l(t)$. $\tau_i(t) = 1$ if node i wakes up at time t, and $\tau_i(t) = 0$ otherwise. Similarly, $\tau_{ji}^l(t) = 1$ if $\tau_i(t) = 1$ and the sent message from node j to i arrives after an effective delay of $l \in [\Gamma_e]$ rounds, and $\tau_{ji}^l(t) = 0$ otherwise. We now can write the update rule of Algorithm 1 as follows Spiridonoff et al. (2020b):

$$
\begin{aligned}
&\mathbf{x}_i(t+\tfrac{1}{2}) := \mathbf{x}_i(t) - \tau_i(t)\,\eta_i(t)\,\nabla \tilde{F}_i^{(u)}(\mathbf{z}_i(t), \vartheta_i^t),\\
&\mathbf{x}_i(t+1) := \left(1 - \tau_i(t) + \frac{\tau_i(t)}{d_i^+ + 1}\right)\mathbf{x}_i(t+\tfrac{1}{2}) + \sum_{j \in \mathcal{N}_i^-} \mathbf{x}_{ji}^1(t),\\
&\hat{\mathbf{x}}_{ji}^l(t+1) := \tau_{ji}^l(t)\left[\tilde{\mathbf{x}}_{ji}(t) + \frac{\mathbf{x}_j(t)}{d_j^+ + 1}\right] + \mathbb{1}_{\{l < \Gamma_e\}}\hat{\mathbf{x}}_{ji}^{l+1}(t+1),\\
&\tilde{\mathbf{x}}_{ji}^l(t+1) := \left(1 - \sum_{l=1}^{\Gamma_d} \tau_{ji}^l(t)\right)\left[\tilde{\mathbf{x}}_{ji}(t) + \tau_i(t)\frac{\mathbf{x}_j(t)}{d_j^+ + 1}\right],
\end{aligned}
\tag{6}
$$

where $\vartheta_i^t = \{\mathcal{D}_{i,r}^t\}_{r=0}^u$ is a set of $u+1$ independent data

batches uniformly drawn from p_i at round t. A similar update rule holds for $y_i(t)$, except for the gradient descent step of (6). Let $\mathbf{X}(t) \in \mathbb{R}^{(n+m')\times d}$ be the concatenation of $\mathbf{x}_i(t)^\top$, $\tilde{\mathbf{x}}_{ij}(t)^\top$, and $\hat{\mathbf{x}}_{ij}^l(t)^\top$ for all $i \in [n]$, $(i,j) \in \mathcal{E}$, and $l \in [\Gamma_e]$, where $m' := (\Gamma_e+1)m$. Similarly, $\mathbf{y}(t) \in \mathbb{R}^{n+m'}$ denotes the vector containing $y_i(t)$, $\tilde{y}_{ij}(t)$, and $\hat{y}_{ij}^l(t)$. Additionally, for all $(i,j) \in \mathcal{E}$ and $l \in [\Gamma_e]$, virtual parameters $\tilde{\mathbf{x}}_{ij}(0)$, $\hat{\mathbf{x}}_{ij}^l(0)$, $\tilde{y}_{ij}(0)$, and $\hat{y}_{ij}^l(0)$ are all initialized to zero. As a consequence, one can show that the following linear system describes Algorithm 1:

$$\begin{aligned} \mathbf{X}(t+1) &:= \mathbf{M}(t)\left(\mathbf{X}(t) - \mathbf{\Delta}(t)\right), \\ \mathbf{y}(t+1) &:= \mathbf{M}(t)\,\mathbf{y}(t), \\ \mathbf{z}_i(t+1) &:= \mathbf{x}_i(t)/y_i(t), \quad \forall i \in [n] \end{aligned} \tag{7}$$

where $\{\mathbf{M}(t)\}_{t\geq 0}$ is a sequence of column stochastic mixing matrices of size $(n+m')\times(n+m')$, and matrices $\mathbf{\Delta}(t) \in \mathbb{R}^{(n+m')\times d}$ with rows as follows:

$$[\mathbf{\Delta}(t)]_i := \begin{cases} \tau_i(t)\,\eta_i(t)\,\nabla \tilde{F}_i^{(u)}(\mathbf{z}_i(t), \vartheta_i^t)^\top, & i \in [n], \\ \mathbf{0}^\top, & i \notin [n]. \end{cases} \tag{8}$$

Note that each matrix $\mathbf{M}(t)$ is associated with an augmented virtual graph over $n+m'$ nodes. According to Assumption 1, one can check that these augmented virtual graphs build a sequence of time-varying Γ_s-strongly connected graphs describing Algorithm 1. For more details on the specific formulation of the robust asynchronous communications, see Spiridonoff et al. (2020a). Next section discusses the convergence of (7).

4. Convergence Result

This section states our main result, which shows the convergence of PARS-Push under standard assumptions for the class of smooth and strongly convex and smooth and non-convex functions. We consider a series of standard assumptions on the function $\ell(.,\xi)$, for almost all $\xi \in \mathcal{S}$. Our assumptions are commonly used in several prior works (Finn et al., 2019; Fallah et al., 2020b; 2021a;b; Ji et al., 2022).

Assumption 2. *The function $\ell(.,\xi)$, for almost all $\xi \in \mathcal{S}$, is twice continuously differentiable and bounded from below. Furthermore, the following properties hold for all $\mathbf{z}, \hat{\mathbf{z}} \in \mathbb{R}^d$:*

(i) there exist constant G such that $\|\nabla \ell(\mathbf{z},\xi)\| \leq G$,

(ii) $\ell(.,\xi)$ is L-smooth, i.e.,

$$\|\nabla \ell(\mathbf{z},\xi) - \nabla \ell(\hat{\mathbf{z}},\xi)\| \leq L\,\|\mathbf{z} - \hat{\mathbf{z}}\|,$$

(iii) the Hessian of $\ell(.,\xi)$ is H-Lipschitz, i.e.

$$\left\|\nabla^2 \ell(\mathbf{z},\xi) - \nabla^2 \ell(\hat{\mathbf{z}},\xi)\right\| \leq H\,\|\mathbf{z} - \hat{\mathbf{z}}\|,$$

(iv) $\ell(.,\xi)$ is μ-strongly convex, i.e.,

$$\|\nabla \ell(\mathbf{z},\xi) - \nabla \ell(\hat{\mathbf{z}},\xi)\| \geq \mu\,\|\mathbf{z} - \hat{\mathbf{z}}\|.$$

Now, we are ready to discuss the convergence results.

◇ **Smooth & Strongly Convex Functions:** Here, we first present the convergence of Algorithm 1 for solving Problem (4) under Assumptions 1 and 2. Before stating the main result, let us introduce the following lemmas.

Lemma 1. *Suppose that Assumptions 2(i)-(iv) hold. Then, for any $0 \leq \alpha \leq \min\left\{\frac{\mu\varepsilon_1(1-\varepsilon_2)}{GHu}, \frac{1-\varepsilon_1^{-1/2u}}{L}\right\}$, $\varepsilon_1, \varepsilon_2 \in (0,1)$, $F_i^{(u)}(\mathbf{z})$ in (4) is $\hat{\mu}(u)$-strongly convex and $\hat{L}(u)$-smooth*

$$\begin{aligned} \hat{\mu}(u) &= \mu(1-\alpha L)^{2u} - \alpha u G H (1-\alpha\mu)^{u-1}, \\ \hat{L}(u) &= L(1-\alpha\mu)^{2u} + \alpha u G H (1-\alpha\mu)^{u-1}, \end{aligned}$$

for all $i \in [n]$, and $\mathbf{z} \in \mathbb{R}^d$.

Proofs can be found in Appendices C and D.

Lemma 1 implies strong convexity and smoothness of (4) under appropriate α. One can check that $\varepsilon_1\varepsilon_2\mu \leq \hat{\mu}(u)$ and $(1+\varepsilon_1-\varepsilon_1\varepsilon_2)L \geq \hat{L}(u)$. Note that a larger value of α implies higher personalization level. Nevertheless, one can infer from Lemma 1 that the condition number $(\hat{L}(u)/\hat{\mu}(u))$ of the surrogate cost $F_i^{(u)}(\mathbf{z})$ increases with larger α.

Lemma 2. *Let Assumptions 2(i)-(iv) hold, and α as in Lemma 1. Then for all $i \in [n]$ and $\mathbf{z} \in \mathbb{R}^d$, we have:*

$$\mathbb{E}_{p_i}\left\|\nabla \tilde{F}_i^{(u)}(\mathbf{z},\vartheta_i) - \nabla F_i^{(u)}(\mathbf{z})\right\|^2 \leq \hat{\sigma}(u)^2,$$

where $\hat{\sigma}(u)^2 := 4(1-\alpha\mu)^{2u}G^2$ and $\vartheta_i = \{\mathcal{D}_{i,r}\}_{r=0}^u$ represents a set of data batches uniformly drawn from p_i.

Lemma 2 indicates a bounded variance between the deterministic and stochastic gradients for each agent $i \in [n]$. It is worth mentioning that Assumptions 2(i)-(iv) can be relaxed if we borrow Assumption 4 in Fallah et al. (2020b).

Proposition 1 (Smooth and Strongly Convex)**.** *Let Assumptions 1 and 2 hold, stepsize α be as in Lemma 1, $\theta(t) = \frac{1}{\hat{\mu}(u)t}$, for $t \geq 1$, and $\theta(0) = 0$. Then, for all $i \in [n]$, the following property holds for the iterates of Algorithm 1:*

$$\mathbb{E}\left[\left\|\mathbf{z}_i(T) - \mathbf{z}^{*(u)}\right\|^2\right] = \mathcal{O}\left(\frac{\Gamma_w\,\hat{\sigma}(u)^2}{\hat{\mu}(u)\,n\,T}\right) + \mathcal{O}\left(\frac{1}{T^{\frac{3}{2}}}\right),$$

where $\hat{\mu}(u)$ and $\hat{\sigma}(u)$ as defined in Lemma 1 and Lemma 2.

Proposition 1 suggests a sublinear convergence rate $\mathcal{O}(1/T)$ for Algorithm 1 to solve (4), under Assumptions 1-(iv), where all of the decentralization terms $(\gamma, \delta, \lambda)$ are associated with $\mathcal{O}(1/T^{3/2})$.

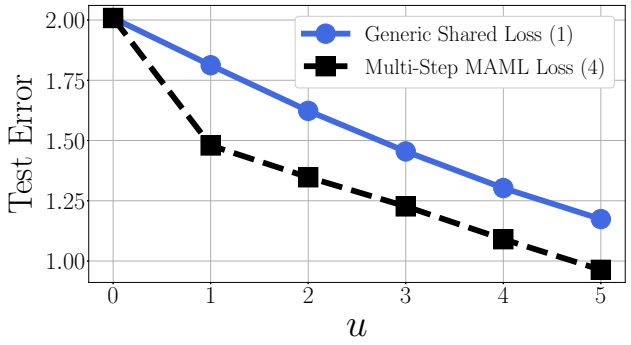

Figure 1: Given a fixed personalization budget u at the test time, optimizing the surrogate loss in (4) yields a lower error compared to (1) in heterogeneous data settings.

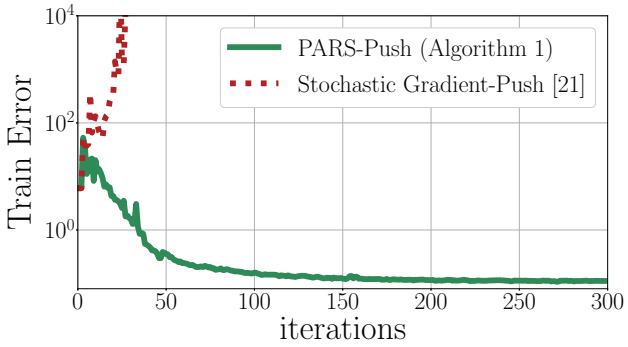

Figure 2: Robustness to asynchronous communications, idle agents, message losses and delays.

We move the result on smooth non-convex functions to Appendix B.

5. Numerical Experiment

In this section, we study a decentralized linear regression problem with regularization on a synthetic heterogeneous data. We consider a random vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$, and a set of vectors $\{\boldsymbol{\beta}_i^*\}_{i\in[n]}$, where each $\boldsymbol{\beta}_i^* \in \mathbb{R}^d$ is obtained by perturbing $\boldsymbol{\beta}^*$ with Gaussian noise. Hence, we define the q-th data sample on agent i as follows:

$$b_{iq} = \mathbf{a}_{iq}^\top \boldsymbol{\beta}_i^* + \zeta_{iq}, \tag{9}$$

where $\xi_{iq} = (b_{iq}, \mathbf{a}_{iq})$ represents a data sample, and $\zeta_{iq} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. We can formally define $f_i(\mathbf{z})$ as follows:

$$f_i(\mathbf{z}) = \mathbb{E}_{\xi_{iq}\sim p_i}\left[\left(b_{iq} - \mathbf{a}_{iq}^\top \mathbf{z}\right)^2 + \frac{1}{2n}\|\mathbf{z}\|^2\right]. \tag{10}$$

We consider a setting with $n = 10$ agents, where each agent maintains data samples in a $d = 30$ dimensional space. We also consider batches with $b = 20$ samples, and select 200 test data points for each agent $i \in [n]$. For the communication, we select a fixed, directed, and strongly connected Erdős-Rényi network with a low connection probability ($p = 0.15$). We simulate the asynchronous setup with $\Gamma_w = 2$, $\Gamma_d = 2$, and $\Gamma_f = 1$, for 6000 iterations.

We compare the average test error of (11) under personalization with that of (1) after personalization with the same budget u in Fig. 1. In words, we consider the surrogate loss in (4), and after obtaining $\mathbf{z}^{*(u)}$ for each $u \in [5]$, we update the personalized parameters under u consecutive stochastic gradient descent steps. We also solve (1) under the same communication setting, and for each $u \in [5]$, we apply u steps of stochastic gradient descent after the train phase to have a fair comparison.

Figure 2 shows the training error and validates that PARS-Push is robust to asynchronous communications with message losses and delays, compared to stochastic gradient-push (Assran et al., 2019), which is state-of-the-art for delayed asynchronous communications.

6. Conclusions

This work studied personalized decentralized optimization under robust asynchronous communications over directed networks. We considered the multi-step MAML problem under heterogeneous data setting and proposed PARS-Push algorithm to solve this problem with arbitrary u personalization budget. We showed the convergence of our method for smooth and strongly convex, and non-convex functions. We proposed a numerical setup to illustrate the convergence and personalization of our method.

References

Assran, M., Loizou, N., Ballas, N., and Rabbat, M. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, pp. 344–353. PMLR, 2019.

Charles, Z. and Konečnỳ, J. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2575–2583. PMLR, 2021.

Chen, Z., Liao, W., Hua, K., Lu, C., and Yu, W. Towards asynchronous federated learning for heterogeneous edge-powered internet of things. *Digital Communications and Networks*, 7(3):317–326, 2021.

Collins, L., Mokhtari, A., and Shakkottai, S. Task-robust model-agnostic meta-learning. *arXiv preprint arXiv:2002.04766*, 2020.

Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. Exploiting shared representations for personalized federated learning. *arXiv preprint arXiv:2102.07078*, 2021.

Dinh, C., Tran, N., and Nguyen, J. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

Fallah, A., Mokhtari, A., and Ozdaglar, A. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1082–1092. PMLR, 2020a.

Fallah, A., Mokhtari, A., and Ozdaglar, A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33, 2020b.

Fallah, A., Georgiev, K., Mokhtari, A., and Ozdaglar, A. On the convergence theory of debiased model-agnostic meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021a.

Fallah, A., Mokhtari, A., and Ozdaglar, A. Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks. *Advances in Neural Information Processing Systems*, 34, 2021b.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. Online meta-learning. In *International Conference on Machine Learning*, pp. 1920–1930. PMLR, 2019.

Gasanov, E., Khaled, A., Horváth, S., and Richtárik, P. Flix: A simple and communication-efficient alternative to local methods in federated learning. *arXiv preprint arXiv:2111.11556*, 2021.

Hadjicostis, C., Vaidya, N., and Domínguez-García, A. Robust distributed average consensus via exchange of running sums. *IEEE Transactions on Automatic Control*, 61 (6):1492–1507, 2015.

Hadjicostis, C., Domínguez-García, A., and Charalambous, T. *Distributed averaging and balancing in network systems*. Now Foundations and Trends, 2018.

Ji, K., Yang, J., and Liang, Y. Theoretical convergence of multi-step model-agnostic meta-learning. *Journal of Machine Learning Research*, 23(29):1–41, 2022.

Kayaalp, M., Vlaski, S., and Sayed, A. Distributed meta-learning with networked agents. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 1361–1365. IEEE, 2021.

Kayaalp, M., Vlaski, S., and Sayed, A. Dif-maml: Decentralized multi-agent meta-learning. *IEEE Open Journal of Signal Processing*, 3:71–93, 2022.

Mojica-Nava, E., Yanguas-Rojas, D., and Uribe, C. A. Robust asynchronous and network-independent cooperative learning. In *2021 American Control Conference (ACC)*, pp. 1619–1624. IEEE, 2021.

Nguyen, J., Malik, K., Zhan, H., Yousefpour, A., Rabbat, M., Esmaeili, M., and Huba, D. Federated learning with buffered asynchronous aggregation. *arXiv preprint arXiv:2106.06639*, 2021.

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Rajeswaran, A., Finn, C., Kakade, S., and Levine, S. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

Spiridonoff, A., Olshevsky, A., and Paschalidis, I. Local sgd with a communication overhead depending only on the number of workers. *arXiv preprint arXiv:2006.02582*, 2020a.

Spiridonoff, A., Olshevsky, A., and Paschalidis, I. Robust asynchronous stochastic gradient-push: Asymptotically optimal and network-independent performance for strongly convex functions. *Journal of machine learning research*, 21(58), 2020b.

Xie, C., Koyejo, S., and Gupta, I. Asynchronous federated optimization. *arXiv preprint arXiv:1903.03934*, 2019.

A. Discussion on Personalization Dynamic

Some formulations of Meta-Learning consider the slightly different scenario where one has access to a set of functions $\{f_i\}$ (Nichol et al., 2018). However, instead of finding the minimizer of their average as in (1), Meta-Learning seeks to find a point that achieves a smaller function value on each f_i after a small number of local steps of a selected optimization algorithm is applied when compared with the same number of steps with random initialization. From the machine learning perspective, one can consider each of the local functions $f_i(\cdot)$ as an individual task to be learned based on some finite dataset. Minimizing the empirical approximation of $f_i(\cdot)$ will correspond to finding an approximate model that performs well on that task. However, in some scenarios, the number of data points for a specific task might be limited. Thus, leveraging the data from other tasks might improve performance even with limited data. In practice, Meta-Learning translates into the task of finding a new objective function $F(\cdot)$ that depends on $\{f_i(\cdot)\}$ such that when minimizing an empirical estimator of $F(\cdot)$ one can achieve good performance locally on $f_i(\cdot)$ with only a few steps of the desired optimization method, or few-shot learning. Usually, the number of available local steps is termed the personalization budget. In (4) when $u = 1$, we have:

$$F_i^{(1)}(\mathbf{z}) := \mathbb{E}_{p_i}\left[f_i\left(\mathbf{z} - \alpha\nabla\tilde{f}_i\left(\mathbf{z}, \mathcal{D}_{i,0}^{\text{test}}\right)\right)\right]. \tag{11}$$

Compared to (1), considering (11) implies an initial point $\mathbf{z}^{*(1)}$ such that each agent i can perform one step of stochastic gradient descent with respect to its local cost function to obtain a personalized model $\mathbf{w}_i^{*(1)} = \mathbf{z}^{*(1)} - \alpha\nabla\tilde{f}_i\left(\mathbf{z}^{*(1)}, \mathcal{D}_{i,0}^{\text{test}}\right)$. Figure 3 illustrates the personalization mechanism for a case of $n = 2$ agents with $u = 2$ steps of local stochastic gradient descent, where after two steps of personalization, the solution to (4) has a lower average cost compared (1). The choice of cost function in (4) enables us to efficiently compute the stochastic gradient as an unbiased estimator of $\nabla F_i^{(u)}(\mathbf{z})$, which we will discuss in Section 3.

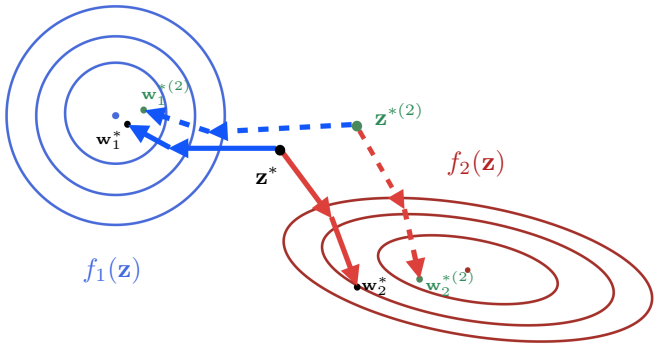

Figure 3: **Personalization Impact:** Two agents seek to minimize (1) (solid line) and (4) (dashed line) with $u = 2$. Given a fixed budget of two stochastic gradient descent steps for personalization, Problem (4) ($\mathbf{z}^{*(2)}$) provides a better initialization point compared to Problem (1) ($\mathbf{z}^*$). For each $i = 1, 2$, $\mathbf{w}_i^*$ and $\mathbf{w}_i^{*(2)}$ indicate the parameters of agent i after applying 2 steps of stochastic gradient descent starting from $\mathbf{z}^*$ and $\mathbf{z}^{*(2)}$, respectively.

B. Convergence Result: Non-Convex

We show first-order stationary convergence of Algorithm 1 under Assumptions 1-(iii) and Γ_w=1. Note that Γ_w=1 also incorporates message losses and asynchronous communications. However, each node must compute a local gradient and perform consensus every round. Before stating the non-convex results, let us state the following lemmas.

Lemma 3. *Let Assumptions 2(i)-(iii) hold. Then, for $\alpha \geq 0$, $F_i^{(u)}(\mathbf{z})$ in (4) is $\hat{L}(u)$-smooth as follows: for all $i \in [n]$ and $\mathbf{z} \in \mathbb{R}^d$, we have:*

$$\hat{L}(u) = (L + \alpha u G H)(1 + \alpha L)^{2u},$$
$$\mathbb{E}_{p_i}\left\|\nabla\tilde{F}_i^{(u)}(\mathbf{z}, \vartheta_i) - \nabla F_i^{(u)}(\mathbf{z})\right\|^2 \leq \hat{\sigma}(u)^2,$$

where $\hat{\sigma}(u)^2 = 4(1 + \alpha L)^{2u} G^2$.

Lemma 3 indicates that (4) is smooth with $\hat{L}^{(u)} \geq L$. Lemma 3 does not require an upper bound on α. However, $\hat{L}(u)$ grows exponentially with α. Moreover, this lemma shows an upper bound on the variance of stochastic gradients. Now, we are ready to present the convergence result for smooth and non-convex functions.

Theorem 1 (Smooth and Non-Convex). *Let Assumptions 1 and 2(i)-(iii) hold with Γ_w:=1, stepsizes $\alpha \geq 0$, $\theta(t) = \frac{\sqrt{n}}{\hat{L}(u)\sqrt{T}}$, and $\mathbf{x}_i(0)$:=$\mathbf{0}$, for all $i \in [n]$. Then, the following property holds for the iterates of Algorithm 1: for any $T \geq n$*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F^{(u)}\left(\frac{\mathbf{X}(t)^\top \mathbf{1}}{n}\right)\right\|^2 = \mathcal{O}\left(\frac{2\hat{L}(u)F^{(u)}(\mathbf{0}) + \hat{\sigma}(u)^2}{(nT)^{\frac{1}{2}}}\right) + \mathcal{O}\left(\frac{1}{T}\right),$$

where $\hat{L}(u)$ and $\hat{\sigma}(u)$ as in Lemma 3.

Theorem 1 shows a sublinear convergence to a first-order stationary point with rate $\mathcal{O}(1/T^{1/2})$ for Algorithm 1 under Assumption 1 with Γ_w=1. It is worth mentioning that this assumption may be relaxed with a uniformly probabilistic wake up once in every $\Gamma_w \geq 1$ rounds, but we leave this for future studies. The terms concerning the network, i.e., λ, δ, appear in $\mathcal{O}(1/T)$.

C. Proof of Lemma 1, Lemma 2, and Proposition 1

Before proving Lemma 1, let us state the following lemma from Finn et al. (2019).

Lemma 4 (Mean Value Inequality). *Let $\mathbf{\Lambda} : \mathbf{x} \in \mathbb{R}^d \to \mathbb{R}^m$ be a differentiable function. Let $\nabla \mathbf{\Lambda}$ be the function Jacobian, such that $\nabla \mathbf{\Lambda}_{i,j} = \frac{\partial \mathbf{\Lambda}_i}{\partial \mathbf{x}_j}$. Let $M = \max_{\mathbf{x}\in\mathbb{R}^d} \|\nabla \mathbf{\Lambda}(\mathbf{x})\|$. Then, we have*

$$\|\mathbf{\Lambda}(\mathbf{y}) - \mathbf{\Lambda}(\mathbf{x})\| \leq M\|\mathbf{y} - \mathbf{x}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Note that $F_i^{(u)}(\mathbf{z})$ is the average of $\tilde{F}_i^{(u)}(\mathbf{z}, \vartheta_i)$ that also implies $\nabla \tilde{F}_i^{(u)}(\mathbf{z}, \vartheta_i)$ is an unbiased stochastic estimator of $\nabla F_i^{(u)}(\mathbf{z})$. Therefore, it is enough to show that $\nabla \tilde{F}_i^{(u)}(., \vartheta_i)$ is $\hat{L}(u)$-smooth and $\hat{\mu}(u)$-strongly convex, because

$$\frac{\hat{\mu}(u)}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq \tilde{F}_i^{(u)}(\mathbf{y}, \vartheta_i) - \tilde{F}_i^{(u)}(\mathbf{x}, \vartheta_i) - \left\langle \nabla \tilde{F}_i^{(u)}(\mathbf{x}, \vartheta_i), \mathbf{y} - \mathbf{x} \right\rangle \leq \frac{\hat{L}(u)}{2}\|\mathbf{y} - \mathbf{x}\|^2, \tag{12}$$

for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and any set ϑ_i of u+1 data batches drawn from distribution p_i. This also implies that:

$$\frac{\hat{\mu}(u)}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq F_i^{(u)}(\mathbf{y}) - F_i^{(u)}(\mathbf{x}) - \left\langle \nabla F_i^{(u)}(\mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle \leq \frac{\hat{L}(u)}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{13}$$

By a similar argument, it can be seen that $\tilde{f}_i(., \mathcal{D}_{i,r})$ is μ-strongly convex and L-smooth, due to Assumption 2, for all $i \in [n]$, and $\mathcal{D}_{i,r}$ data batches drawn uniformly from p_i. Before proving the smoothness and strong-convexity for $\tilde{F}_i^{(u)}(., \vartheta_i)$, let us discuss the Mean Value Inequality for $\Psi_i(\mathbf{w}, \mathcal{D}_i)$. Note that by definition, we know that all properties in Assumption 2 hold for $\tilde{f}_i(., \mathcal{D}_i)$, for all $i \in [n]$ and data batches $\mathcal{D}_i$ uniformly sampled from p_i. Therefore we have:

$$\mu\, \mathbf{I} \leq \|\nabla^2 \tilde{f}_i(\mathbf{w}, \mathcal{D}_i)\| \leq L\, \mathbf{I}, \tag{14}$$

for all $\mathbf{z} \in \mathbb{R}^d$, and data batches $\mathcal{D}_i$ drawn from distribution p_i. This indicates that the following property holds for $\nabla \Psi_i(\mathbf{w}, \mathcal{D}_i) = \mathbf{I} - \alpha \nabla^2 \tilde{f}_i(\mathbf{w}, \mathcal{D}_i)$ for $\Psi_i(\mathbf{w}, \mathcal{D}_i)$:

$$(1 - \alpha L)\, \mathbf{I} \leq \nabla \Psi_i(\mathbf{w}, \mathcal{D}_i) \leq (1 - \alpha\mu)\, \mathbf{I}, \tag{15}$$

where due to Lemma 4, the following property holds:

$$\|\Psi_i(\mathbf{w}, \mathcal{D}_i) - \Psi_i(\mathbf{v}, \mathcal{D}_i)\| \leq (1 - \alpha\mu)\|\mathbf{w} - \mathbf{v}\|, \tag{16}$$

for all $i \in [n]$ and uniformly sample data batches $\mathcal{D}_i$ with respect to p_i.

According to what we discussed so far, we are now ready to prove the smoothness and strong convexity of $\tilde{F}_i^{(u)}$. First note that, for all $\mathbf{w} \in \mathbb{R}^d$,

$$\tilde{F}_i^{(u)}(\mathbf{w}, \vartheta_i) = \tilde{f}_i(\mathbf{w}_i^{(u)}, \mathcal{D}_{i,u}), \tag{17}$$

$$\text{where} \quad \mathbf{w}_i^{(r+1)} = \mathbf{w}_i^{(r)} - \alpha \nabla \tilde{f}_i(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r}), \quad r = 0, 1, \ldots, u-1, \tag{18}$$

$$\text{and} \quad \mathbf{w}_i^{(0)} = \mathbf{w}, \tag{19}$$

where $\vartheta_i = \{\mathcal{D}_{i,r}\}$. We also know that:

$$\nabla \tilde{F}_i^{(u)}(\mathbf{w}, \vartheta_i) = \left[\prod_{r=0}^{u-1} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r})\right)\right] \nabla \tilde{f}_i(\mathbf{w}_i^{(u)}, \mathcal{D}_{i,u}). \tag{20}$$

Now, consider two points $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$. According to the definition in (17) as well as the property in (16), we can see that for any two points $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$:

$$\left\|\mathbf{w}_i^{(u)} - \mathbf{v}_i^{(u)}\right\| \leq (1 - \alpha\mu) \left\|\mathbf{w}_i^{(u-1)} - \mathbf{v}_i^{(u-1)}\right\| \tag{21}$$

$$\leq (1 - \alpha\mu)^2 \left\|\mathbf{w}_i^{(u-2)} - \mathbf{v}_i^{(u-2)}\right\| \tag{22}$$

$$\vdots$$

$$\leq (1 - \alpha\mu)^u \left\|\mathbf{w}_i^{(0)} - \mathbf{v}_i^{(0)}\right\| \tag{23}$$

$$= (1 - \alpha\mu)^u \left\|\mathbf{w} - \mathbf{v}\right\|. \tag{24}$$

We first show that $\tilde{F}_i^{(u)}(., \vartheta_i)$is $\hat{L}(u)$-smooth, by providing an upper bound on the following expression:

$$\left\|\nabla \tilde{F}_i^{(u)}(\mathbf{w}, \vartheta_i) - \nabla \tilde{F}_i^{(u)}(\mathbf{v}, \vartheta_i)\right\| \tag{25}$$

$$= \Bigg\| \left[\prod_{r=0}^{u-1} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r}\right)\right)\right] \nabla \tilde{f}_i \left(\mathbf{w}_i^{(u)}, \mathcal{D}_{i,u}\right)$$

$$- \left[\prod_{r=0}^{u-1} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{v}_i^{(r)}, \mathcal{D}_{i,r}\right)\right)\right] \nabla \tilde{f}_i \left(\mathbf{v}_i^{(u)}, \mathcal{D}_{i,u}\right) \Bigg\| \tag{26}$$

$$\overset{\text{tri. ineq.}}{\leq} \left\| \left[\prod_{r=0}^{u-1} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r}\right)\right)\right] \left(\nabla \tilde{f}_i \left(\mathbf{w}_i^{(u)}, \mathcal{D}_{i,u}\right) - \nabla \tilde{f}_i \left(\mathbf{v}_i^{(u)}, \mathcal{D}_{i,u}\right)\right) \right\| \tag{27}$$

$$+ \left\| \left[\underbrace{\prod_{r=0}^{u-1} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r}\right)\right) - \prod_{r=0}^{u-1} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{v}_i^{(r)}, \mathcal{D}_{i,r}\right)\right)}_{\mathcal{Q}_{i,u-1}}\right] \nabla \tilde{f}_i \left(\mathbf{v}_i^{(u)}, \mathcal{D}_{i,u}\right) \right\| \tag{28}$$

$$\overset{\text{Assump. 2}}{\leq} L \left\| \left[\prod_{r=0}^{u-1} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r}\right)\right)\right] \right\| \left\|\mathbf{w}_i^{(u)} - \mathbf{v}_i^{(u)}\right\| + \left\|\mathcal{Q}_{i,u-1}\right\| \left\|\nabla \tilde{f}_i \left(\mathbf{v}_i^{(u)}, \mathcal{D}_{i,u}\right)\right\| \tag{29}$$

$$\overset{(15)}{\leq} L(1 - \alpha\mu)^u \left\|\mathbf{w}_i^{(u)} - \mathbf{v}_i^{(u)}\right\| + G \left\|\mathcal{Q}_{i,u-1}\right\| \tag{30}$$

$$\overset{(21)}{\leq} L(1 - \alpha\mu)^{2u} \left\|\mathbf{w} - \mathbf{v}\right\| + G \left\|\mathcal{Q}_{i,u-1}\right\|. \tag{31}$$

To complete the proof for smoothness, we now propose a recursive upper bound on $\|\mathcal{Q}_{i,u-1}\|$:

$$\left\|\mathcal{Q}_{i,u-1}\right\| = \left\| \prod_{r=0}^{u-1} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r}\right)\right) - \prod_{r=0}^{u-1} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{v}_i^{(r)}, \mathcal{D}_{i,r}\right)\right) \right\| \tag{32}$$

$$= \left\| \prod_{r=0}^{u-1} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r} \right) \right) - \left[\prod_{r=0}^{u-2} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r} \right) \right) \right] \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{v}_i^{(u-1)}, \mathcal{D}_{i,r-1} \right) \right) \right. \tag{33}$$

$$\left. + \left[\prod_{r=0}^{u-2} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r} \right) \right) \right] \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{v}_i^{(u-1)}, \mathcal{D}_{i,r-1} \right) \right) - \prod_{r=0}^{u-1} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{v}_i^{(r)}, \mathcal{D}_{i,r} \right) \right) \right\| \tag{34}$$

$$\overset{\text{tri. ineq.}}{\leq} \alpha \left\| \prod_{r=0}^{u-2} \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{w}_i^{(r)}, \mathcal{D}_{i,r} \right) \right) \right\| \left\| \nabla^2 \tilde{f}_i \left(\mathbf{w}_i^{(u-1)}, \mathcal{D}_{i,r-1} \right) - \nabla^2 \tilde{f}_i \left(\mathbf{v}_i^{(u-1)}, \mathcal{D}_{i,r-1} \right) \right\| \tag{35}$$

$$+ \left\| \mathcal{Q}_{i,u-2} \right\| \left\| \mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{v}_i^{(u-1)}, \mathcal{D}_{i,r} \right) \right\| \tag{36}$$

$$\overset{\text{Assump. 2}}{\leq} \alpha H (1 - \alpha\mu)^{u-1} \left\| \mathbf{w}_{u-1} - \mathbf{v}_{u-1} \right\| + (1 - \alpha\mu) \left\| \mathcal{Q}_{i,u-2} \right\| \tag{37}$$

$$\overset{(21)}{\leq} \alpha H (1 - \alpha\mu)^{2u-2} \left\| \mathbf{w} - \mathbf{v} \right\| + (1 - \alpha\mu) \left\| \mathcal{Q}_{i,u-2} \right\|. \tag{38}$$

Consequently, we can infer the following recursion:

$$\left\| \mathcal{Q}_{i,r} \right\| \leq \alpha H (1 - \alpha\mu)^{2u-2} \left\| \mathbf{w} - \mathbf{v} \right\| + (1 - \alpha\mu) \left\| \mathcal{Q}_{i,r-1} \right\|, \quad \forall r \geq 1, \tag{39}$$

and we also have:

$$\left\| \mathcal{Q}_{i,0} \right\| = \left\| \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{w}_i^{(0)}, \mathcal{D}_{i,0} \right) \right) - \left(\mathbf{I} - \alpha \nabla^2 \tilde{f}_i \left(\mathbf{v}_i^{(0)}, \mathcal{D}_{i,0} \right) \right) \right\| \overset{\text{Assump. 2}}{\leq} \alpha H \| \mathbf{w} - \mathbf{v} \|. \tag{40}$$

According to (39) and (40), we can see that:

$$\left\| \mathcal{Q}_{i,u} \right\| \leq \alpha H (1 - \alpha\mu)^{2u} \left\| \mathbf{w} - \mathbf{v} \right\| + (1 - \alpha\mu) \left\| \mathcal{Q}_{i,u-1} \right\| \tag{41}$$

$$\leq \alpha H \left[(1 - \alpha\mu)^{2u} + (1 - \alpha\mu)^{2u-1} \right] \left\| \mathbf{w} - \mathbf{v} \right\| + (1 - \alpha\mu)^2 \left\| \mathcal{Q}_{i,u-2} \right\| \tag{42}$$

$$\vdots \tag{43}$$

$$\leq \alpha H \left[\sum_{r=0}^{u-1} (1 - \alpha\mu)^{2u-r} \right] \left\| \mathbf{w} - \mathbf{v} \right\| + (1 - \alpha\mu)^u \left\| \mathcal{Q}_{i,0} \right\| \tag{44}$$

$$\leq \alpha H (1 - \alpha\mu)^u (u+1) \| \mathbf{w} - \mathbf{v} \|. \tag{45}$$

As a result of (25) and (41), we can conclude the smoothness of $\tilde{F}_i^{(u)}(., \vartheta_i)$ as follows:

$$\left\| \nabla \tilde{F}_i^{(u)}(\mathbf{w}, \vartheta_i) - \nabla \tilde{F}_i^{(u)}(\mathbf{v}, \vartheta_i) \right\| \leq \left(\underbrace{L(1 - \alpha\mu)^{2u} + \alpha u G H (1 - \alpha\mu)^{u-1}}_{\hat{L}(u)} \right) \left\| \mathbf{w} - \mathbf{v} \right\|. \tag{46}$$

Before showing the strong convexity result, let us show a similar property to 21:

$$\left\| \mathbf{w}_i^{(u)} - \mathbf{v}_i^{(u)} \right\| = \left\| \mathbf{w}_i^{(u-1)} - \alpha \nabla \tilde{f}_i (\mathbf{w}_i^{(u-1)}, \mathcal{D}_{i,r}) - \mathbf{v}_i^{(u-1)} + \alpha \nabla \tilde{f}_i (\mathbf{v}_i^{(u-1)}, \mathcal{D}_{i,r}) \right\| \tag{47}$$

$$\overset{\text{tri. ineq.}}{\geq} \left\| \mathbf{w}_i^{(u-1)} - \mathbf{v}_i^{(u-1)} \right\| - \alpha \left\| \nabla \tilde{f}_i (\mathbf{w}_i^{(u-1)}, \mathcal{D}_{i,r}) - \alpha \nabla \tilde{f}_i (\mathbf{v}_i^{(u-1)}, \mathcal{D}_{i,r}) \right\| \tag{48}$$

$$\overset{\text{Assump. 2}}{\geq} \left\| \mathbf{w}_i^{(u-1)} - \mathbf{v}_i^{(u-1)} \right\| - \alpha L \left\| \mathbf{w}_i^{(u-1)} - \mathbf{v}_i^{(u-1)} \right\| \tag{49}$$

$$\geq (1 - \alpha L) \left\| \mathbf{w}_i^{(u-1)} - \mathbf{v}_i^{(u-1)} \right\| \tag{50}$$

$$\geq (1 - \alpha L)^2 \left\| \mathbf{w}_i^{(u-2)} - \mathbf{v}_i^{(u-2)} \right\| \tag{51}$$

$$\vdots$$

$$\geq (1-\alpha L)^u \left\|\mathbf{w}_i^{(0)} - \mathbf{v}_i^{(0)}\right\| \tag{52}$$

$$= (1-\alpha L)^u \left\|\mathbf{w} - \mathbf{v}\right\|. \tag{53}$$

Now, we are ready to show that $\tilde{F}_i^{(u)}(.,\vartheta_i)$ is $\hat{\mu}(u)$-strongly convex, by providing a lower bound on the following expression:

$$\left\|\nabla\tilde{F}_i^{(u)}(\mathbf{w},\vartheta_i) - \nabla\tilde{F}_i^{(u)}(\mathbf{v},\vartheta_i)\right\| \tag{54}$$

$$= \left\| \left[\prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right]\nabla\tilde{f}_i\left(\mathbf{w}_i^{(u)},\mathcal{D}_{i,u}\right)\right.$$

$$\left. - \left[\prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{v}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right]\nabla\tilde{f}_i\left(\mathbf{v}_i^{(u)},\mathcal{D}_{i,u}\right)\right\| \tag{55}$$

$$\overset{\text{tri. ineq.}}{\geq} \left\| \left[\prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right]\left(\nabla\tilde{f}_i\left(\mathbf{w}_i^{(u)},\mathcal{D}_{i,u}\right) - \nabla\tilde{f}_i\left(\mathbf{v}_i^{(u)},\mathcal{D}_{i,u}\right)\right)\right\| \tag{56}$$

$$- \left\| \left[\underbrace{\prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right) - \prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{v}_i^{(r)},\mathcal{D}_{i,r}\right)\right)}_{\mathcal{Q}_{i,u-1}}\right]\nabla\tilde{f}_i\left(\mathbf{v}_i^{(u)},\mathcal{D}_{i,u}\right)\right\| \tag{57}$$

$$\overset{\text{Assump. 2}}{\geq} \mu\left\| \left[\prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right]\right\| \left\|\mathbf{w}_i^{(u)} - \mathbf{v}_i^{(u)}\right\| - \left\|\mathcal{Q}_{i,u-1}\right\| \left\|\nabla\tilde{f}_i\left(\mathbf{v}_i^{(u)},\mathcal{D}_{i,u}\right)\right\| \tag{58}$$

$$\overset{(15)}{\geq} \mu(1-\alpha L)^u \left\|\mathbf{w}_i^{(u)} - \mathbf{v}_i^{(u)}\right\| - G\left\|\mathcal{Q}_{i,u-1}\right\| \tag{59}$$

$$\overset{(47)}{\geq} \mu(1-\alpha L)^{2u}\left\|\mathbf{w}-\mathbf{v}\right\| - G\left\|\mathcal{Q}_{i,u-1}\right\|, \tag{60}$$

where by applying (41), we can see that:

$$\left\|\nabla\tilde{F}_i^{(u)}(\mathbf{w},\vartheta_i) - \nabla\tilde{F}_i^{(u)}(\mathbf{v},\vartheta_i)\right\| \geq \left(\underbrace{\mu(1-\alpha L)^{2u} - \alpha u G H(1-\alpha\mu)^{u-1}}_{\hat{\mu}(u)}\right)\left\|\mathbf{w}-\mathbf{v}\right\|. \tag{61}$$

Equations (46) and (61) conclude the proof of Lemma 1. The choice of personalized stepsize α, ensures that $\hat{\mu}(u)$ is a positive constant. Indeed, given arbitrary integer $u>0$, $\varepsilon_1, \varepsilon_2 \in (0,1)$ and personalized stepsize α as in Lemma 1, we have $\hat{\mu}(u) \geq \varepsilon_1\varepsilon_2\mu$, and $\hat{L}(u) \leq (1+\varepsilon_1-\varepsilon_1\varepsilon_2)L$.

Now, note that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $i \in [n]$, and ϑ_i, we have:

$$\left\|\nabla\tilde{F}_i^{(u)}(\mathbf{w},\vartheta_i)\right\| = \left[\prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r})\right)\right]\nabla\tilde{f}_i(\mathbf{w}_i^{(u)},\mathcal{D}_{i,u}) \tag{62}$$

$$\leq \left[\prod_{r=0}^{u-1}\left\|\mathbf{I}-\alpha\nabla^2\tilde{f}_i(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r})\right\|\right]\left\|\nabla\tilde{f}_i(\mathbf{w}_i^{(u)},\mathcal{D}_{i,u})\right\| \tag{63}$$

$$\overset{(15),\text{Assump. 2}}{\leq} (1-\alpha\mu)^u G. \tag{64}$$

Therefore, we have:

$$\left\|\nabla\tilde{F}_i^{(u)}(\mathbf{z},\vartheta_i) - \nabla F_i^{(u)}(\mathbf{z})\right\| \leq \left\|\nabla\tilde{F}_i^{(u)}(\mathbf{z},\vartheta_i)\right\| + \left\|\nabla F_i^{(u)}(\mathbf{z})\right\| \tag{65}$$

$$\leq \left\|\nabla\tilde{F}_i^{(u)}(\mathbf{z},\vartheta_i)\right\| + \mathbb{E}_{p_i}\left\|\nabla\tilde{F}_i^{(u)}(\mathbf{z},\vartheta_i)\right\| \tag{66}$$

$$\overset{(62)}{\leq} 2(1-\alpha\mu)^u G, \tag{67}$$

where we can conclude the lemma by taking expectation from the square of (65).

Finally, Lemma 1 implies the strong convexity and smoothness parameters for the surrogate loss (4). Lemma 2 states a bounded variance for the stochastic gradients of the surrogate cost. Therefore, the proof of Proposition 1 is an immediate consequence of Spiridonoff et al. (2020b, Theorem 15) under these two lemmas.

D. Proof of Lemma 3 and Theorem 1

.

The proof of Lemma 3 is similar to the proof of Lemma 1 and Lemma 2. First note that, according to Assumptions 2(i)-(iii), $\tilde{f}_i(.,\mathcal{D}_{i,r})$ is L-smooth, for all $i \in [n]$, and $\mathcal{D}_{i,r}$ data batches drawn uniformly from p_i. Therefore we have:

$$-L\,\mathbf{I} \leq \nabla^2 \tilde{f}_i(\mathbf{w},\mathcal{D}_i) \leq L\,\mathbf{I}, \tag{68}$$

for all $\mathbf{z} \in \mathbb{R}^d$, and data batches $\mathcal{D}_i$ drawn from distribution p_i. This suggests that:

$$\nabla\Psi_i(\mathbf{w},\mathcal{D}_i) \leq (1+\alpha L)\,\mathbf{I}, \tag{69}$$

where due to Lemma 4, the following property holds:

$$\|\Psi_i(\mathbf{w},\mathcal{D}_i) - \Psi_i(\mathbf{v},\mathcal{D}_i)\| \leq (1+\alpha L)\|\mathbf{w}-\mathbf{v}\|, \tag{70}$$

for all $i \in [n]$ and uniformly sample data batches $\mathcal{D}_i$ with respect to p_i. Then, for each two points $\mathbf{w},\mathbf{v} \in \mathbb{R}^d$, we have:

$$\begin{aligned}
\left\|\mathbf{w}_i^{(u)} - \mathbf{v}_i^{(u)}\right\| &\leq (1+\alpha L)\left\|\mathbf{w}_i^{(u-1)} - \mathbf{v}_i^{(u-1)}\right\| && (71)\\
&\leq (1+\alpha L)^2\left\|\mathbf{w}_i^{(u-2)} - \mathbf{v}_i^{(u-2)}\right\| && (72)\\
&\vdots \\
&\leq (1+\alpha L)^u\left\|\mathbf{w}_i^{(0)} - \mathbf{v}_i^{(0)}\right\| && (73)\\
&= (1+\alpha L)^u\left\|\mathbf{w}-\mathbf{v}\right\|. && (74)
\end{aligned}$$

We now show that $\tilde{F}_i^{(u)}(.,\vartheta_i)$is $\hat{L}(u)$-smooth, as follows:

$$\begin{aligned}
&\left\|\nabla\tilde{F}_i^{(u)}(\mathbf{w},\vartheta_i) - \nabla\tilde{F}_i^{(u)}(\mathbf{v},\vartheta_i)\right\| && (75)\\
&= \left\|\left[\prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right]\nabla\tilde{f}_i\left(\mathbf{w}_i^{(u)},\mathcal{D}_{i,u}\right)\right. \\
&\quad \left. - \left[\prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{v}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right]\nabla\tilde{f}_i\left(\mathbf{v}_i^{(u)},\mathcal{D}_{i,u}\right)\right\| && (76)\\
&\overset{\text{tri. ineq.}}{\leq} \left\|\left[\prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right]\left(\nabla\tilde{f}_i\left(\mathbf{w}_i^{(u)},\mathcal{D}_{i,u}\right) - \nabla\tilde{f}_i\left(\mathbf{v}_i^{(u)},\mathcal{D}_{i,u}\right)\right)\right\| && (77)\\
&\quad + \left\|\left[\underbrace{\prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right) - \prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{v}_i^{(r)},\mathcal{D}_{i,r}\right)\right)}_{\mathcal{Q}_{i,u-1}}\right]\nabla\tilde{f}_i\left(\mathbf{v}_i^{(u)},\mathcal{D}_{i,u}\right)\right\| && (78)\\
&\overset{\text{Assump. 2(i)}-(iii)}{\leq} L\left\|\left[\prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right]\right\|\left\|\mathbf{w}_i^{(u)} - \mathbf{v}_i^{(u)}\right\| + \|\mathcal{Q}_{i,u-1}\|\left\|\nabla\tilde{f}_i\left(\mathbf{v}_i^{(u)},\mathcal{D}_{i,u}\right)\right\| && (79)
\end{aligned}$$

$$\overset{(69)}{\leq} L(1+\alpha L)^u \left\|\mathbf{w}_i^{(u)} - \mathbf{v}_i^{(u)}\right\| + G\left\|\mathcal{Q}_{i,u-1}\right\| \tag{80}$$

$$\overset{(71)}{\leq} L(1+\alpha L)^{2u}\left\|\mathbf{w}-\mathbf{v}\right\| + G\left\|\mathcal{Q}_{i,u-1}\right\|. \tag{81}$$

Moreover, we have:

$$\left\|\mathcal{Q}_{i,u-1}\right\| = \left\| \prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right) - \prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{v}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right\| \tag{82}$$

$$= \left\| \prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right) - \left[\prod_{r=0}^{u-2}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right]\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{v}_i^{(u-1)},\mathcal{D}_{i,r-1}\right)\right)\right. \tag{83}$$

$$\left. + \left[\prod_{r=0}^{u-2}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right]\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{v}_i^{(u-1)},\mathcal{D}_{i,r-1}\right)\right) - \prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{v}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right\| \tag{84}$$

$$\overset{\text{tri. ineq.}}{\leq} \alpha\left\|\prod_{r=0}^{u-2}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r}\right)\right)\right\|\left\|\nabla^2\tilde{f}_i\left(\mathbf{w}_i^{(u-1)},\mathcal{D}_{i,r-1}\right) - \nabla^2\tilde{f}_i\left(\mathbf{v}_i^{(u-1)},\mathcal{D}_{i,r-1}\right)\right\| \tag{85}$$

$$+ \left\|\mathcal{Q}_{i,u-2}\right\|\left\|\mathbf{I}-\alpha\nabla^2\tilde{f}_i\left(\mathbf{v}_i^{(u-1)},\mathcal{D}_{i,r}\right)\right\| \tag{86}$$

$$\overset{\text{Assump. } 2(i)-(iii)}{\leq} \alpha H(1+\alpha L)^{u-1}\left\|\mathbf{w}_{u-1}-\mathbf{v}_{u-1}\right\| + (1+\alpha L)\left\|\mathcal{Q}_{i,u-2}\right\| \tag{87}$$

$$\overset{(71)}{\leq} \alpha H(1+\alpha L)^{2u-2}\left\|\mathbf{w}-\mathbf{v}\right\| + (1+\alpha L)\left\|\mathcal{Q}_{i,u-2}\right\|. \tag{88}$$

Thus, we can infer that:

$$\left\|\mathcal{Q}_{i,r}\right\| \leq \alpha H(1+\alpha L)^{2u-2}\left\|\mathbf{w}-\mathbf{v}\right\| + (1+\alpha L)\left\|\mathcal{Q}_{i,r-1}\right\|, \quad \forall r \geq 1, \tag{89}$$

therefore,

$$\left\|\mathcal{Q}_{i,u}\right\| \leq \alpha H(1+\alpha L)^{2u}\left\|\mathbf{w}-\mathbf{v}\right\| + (1+\alpha L)\left\|\mathcal{Q}_{i,u-1}\right\| \tag{90}$$

$$\leq \alpha H\left[(1+\alpha L)^{2u} + (1+\alpha L)^{2u-1}\right]\left\|\mathbf{w}-\mathbf{v}\right\| + (1+\alpha L)^2\left\|\mathcal{Q}_{i,u-2}\right\| \tag{91}$$

$$\vdots \tag{92}$$

$$\leq \alpha H\left[\sum_{r=0}^{u-1}(1+\alpha L)^{2u-r}\right]\left\|\mathbf{w}-\mathbf{v}\right\| + (1+\alpha L)^u\left\|\mathcal{Q}_{i,0}\right\| \tag{93}$$

$$\leq \alpha H(1+\alpha L)^{2u}(u+1)\|\mathbf{w}-\mathbf{v}\|. \tag{94}$$

As a result of (75) and (90), we can conclude the smoothness of $\tilde{F}_i^{(u)}(.,\vartheta_i)$ as follows:

$$\left\|\nabla\tilde{F}_i^{(u)}(\mathbf{w},\vartheta_i) - \nabla\tilde{F}_i^{(u)}(\mathbf{v},\vartheta_i)\right\| \leq \left(L(1+\alpha L)^{2u} + \alpha u G H(1+\alpha L)^{2u-2}\right)\left\|\mathbf{w}-\mathbf{v}\right\| \tag{95}$$

$$\leq \left(\underbrace{(L+\alpha u G H)(1+\alpha L)^{2u}}_{\hat{L}(u)}\right)\left\|\mathbf{w}-\mathbf{v}\right\| \tag{96}$$

Moreover, note that for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, $i\in[n]$, and ϑ_i, we have:

$$\left\|\nabla\tilde{F}_i^{(u)}(\mathbf{w},\vartheta_i)\right\| = \left[\prod_{r=0}^{u-1}\left(\mathbf{I}-\alpha\nabla^2\tilde{f}_i(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r})\right)\right]\nabla\tilde{f}_i(\mathbf{w}_i^{(u)},\mathcal{D}_{i,u}) \tag{97}$$

$$\leq \left[\prod_{r=0}^{u-1}\left\|\mathbf{I}-\alpha\nabla^2\tilde{f}_i(\mathbf{w}_i^{(r)},\mathcal{D}_{i,r})\right\|\right]\left\|\nabla\tilde{f}_i(\mathbf{w}_i^{(u)},\mathcal{D}_{i,u})\right\| \tag{98}$$

$$\overset{(69),\text{Assump. } 2(i)-(iii)}{\leq} (1+\alpha L)^u G. \tag{99}$$

Therefore, we have:

$$\left\|\nabla \tilde{F}_i^{(u)}(\mathbf{z},\vartheta_i) - \nabla F_i^{(u)}(\mathbf{z})\right\| \leq \left\|\nabla \tilde{F}_i^{(u)}(\mathbf{z},\vartheta_i)\right\| + \left\|\nabla F_i^{(u)}(\mathbf{z})\right\| \tag{100}$$

$$\leq \left\|\nabla \tilde{F}_i^{(u)}(\mathbf{z},\vartheta_i)\right\| + \mathbb{E}_{p_i}\left\|\nabla \tilde{F}_i^{(u)}(\mathbf{z},\vartheta_i)\right\| \tag{101}$$

$$\overset{(62)}{\leq} 2(1+\alpha L)^u G, \tag{102}$$

Equations (95), and (100) conclude the proof of Lemma 3.

Now that we have shown the smoothness of $F_i^{(u)}(\cdot)$ and bounded variance for the stochastic gradients, we are ready to state the convergence result for the smooth non-convex functions.

Before stating the proof of Theorem 1, note that under Γ_w=1, matrices $\mathbf{\Delta}(t) \in \mathbb{R}^{(n+m')\times d}$ in the linear system (7), turn into:

$$[\mathbf{\Delta}(t)]_i := \begin{cases} \theta\, \nabla \tilde{F}_i^{(u)}(\mathbf{z}_i(t),\vartheta_i^t)^\top, & i \in [n], \\ \mathbf{0}^\top, & i \notin [n], \end{cases} \tag{103}$$

where $\theta(t) = \theta$ is a constant stepsize.

Moreover, the following lemma helps us to provide a bound between $\mathbf{z}_i(t)$ and $\frac{\mathbf{X}(t)^\top \mathbf{1}}{n}$, for all $i \in [n]$, and $t \geq 0$.

Lemma 5. *Suppose Assumption 1 holds. Consider the sequence of $\mathbf{z}_i(t)$, $i \in [n]$, generated by (7). Then the following property holds: for all $t > 0$*

$$\left\|\mathbf{z}_i(t) - \frac{\mathbf{X}(t)^\top \mathbf{1}}{n}\right\|_1 \leq \delta\lambda^t \left\|\mathbf{X}(0)\right\|_1 + \sum_{k=0}^{t-1} \delta\lambda^{t-k} \left\|\mathbf{\Delta}(k)\right\|_1,$$

where $\delta := \frac{1}{1-n\gamma^6}$, $\lambda := (1-n\gamma^6)^{1/(2n\Gamma_s)}$, $\gamma := (1/n)^{n\Gamma_s}$, and $\|\mathbf{X}\|_1$ denotes the sum of 1-norm of matrix $\mathbf{X}$'s rows, for any matrix $\mathbf{X}$.

According to (7), and the fact that $M(t)$ is a column stochastic matrix, we have:

$$\frac{1}{n}\mathbf{1}^\top \mathbf{X}(t+1) = \frac{1}{n}\mathbf{1}^\top \mathbf{X}(t) - \frac{1}{n}\mathbf{1}^\top \mathbf{\Delta}(t) \tag{104}$$

Due to Lemma 3, we can infer that $F^{(u)}$ is $\hat{L}(u)$-smooth. Therefore, the following property holds:

$$\mathbb{E}F^{(u)}\left(\frac{\mathbf{X}(t+1)^\top \mathbf{1}}{n}\right) = \mathbb{E}F^{(u)}\left(\frac{\mathbf{X}(t)^\top \mathbf{1}}{n} - \frac{\mathbf{\Delta}(t)^\top \mathbf{1}}{n}\right) \tag{105}$$

$$\leq \mathbb{E}F^{(u)}\left(\frac{\mathbf{X}(t)^\top \mathbf{1}}{n}\right) - \theta\mathbb{E}\left\langle \nabla F^{(u)}\left(\frac{\mathbf{X}(t)^\top \mathbf{1}}{n}\right), \frac{1}{n}\sum_{i=1}^n \nabla \tilde{F}_i^{(u)}(\mathbf{z}_i(t),\vartheta_i^t)\right\rangle \tag{106}$$

$$+ \frac{\theta^2 \hat{L}(u)}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla \tilde{F}_i^{(u)}(\mathbf{z}_i(t),\vartheta_i^t)\right\|^2, \tag{107}$$

where using $2\langle a,b\rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a}-\mathbf{b}\|^2$, and taking the expectation, we have:

$$\mathbb{E}\left\|\nabla F^{(u)}\left(\frac{\mathbf{X}(t)^\top \mathbf{1}}{n}\right)\right\|^2 \leq \frac{2\mathbb{E}F^{(u)}\left(\frac{\mathbf{X}(t)^\top \mathbf{1}}{n}\right)}{\theta} - \frac{2\mathbb{E}F^{(u)}\left(\frac{\mathbf{X}(t+1)^\top \mathbf{1}}{n}\right)}{\theta} \tag{108}$$

$$+ \frac{\theta\,\hat{\sigma}(u)^2 \hat{L}(u)}{n} + \frac{2\hat{L}(u)^2}{n}\Psi_x(t),$$

where $\Psi_x(t) = \|\mathbf{Z} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\mathbf{X}(t)\|_F^2$. Note that according to Lemma 5, we know that $\Psi_x(t) = \mathcal{O}(\theta^2)$. Thus, summing (108) from $t = 0$ to $T-1$ and dividing by T leads to the proof.