FedPD: A Federated Learning Framework with Optimal Rates and Adaptivity to Non-IID Data

Xinwei Zhang†, Mingyi Hong†, Sairaj Dhople†, Wotao Yin‡ and Yang Liu$^\#$ *

July 9, 2020

Abstract

Federated Learning (FL) is popular for communication-efficient learning from distributed data. To utilize data at different clients without moving them to the cloud, algorithms such as the Federated Averaging (FedAvg) have adopted a "computation then aggregation" (CTA) model, in which multiple local updates are performed using local data, before sending the local models to the cloud for aggregation. These algorithms fail to work when facing practical challenges, e.g., the local data being non-identically independent distributed. In this paper, we first characterize the behavior of the FedAvg algorithm, and show that without strong and unrealistic assumptions on the problem structure, it can behave erratically (e.g., diverge to infinity). Aiming at designing FL algorithms that are provably fast and require as few assumptions as possible, we propose a new algorithm design strategy from the primal-dual optimization perspective. Our strategy yields algorithms that can deal with non-convex objective functions, achieve the best possible optimization and communication complexity (in certain sense), and deal with full-batch and mini-batch local computation models. Importantly, the proposed algorithms are *communication efficient*, in that the communication effort can be reduced when the level of heterogeneity among the local data also reduces. To our knowledge, this is the first algorithmic framework for FL that achieves all the above properties.

1 Introduction

Federated learning (FL)—a distributed machine learning approach proposed in [1]—has gained popularity for applications involving learning from distributed data. In FL, a cloud server (the "server") can communicate with distributed data sources (the "agents"). The goal is to train a global model that works well for all the distributed data, but without requiring the agents to reveal too much local information. Since its inception, the broad consensus on FL's implementation appears to involve a generic "computation then aggregation" (CTA) protocol. This involves the following steps: S1) the server sends the model $\mathbf{x}$ to the agents; S2) the agents update their local models $\mathbf{x}_i$'s for several iterations based on their local data; S3) the server aggregates $\mathbf{x}_i$'s to obtain a new global model $\mathbf{x}$. It is widely acknowledged that multiple local steps save communication efforts, while only transmitting local models protects data privacy [2].

Even though the FL paradigm has attracted significant research from both academia and industry, and many algorithms such as Federated Averaging (FedAvg) have been proposed, several attributes are not clearly established. In particular, the commonly adopted CTA protocol poses significant theoretical

---

*† University of Minnesota, email: {zhan6234,mhong,sdhople}@umn.edu; ‡ University of California, Los Angeles, email: wotaoyin@math.ucla.edu; $^\#$ Webank, Co. Ltd, email: yangliu@webank.com.

and practical challenges to designing effective FL algorithms. This work attempts to provide a deeper understanding of FL, by raising and resolving a few theoretical questions, as well as by developing an effective algorithmic framework with several desirable features.

Problem Formulation. Consider the vanilla FL solving the following problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{N} \sum_{i=1}^N f_i(\mathbf{x}), \text{ with } f_i(\mathbf{x}) \triangleq w_i \sum_{\xi_i \in \mathcal{D}_i} F(\mathbf{x}; \xi_i), \tag{1}$$

where N is the number of agents; ξ_i denotes one sample in data set $\mathcal{D}_i$ stored on the i-th agent; $F : \mathbb{R}^d \to \mathbb{R}$ is the "loss function" for the i-th data point; and $w_i > 0$ is a "weight coefficient" (a common choice is $w_i = 1/|\mathcal{D}_i|$ [2]). We assume that the loss function takes the same form across different agents, and furthermore, we denote $M := \sum_{i=1}^N |\mathcal{D}_i|$ to be the total number of samples. One can also consider a related setting, where each $f_i(\mathbf{x})$ represents the expected loss [3]

$$f_i(\mathbf{x}) \triangleq \mathbb{E}_{\xi_i \in \mathcal{P}_i} F(\mathbf{x}; \xi_i), \tag{2}$$

where $\mathcal{P}_i$ denotes the data distribution on the i-th agent. Throughout the paper, we will make the following blanket assumptions for problem (1):

A 1. *Each $f_i(\cdot)$, as well as f in (1) is L–smooth:*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \le L \|\mathbf{x} - \mathbf{y}\|, \quad \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L \|\mathbf{x} - \mathbf{y}\|, \ \forall \, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, i = 1, \dots, N.$$

A 2. *The objective of problem (1) is lower bounded: $f(\mathbf{x}) \ge c > -\infty, \ \forall \ \mathbf{x} \in \mathbb{R}^d$.*

In addition to these standard assumptions, state-of-the-art efforts on analysis of FL algorithms oftentimes invoke a number of more *restrictive* assumptions.

A 3. (Bounded Gradient (BG)) *The gradients ∇f_i's are upper bounded (by a constant $G > 0$)*

$$\|\nabla f_i(\mathbf{x})\|^2 \le G^2, \quad \forall \, \mathbf{x} \in \mathbb{R}^d, \forall \, i = 1, \dots, N. \tag{3}$$

A 4. (I.I.D. Data) *Either one of the following holds:*

$$\mathbb{E} \, \nabla[f_i(\mathbf{x})] = \nabla f(\mathbf{x}), \quad \forall \, \mathbf{x} \in \mathbb{R}^d, \ \forall \, i = 1, \dots, N, \tag{4}$$

$$\sum_{i=1}^N \|\nabla f_i(\mathbf{x})\|^2 \le B^2 \|\nabla f(\mathbf{x})\|^2, \quad \forall \, \mathbf{x} \ \in \ \{\mathbf{x} \in \mathbb{R}^d \, | \, \|\nabla f(\mathbf{x})\|^2 > \epsilon\}. \tag{5}$$

Let us comment on the above assumptions.

First, the BG assumption typically does not hold for (1), in particular, $f_i(\mathbf{x}) = \|\mathbf{A}_i\mathbf{x} - \mathbf{b}_i\|^2$ (where $\mathbf{A}_i$ and $\mathbf{b}_i$ are related to data). However, the BG assumption is critical for analyzing FedAvg-type algorithms because it bounds the distance traveled after multiple local iterates. Second, (4) is typically used in FL to characterize homogeneity about local data [4, 5]. However, an assumption of this type does not hold for FL applications where the data (such as medical records, keyboard input data) are generated by the individual agents [1, 3, 6, 7, 8, 9]. A reasonable relaxation to this i.i.d. assumption is the following notion of δ-non-i.i.d.-ness of the data distribution.

A 5. (δ-Non-I.I.D. Data) *The local functions are called δ-non-i.i.d. if either one of the equivalent conditions below holds:*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_j(\mathbf{x})\| \le \delta, \ \forall \ \mathbf{x} \in \mathbb{R}^d, \ \forall \ i \ne j, \text{ or } \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\| \le \delta \ \forall \ \mathbf{x} \in \mathbb{R}^d, \ \forall \ i. \tag{6}$$

By varying δ from 0 to ∞, (6) provides a characterization of data non-i.i.d.-ness. In Appendix A, we give a few examples of loss functions with different values of δ. Note that the second inequality in (6) is often used in decentralized optimization to quantify the *similarity* of local problems [10, 11]. Third, (5) does not hold for many practical problems. To see this, note that this condition is parameterized by ϵ, which is typically the desired optimization accuracy [12]. Since ϵ can be chosen arbitrarily small, (5) essentially requires that the problem is *realizable*, that is, $\|\nabla f(\mathbf{x})\|$ approaches zero only when all the local gradients approach zero at $\mathbf{x}$, that is, when the local data are "similar".

Finally, we mention that our objective is to understand FL algorithm from an optimization perspective. So we say that a solution $\mathbf{x}$ is an ϵ-stationary solution if the following holds:

$$\|\nabla f(\mathbf{x})\|^2 \le \epsilon. \tag{7}$$

We are interested in finding the *minimum* system resources required, such as the number of local updates, the number of times local variables are transmitted to the server, and the number of times local samples $F(\mathbf{x}; \xi_i)$'s are accessed, before computing an ϵ-solution (7). These quantities are referred to as *local computation*, *communication* complexity, and *sample* complexities, respectively.

Questions to address Despite extensive recent research, the FL framework, and in particular, the CTA protocol is not yet well understood. Below, we list four questions regarding the CTA protocol.

Q1 (local updates). What are the best local update directions for the agents to take so as to achieve the best overall system performance (stability, sample complexity, etc.)?

Q2 (global aggregation). Can we use more sophisticated processing in the aggregation step to help improve the system performance (sample or communication complexity)?

Q3 (communication efficiency). If multiple local updates are preformed between two aggregation steps, will it reduce the communication overhead?

Q4 (assumptions). What is the best performance that the CTA type algorithms can achieve while relying on a minimum set of assumptions about the problem?

Although these questions are not directly related to data privacy, another important aspect of FL, we argue that answering these fundamental questions can provide much needed understanding on algorithms following the CTA, and thus the FL approach. A few recent works (to be reviewed shortly) have touched upon those questions, but to our knowledge, none of them has conducted a thorough investigation of the questions listed above.

Related Works. We start with a popular method following the CTA protocol, the FedAvg in Algorithm 1, which covers the original FedAvg [1], the Local SGD [4], PR-SGD [13, 3] and the RI-SGD [14] among others.

In FedAvg, T is the total *stage* number, Q the number of local updates, r the index of the *stage*, q the index of the *inner iteration*, and $\eta^{r,q}$'s are the stepsizes. It has two options for local updates:

Option 1: Sample $\xi_i^{r,q}$ form $\mathcal{D}_i$,

Set $\mathbf{x}_i^{r,q+1} \triangleq \mathbf{x}_i^{r,q} - \eta^{r,q}\nabla F(\mathbf{x}_i^{r,q}; \xi_i^{r,q})$. $\quad$ (8)

Option 2 : $\mathbf{x}_i^{r,q+1} \triangleq \mathbf{x}_i^{r,q} - \eta^{r,q}\nabla f_i(\mathbf{x}_i^{r,q})$. $\quad$ (9)

Algorithm 1 FedAvg Algorithm

Initialize: $\mathbf{x}_i^0 \triangleq \mathbf{x}^0, i = 1, \dots, N$
for $r = 0, \dots, T-1$ *(stage)* **do**
$\quad$ **for** $q = 0, \dots, Q-1$ *(iteration)* **do**
$\quad\quad$ either **Option 1** (Local SGD), for all i
$\quad\quad$ or **Option 2** (Local GD), for all i
$\quad$ **end for**
$\quad$ Global averaging: $\mathbf{x}^{r+1} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i^{r,Q}$
$\quad$ Update agents' $\mathbf{x}_i^{r+1,0} = \mathbf{x}^{r+1}, \ i = 1, \dots, N$
end for

3

Table 1: Convergence rates of FL algorithms, measured by total rounds of communication (RC), number of local updates (LC), and number of accessed sample (AS), before reaching ϵ-stationary solution. DN refers to degree of non-i.i.d, BG refers to bounded gradient, NC is non-convex, μSC means μ-Strongly Convex. p is the function of $\mathcal{O}(\frac{\epsilon}{\delta^2})$ illustrated in Fig. 1. $\star$The i.i.d. assumption of FedProx is described in (5); VRL-SGD needs assumption of bounded variance of the stochastic gradient, which in our finite sum setting implies the BG.

Algorithm	Convexity	DN	BG	RC (T)	LC (QT)	AS
FedAvg [4]	μSC	0	No	$\mathcal{O}\left(1/\epsilon^{1/2}\right)$	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(1/\epsilon)$
FedAvg [9]	μSC	-	Yes	$\mathcal{O}\left(1/\epsilon\right)$	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(1/\epsilon)$
Coop-SGD [5]	NC	0	No	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(1/\epsilon^2)$
MPR-SGD [3]	NC	-	Yes	$\mathcal{O}(1/\epsilon^{3/2})$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(1/\epsilon^2)$
Local-GD [15]	C	-	No	$\mathcal{O}(1/\epsilon^{3/2})$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(M/\epsilon^2)$
FedProx [12]	NC	-$\star$	No	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(1/\epsilon^2)$
F-SVRG[17]	NC	-	No	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(Q/\epsilon)$	$\mathcal{O}(M/\epsilon + Q/\epsilon)$
VRL-SGD[16]	NC	-	Yes$\star$	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(1/\epsilon^2)$
Fed-PD-GD	NC	$\delta > 0$	No	$\mathcal{O}((1-p)/\epsilon)$	$\mathcal{O}(\log(1/\epsilon)/\epsilon)$	$\mathcal{O}(M\log(1/\epsilon)/\epsilon)$
Fed-PD-SGD	NC	$\delta > 0$	No	$\mathcal{O}((1-p)/\epsilon)$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(1/\epsilon^2)$
Fed-PD-VR	NC	-	No	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(Q/\epsilon)$	$\mathcal{O}(M + \sqrt{M}/\epsilon)$

Many recent works are extensions of FedAvg. The algorithm proposed in [3] adds momentum to the algorithm. In [14], the data on the local agents are separated into blocks and shared with other agents. In [15] the *local GD* version (9) is studied. In [5], a cooperative-SGD is considered; it includes virtual agents, extra variables, and relaxes the parameter server topology.

It is pertinent to consider how these algorithms address questions Q1–Q4. For Q1, most FedAvg-type algorithms perform multiple local (stochastic) GD steps to minimize the local loss function. However, we will see shortly that in many cases, successive local GD steps lead to algorithm divergence. For Q2, most algorithms use simple averaging, and there is little discussion on whether other types of (linear) processing will be helpful. For Q3, a number of recent works such as [3, 15] show that, for non-convex problems, to achieve ϵ-solution (7), a total of $O(1/\epsilon^{3/2})$ aggregation steps are needed. However, it is not clear if this achieves the best communication complexity. As for Q4, the FedAvg-type algorithm typically requires either some variant of the BG assumption, or some i.i.d. assumption, or both; See Table 1 rows 1–5 for detailed discussions.

A number of more recent works have improved upon FedAvg in various aspects. FedProx [12] addresses Q1 and Q4 by perturbing the update direction. This algorithm does not need the BG, but it still requires the i.i.d. assumption (5). The VRL-SGD proposed in [16] addresses Q1 and Q4 by using the variance reduction (VR) technique to update the directions for local agents and achieves $\mathcal{O}(1/\epsilon)$ communication complexity without the i.i.d. assumption. F-SVRG [17] is another recent algorithm that uses VR. This algorithm *does not* follow the CTA protocol as the agents have to transmit the local gradients, but it does not require A3 and A4. The PR-SPIDER [18] further improves upon FSVRG by reducing sample complexity (SC) from $\mathcal{O}(M/\epsilon)$ to $\mathcal{O}(\sqrt{M}/\epsilon)$ (where M is typically larger than $1/\epsilon$). Although FSVRG and PR-SPIDER neither require the BG or the i.i.d. assumptions, they require the agents to transmit local gradients to the server and thus do not follow the CTA protocol. This is undesirable, as it has been shown that local gradient information can leak private data [19]. Additionally, questions Q2-Q3 are *not* addressed in these works.

Our Main Contributions. First, we address Q1-Q4 and provide an in-depth examination of the CTA protocol. We show that algorithms following the CTA protocol that are based on successive local gradient updates, the best possible communication efficiency is $\mathcal{O}(1/\epsilon)$; neither additional local processing nor

general linear processing can improve this rate. We then show that the BG and/or i.i.d. data assumption is important for the popular FedAvg to work as intended.

Our investigation suggests that the existing FedAvg-based algorithms are (provably) insufficient in dealing with many practical problems, calling for a new design strategy. We then propose a meta-algorithm called Federated Primal-Dual (FedPD), which also follows the CTA protocol and can be implemented in several different forms with desirable properties. In particular, it *i)* can deal with the general non-convex problem, *ii)* achieve the best possible optimization and communication complexity when data is non-i.i.d., *iii)* achieve convergence under only Assumptions A1– A2.

Most importantly, the communication pattern of the proposed algorithm can be adapted to the degree of non-i.i.d.-ness of the local data. That is, we show that under the δ-non-i.i.d. condition (6), communication saving and data heterogeneity interestingly exhibit a linear-logarithmic relationship; see Fig. 1 for an illustration. To our knowledge, this is the first algorithm for FL that achieves all the above properties.

Figure 1: Relation of the percentage of comm. savings, accuracy ϵ, heterogeneity δ; Fixing ϵ, when the data is heterogeneous (left) the curve is linear, while the data is homogeneous (right) the curve is logarithmic. Details in Sec. 3.2.

2 Addressing Open Questions

We first address Q2–Q3. Specifically, for problems satisfying A1–A2, does performing multiple local updates or using different ways to combine local models reduce communication complexity? We show that such a saving is impossible; there exist problems satisfying A1–A2, yet no matter what types of linear combinations the server performs, as long as the agents use local gradients to update the model, it takes at least $\mathcal{O}(1/\epsilon)$ communication rounds to achieve an ϵ-stationary solution (7). Consider the following generic CTA protocol. Let t denote the index for communication rounds. Between two rounds $t-1$ and t, each agent performs Q local updates. Denote $x_i^{t-1,q}$ to be the q-th local update. Then, $x_i^{t-1,Q}$'s are sent to the server, combined through a (possibly time-varying) function $V^t(\cdot) : \mathbb{R}^{Nd} \to \mathbb{R}^d$, and sent back. The agents then generate a new iterate, by combining the received message with past gradients using a (possibly time-varying) function $W_i^t(\cdot)$:

$$x^t = V^t(\{x_i^{t-1,Q}\}_{i=1}^N), \; x_i^{t,0} = x^t, \;\; \forall \, i \in [N], \tag{10a}$$

$$x_i^{t,q} \in W_i^t\left(\{x_i^{r,k}, \{\nabla F(x_i^{r,q};\xi_i)\}_{\xi_i \in D_i}\}_{k \in [q-1], r \in [t]}\right), \; \forall \, q \in [Q], \; \forall \, i \in [N]. \tag{10b}$$

We focus on the case where the $V^t(\cdot)$'s and $W_i^t(\cdot)$'s are *linear* operators, which implies that $x_i^{t,q}$ can use all past iterates and (sample) gradients for its update. Clearly, (10) covers both the local-GD and local-SGD versions of FedAvg as special cases. In the following, we provide an informal statement of the result. The formal statement and the full proof are given in Appendix B and Theorem 2.

Claim 2.1. (Informal) *Consider any algorithm A that belongs to the class described in* (10)*, with* $V^t(\cdot)$ *and* $W_i^t(\cdot)$*'s being linear and possibly time-varying operators. Then, there exists a non-convex problem instance satisfying Assumptions 1–2 such that for any $Q > 0$, algorithm A takes at least $\mathcal{O}(1/\epsilon)$ communication rounds to reach an ϵ-stationary solution satisfying* (7)*.*

Remark 1. The proof technique is related to those developed from both classical and recent works that characterize lower bounds for first-order methods, in both centralized [20, 21] and decentralized [22, 23]
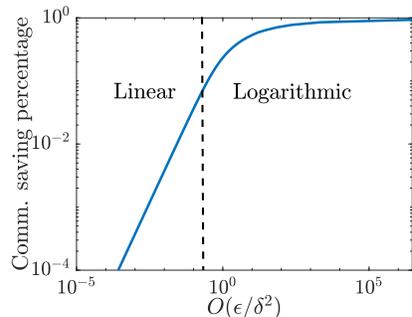
5

settings. The main technical difference is that our processing model (10) additionally allows local processing iterations, and there is a central aggregator. Our goal is *not* to establish lower bounds on the number of (centralized) gradient access, nor to show the optimal graph dependency, but to characterize (potential) communication savings when allowing multiple steps of local processing.

In the proof, we construct difficult problem instances in which f_i's are not i.i.d. (more precisely, δ in assumption (4) grows with the total number of iterations T). Then we show that it is necessary to aggregate (thus communicate) to make any progress. On the other hand, it is obvious that in another extreme case where the data are 0-non-i.i.d., only $\mathcal{O}(1)$ communication rounds are needed. An *open question* is: when the local data are *related* to each other, i.e., δ lies between 0 and infinity, is it possible to reduce the total communication rounds? This question is addressed below in Sec. 3. ∎

We now address Q1 and Q4. We consider the FedAvg Algorithm 1, and show that BN and/or i.i.d. assumptions are critical for them to perform well. Our result suggests that, despite its popularity, components in FedAvg, such as the pure local (stochastic) gradient directions and linear aggregation are not compatible with each other. The proof of the results below are given in Appendix C.

Claim 2.2. *Fix any constant $\eta > 0$, $Q > 1$ for Algorithm 1. There exists a problem that satisfies A1 and A2 but fails to satisfy A3 and A4, on which FedAvg diverges to infinity.*

Remark 2. A recent work [15] has shown that FedAvg with *constant* stepsize $\eta > 0$ can only converge to a neighborhood of the global minimizer for *convex* problems. Moreover, the error to the global optima is related to Q and the degree of non-i.i.d.-ness as measured by the size of $\sum_{i=1}^{N} \|\nabla f_i(\mathbf{x}^\star)\|^2$ where $\mathbf{x}^\star$ is the global optimal solution. On the other hand, our result indicates that when f_i's are *non-convex*, FedAvg can perform much worse without the BN and the i.i.d. assumption. Even if $Q = 2$ and there exists a solution such that $\sum_{i=1}^{N} \|f_i(\hat{\mathbf{x}})\|^2 = 0$, FedAvg (with constant stepsize η) diverges and the iteration can go to ∞. ∎

One may think that using a constant stepsize is the culprit for the divergence in Claim 2.2. In fact, we can show that having BG or not can still impact the performance of FedAvg, even when *diminishing* stepsize is used. In particular, we show in Appendix D, that FedAvg converges under the BG assumption for any *diminishing* stepsize, but without it, the choice of the stepsize can be *significantly restricted.*

3   The FedPD Framework

The previous section reveals a number of properties about FedAvg and, broadly speaking, the CTA protocol. But *why* does FedAvg only work under very restrictive conditions? Is it because the local update directions are not chosen correctly? Is it possible to make it work without any additional assumptions? Can we reduce communication effort when the local data becomes more i.i.d.?

In this section, we propose a meta-algorithm called Federated Primal-Dual (FedPD), which can be specialized into different sub-variants to address the above questions. These algorithms possess a few desirable features: They can achieve the best optimization and communication complexity when data is non-i.i.d. (i.e., achieving the bounds in Claim 2.1); they only require A1 –A2, while being able to utilize both full or sampled local gradients. Most importantly, the communication pattern of the proposed algorithm can be made adaptive to the degree of data non-i.i.d.-ness across the agents.

3.1 The Proposed Algorithm

Our algorithm is based on the following *global consensus* reformulation of the original problem (1):

$$\min_{\mathbf{x}_0, \mathbf{x}_i} \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{x}_i), \quad \text{s.t.} \ \ \mathbf{x}_i = \mathbf{x}_0, \ \forall i \in [N]. \tag{11}$$

Similar to traditional primal-dual based algorithms [24], the idea is that, when relaxing the equality constraints, the resulting problem is *separable* across different nodes. However, different from ADMM, the agents can now perform either a *single* (or *multiple*) local update(s) between two communication rounds. Importantly, such flexibility makes it possible to adapt the communication frequency to the degree of δ-non-i.i.d.-ness of the local data. In particular, we identify that under δ-non-i.i.d. (4), the fraction ϵ/δ^2 is the key quantity that determines communication saving; see Fig. 1. Intuitively, significant reduction can be achieved when δ is smaller than ϵ; otherwise, the reduction goes to zero linearly as δ increases. To our knowledge, none of the existing ADMM based algorithms, nor FL based algorithms, are able to provably achieve such a reduction.

To present our algorithm, let us define the augmented Lagrangian (AL) function of (11) as

$$\mathcal{L}(\mathbf{x}_{0:N}, \lambda) \triangleq \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_i(\mathbf{x}_0, \mathbf{x}_i, \lambda_i), \ \mathcal{L}_i(\mathbf{x}_i, \mathbf{x}_0, \lambda_i) \triangleq f_i(\mathbf{x}_i) + \langle \lambda_i, \mathbf{x}_i - \mathbf{x}_0 \rangle + \frac{1}{2\eta} \|\mathbf{x}_i - \mathbf{x}_0\|^2.$$

Fixing $\mathbf{x}_0$, the AL is separable over all local pairs $\{(\mathbf{x}_i, \lambda_i)\}$. The key technique in the design is to specify *how* each local AL $\mathcal{L}_i(\cdot)$ should be optimized, and *when* to perform model aggregation.

Federated primal-dual algorithm (FedPD) captures the main idea of the classical primal-dual based algorithm while meeting the flexibility need of FL; see Algorithm 2. In particular, its update rules share a similar pattern as ADMM, but it does not specify how the local models are updated. Instead, an *oracle* $\text{Oracle}_i(\cdot)$ is used as a placeholder for local processing, and we will see that careful instantiations of these oracles lead to algorithms with different properties. Importantly, we introduce a critical constant $p \in [0, 1)$, which determines the frequency at which the aggregation and communication steps are skipped. In Algorithm 3 and Algorithm 4, we provide two useful examples of the local oracles.

In Algorithm 3, Q_i's are chosen so that the local problems are solved accurately enough to satisfy:

$$\left\| \nabla_{\mathbf{x}_i} \mathcal{L}(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2 \le \epsilon_1. \tag{12}$$

We provide two ways for solving this subproblem by using GD and SGD, but any other solver that achieves (12) can be used. For the SGD version, the stochastic gradient is defined as

$$h_i(\mathbf{x}_i^{r,q}; \xi_i^{r,q}) \triangleq \nabla F(\mathbf{x}_i^{r,q}; \xi_i^{r,q}), \ \text{with} \ \xi_i^{r,q} \sim \mathcal{D}_i, \tag{13}$$

where $\sim$ denotes uniform sampling. Despite the simplicity of the local updates, we will show that using Oracle I makes FedPD adaptive to the non-i.i.d. parameter δ.

In Algorithm 4, the oracle applies the variance reduction technique to reduce the sample complexity. The detailed descriptions and the analyses are given in Appendix G due to the space limitation.

3.2 Convergence and Complexity Analysis

We analyze the convergence of FedPD with Oracle I. The detailed proofs are given in Appendix F. The convergence analysis of FedPD with an alternative Oracle is given in Appendix G

Algorithm 2 Federated Primal-Dual Algorithm

Input: $\mathbf{x}^0, \eta, p, T, Q_1, \ldots, Q_N$
Initialize: $\mathbf{x}_0^0 = \mathbf{x}^0$,
for $r = 0, \ldots, T-1$ **do**
   for $i = 1, \ldots, N$ in parallel **do**
      Local Updates:
      $\mathbf{x}_i^{r+1} = \text{Oracle}_i(\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r), Q_i)$
      $\lambda_i^{r+1} = \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r)$
      $\mathbf{x}_{0,i}^{r+} = \mathbf{x}_i^{r+1} + \eta\lambda_i^{r+1}$
   end for
   With probability $1-p$:
      Global Communicate:
      $\mathbf{x}_0^{r+1} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_{0,i}^{r+}$
      $\mathbf{x}_{0,i}^{r+1} = \mathbf{x}_0^{r+1}, \; i = 1, \ldots, N$
   With probability p:
      Local Update: $\mathbf{x}_{0,i}^{r+1} \triangleq \mathbf{x}_{0,i}^{r+}$
end for

Algorithm 3 Oracle Choice I

Input: $\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r), Q_i$
Initialize: $\mathbf{x}_{i,0}^r = \mathbf{x}_i^r$,
Option I (GD)
for $q = 0, \ldots, Q_i - 1$ **do**
  $\mathbf{x}_i^{r,q+1} = \mathbf{x}_i^{r,q} - \eta_1 \nabla_{\mathbf{x}_i}\mathcal{L}(\mathbf{x}_i^{r,q}, \mathbf{x}_{0,i}^r, \lambda_i^r)$
end for
Option II (SGD)
for $q = 0, \ldots, Q_i - 1$ **do**
  $\mathbf{x}_i^{r,q+1} = \mathbf{x}_i^{r,q} - \eta_1(h_i(\mathbf{x}_i^{r,q}; \xi_i^{r,q}) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r,q} - \mathbf{x}_{0,i}^r))$
end for
Output: $\mathbf{x}_i^{r+1} \triangleq \mathbf{x}_i^{r,Q_i}$

Theorem 1. *Suppose A1 –A2 hold. Consider FedPD with Oracle I, where Q_i are selected by (12).*
Case I) *Suppose A5 holds with $\delta = \infty$. Set $0 < \eta < \frac{\sqrt{5}-1}{4L}$, $p = 0$. Then we have:*

$$\frac{1}{T}\sum_{r=0}^{T}\|\nabla f(\mathbf{x}_0^r)\|^2 \leq \frac{C_2}{T}D_0 + C_4\epsilon_1, \;\; \text{with} \;\; D_0 := f(\mathbf{x}_0^0) - f(\mathbf{x}^\star).$$

Case II) *Suppose $0 < \eta < \frac{\sqrt{5}-1}{4L}$, $0 \leq p < 1$, and A5 holds with a finite δ. Then we have:*

$$\frac{1}{T}\sum_{r=0}^{T}\mathbb{E}\|\nabla f(\mathbf{x}_0^r)\|^2 \leq \frac{C_2}{T}D_0 + \frac{\eta(N-1)C_5(1-C_3^{1/(1-p)})^2 p(p^2(3+L\eta)^2+4)}{N(1-2L\eta-p(1+L\eta))^2}(\delta^2+\epsilon_1) + C_4\epsilon_1. \tag{14}$$

Here $C_2, C_4, C_5 > 0$ are constants independent of T, δ, p; $C_3 := \frac{p(1+L\eta)+L\eta}{1-L\eta} \geq 0$.

Remark 3. (**Communication complexity**) Case I says if one does not skip communication ($p = 0$), then to achieve ϵ-stationarity (i.e., $\|\nabla f(\mathbf{x}_0^t)\|^2 \leq \epsilon$ for some $r \in (1,T)$), they need to set $T = 1/(2C_2D_0\epsilon)$, $\epsilon_1 = \epsilon/(2C_4)$, and the total communication rounds is T.

In Case II, the second term on the right hand side of (14) involves both δ and p, so we can select them appropriately to reduce communication while maintaining the same accuracy. Specifically, we choose $\epsilon_1 = \min\{\epsilon/(4C_4), \delta^2\}$ and $T = 1/(2C_2 D_0 \epsilon)$. Then we have

$$C(p) \triangleq \eta \frac{(1 - C_3^{1/(1-p)})^2 p(p^2(3 + L\eta)^2 + 4)}{(1 - 2L\eta - p(1 + L\eta))^2} \leq \frac{\epsilon}{3\delta^2}.$$

Then we can achieve the same ϵ-accuracy as Case I. The communication rounds here is $T(1 - p)$.

The relation between the saving p and $\frac{\epsilon}{\delta^2}$ is showed in Table 2. Roughly speaking, p is inversely proportional to degree of non-i.i.d-ness δ^2 when $\delta^2 \in (\mathcal{O}(\epsilon), \infty)$; further, $p \to 1$ at a log-rate when $\delta^2 \to 0$. Our result also indicates that, when using stage-wise training for neural networks, the algorithm can communicate less at the early stages since they typically have lower accuracy target to enable larger stepsizes [25].

Table 2: The relation between p and $\frac{\epsilon}{\delta^2}$ with fixed $\eta = \frac{\sqrt{5}-1}{8L}$.

Range of p	C_3	$C(p)$	p as function of $\frac{\epsilon}{\delta^2}$	Relation
$[0, \frac{1-2L\eta}{1+L\eta})$	< 1	$\approx 12\eta p$	$\frac{1}{36\eta}\frac{\epsilon}{\delta^2}$	**Linear**
$[\frac{1-2L\eta}{1+L\eta}, 1)$	≥ 1	$\approx 14\eta C_3^{2/(1-p)}$	$1 - 2/\log(\frac{1}{42\eta}\frac{\epsilon}{\delta^2})$	**Log**

Remark 4. (**Computation complexity**) To achieve ϵ accuracy, we need both $T = \mathcal{O}(1/\epsilon)$ and $\epsilon_1 = \mathcal{O}(\epsilon)$. As the local AL is strongly convex with respect to $\mathbf{x}_i$, optimizing it to ϵ accuracy requires $\mathcal{O}(\log(\epsilon))$ iterations for GD and $\mathcal{O}(1/\epsilon)$ for SGD [26]. So the total number of times that the local gradients (respectively, stochastic gradients) are accessed is given by $\mathcal{O}(1/\epsilon \times \log(1/\epsilon))$ (respectively, $\mathcal{O}(1/\epsilon^2)$).

We conclude this section by noting that the above communication and computation complexity results we have obtained are the best so far among all FL algorithms for non-convex problems satisfying A1 − A2. Please see the last three rows of Table 1 for a summary of the results.

3.3 Connection with Other Algorithms

Before we close this section, we discuss the relation of FedPD with a few existing algorithms.

The FedProx/FedDANE In FedProx [12] the agents optimize the following local objective: $f_i(\mathbf{x}_i) + \frac{\rho}{2}\|\mathbf{x}_i - \mathbf{x}_0^r\|^2$. This fails to converge to the global stationary solution. In contrast, FedPD introduces extra local dual variables $\{\lambda_i\}$ that record the gap between the local model $\mathbf{x}_i$ and the global model $\mathbf{x}_0$ which help the global convergence. FedDANE [27] also proposes a way of designing the subproblem by using the global gradient, but this violates the CTA protocol. Compared with these two algorithms, the proposed FedPD has weaker assumptions, and it achieves better sample and/or communication complexity.

Event Triggering Algorithms. A number of recent works such as Lazily Aggregated Gradient (LAG) [28] and COLA [29] have been proposed to occasionally skip message exchanges among the agents to save communication. In LAG, each agent receives the global model every iteration, and decides whether to send some local *gradient differences* by checking certain conditions. Since gradient information is transmitted, LAG does not belong to the algorithm class (10). When the local problems are *unbalanced*, in the sense that the discrepancy between the local Lipschitz gradients L_i's is large, then the agents with smaller L_i's can benefit from the lazy aggregation. Meanwhile, instead of measuring whether the local problems are balanced, the δ-non-i.i.d. criteria characterizes if local problems are *similar* by measuring the uniform difference between arbitrary pairs of the local problems. If the data is i.i.d., then the agents benefit equally from the communication reduction.
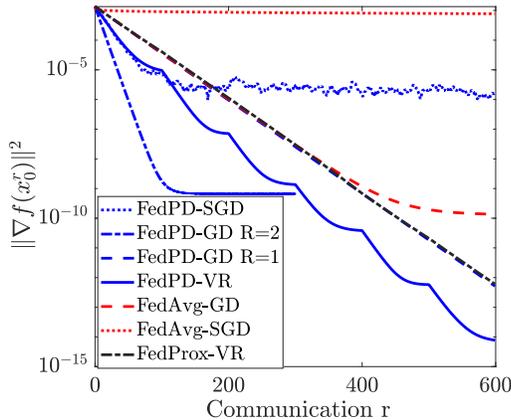
(a) Stationary gap of FedAvg, FedProx and FedPD; weakly non-i.i.d. data.
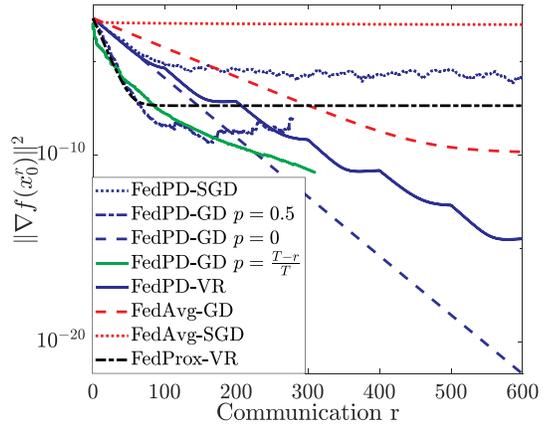
(b) Stationary gap of FedAvg, FedProx and FedPD; strongly non-i.i.d. data.

Figure 2: The convergence result of the algorithms on penalized logistic regression with weakly and strongly non-i.i.d. data with respect to the number of communication rounds.

4   Numerical Experiments

In the first experiment, we show the convergence of the proposed algorithms on synthetic data with FedAvg and FedProx as baselines. We use the non-convex penalized logistic regression [30] as the loss function. We use two ways to generate the dataset, in the first case (referred to as the "weakly non-i.i.d" case), the data is generated in an i.i.d. way. In the second case (referred to as the "strongly non-i.i.d" case), we generate the data using non-i.i.d. distribution. In both cases there are 400 samples on each agent with total 100 agents.

We run FedPD with Oracle I (FedPD-SGD and FedPD-GD) and Oracle II (FedPD-VR). For FedPD-SGD, we set $Q = 600$, and for FedPD-GD and FedPD-VR we set $Q = 8$. For FedPD-GD we set $p = 0$ and $p = 0.5$, wherein the later case the agents skips half of the communication rounds. For FedPD-VR, we set mini-batch size $B = 1$ and gradient computation frequency $I = 20$. For comparison, we also run FedAvg with local GD/SGD and FedProx. For FedAvg with GD, $Q = 8$, and for FedAvg with SGD, $Q = 600$. For FedProx, we solve the local problem using variance reduction for $Q = 8$ iterations. The total number of iterations T is set as 600 for all algorithms.

Fig. 2 shows the results with respect to the number of communication rounds. In Fig. 2(a), we compare the convergence of the tested algorithms on weakly non-i.i.d. data set. It is clear that FedProx and FedPD with $p = 0$ (i.e., no communication skipping) are comparable. Meanwhile, FedAvg with local GD will not converge to the stationary point with a constant stepsize when local update step $Q > 1$. By skipping half of the communication, FedPD with local GD can still achieve a similar error as FedAvg, but using fewer communication rounds. In Fig. 2(b), we compare the convergence results of different algorithms with the strongly non-i.i.d. data set. We can see that the algorithms using stochastic solvers become less stable compared with the case when the data sets are weakly non-i.i.d. Further, FedPD-VR and FedPD-GD with $p = 0$ are able still to converge to the global stationary point while FedProx will achieve a similar error as the FedAvg with local GD.

We included more details on the experimental results and additional experiments in Appendix H.

10

5 Conclusion

We study federated learning under the CTA protocol. We explore a number of theoretical properties of this protocol, and design a meta-algorithm called FedPD, which contains various algorithms with desirable properties, such achieving the best communication/computation complexity, as being able to adapt its communication pattern with data heterogeneity.

References

[1] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[2] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith, "Federated learning: Challenges, methods, and future directions," *arXiv preprint arXiv:1908.07873*, 2019.

[3] Hao Yu, Rong Jin, and Sen Yang, "On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization," in *Proceedings of the 36th International Conference on Machine Learning*, Kamalika Chaudhuri and Ruslan Salakhutdinov, Eds., Long Beach, California, USA, 09–15 Jun 2019, vol. 97 of *Proceedings of Machine Learning Research*, pp. 7184–7193, PMLR.

[4] Sebastian Urban Stich, "Local sgd converges fast and communicates little," *ICLR 2019 - International Conference on Learning Representations*, p. 17, 2019.

[5] Jianyu Wang and Gauri Joshi, "Cooperative SGD: A unified framework for the design and analysis of communication-efficient sgd algorithms," *arXiv preprint arXiv:1808.07576*, 2018.

[6] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Network*, vol. 33, no. 5, pp. 156–165, Sep. 2019.

[7] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar, "Federated multi-task learning," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., pp. 4424–4434. Curran Associates, Inc., 2017.

[8] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konecny, Stefano Mazzocchi, H Brendan McMahan, et al., "Towards federated learning at scale: System design," *arXiv preprint arXiv:1902.01046*, 2019.

[9] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.

[10] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2017.

[11] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan, "When edge meets learning: Adaptive control for resource-constrained distributed machine learning," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 63–71.

[12] Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith, "On the convergence of federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.

[13] Hao Yu, Sen Yang, and Shenghuo Zhu, "Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 5693–5700.

[14] Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe, "Trading redundancy for communication: Speeding up distributed SGD for non-convex optimization," in *Proceedings of the 36th International Conference on Machine Learning*, Kamalika Chaudhuri and Ruslan Salakhutdinov, Eds., Long Beach, California, USA, 09–15 Jun 2019, vol. 97 of *Proceedings of Machine Learning Research*, pp. 2545–2554, PMLR.

[15] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik, "First analysis of local GD on heterogeneous data," *arXiv preprint arXiv:1909.04715*, 2019.

[16] Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng, "Variance reduced local SGD with lower communication complexity," *arXiv preprint arXiv:1912.12844*, 2019.

[17] Shicong Cen, Huishuai Zhang, Yuejie Chi, Wei Chen, and Tie-Yan Liu, "Convergence of distributed stochastic variance reduced methods without sampling extra data," *arXiv preprint arXiv:1905.12648*, 2019.

[18] Pranay Sharma, Prashant Khanduri, Saikiran Bulusu, Ketan Rajawat, and Pramod K Varshney, "Parallel restarted spider–communication efficient distributed nonconvex optimization with optimal computation complexity," *arXiv preprint arXiv:1912.06036*, 2019.

[19] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen, "iDLG: Improved deep leakage from gradients," 2020.

[20] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, Springer, 2004.

[21] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Lower bounds for finding stationary points i," *Mathematical Programming*, Jun 2019.

[22] K. Scaman, F. Bach, S. Bubeck, Y. Lee, and L. Massoulié, "Optimal algorithms for smooth and strongly convex distributed optimization in networks," *arXiv preprint arXiv:1702.08704*, 2017.

[23] H. Sun and M. Hong, "Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms," *IEEE Transactions on Signal processing*, July 2019, accepted for publication.

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

[25] Andrew Senior, Georg Heigold, Marc'Aurelio Ranzato, and Ke Yang, "An empirical study of learning rates in deep neural networks for speech recognition," in *2013 IEEE international conference on acoustics, speech and signal processing.* IEEE, 2013, pp. 6724–6728.

[26] K. Yuan, B. Ying, S. Vlaski, and A. H. Sayed, "Stochastic gradient descent with finite samples sizes," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2016, pp. 1–6.

[27] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith, "FedDANE: A federated newton-type method," *arXiv preprint arXiv:2001.01920*, 2020.

[28] Tianyi Chen, Georgios Giannakis, Tao Sun, and Wotao Yin, "LAG: Lazily aggregated gradient for communication-efficient distributed learning," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., pp. 5050–5060. Curran Associates, Inc., 2018.

[29] W. Li, Y. Liu, Z. Tian, and Q. Ling, "COLA: Communication-censored linearized admm for decentralized consensus optimization," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5237–5241.

[30] Anestis Antoniadis, Irène Gijbels, and Mila Nikolova, "Penalized likelihood regression for generalized linear models with non-quadratic penalties," *Annals of the Institute of Statistical Mathematics*, vol. 63, no. 3, pp. 585–615, 2011.

[31] Sebastian Caldas, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar, "Leaf: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2018.

A Examples of Cost Functions Satisfy A5

In this part, we provide a commonly used function that satisfies A5.

Logistic Regression

Consider the case where the kth sample $\xi_{i,k}$ in data set $\mathcal{D}_i$ consist of a feature vector $\mathbf{a}_k$ and a scalar label b_k. The feature vector $\mathbf{a}_k$ has the same length as $\mathbf{x}$ and b_k is a scalar in $\mathbb{R}$. Then the loss function of a logistic regression problem is expressed as

$$f_i(\mathbf{x}) = \frac{1}{|\mathcal{D}_i|} \sum_{(\mathbf{a}_k, b_k) \in \mathcal{D}_i} \frac{1}{1 + \exp(b_k - \mathbf{a}_k^T \mathbf{x})}. \tag{15}$$

The gradient of this loss function is

$$\nabla f_i(\mathbf{x}) = \frac{1}{|\mathcal{D}_i|} \sum_{(\mathbf{a}_k, b_k) \in \mathcal{D}_i} \frac{\mathbf{a}_k \exp(b_k - \mathbf{a}_k^T \mathbf{x})}{(1 + \exp(b_k - \mathbf{a}_k^T \mathbf{x}))^2}. \tag{16}$$

Define the scalar $\frac{\exp(b_k - \mathbf{a}_k^T \mathbf{x})}{(1 + \exp(b_k - \mathbf{a}_k^T \mathbf{x}))^2}$ as $v(\mathbf{a}_k, b_k, \mathbf{x})$, we have $v(\mathbf{a}_k, b_k, \mathbf{x}) \in (0, 1)$, $\forall x, \mathbf{a}_k, b_k$. Further stack $v(\mathbf{a}_k, b_k, \mathbf{x})$ as $\mathbf{v}(\mathcal{D}_i, \mathbf{x})$, that is $\mathbf{v}(\mathcal{D}_i, \mathbf{x}) = [v(\mathbf{a}_1, b_1, \mathbf{x}); \ldots, ; v(\mathbf{a}_{|\mathcal{D}_i|}, b_{|\mathcal{D}_i|}, \mathbf{x})]$. Further we define A_i as the stacked matrix of all $\mathbf{a}_k \in \mathcal{D}_i$ (i.e., $A_i = [\mathbf{a}_1, \ldots, \mathbf{a}_{|\mathcal{D}_i|}]$), then we can express $\nabla f_i(\mathbf{x})$ as

$$\nabla f_i(\mathbf{x}) = \frac{1}{|\mathcal{D}_i|} A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x}). \tag{17}$$

The difference between the gradients of f_i and f_j is

$$
\begin{aligned}
\|\nabla f_i(\mathbf{x}) - \nabla f_j(\mathbf{x})\| &= \left\| \frac{1}{|\mathcal{D}_i|} A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x}) - \frac{1}{|\mathcal{D}_j|} A_j \mathbf{v}(\mathcal{D}_j, \mathbf{x}) \right\| \\
&\leq \frac{1}{|\mathcal{D}_i|} \|A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x})\| + \frac{1}{|\mathcal{D}_j|} \|A_j \mathbf{v}(\mathcal{D}_j, \mathbf{x})\|.
\end{aligned} \tag{18}
$$

As $v(\mathbf{a}, b, \mathbf{x}) \in (0, 1)$, we know $\|\mathbf{v}(\mathcal{D}_i, \mathbf{x})\| \leq \|[1, \ldots, 1]\| = \sqrt{|\mathcal{D}_i|}$, which implies:

$$\|A_i\| \geq \frac{\|A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x})\|}{\|\mathbf{v}(\mathcal{D}_i, \mathbf{x})\|} \geq \frac{\|A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x})\|}{\sqrt{|\mathcal{D}_i|}}.$$

Utilizing the above inequality in (18), we obtain:

$$
\begin{aligned}
\|\nabla f_i(\mathbf{x}) - \nabla f_j(\mathbf{x})\| &\leq \frac{1}{|\mathcal{D}_i|} \|A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x})\| + \frac{1}{|\mathcal{D}_j|} \|A_j \mathbf{v}(\mathcal{D}_j, \mathbf{x})\| \\
&\leq \frac{1}{\sqrt{|\mathcal{D}_i|}} \|A_i\| + \frac{1}{\sqrt{|\mathcal{D}_j|}} \|A_j\|.
\end{aligned} \tag{19}
$$

So we can define $\delta = \max_{i,j} \left\{ \frac{1}{\sqrt{|\mathcal{D}_i|}} \|A_i\| + \frac{1}{\sqrt{|\mathcal{D}_j|}} \|A_j\| \right\}$ which is a finite constant. Note that the above analysis holds true for any D_i and $\mathbf{x}$. Note that with finer analysis we can obtain better bounds for δ.

Hyperbolic Tangent

Similar to logistic regression, we can also show that A5 holds for hyperbolic tangent function which is commonly used in neural network models. First, notice that the hyperbolic tangent is a rescaled version of logistic regression:

$$\tanh(b_k - \mathbf{a}_k^T \mathbf{x}) = \frac{\exp(b_k - \mathbf{a}_k^T \mathbf{x}) - \exp(\mathbf{a}_k^T \mathbf{x} - b_k)}{\exp(b_k - \mathbf{a}_k^T \mathbf{x}) + \exp(\mathbf{a}_k^T \mathbf{x} - b_k)} = \frac{2}{1 + \exp(2(b_k - \mathbf{a}_k^T \mathbf{x}))} - 1,$$

Therefore we have

$$\nabla_{\mathbf{x}} \tanh(b_k - \mathbf{a}_k^T \mathbf{x}) = 4 \nabla_{\mathbf{x}} \frac{1}{1 + \exp(2(b_k - \mathbf{a}_k^T \mathbf{x}))}.$$

So, δ for tanh is 4 times that applicable to the logistic regression problem. Note that this analysis can further cover a wide range of neural network training problems that uses cross entropy loss and sigmoidal activation functions (e.g. MLP, CNN and RNN).

Special Case in Linear Regression

Consider the linear regression problem

$$f_i(x) = \frac{1}{2} \|A_i \mathbf{x} + \mathbf{b}_i\|^2, i = 1, \ldots, N.$$

We have

$$\nabla f_i(\mathbf{x}) = A_i^T A_i \mathbf{x} + A_i^T \mathbf{b}_i.$$

Then if the feature A_i's satisfy $A_i^T A_i = A_j^T A_j, \forall\, i \neq j$, we have

$$\delta = \max_{i,j} \left| A_i^T \mathbf{b}_i - A_j^T \mathbf{b}_j \right|.$$

B   Proof of Claim 2.1

The proof is related to techniques developed in classical and recent works that characterize lower bounds for first-order methods in centralized [20, 21] and decentralized [22, 23] settings. Technically, our computational / communication model is *different* compared to the aforementioned works, since we allow arbitrary number of local processing iterations, and we have a central aggregator. The difference here is that our goal is *not* to show the lower bounds on the number of total (centralized) gradient access, nor to show the optimal graph dependency. The main point we would like to make is that there exist constructions of *local* functions f_i's such that *no matter* how many times that local first-order processing is performed, without *communication* and *aggregation*, no significant progress can be made in reducing the stationarity gap of the original problem.

For notational simplicity, we will assume that the full local gradients $\{\nabla f_i(x_i^k)\}$ can be evaluated. Later we will comment on how to extend this result to enable access to the sample gradients $\nabla F(x_i^k; \xi_i)$. In particular, we consider the following slightly simplified model for now:

$$x^t = V^t(\{x_i^{t-1,Q}\}_{i=1}^N), \ x_i^{t,0} = x^t, \ \forall\, i \in [N], \tag{20a}$$

$$x_i^{t,q} \in W_i^t \left(\left\{ x_i^{r,k}, \left\{ \nabla f_i(x_i^{r,k}) \right\} \right\}_{r=0:t}^{k=0:q-1} \right), q \in [Q], \ \forall\, i. \tag{20b}$$

15

B.1 Notation.

In this section, we will call each t a "stage," and call each local iteration q an "iteration." We use x to denote the variable located at the server. We use x_i (and sometimes x_q) to denote the local variable at node i, and use $x_i[j]$ and $x_i[k]$ to denote its jth and kth elements, respectively. We use $g_i(\cdot)$ and $f_i(\cdot)$ to denote some functions related to node i, and $g(\cdot)$ and $f(\cdot)$ to denote the average functions of g_i's and f_i's, respectively. We use N to denote the total number of nodes.

B.2 Main Constructions.

Suppose there are N distributed nodes in the system, and they can all communicate with the server. To begin, we construct the following two non-convex functions

$$g(x) := \frac{1}{N} \sum_{i=1}^{N} g_i(x), \quad f(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x). \tag{21}$$

Here we have $x \in \mathbb{R}^{T+1}$. We assume N is constant, and T is the total number of stages (a large number and one that can potentially increase). For notational simplicity, and without loss of generality, we assume that $T \geq N$, and it is divisible by N. Let us define the component functions g_i's in (21) as follows.

$$g_i(x) = \Theta(x, 1) + \sum_{j=1}^{T/N} \Theta(x, (j-1)N + i + 1), \tag{22}$$

where we have defined the following functions

$$
\begin{aligned}
\Theta(x, j) &:= \Psi(-x[j-1])\Phi(-x[j]) - \Psi(x[j-1])\Phi(x[j]), \ \forall \ j = 2, \cdots, T+1, \\
\Theta(x, 1) &:= -\Psi(1)\Phi(x[1]).
\end{aligned} \tag{23a}
$$

Clearly, each $\Theta(x, j)$ is only related to two components in x, i.e., $x[j-1]$ and $x[j]$.

The component functions $\Psi, \Phi : \mathbb{R} \to \mathbb{R}$ are given as below

$$
\Psi(w) := \begin{cases} 0 & w \leq 0 \\ 1 - e^{-w^2} & w > 0, \end{cases}
$$

$$\Phi(w) := 4 \arctan w + 2\pi.$$

By the above definition, the average function becomes:

$$g(x) := \frac{1}{M} \sum_{j=1}^{M} g_i(x) = \Theta(x, 1) + \sum_{j=2}^{T+1} \Theta(x, j) \tag{24}$$

$$= -\Psi(1)\Phi(x[1]) + \sum_{j=2}^{T+1} \left[\Psi(-x[j-1]) \Phi(-x[j]) - \Psi(x[j-1]) \Phi(x[j]) \right].$$

See Fig. 3 for an illustration of the construction discussed above.

Further, for a given error constant $\epsilon > 0$ and a given the Lipschitz constant L, let us define

$$f_i(x) := \frac{2\pi\epsilon}{L} g_i\left(\frac{xL}{\pi\sqrt{2\epsilon}}\right). \tag{25}$$
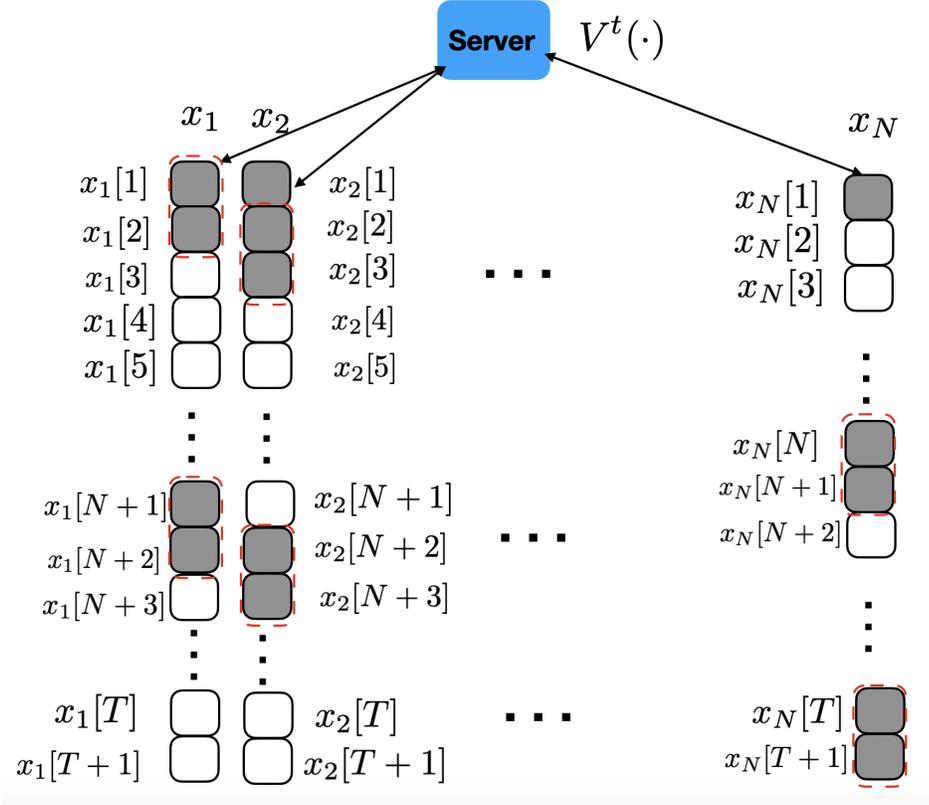
Figure 3: The example constructed for proving Claim 2.1. Here each agent has a local length $T+1$ vector x_i; each block in the figure represents one dimension of the local vector. If for agent i, its jth block is white it means that f_i is not a function of $x_i[j]$, while if jth block is shaded means f_i is a function of $x_i[j]$. Each dashed red box contains two variables that are coupled together by a function $\Theta(\cdot)$.

Therefore, we also have

$$f(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x) = \frac{2\pi\epsilon}{L} g\left(\frac{xL}{\pi\sqrt{2\epsilon}}\right). \tag{26}$$

B.3 Properties.

First we present some properties of the component functions h_i's.

Lemma 1. *The functions Ψ and Φ satisfy the following:*

1. *For all $w \leq 0$, $\Psi(w) = 0$, $\Psi'(w) = 0$.*

2. *The following bounds hold for the functions and their first- and second-order derivatives:*

$$0 \leq \Psi(w) < 1, \quad 0 \leq \Psi'(w) \leq \sqrt{\frac{2}{e}},$$

$$-\frac{4}{e^{\frac{3}{2}}} \leq \Psi''(w) \leq 2, \quad \forall w > 0.$$

17

$$0 < \Phi(w) < 4\pi, \quad 0 < \Phi'(w) \le 4,$$

$$-\frac{3\sqrt{3}}{2} \le \Phi''(w) \le \frac{3\sqrt{3}}{2}, \quad \forall w \in \mathbb{R}.$$

3. The following key property holds:

$$\Psi(w)\Phi'(v) > 1, \quad \forall \, w \ge 1, \, |v| < 1. \tag{27}$$

4. The function h is lower bounded as follows:

$$g_i(0) - \inf_x g_i(x) \le 5\pi T/N,$$

$$g(0) - \inf_x g(x) \le 5\pi T/N.$$

5. The first-order derivative of g (respectively, g_i) is Lipschitz continuous with constant $\ell = 27\pi$ (respectively, $\ell_i = 27\pi, \, \forall \, i$).

Proof. Property 1) is easy to check.

To prove Property 2), note that following holds for $w > 0$:

$$\Psi(w) = 1 - e^{-w^2}, \quad \Psi'(w) = 2e^{-w^2}w, \quad \Psi''(w) = 2e^{-w^2} - 4e^{-w^2}w^2, \, \forall \, w > 0. \tag{28}$$

Obviously, $\Psi(w)$ is an increasing function over $w > 0$, therefore the lower and upper bounds are $\Psi(0) = 0, \Psi(\infty) = 1$; $\Psi'(w)$ is increasing on $[0, \frac{1}{\sqrt{2}}]$ and decreasing on $[\frac{1}{\sqrt{2}}, \infty]$, where $\Psi''(\frac{1}{\sqrt{2}}) = 0$, therefore the lower and upper bounds are $\Psi'(0) = \Psi'(\infty) = 0, \Psi'(\frac{1}{\sqrt{2}}) = \sqrt{\frac{2}{e}}$; $\Psi''(w)$ is decreasing on $(0, \sqrt{\frac{3}{2}}]$ and increasing on $[\sqrt{\frac{3}{2}}, \infty)$ (this can be verified by checking the signs of $\Psi'''(w) = 4e^{-w^2}w(2w^2 - 3)$ in these intervals). Therefore the lower and upper bounds are $\Psi''(\sqrt{\frac{3}{2}}) = -\frac{4}{e^{\frac{3}{2}}}, \Psi''(0^+) = 2$, i.e.,

$$0 \le \Psi(w) < 1, \quad 0 \le \Psi'(w) \le \sqrt{\frac{2}{e}}, \quad -\frac{4}{e^{\frac{3}{2}}} \le \Psi''(w) \le 2, \quad \forall w > 0.$$

Further, for all $w \in \mathbb{R}$, the following holds:

$$\Phi(w) = 4\arctan w + 2\pi, \quad \Phi'(w) = \frac{4}{w^2 + 1}, \quad \Phi''(w) = -\frac{8w}{(w^2 + 1)^2}. \tag{29}$$

Similarly, as above, we can obtain the following bounds:

$$0 < \Phi(w) < 4\pi, \quad 0 < \Phi'(w) \le 4, \quad -\frac{3\sqrt{3}}{2} \le \Phi''(w) \le \frac{3\sqrt{3}}{2}, \quad \forall w \in \mathbb{R}.$$

To show Property 3), note that for all $w \ge 1$ and $|v| < 1$,

$$\Psi(w)\Phi'(v) > \Psi(1)\Phi'(1) = 2(1 - e^{-1}) > 1$$

where the first inequality is true because $\Psi(w)$ is strictly increasing and $\Phi'(v)$ is strictly decreasing for all $w > 0$ and $v > 0$, and that $\Phi'(v) = \Phi'(|v|)$.

Next we show Property 4). Note that $0 \leq \Psi(w) < 1$ and $0 < \Phi(w) < 4\pi$. Therefore we have $g(0) = -\Psi(1)\Phi(0) < 0$ and using the construction in (22)

$$\inf_x g_i(x) \geq -\Psi(1)\Phi(x[1]) - \sum_{j=1}^{T/N} \sup_{w,v} \Psi(w)\Phi(v) \tag{30}$$

$$\geq -4\pi - 4(T/N)\pi \geq -5\pi T/N, \tag{31}$$

where the first inequality follows from $\Psi(w)\Phi(v) > 0$, the second follows from $\Psi(w)\Phi(v) < 4\pi$, and the last is true because $T/N \geq 1$.

Finally, we show Property 5), using the fact that a function is Lipschitz if it is piecewise smooth with bounded derivative. Before proceeding, let us note a few properties of the construction in (24) (also see Fig. 3). First, for a given node q, its local function h_q is only related to the following $x[j]$'s

$$j = 1 + q + \ell \times N \geq 1, \; \ell = 0, \cdots, (N-1),$$
$$j = q + \ell \times N \geq 1, \; \ell = 0, \cdots, (N-1),$$

or equivalently

$$q = j - 1 - \ell \times N \geq 1, \; \ell = 0, \cdots, (N-1),$$
$$q = j - \ell \times N \geq 1, \; \ell = 0, \cdots, (N-1).$$

Then the first-order partial derivative of $g_q(y)$ can be expressed below.
Case I) If $j \neq 1$ we have

$$\frac{\partial g_q}{\partial x[j]} = \begin{cases} (-\Psi(-x[j-1])\Phi'(-x[j]) - \Psi(x[j-1])\Phi'(x[j])), \\ \qquad q = j-1-N(\ell) \geq 1, \; \ell = 0, \cdots, \frac{T}{N}-1, j = 2,3,\cdots,T+1 \\ (-\Psi'(-x[j])\Phi(-x[j+1]) - \Psi'(x[j])\Phi(x[j+1])), \\ \qquad q = j-N(\ell) \geq 1, \; \ell = 0, \cdots, \frac{T}{N}-1, j = 3,4,\cdots T \\ 0 \\ \qquad \text{otherwise.} \end{cases} \tag{32}$$

Case II) If $j = 1$, we have

$$\frac{\partial g_q}{\partial x[1]} = \begin{cases} -\Psi(1)\Phi'(x[1]) + (-\Psi'(-x[1])\Phi(-x[2]) - \Psi'(x[1])\Phi(x[2])), & q = 1 \\ -\Psi(1)\Phi'(x[1]), & q \neq 1 \end{cases} \tag{33}$$

From the above derivation, it is clear that for any j, q, $\frac{\partial g_q}{\partial x[j]}$ is either zero or is a piecewise smooth function separated at the non-differentiable point $x[j] = 0$, because the function $\Psi'(\cdot)$ is not differentiable at 0.

Further, fix a point $x \in \mathbb{R}^{T+1}$ and a unit vector $v \in \mathbb{R}^{T+1}$ where $\sum_{j=1}^{T+1} v[j]^2 = 1$. Define

$$\ell_q(\theta; x, v) := g_q(x + \theta v)$$

to be the directional projection of g_q on to the direction v at point x. We will show that there exists $C > 0$ such that $|\ell_q''(0; x, v)| \leq C$ for all $x \neq 0$ (where the second-order derivative is taken with respect to θ).

First, by noting the fact that each if $x[j]$ appears in $g_q(x)$, then it must also be *coupled with* either $x[j+1]$ or $x[j-1]$, but not other $x[k]$'s for $k \neq j-1, j+1$. This means that $\frac{\partial^2 g_q(x)}{\partial x[j_1]\partial x[j_2]} = 0, \forall j_2 \neq \{j_1, j_1+1, j_1-1\}$. Using this fact, we can compute $\ell_q''(0; x, v)$ as follows:

$$\ell_q''(0; x, v) = \sum_{j_1, j_2=1}^{T} \frac{\partial^2 g_q(x)}{\partial x[j_1]\partial x[j_2]} v[j_1]v[i_2]$$

$$= \sum_{\delta \in \{0,1,-1\}} \sum_{j=1}^{T} \frac{\partial^2 g_q(x)}{\partial x[j]\partial x[j+\delta]} v[j]v[j+\delta],$$

where we take $v[0] := 0$ and $v[T+1] := 0$.

By using (32) and the above facts, the second-order partial derivative of $g_q(x)$ ($\forall x \neq 0$) is given as follows when $j \neq 1$:

$$\frac{\partial^2 g_q}{\partial x[j]\partial x[j]} = \begin{cases} (\Psi(-x[j-1])\Phi''(-x[j]) - \Psi(x[j-1])\Phi''(x[j])), \\ \qquad\qquad q = j-1-N(\ell) \geq 1, \; \ell = 0, \cdots, \frac{T}{N}-1, j = 2, 3, \cdots, T+1 \\ (\Psi''(-x[j])\Phi(-x[j+1]) - \Psi''(x[j])\Phi(x[j+1])), \\ \qquad\qquad q = j-N(\ell) \geq 1, \; \ell = 0, \cdots, \frac{T}{N}-1, j = 3, 4, \cdots, T \\ 0, \\ \qquad\qquad \text{otherwise} \end{cases}$$

(34)

$$\frac{\partial^2 g_q}{\partial x[j]\partial x[j+1]} = \begin{cases} (\Psi'(-x[j])\Phi'(-x[j+1]) - \Psi'(x[j])\Phi'(x[j+1])), \\ \qquad\qquad q = j-N(\ell) \geq 1, \; \ell = 0, \cdots, \frac{T}{N}-1, j = 3, 4, \cdots, T \\ 0, \\ \qquad\qquad \text{otherwise} \end{cases}$$

(35)

$$\frac{\partial^2 g_q}{\partial x[j]\partial x[j-1]} = \begin{cases} (\Psi'(-x[j-1])\Phi'(-x[j]) - \Psi'(x[j-1])\Phi'(x[j])), \\ \qquad\qquad q = j-N(\ell) \geq 1, \; \ell = 0, \cdots, \frac{T}{N}-1, j = 2, 3, \cdots, T+1 \\ 0, \\ \qquad\qquad \text{otherwise} \end{cases}$$

(36)

By applying Lemma 1 − i) [i.e., $\Psi(w) = \Psi'(w) = \Psi''(w) = 0$ for $\forall\, w \leq 0$], we can obtain that at least one of the terms $\Psi(-x[j-1])\Phi''(-x[j])$ or $-\Psi(x[j-1])\Phi''(x[j])$ is zero. It follows that

$$\Psi(-x[j-1])\Phi''(-x[j]) - \Psi(x[j-1])\Phi''(x[j]) \leq \sup_w |\Psi(w)| \sup_v |\Phi''(v)|.$$

Taking the maximum over equations (34) to (36) and plug in the above inequalities, we obtain

$$\left|\frac{\partial^2 g_q}{\partial x[j_1]\partial x[j_2]}\right| \leq \max\{\sup_w |\Psi''(w)| \sup_v |\Phi(v)|, \sup_w |\Psi(w)| \sup_v |\Phi''(v)|, \sup_w |\Psi'(w)| \sup_v |\Phi'(v)|\}$$

$$= \max\left\{8\pi, \frac{3\sqrt{3}}{2}, 4\sqrt{\frac{2}{e}}\right\} < 8\pi, \quad \forall\, j_1 \neq 1,$$

20

where the equality comes from Lemma 1 – ii).

When $j = 1$, by using (33), we have the following:

$$\frac{\partial^2 g_q(x)}{\partial x[1]\partial x[1]} = \begin{cases} -\Psi(1)\Phi''(x[1]) + (-\Psi''(-x[1])\Phi(-x[2]) - \Psi''(x[1])\Phi(x[2])), & q = 1 \\ -\Psi(1)\Phi''(x[1]), & \text{otherwise} \end{cases},$$

$$\frac{\partial^2 g_q(x)}{\partial x[1]\partial x[2]} = \begin{cases} (-\Psi'(-x[1])\Phi'(-x[2]) - \Psi'(x[1])\Phi'(x[2])), & q = 1 \\ 0, & \text{otherwise} \end{cases}.$$

Again by applying Lemma 1 – i) and ii),

$$\left|\frac{\partial^2 g_q(x)}{\partial x[1]\partial x[j_2]}\right| \leq \max\{\sup_w |\Psi(1)\Phi''(w)| + \sup_w |\Psi''(w)|\sup_v |\Phi(v)|, \sup_w |\Psi'(w)|\sup_v |\Phi'(v)|\}$$

$$= \max\left\{\frac{3\sqrt{3}}{2}(1 - e^{-1}) + 8\pi, 4\sqrt{\frac{2}{e}}\right\} < 9\pi, \ \forall \ j_2.$$

Summarizing the above results, we obtain:

$$|\ell_q''(0; x, v)| = |\sum_{\delta \in \{0,1,-1\}} \sum_{j=1}^{T} \frac{\partial^2 g_q(y)}{\partial x[j]\partial x[j+\delta]} v[j]v[j+\delta]|$$

$$\leq 9\pi \sum_{\delta \in \{0,1,-1\}} |\sum_{j=1}^{T} v[j]v[j+\delta]|$$

$$\leq 9\pi \left(|\sum_{j=1}^{T} v[j]^2| + 2|\sum_{j=1}^{T} v[j]v[j+1]|\right)$$

$$\leq 27\pi \sum_{j=1}^{T} |v[j]^2| = 27\pi.$$

Overall, the first-order derivatives of h_q are Lipsschitz continuous for any q with constant at most $\ell = 27\pi$.
∎

The following lemma is a simple extension of the previous result.

Lemma 2. *We have the following properties for the functions f defined in (26) and (25):*

1. *We have $\forall \ x \in \mathbb{R}^{T+1}$*

$$f(0) - \inf_x f(x) \leq \frac{10\pi^2\epsilon}{LN}T.$$

2. *We have*

$$\|\nabla f(x)\| = \sqrt{2\epsilon}\left\|\nabla g\left(\frac{xL}{\pi\sqrt{2\epsilon}}\right)\right\|, \ \forall \ x \in \mathbb{R}^{T+1}. \tag{37}$$

3. *The first-order derivatives of f and that for each $f_i, i \in [N]$ are Lipschitz continuous, with the same constant $U > 0$.*

21

Proof. To show that property 1) is true, note that we have the following:

$$f(0) - \inf_x f(x) = \frac{2\pi\epsilon}{L}\left(g(0) - \inf_x g(x)\right).$$

Then by applying Lemma 1 we have that for any $T \geq 1$, the following holds

$$f(0) - \inf_x f(x) \leq \frac{2\pi\epsilon}{L} \times \frac{5\pi T}{N}.$$

Property 2) is true is due to the definition of f_i, so that we have:

$$\nabla f_i(x) = \sqrt{2\epsilon} \times \nabla g_i\left(\frac{xL}{\pi\sqrt{2\epsilon}}\right).$$

Property 3) is true because the following:

$$\|\nabla f(z) - \nabla f(y)\| = \sqrt{2\epsilon}\left\|\nabla g\left(\frac{zU}{\pi\sqrt{2\epsilon}}\right) - \nabla g\left(\frac{yU}{\pi\sqrt{2\epsilon}}\right)\right\| \leq U\|z - y\|$$

where the last inequality comes from Lemma 1 – (5). This completes the proof. ∎

Next let us analyze the size of ∇g. We have the following result.

Lemma 3. *If there exists $k \in [T]$ such that $|x[k]| < 1$, then*

$$\|\nabla g(x)\| = \left\|\frac{1}{N}\sum_{i=1}^{N}\nabla g_i(x)\right\| \geq \left|\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g_i(x)}{\partial x[k]}\right| > 1/N.$$

Proof. The first inequality holds for all $k \in [T]$, since $\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial y[k]}g_i(x)$ is one element of $\frac{1}{N}\sum_{i=1}^{N}\nabla g_i(x)$.
We divide the proof for the second inequality into two cases.
Case 1. Suppose $|x[j-1]| < 1$ for all $2 \leq j \leq k$. Therefore, we have $|x[1]| < 1$. Using (33), we have the following inequalities:

$$\frac{\partial g_i(x)}{\partial x[1]} \overset{(i)}{\leq} -\Psi(1)\Phi'(x[1]) \overset{(ii)}{<} -1, \forall i \tag{38}$$

where (i) is true because $\Psi'(w), \Phi(w)$ are all non-negative from Lemma 1 -(2); (ii) is true due to Lemma $1 - (3)$. Therefore, we have the following

$$\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla g_i(x)\right\| \geq \left|\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial x[1]}g_i(x)\right| > 1.$$

Case 2) Suppose there exists $2 \leq j \leq k$ such that $|x[j-1]| \geq 1$.
We choose j so that $|x[j-1]| \geq 1$ and $|x[j]| < 1$. Therefore, depending on the choices of (i, j) we have three cases:

$$\frac{\partial g_i(x)}{\partial x[j]} = \begin{cases} \left(-\Psi\left(-x[j-1]\right)\Phi'\left(-x[j]\right) - \Psi\left(x[j-1]\right)\Phi'\left(x[j]\right)\right), \\ \qquad i = j-1-N(\ell) \geq 1, \ \ell = 0, \cdots, \frac{T}{N}-1, j = 2, 3, \cdots, T+1 \\ \left(-\Psi'\left(-x[j]\right)\Phi\left(-x[j+1]\right) - \Psi'\left(x[j]\right)\Phi\left(x[j+1]\right)\right), \\ \qquad i = j-1-N(\ell) \geq 1, \ \ell = 0, \cdots, \frac{T}{N}-1, j = 3, 4, \cdots, T \\ 0 \\ \qquad \text{otherwise} \end{cases} \tag{39}$$

First, note that $\frac{\partial g_i(x)}{\partial x[j]} \leq 0$, for all i, j, by checking the definitions of $\Psi(\cdot), \Phi'(\cdot), \Psi'(\cdot), \Phi(\cdot)$.

Then for (i, j) satisfying the first condition, because $|x[j-1]| \geq 1$ and $|x[j]| < 1$, using Lemma 1 – (3), and the fact that the negative part is zero for Ψ, and Φ' is even function, the expression further simplifies to:

$$-\Psi(|x[j-1]|)\Phi'\left(|x[j]|\right) \overset{(27)}{<} -1. \tag{40}$$

If the second condition holds true, the expression is obviously non-positive because both Ψ' and Φ are non-negative. Overall, we have

$$\left| \frac{1}{N} \sum_{i=1}^{N} \frac{\partial g_i(x)}{\partial x[j]} \right| > \frac{1}{N}.$$

This completes the proof. ∎

Lemma 4. *Consider using an algorithm of the form* (20) *to solve the following problem:*

$$\min_{x \in \mathbb{R}^{T+1}} \ g(x) = \frac{1}{N} \sum_{i=1}^{N} g_i(x). \tag{41}$$

Assume the initial solution: $x_i = 0, \ \forall \ i \in [N]$. *Let* $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} \alpha_i x_i$ *denote some linear combination of local variables, where* $\{\alpha_i > 0\}$ *are the coefficients (possibly time-varying and dependent on t). Then no matter how many local computation steps* (20b) *are performed, at least* T *communication steps* (20a) *are needed to ensure* $\bar{x}[T] \neq 0$.

Proof. For a given $j \geq 2$, suppose that $x_i[j], x_i[j+1], ..., x_i[T] = 0, \ \forall i$, that is, support$\{x_i\} \subseteq \{1, 2, 3, ..., j-1\}$ for all i. Then $\Psi'(x_i[j]) = \Psi'(-x_i[j]) = 0$ for all i, and g_i has the following partial derivative (see (32))

$$\frac{\partial g_i(x_i)}{\partial x_i[j]} = -\left(\Psi\left(-x_i[j-1]\right)\Phi'\left(-x_i[j]\right)\right) + \left(\Psi\left(x_i[j-1]\right)\Phi'\left(x_i[j]\right)\right), \tag{42}$$

$$i = j - 1 - N(\ell) \geq 1, \ \ell = 0, \cdots, \frac{T}{N} - 1, j = 2, 3, \cdots, T+1. \tag{43}$$

Clearly, if $x_i[j-1] = 0$, then by the definition of $\Psi(\cdot)$, the above partial gradient is also zero. In other words, the above partial gradient is only non-zero if $x_i[j-1] \neq 0$.

Recall that we have assumed that the server aggregation is performed using a linear combination $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} \alpha_i x_i$, with the coefficients α_i's possibly depending on the stage t (but such a dependency will be irrelevant for our purpose, as will be see shortly). Therefore, at a given stage t, for a given node i, when $j \geq 3$, its jth element will become *nonzero* only if one of the following two cases hold true:

- If before the aggregation step (i.e., at stage $t-1$), some other node q has $x_q[j]$ being nonzero.

- If $\frac{\partial g_i(x_i)}{\partial x_i[j]}$ is nonzero at stage t.

Now suppose that the initial solution is $x_i[j] = 0$ for all (i, j). Then at the first iteration only $\frac{\partial g_i(x_i)}{\partial x_i[1]}$ is non-zero for all i, due to the fact that $\frac{\partial g_i(x_i)}{\partial x_i[1]} = \Psi(1)\Phi'(0) = 4(1 - e^{-1})$ for all i from (33). It is also important to observe that, if all nodes $i \neq 1$ were to perform subsequent local updates (20b), the local

variable x_j will have the same support (i.e., only the first element is non-zero). To see this, suppose $k = 2$, then for $i = 2$, we have

$$\frac{\partial g_i(x_i)}{\partial x_i[2]} = \left(-\Psi'\left(-x[2]\right)\Phi\left(-x[3]\right) - \Psi'\left(x[2]\right)\Phi\left(x[3]\right)\right) = 0, \tag{44}$$

since $x[2] = 0$ implies $\Psi'\left(-x[2]\right) = 0$. Similarly reasoning applies when $i = 2$, $k \geq 3$.

If $i \geq 3$, then these local functions are not related to $x_i[2]$, so the partial derivative is also zero.

Now let us look at node $i = 1$. For this node, according to (42), we have

$$\frac{\partial g_1(x_1)}{\partial x_1[2]} = -\left(\Psi\left(-x_1[1]\right)\Phi'\left(-x_1[2]\right)\right) + \left(\Psi\left(x_1[1]\right)\Phi'\left(x_1[2]\right)\right). \tag{45}$$

Since $x_1[1]$ can be non-zero, then this partial gradient can also be non-zero. Further, with a similar argument as above, we can also confirm that no matter how many local computation steps that node 1 performs, only the first two elements of x_1 can be non-zero.

So for the first stage $t = 1$, we conclude that, no matter how many local computation that the nodes perform (in the form of the computation step given in (20b)), only x_1 can have two non-zero entries, while the rest of the local variables only have one non-zero entries.

Then suppose that the communication and aggregation step is performed once. It follows that after broadcasting $\bar{x} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i x_i$ to all the nodes, everyone can have two non-zero entries. Then the nodes proceed with local computation, and by the same argument as above, one can show that this time only x_2 can have three non-zero entries. Following the above procedure, it is clear that each aggregation step can advance the non-zero entry of $\bar{x}$ by one, while performing multiple local updates does not advance the non-zero entry. Then we conclude that we need at least T communication steps, and local gradient computation steps, to make $x_i[T]$ possibly non-zero. ∎

B.4   Main Result for Claim 2.1.

Below we state and prove a formal version of Claim 2.1 given in the main text.

Theorem 2. *Let ϵ be a positive number. Let $x_i^0[j] = 0$ for all $i \in [N]$, and all $j = 1, \cdots, T+1$. Consider any algorithm obeying the rules given in (10), where the $V^t(\cdot)$ and $W_i^t(\cdot)$'s are linear operators. Then regardless of the number of local updates there exists a problem satisfying Assumption 1 – 2, such that it requires at least the following number of stages t (and equivalently, aggregation and communications rounds in (20a))*

$$t \geq \frac{(f(0) - \inf_x f(x))LN}{10\pi^2}\epsilon^{-1} \tag{46}$$

to achieve the following error

$$h_t^* = \left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x^t)\right\|^2 < \epsilon. \tag{47}$$

Proof of Claim 2.1. First, let us show that the algorithm obeying the rules given in (20) has the desired property. Note that the difference between two rules is whether the *sampled* local gradients are used for the update, or the full local gradients are used.

By Lemma 4 we have $\bar{x}[T] = 0$ for all $t < T$. Then by applying Lemma 2 – (2) and Lemma 3, we can conclude that the following holds

$$\|\nabla f(\bar{x}[T])\| = \sqrt{2\epsilon} \left\| \nabla h \left(\frac{\bar{x}[T]U}{\pi\sqrt{2\epsilon}} \right) \right\| > \sqrt{2\epsilon}/N, \tag{48}$$

where the second inequality follows that there exists $k \in [T]$ such that $|\frac{\bar{x}[k]U}{\pi\sqrt{2\epsilon}}| = 0 < 1$, then we can directly apply Lemma 3.

The third part of Lemma 2 ensures that f_i's are L-Lipschitz continuous gradient, and the first part shows

$$f(0) - \inf_x f(x) \leq \frac{10\pi^2\epsilon}{LN}T,$$

Therefore we obtain

$$T \geq \frac{(f(0) - \inf_x f(x)) LN}{10\pi^2}\epsilon^{-1}. \tag{49}$$

This completes the proof.

Second, consider the algorithm obeying the rules give in (10), in which local *sampled* gradients are used. By careful inspection, the result for this case can be trivially extended from the previous case. We only need to consider the following local functions

$$\hat{f}_i(x) = \sum_{\xi_i \in D_i} F(x; \xi_i) \tag{50}$$

where each sampled loss function $F(x; \xi_i)$ is defined as

$$F(\mathbf{x}; \xi_i) = \delta(\xi_i)f_i(x), \quad \text{where } f_i(x) \text{ is defined in (25)}. \tag{51}$$

where $\delta(\xi_i)$'s satisfy $\delta(\xi_i) > 0$ and $\sum_{\xi_i \in D_i} \delta(\xi_i) = 1$. It is easy to see that, the local sampled gradients have the same dependency on x as their averaged version (by dependency we meant the structure that is depicted in Fig. 3). Therefore, the progression of the non-zero pattern of the average $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$ is exactly the same as the batch gradient version. Additionally, since the local function $\hat{f}(x)$ is exactly the same as the previous local function $f(x)$, so other estimates, such as the one that bounds $f(0) - \inf f(x)$, also remain the same.

C   Proof of Claim 2.2

First let us consider FedAvg with local-GD update (9). We consider the following problem with $N = 2$, which satisfies both Assumptions 1 and 2, with $f(\mathbf{x}) = 0, \forall \mathbf{x}$

$$f_1(\mathbf{x}) = \frac{1}{2}\mathbf{x}^2, \quad f_2(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^2. \tag{52}$$

Each local iteration of the FedAvg is given by

$$\mathbf{x}_1^{r+1} = (1 - \eta^{r+1})\mathbf{x}_1^r, \quad \mathbf{x}_2^{r+1} = (1 + \eta^{r+1})\mathbf{x}_2^r. \tag{53}$$

For simplicity, let us define $\mathbf{y} = [\mathbf{x}_1, \mathbf{x}_2]^T$, and define the matrix $\mathbf{D} = [1 - \eta, 0; 0, 1 + \eta]$. Then running Q rounds of the FedAvg algorithm starting with $r = kQ$ for some non-negative integer $k \geq 0$, can be expressed as

$$\mathbf{y}^{(k+1)Q} = \mathbf{D}^{Q-1}\mathbf{y}^{kQ+1}, \quad \mathbf{y}^{kQ+1} = \frac{1}{2}\mathbf{1}\mathbf{1}^T\mathbf{D}\mathbf{y}^{kQ}. \tag{54}$$

Therefore overall we have

$$\mathbf{y}^{(k+1)Q} = \frac{1}{2}\mathbf{D}^{Q-1}\mathbf{1}\mathbf{1}^T\mathbf{D}\mathbf{y}^{kQ}. \tag{55}$$

It is easy to show that for any $Q > 1$, the eigenvalues of the matrix $\frac{1}{2}\mathbf{D}^{Q-1}\mathbf{1}\mathbf{1}^T\mathbf{D}$ are 0 and $\frac{(1+\eta)^Q + (1-\eta)^Q}{2} > 1$.

It follows that the above iteration will diverge for any $Q > 1$ starting from any non-zero initial point.

Moreover, when the sample on one agent are the same (e.g., agent 1 has two samples that both has loss function x^2), then using SGD as local update will be identical to the update of GD.

D   Results showing the role of GB for FedAvg with diminishing stepsizes

This section is used to show the role of A3 in FedAvg. First we prove that with A3, FedAvg can use arbitrary stepsize as long as it diminishes to zero. Next, we show by a simple example that without A3, FedAvg will diverge with the same diminishing stepsize choice.

Claim D.1. *Suppose A1–A3 hold and the stepsizes satisfy: 1) $\eta^{r,0} = \eta \in (0, 1/L)$ for all r; 2) set $0 < \eta^{r,q} \leq \min\{\frac{1}{2(Q-1)L}, \frac{\eta}{Q}\}$, $\lim_{r\to\infty} \eta^{r,q} = 0$, $q \neq 0$. Then the average gradient converges to zero for FedAvg with local-GD update (9):*

$$\frac{1}{T}\sum_{r=0}^{T}\|\nabla f(\mathbf{x}^r)\|^2 \leq \frac{2(f(\mathbf{x}^0) - f(\mathbf{x}^\star))}{C_1 T} + \frac{2QG^2\eta^2}{C_1 T}\sum_{r=0}^{T}\sum_{q=1}^{Q-1}\eta^{r,q}, \quad \text{for some } C_1 := \eta(1 - L\eta).$$

Claim D.2. *Suppose that all the assumptions made in Claim D.1 hold, except that A3 does not hold. Then FedAvg with local-GD can diverge for any $Q > 1$.*

Before we prove Claim D.1, the following lemma is needed.

Lemma 5. *Under A1 and A3, following the update steps in Algorithm 1, between each outer iterations we have:*

$$f(\mathbf{x}^{r+1}) - f(\mathbf{x}^r) \leq - (\eta^{r,0}(1 - L\eta^{r,0}) + \sum_{q=1}^{Q-1}\frac{\eta^{r,q}}{2})\|\nabla f(\mathbf{x}^r)\|^2$$

$$- \sum_{q=1}^{Q-1}(\frac{\eta^{r,q}}{2} - 2L(Q-1)(\eta^{r,q})^2)\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q})\right\|^2 \tag{56}$$

$$+ \frac{QG^2}{2}((\eta^{r,0})^2 + \sum_{q=1}^{Q-1}(\eta^{r,q})^2)\sum_{q=1}^{Q-1}\eta^{r,q},$$

where $r_0 + 1 \mod Q = 0$.

Proof: By using A1 we have:

$$f(\mathbf{x}^{r+1}) - f(\mathbf{x}^r)$$

$$\leq \left\langle \nabla f(\mathbf{x}^r), \mathbf{x}^{r+1} - \mathbf{x}^r \right\rangle + \frac{L}{2} \left\| \mathbf{x}^{r+1} - \mathbf{x}^r \right\|^2$$

$$\stackrel{(a)}{=} - \left\langle \nabla f(\mathbf{x}^r), \frac{1}{N} \sum_{i=1}^{N} \sum_{q=0}^{Q-1} \eta^{r,q} \nabla f_i(\mathbf{x}_i^{r,q}) \right\rangle + \frac{L}{2} \left\| \frac{1}{N} \sum_{i=1}^{N} \sum_{q=0}^{Q-1} \eta^{r,q} \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2$$

$$\stackrel{(b)}{\leq} - \sum_{q=1}^{Q-1} \eta^{r,q} \left\langle \nabla f(\mathbf{x}^r), \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{r,q}) \right\rangle + L(\eta^{r,0})^2 \left\| \nabla f(\mathbf{x}^r) \right\|^2 \tag{57}$$

$$+ (Q-1)L \sum_{q=1}^{Q-1} (\eta^{r,q})^2 \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2$$

$$\stackrel{(c)}{=} -\eta^{r,0} \left\| \nabla f(\mathbf{x}^r) \right\|^2 - \sum_{q=1}^{Q-1} \eta^{r,q} \left\langle \nabla f(\mathbf{x}^r), \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{r,q}) \right\rangle$$

$$+ L(\eta^{r,0})^2 \left\| \nabla f(\mathbf{x}^r) \right\|^2 + (Q-1)L \sum_{q=1}^{Q-1} (\eta^{r,q})^2 \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2,$$

where (a) comes from the update rule in Algorithm 1, in (b) we use Jensen's inequality and notice $\mathbf{x}_i^{r,0} = \mathbf{x}^r$ so in (c) we extract the terms with index $(r, 0)$ from the inner product.

Note that for any vector a, b of the same length, the equality $2 \langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$, holds, we have

$$- \eta^{r,q} \left\langle \nabla f(\mathbf{x}^r), \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{r,q}) \right\rangle + (Q-1)L(\eta^{r,q})^2 \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2$$

$$= -\frac{\eta^{r,q}}{2} \left\| \nabla f(\mathbf{x}^r) \right\|^2 - \frac{\eta^{r,q}}{2} \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2 + \frac{\eta^{r,q}}{2} \left\| \nabla f(\mathbf{x}^r) - \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2$$

$$+ (Q-1)L(\eta^{r,q})^2 \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2 \tag{58}$$

$$\stackrel{(a)}{\leq} -\frac{\eta^{r,q}}{2} \left\| \nabla f(\mathbf{x}^r) \right\|^2 + \frac{\eta^{r,q}}{2N} \sum_{i=1}^{N} \left\| \nabla f_i(\mathbf{x}^r) - \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2$$

$$- \frac{\eta^{r,q}}{2} ((1 - 2(Q-1)L\eta^{r,q})) \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2$$

$$\stackrel{(b)}{\leq} -\frac{\eta^{r,q}}{2} \left\| \nabla f(\mathbf{x}^r) \right\|^2 + \frac{L^2 \eta^{r,q}}{2N} \sum_{i=1}^{N} \left\| \mathbf{x}^r - \mathbf{x}_i^{r,q} \right\|^2 - \frac{\eta^{r,q}}{2} ((1 - 2(Q-1)L\eta^{r,q})) \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2,$$

where we use Jensen's inequality in (a) and A1 in (b).

Further note that

$$
\begin{aligned}
\|\mathbf{x}^r - \mathbf{x}_i^{r,q}\|^2 &= \left\| \mathbf{x}^r - \mathbf{x}^r + \sum_{\tau=0}^{q-1} \eta^{r,\tau} \nabla f_i(\mathbf{x}_i^{r,\tau}) \right\|^2 \\
&= \left\| \sum_{\tau=0}^{q-1} \eta^{r,\tau} \nabla f_i(\mathbf{x}_i^{r,\tau}) \right\|^2 \\
&\overset{(a)}{\le} 2(q-1) \sum_{\tau=1}^{q-1} (\eta^{r,\tau})^2 \|\nabla f_i(\mathbf{x}_i^{r,\tau})\|^2 + 2(\eta^{r,0})^2 \left\| \nabla f_i(\mathbf{x}_i^{r,0}) \right\|^2 \\
&\overset{(b)}{\le} 2 \left((q-1) \sum_{\tau=1}^{q-1} (\eta^{r,\tau})^2 + (\eta^{r,0})^2 \right) G^2.
\end{aligned}
\tag{59}
$$

The first equality comes from the update rule of $\mathbf{x}_i^{r,q}$, which basically performs q steps of updates on $\mathbf{x}^r$; (a) comes from Jensen's inequality; in (b) we use A3.

Substitute (59) to (58) and then to (57), rearrange the terms we obtain (56), which ends the proof of the lemma. ∎

D.1   Proof of Claim D.1

Next we prove Claim D.1

Proof: By choosing $\eta^{r,0} = \eta_1 = \in (0, 1/L)$ as constant and $\eta^{r,q} \le 1/(2QL)$, $\forall\, q \ne 0$ then applying Lemma 5 we have

$$
\begin{aligned}
f(\mathbf{x}^{r+1}) - f(\mathbf{x}^r) \le & -(C_1 + \sum_{q=1}^{Q-1} \frac{\eta^{r,q}}{2}) \|\nabla f(\mathbf{x}^r)\|^2 \\
& + \frac{QG^2}{2} ((\eta_1)^2 + \sum_{q=1}^{Q-1} (\eta^{r,q})^2) \sum_{q=1}^{Q-1} \eta^{r,q},
\end{aligned}
\tag{60}
$$

where $C_1 = \eta_1(1 - L\eta_1) > 0$. Using telescope sum from $r = 0$ to $r = T - 1$ we have

$$
\begin{aligned}
f(\mathbf{x}^T) - f(\mathbf{x}^0) \le & -\sum_{r=0}^{T-1} (C_1 + \sum_{q=1}^{Q-1} \frac{\eta^{r,q}}{2}) \|\nabla f(\mathbf{x}^r)\|^2 \\
& + \frac{QG^2}{2} \sum_{r=0}^{T-1} ((\eta_1)^2 + \sum_{q=1}^{Q-1} (\eta^{r,q})^2) \sum_{q=1}^{Q-1} \eta^{r,q}.
\end{aligned}
\tag{61}
$$

Rearrange the terms and multiply both side by $2/(TC_1)$, then we have

$$
\begin{aligned}
(\frac{1}{T} + \frac{\sum_{r=0}^{T-1} \sum_{q=1}^{Q-1} \eta^{r,q}}{TC_1}) \sum_{r=0}^{T} \|\nabla f(\mathbf{x}^r)\|^2 \le & \frac{2(f(\mathbf{x}^0) - f(\mathbf{x}^\star))}{C_1 T} \\
& + \frac{QG^2}{C_1 T} \sum_{r=0}^{T-1} ((\eta_1)^2 + \sum_{q=1}^{Q-1} (\eta^{r,q})^2) \sum_{q=1}^{Q-1} \eta^{r,q}.
\end{aligned}
\tag{62}
$$

Choose $\eta^{r,q} \leq \eta_1/Q$, then $(\eta_1)^2 + \sum_{q=1}^{Q-1}(\eta^{r,q})^2 \leq 2(\eta_1)^2$. Choose $\{\eta^{r,q}\}$ as a sequence that diminishes to 0, then for all $q \neq 0$, as $T \to \infty$, $\frac{2\eta_1 Q^2 G^2}{C_1} \frac{1}{QT} \sum_{r=0}^{T-1} \sum_{q=1}^{Q-1} \eta^{r,q} \to 0$. Therefore the right hand side converges to 0, Claim D.1 is proved.

E   Proof of Claim D.2

Proof. We consider the following problem with $N = 2$, which satisfies both Assumptions 1 and 2, with $f(\mathbf{x}) = 0, \forall \mathbf{x}$

$$f_1(\mathbf{x}) = \mathbf{x}^2, \quad f_2(\mathbf{x}) = -\mathbf{x}^2. \tag{63}$$

Each local iteration of the FedAvg is given by

$$\mathbf{x}_1^{r+1} = (1 - \eta^r)\mathbf{x}_1^r, \quad \mathbf{x}_2^{r+1} = (1 + \eta^r)\mathbf{x}_2^r. \tag{64}$$

For simplicity, let us define $\mathbf{y} = [\mathbf{x}_1, \mathbf{x}_2]^T$, and define the matrix $\mathbf{D}_r = [1 - \eta^r, 0; 0, 1 + \eta^r]$. Then running Q rounds of the FedAvg algorithm starting with $r = kQ$ for some non-negative integer $k \geq 0$, can be expressed as

$$\mathbf{y}^{(k+1)Q} = \prod_{r=kQ+1}^{(k+1)Q-1} \mathbf{D}_r \mathbf{y}^{kQ+1}, \quad \mathbf{y}^{kQ+1} = \frac{1}{2}\mathbf{1}\mathbf{1}^T \mathbf{D}_{kQ}\mathbf{y}^{kQ}. \tag{65}$$

Therefore overall we have

$$\mathbf{y}^{(k+1)Q} = \frac{1}{2} \prod_{r=kQ+1}^{(k+1)Q-1} \mathbf{D}_r \mathbf{1}\mathbf{1}^T \mathbf{D}_{kQ}\mathbf{y}^{kQ}. \tag{66}$$

In particular, we pick $\eta^r = \frac{1}{\sqrt{r}}$ when $r \neq kQ + 1$ and $\eta^{kQ+1} = 1/2$. Then for $Q > 1$, it is easy to compute the eigenvalues of the matrix $\frac{1}{2}\prod_{r=kQ+1}^{(k+1)Q-1} \mathbf{D}_r \mathbf{1}\mathbf{1}^T \mathbf{D}_{kQ}$ to be:

$$\lambda_1 = 0, \ \lambda_2 = \frac{1}{4} \prod_{r=kQ+2}^{(k+1)Q-1} (1 - \frac{1}{\sqrt{r}})(1 - \frac{1}{\sqrt{kQ}}) + \frac{3}{4} \prod_{r=kQ+2}^{(k+1)Q-1} (1 + \frac{1}{\sqrt{r}})(1 + \frac{1}{\sqrt{kQ}}).$$

It is clear that λ_2 is strictly larger than one which indicates that the algorithm will diverge. ∎

F   Proofs for Results in Section 3

F.1   Proof of Theorem 1

First let us prove Theorem 1 about the FedPD algorithm with Oracle I.

Towards this end, let us first introduce some notations. First recall that when Oracle I is used, the local problem is solved such that the following holds true:

$$\left\|\nabla_{\mathbf{x}_i}\mathcal{L}(\mathbf{x}_i^{r+1}, \mathbf{x}_0^r, \lambda_i^r)\right\|^2 \leq \epsilon_1. \tag{67}$$

Note that if SGD is applied in Oracle I to solve the local problem, then this condition (67) is replaced with the following

$$\mathbb{E}[\|\nabla_{\mathbf{x}_i}\mathcal{L}(\mathbf{x}_i^{r+1},\mathbf{x}_0^r,\lambda_i^r)\|^2] \le \epsilon_1. \tag{68}$$

The difference does not significantly change the proofs and the results. So throughout the proof of Theorem 1, we use (67) as the condition.

Throughout the proof, we denote the expectation taken on the communication r^{th} iteration to the $r+1^{\text{th}}$ iteration conditioning on all the previous knowledge as $\mathbb{E}_{r+1}$.

Then we define the error between different nodes as

$$\triangle^r \triangleq [\triangle\mathbf{x}_0^r;\triangle\mathbf{x}^r], \quad \text{with } \triangle\mathbf{x}_0^r \triangleq \max_{i,j}\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,j}^r\|, \quad \triangle\mathbf{x}^r \triangleq \max_{i,j}\|\mathbf{x}_i^r - \mathbf{x}_j^r\|. \tag{69}$$

Here, $\triangle\mathbf{x}_0^r$ denotes the maximum difference of estimated center model among all the nodes and $\triangle\mathbf{x}^r$ denotes the maximum difference of local models among all nodes.

From the termination condition that generates $\mathbf{x}_i^{r+1}$ (given in (67)), we have

$$\nabla f_i(\mathbf{x}_i^{r+1}) + \lambda_i^{r+1} = \nabla f_i(\mathbf{x}_i^{r+1}) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r) = \mathbf{e}_i^{r+1}, \quad \text{where } \|\mathbf{e}_i^{r+1}\|^2 \le \epsilon_1, \tag{70}$$

where the first equality holds because of the update rule of λ_i. Furthermore, from the update step of λ_i^{r+1}, we can explicitly write down the following expression

$$\lambda_i^{r+1} = \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r) = -\nabla f_i(\mathbf{x}_i^{r+1}) + \mathbf{e}_i^{r+1}.$$

The main lemmas that we need are outlined below. Their proofs can be found in Sec. F.1.1– F.1.4. The first lemma shows the sufficient descent of the local AL function.

Lemma 6. *Suppose A1 holds true. Consider FedPD with Algorithm 4 (Oracle I) as the update rule. When the local problem is solved such that (67) is satisfied, the difference of the local augmented Lagrangian is bounded by*

$$\mathcal{L}_i(\mathbf{x}_i^{r+1},\mathbf{x}_{0,i}^{r+},\lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^r,\mathbf{x}_{0,i}^r,\lambda_i^r)$$
$$\le -\frac{1-2L\eta}{2\eta}\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\|^2 - \frac{1}{2\eta}\|\mathbf{x}_{0,i}^{r+} - \mathbf{x}_{0,i}^r\|^2 + \eta\|\lambda_i^{r+1} - \lambda_i^r\|^2 + \frac{\epsilon_1}{2L}. \tag{71}$$

Then we derive a key lemma about how the error propagates if the communication step is skipped.

Lemma 7. *Suppose A1 and A5 hold. Consider FedPD with Algorithm 4 (Oracle I) as the update rule. When the local problem is solved such that (67) is satisfied, the difference between the local models $\mathbf{x}_i^r$'s and the difference between local copies of the global models $\mathbf{x}_{0,i}^r$'s are bounded by*

$$\mathbb{E}_{r+1}\triangle^{r+1} \le \frac{1}{1-L\eta}(A\triangle^r + \eta B(\delta + 2\sqrt{\epsilon_1}). \tag{72}$$

where

$$A = \begin{bmatrix} p(1+L\eta) & L\eta(1-L\eta) \\ 1 & L\eta \end{bmatrix}$$

is a rank 1 matrix with eigenvalues $(0, L\eta + p(1+L\eta))$ and $B = [p(3+L\eta), 2]^T$.

30

We define a virtual sequence $\{\overline{\mathbf{x}}_0^r\}$ where $\overline{\mathbf{x}}_0^r \triangleq \frac{1}{N}\sum_{i=1}^N \mathbf{x}_{0,i}^r$ which is the average of the local $\mathbf{x}_{0,i}^r$ and we know that $\mathbf{x}_{0,i}^r = \mathbf{x}_0^r$ when $r \mod R = 1$, that is, when the communication and aggregation step is performed. Next, we bound the error between the local AL and the global AL evaluated at the virtual sequence.

Lemma 8. *Suppose A1 holds. Consider FedPD with Algorithm 4 (Oracle I) as the update rule. When the local problem is solved such that (67) is satisfied, the difference between local AL and the global AL is bounded as below:*

$$\frac{1}{N}\sum_{i=1}^N [\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) - cL_i(\mathbf{x}_i^{r+1}, \overline{\mathbf{x}}_0^{r+1}, \lambda_i^{r+1})] \geq -\frac{(N-1)}{2N\eta}(\triangle \mathbf{x}_0^{r+1})^2. \tag{73}$$

Lastly we bound the original objective function using the global AL.

Lemma 9. *Under A1 and A2, when the local problem is solved to ϵ_1 accuracy, the difference between the original loss and the augmented Lagrangian is bounded.*

$$f(\mathbf{x}_0^r) \leq \mathcal{L}(\mathbf{x}_0^r, \mathbf{x}_1^r, \dots, \mathbf{x}_N^r, \lambda_1^r, \dots, \lambda_N^r) - \frac{1-2L\eta}{N\eta}\sum_{i=1}^N \|\mathbf{x}_i^r - \mathbf{x}_0^r\|^2 + \frac{\epsilon_1}{2L}. \tag{74}$$

Using the previous lemmas, we can then prove Theorem 1.

F.1.1   Proof of Lemma 6

We divide the left hand side (LHS) of (71), i.e., $\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)$, into the sum of three parts:

$$\begin{aligned}
\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) &= \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r \lambda_i^r) \\
&+ \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) \\
&+ \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1}),
\end{aligned} \tag{75}$$

which correspond to the three steps in the algorithm's update steps.

We bound the first difference by first applying A1 to $-f(\cdot)$:

$$-f_i(\mathbf{x}_i^r) \leq -f_i(\mathbf{x}_i^{r+1}) + \langle -\nabla f_i(\mathbf{x}_i^{r+1}), \mathbf{x}_i^r - \mathbf{x}_i^{r+1}\rangle + \frac{L}{2}\left\|\mathbf{x}_i^r - \mathbf{x}_i^{r+1}\right\|^2,$$

and obtain the following series of inequalities:

$$
\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \leq \langle \nabla f_i(\mathbf{x}_i^{r+1}), \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \rangle + \frac{L}{2} \left\| \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\|^2
$$

$$
+ \langle \lambda_i^r, \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \rangle + \frac{1}{2\eta} \left\| \mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r \right\|^2 - \frac{1}{2\eta} \left\| \mathbf{x}_i^r - \mathbf{x}_{0,i}^r \right\|^2
$$

$$
\overset{(a)}{=} \langle \nabla f_i(\mathbf{x}_i^{r+1}) + \lambda_i^r, \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \rangle + \frac{L}{2} \left\| \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\|^2
$$

$$
+ \frac{1}{2\eta} \langle \mathbf{x}_i^{r+1} + \mathbf{x}_i^r - 2\mathbf{x}_{0,i}^r, \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \rangle
$$

$$
\overset{(b)}{=} \left\langle \nabla f_i(\mathbf{x}_i^{r+1}) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r), \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\rangle + \frac{L}{2} \left\| \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\|^2 \tag{76}
$$

$$
- \frac{1}{2\eta} \left\| \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\|^2
$$

$$
\overset{(c)}{\leq} \frac{1}{2L} \left\| \nabla f_i(\mathbf{x}_i^{r+1}) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r) \right\|^2 + \frac{L}{2} \left\| \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\|^2
$$

$$
- \frac{1 - L\eta}{2\eta} \left\| \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\|^2
$$

$$
\overset{(d)}{\leq} - \frac{1 - 2L\eta}{2\eta} \left\| \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\|^2 + \frac{\epsilon_1}{2L}.
$$

In the above equation, in (a) we use the fact that $\|a\|^2 - \|b\|^2 = \langle a + b, a - b \rangle$ when vector a, b has the same length to the last two terms; in (b) we split the last term into $2\mathbf{x}_i^{r+1} - 2\mathbf{x}_{0,i}^r$ and $-\mathbf{x}_i^{r+1} + \mathbf{x}_i^r$; in (c) we use the fact that $\langle a, b \rangle \leq \frac{L}{2} \|a\|^2 + \frac{1}{2L} \|b\|^2)$; in (d) we apply the fact that $\mathbf{x}_i^{r+1}$ is the inexact solution; see (70).

Then we bound the second difference in (75) by the following:

$$
\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) = \langle \lambda_i^{r+1} - \lambda_i^r, \mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r \rangle
$$

$$
\overset{(a)}{=} \langle \lambda_i^{r+1} - \lambda_i^r, \eta(\lambda_i^{r+1} - \lambda_i^r) \rangle \tag{77}
$$

$$
= \eta \left\| \lambda_i^{r+1} - \lambda_i^r \right\|^2,
$$

where (a) directly comes from the update rule of λ_i^{r+1}.

Further we bound the third difference in (75) by the following:

$$
\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1})
$$

$$
= \left\langle \lambda_i^{r+1}, \mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+} \right\rangle - \left\langle \lambda_i^{r+1}, \mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r \right\rangle + \frac{1}{2\eta} \left\| \mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+} \right\|^2 - \frac{1}{2\eta} \left\| \mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r \right\|^2
$$

$$
\overset{(a)}{=} \left\langle \lambda_i^{r+1}, \mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r+} \right\rangle + \frac{1}{2\eta} \left\langle 2\mathbf{x}_i^{r+1} - 2\mathbf{x}_{0,i}^{r+} + \mathbf{x}_{0,i}^{r+} - \mathbf{x}_{0,i}^r, \mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r+} \right\rangle \tag{78}
$$

$$
= \left\langle \frac{1}{\eta}(\eta\lambda_i^{r+1} + \mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+}), \mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r+} \right\rangle - \frac{1}{2\eta} \left\| \mathbf{x}_{0,i}^{r+} - \mathbf{x}_{0,i}^r \right\|^2
$$

$$
\overset{(b)}{=} - \frac{1}{2\eta} \left\| \mathbf{x}_{0,i}^{r+} - \mathbf{x}_{0,i}^r \right\|^2,
$$

where, in (a), we use the same reasoning as in (76) (a) and (b); in (b) we apply the update rule of $\mathbf{x}_{0,i}^{r+}$ in the FedPD algorithm, which implies that the first term becomes zero.

Finally we sum up (76), (77), (78) and Lemma 6 is proved.

F.1.2   Proof of Lemma 7

First we derive the relation between $\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_j^{r+1}\right\|$ for arbitrary $i \neq j$ and $\triangle^r$ by using the definition of ϵ_1 (70):

$$
\begin{aligned}
\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_j^{r+1}\right\| &\overset{(70)}{=} \left\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,j}^r - \eta(\nabla f_i(\mathbf{x}_i^{r+1}) + \lambda_i^r - \mathbf{e}_i^{r+1} - \nabla f_j(\mathbf{x}_j^{r+1}) - \lambda_j^r + \mathbf{e}_j^{r+1})\right\| \\
&\leq \left\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,j}^r\right\| + \eta \left\|\nabla f_i(\mathbf{x}_i^{r+1}) - \nabla f_j(\mathbf{x}_j^{r+1})\right\| \\
&\quad + \eta \left\|\lambda_i^r - \lambda_j^r\right\| + \eta(\left\|\mathbf{e}_i^{r+1}\right\| + \left\|\mathbf{e}_j^{r+1}\right\|) \\
&\overset{(a)}{\leq} \triangle\mathbf{x}_0^r + \eta \left\|\nabla f_i(\mathbf{x}_i^{r+1}) - \nabla f_i(\mathbf{x}_j^{r+1}) + \nabla f_i(\mathbf{x}_j^{r+1}) - \nabla f_j(\mathbf{x}_j^{r+1})\right\| \\
&\quad + \eta \left\|\lambda_i^r - \lambda_j^r\right\| + 2\eta\sqrt{\epsilon_1} \\
&\overset{(b)}{\leq} \triangle\mathbf{x}_0^r + L\eta \left\|\mathbf{x}_i^{r+1} - \mathbf{x}_j^{r+1}\right\| + \eta \left\|\nabla f_i(\mathbf{x}_j^{r+1}) - \nabla f_j(\mathbf{x}_j^{r+1})\right\| \\
&\quad + \eta \left\|\lambda_i^r - \lambda_j^r\right\| + 2\eta\sqrt{\epsilon_1} \\
&\overset{(c)}{\leq} \triangle\mathbf{x}_0^r + L\eta \left\|\mathbf{x}_i^{r+1} - \mathbf{x}_j^{r+1}\right\| + \eta\delta + \eta \left\|\lambda_i^r - \lambda_j^r\right\| + 2\eta\sqrt{\epsilon_1} \\
&\overset{(d)}{=} \frac{1}{1 - L\eta}\triangle\mathbf{x}_0^r + \frac{\eta}{1 - L\eta}\delta + \frac{\eta}{1 - L\eta} \left\|\lambda_i^r - \lambda_j^r\right\| + \frac{2\eta}{1 - L\eta}\sqrt{\epsilon_1},
\end{aligned}
\tag{79}
$$

where in (a) we plug the definition of $\triangle\mathbf{x}_0^r$ and $\mathbf{e}_i^{r+1}$; in (b) we use A1; (c) comes form A5; in (d) we move the second term to the left and divide both side by $1 - L\eta$.

Then we bound the difference $\left\|\lambda_i^r - \lambda_j^r\right\|$ by plugging in the expression of λ_i^r in (70), and note that $\lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r) = \lambda_i^{r+1}$:

$$
\begin{aligned}
\left\|\lambda_i^r - \lambda_j^r\right\| &= \left\|-\nabla f_i(\mathbf{x}_i^r) + \mathbf{e}_i^r + \nabla f_j(\mathbf{x}_j^r) - \mathbf{e}_j^r\right\| \\
&\overset{(a)}{\leq} \left\|\nabla f_i(\mathbf{x}_i^r) - \nabla f_i(\mathbf{x}_j^r)\right\| + \left\|\nabla f_i(\mathbf{x}_j^r) - \nabla f_j(\mathbf{x}_j^r)\right\| + 2\sqrt{\epsilon_1} \\
&\overset{(b)}{\leq} L \left\|\mathbf{x}_i^r - \mathbf{x}_j^r\right\| + \delta + 2\sqrt{\epsilon_1} \\
&\overset{(c)}{\leq} L\triangle\mathbf{x}^r + \delta + 2\sqrt{\epsilon_1},
\end{aligned}
\tag{80}
$$

where (a) and (b) follow the same argument in (a), (b) and (c) of (79) ; in (c) we plug in the definition of $\triangle\mathbf{x}^r$.

Next we bound the difference $\left\|\mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,j}^{r+1}\right\|$. With probability $1-p$ the aggregation step has just been done at iteration r, $\mathbf{x}_{0,i}^{r+1} = \mathbf{x}_{0,j}^{r+1}$. With probability p, they are not equal, then we take expectation with

communication probability p, and get

$$
\mathbb{E}_{r+1}\left\|\mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,j}^{r+1}\right\| = p\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_j^{r+1} + \eta(\lambda_i^{r+1} - \lambda_j^{r+1})\right\|
$$

$$
\leq p\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_j^{r+1}\right\| + p\eta\left\|\lambda_i^{r+1} - \lambda_j^{r+1}\right\| \tag{81}
$$

$$
\overset{(a)}{\leq} p(1 + L\eta)\triangle\mathbf{x}^{r+1} + p\eta(\delta + 2\sqrt{\epsilon_1}),
$$

where in (a) we plug in the definition of $\triangle\mathbf{x}^{r+1}$ and (80). As these relations hold true for arbitrary (i,j) pairs, they are also true for the maximum of $\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_j^{r+1}\right\|$ and $\left\|\mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,j}^{r+1}\right\|$.

Therefore stacking (79) and (81) and plug in (80), we have

$$
\triangle\mathbf{x}^{r+1} \leq \frac{1}{1 - L\eta}(L\eta\triangle\mathbf{x}^r + \triangle\mathbf{x}_0^r) + \frac{2\eta}{1 - L\eta}(\delta + 2\sqrt{\epsilon_1}),
$$
$$
\mathbb{E}_{r+1}\triangle\mathbf{x}_0^{r+1} \leq p\frac{1 + L\eta}{1 - L\eta}(L\eta\triangle\mathbf{x}^r + \triangle\mathbf{x}_0^r) + p\frac{\eta(3 + L\eta)}{1 - L\eta}(\delta + 2\sqrt{\epsilon_1}). \tag{82}
$$

Rewrite it into matrix form then we complete the proof of Lemma 7.

F.1.3 Proof of Lemma 8

Let us first recall that the definition of local AL is given below:

$$
\mathcal{L}_i(\mathbf{x}_i, \mathbf{x}_0, \lambda_i) \triangleq f_i(\mathbf{x}_i) + \langle\lambda_i, \mathbf{x}_i - \mathbf{x}_0\rangle + \frac{1}{2\eta}\|\mathbf{x}_i - \mathbf{x}_0\|^2.
$$

Similar to (78), we have

$$
\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \bar{\mathbf{x}}_0^{r+1}, \lambda_i^{r+1}) = \left\langle\lambda_i^{r+1}, \mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+}\right\rangle - \left\langle\lambda_i^{r+1}, \mathbf{x}_i^{r+1} - \bar{\mathbf{x}}_0^{r+1}\right\rangle
$$

$$
+ \frac{1}{2\eta}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+}\right\|^2 - \frac{1}{2\eta}\left\|\mathbf{x}_i^{r+1} - \bar{\mathbf{x}}_0^{r+1}\right\|^2
$$

$$
\overset{(a)}{=} -\frac{1}{2\eta}\left\|\mathbf{x}_{0,i}^{r+} - \bar{\mathbf{x}}_0^{r+1}\right\|^2
$$

$$
\overset{(b)}{=} -\frac{1}{2\eta}\left\|\mathbf{x}_{0,i}^{r+} - \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_{0,j}^{r+}\right\|^2 \tag{83}
$$

$$
= -\frac{1}{2\eta}\left\|\frac{1}{N}\sum_{j=1}^{N}(\mathbf{x}_{0,i}^{r+} - \mathbf{x}_{0,j}^{r+})\right\|^2
$$

$$
\overset{(c)}{\geq} -\frac{1}{2\eta N}\sum_{j\neq i}\left\|\mathbf{x}_{0,i}^{r+} - \mathbf{x}_{0,j}^{r+}\right\|^2
$$

$$
\overset{(d)}{\geq} -\frac{N-1}{2\eta N}(\triangle\mathbf{x}_0^{r+1})^2,
$$

where (a) follows the same argument in (78); in (b), we plug in the definition of $\bar{\mathbf{x}}_0^{r+1}$; in (c) we use Jensen's inequality and we bound the term with $\triangle\mathbf{x}_0^{r+1}$. Then the lemma is proved.

34

F.1.4 Proof of Lemma 9

Applying A1, we have

$$f_i(\mathbf{x}_0^r) \leq f_i(\mathbf{x}_i^r) + \langle \nabla f_i(\mathbf{x}_i^r), \mathbf{x}_0^r - \mathbf{x}_i^r \rangle + \frac{L}{2} \|\mathbf{x}_0^r - \mathbf{x}_i^r\|^2$$

$$\stackrel{(70)}{=} \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_0^r, \lambda_i^r) - \langle \mathbf{e}_i^r, \mathbf{x}_0^r - \mathbf{x}_i^r \rangle - \frac{1 - L\eta}{2\eta} \|\mathbf{x}_0^r - \mathbf{x}_i^r\|^2 \tag{84}$$

$$\leq \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_0^r, \lambda_i^r) + \frac{\epsilon_1}{2L} - \frac{1 - 2L\eta}{2\eta} \|\mathbf{x}_0^r - \mathbf{x}_i^r\|^2 .$$

Taking an average over N agents we are able to prove Lemma 9.

F.1.5 Proof of Theorem 1

First notice that from the optimality condition (70), the following holds:

$$\left\|\lambda_i^r - \lambda_i^{r-1}\right\|^2 \leq 2L^2 \left\|\mathbf{x}_i^r - \mathbf{x}_i^{r-1}\right\|^2 + 4\epsilon_1. \tag{85}$$

Then we bound the gradients of $\mathcal{L}(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)$.

$$\left\|\nabla_{\mathbf{x}_i} \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\| = \left\|\nabla f_i(\mathbf{x}_i^r) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^r - \mathbf{x}_{0,i}^r)\right\|$$

$$\stackrel{(70)}{=} \left\|\nabla f_i(\mathbf{x}_i^r) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^r - \mathbf{x}_{0,i}^r) - \nabla f_i(\mathbf{x}_i^{r+1}) - \lambda_i^r - \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r) + \mathbf{e}_i^{r+1}\right\| \tag{86}$$

$$\leq \frac{1 + L\eta}{\eta} \left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\| + \sqrt{\epsilon_1}.$$

Further, we note that, when no aggregation has been performed at iteration r, then $\mathbf{x}_{0,i}^r = \mathbf{x}_i^r + \eta\lambda_i^r$, so the following holds

$$\left\|\nabla_{\mathbf{x}_0} \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\| = \left\|\lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^r - \mathbf{x}_{0,i}^r)\right\| = 0. \tag{87}$$

When there the aggregation has been performed at iteration r, then $\mathbf{x}_{0,i}^r = \frac{1}{N}\sum_{j=1}^N (\mathbf{x}_j^r + \eta\lambda_j^r)$, $\forall i$, so we have

$$\|\nabla_{\mathbf{x}_0} \mathcal{L}(\mathbf{x}_0^r, \mathbf{x}_1^r, \dots, \mathbf{x}_N^r, \lambda_1^r, \dots, \lambda_N^r)\| = \left\|\frac{1}{N}\sum_{i=1}^N (\lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^r - \mathbf{x}_{0,i}^r))\right\| = 0. \tag{88}$$

Further we have:

$$\begin{aligned}
\left\|\nabla_{\lambda_i} \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\| &= \left\|\mathbf{x}_i^r - \mathbf{x}_{0,i}^r\right\| \\
&\leq \left\|\mathbf{x}_i^r - \mathbf{x}_{0,i}^{r-1}\right\| + \left\|\mathbf{x}_{0,i}^{r-1} - \mathbf{x}_{0,i}^r\right\| \\
&\leq \eta\left\|\lambda_i^r - \lambda_i^{r-1}\right\| + \left\|\mathbf{x}_{0,i}^{r-1} - \mathbf{x}_{0,i}^r\right\| \\
&\leq \eta(L\left\|\mathbf{x}_i^r - \mathbf{x}_i^{r-1}\right\| + 2\sqrt{\epsilon_1}) + \left\|\mathbf{x}_{0,i}^{r-1} - \mathbf{x}_{0,i}^r\right\| .
\end{aligned} \tag{89}$$

Summing (86) and (89), denote $\left\|\nabla_{\mathbf{x}_i} \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\| + \left\|\nabla_{\lambda_i} \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|$ as $\left\|\nabla \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|$ we have

$$\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\| \leq \left\|\mathbf{x}_{0,i}^{r-1} - \mathbf{x}_{0,i}^r\right\| + \frac{1+L\eta}{\eta}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\| + L\eta\left\|\mathbf{x}_i^r - \mathbf{x}_i^{r-1}\right\| + (1+2\eta)\sqrt{\epsilon_1}. \qquad (90)$$

Squaring both sides of the above inequality, we obtain:

$$\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 \leq C_6\left(\left\|\mathbf{x}_{0,i}^{r-1} - \mathbf{x}_{0,i}^r\right\|^2 + \left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 + \left\|\mathbf{x}_i^r - \mathbf{x}_i^{r-1}\right\|^2 + \epsilon_1\right), \qquad (91)$$

where $C_6 \geq \max\{(\frac{1+L\eta}{\eta})^2, (1+2\eta)^2, L^2\eta^2\}$.

Apply (85) to Lemma 6 we have

$$\frac{1 - 2L\eta - 4L^2\eta^2}{2\eta}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 + \frac{1}{2\eta}\left\|\mathbf{x}_{0,i}^{r+} - \mathbf{x}_{0,i}^r\right\|^2 + \frac{1+8L\eta}{2L}\epsilon_1$$
$$\leq \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) + \frac{1+8L\eta}{L}\epsilon_1. \qquad (92)$$

Notice that when communication is not performed $\left\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r+1}\right\|^2 \leq \left\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r+}\right\|^2$, and when communication is performed

$$\frac{1}{N}\sum_{i=1}^N\left\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r+1}\right\|^2 = \frac{2}{N}\sum_{i=1}^N\left\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r+}\right\|^2 + \frac{2}{N}\sum_{i=1}^N\left\|\mathbf{x}_{0,i}^{r+} - \mathbf{x}_{0,i}^{r+1}\right\|^2$$
$$\leq \frac{2}{N}\sum_{i=1}^N\left\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r+}\right\|^2 + \frac{N-1}{\eta N}(\triangle\mathbf{x}_0^{r+1})^2, \qquad (93)$$

where the last inequality holds due to the use of Jensen's inequality, and the definition of $\triangle\mathbf{x}_0^{r+1}$ in (69). It follows that summing both sides of (92) over i, we have

$$\frac{1 - 2L\eta - 4L^2\eta^2}{2\eta}\sum_{i=1}^N\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 + \sum_{i=1}^N(\frac{1}{4\eta}\left\|\mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^r\right\|^2 - \frac{N-1}{4\eta}(\triangle\mathbf{x}_0^{r+1})^2) + \frac{N(1+8L\eta)}{2L}\epsilon_1$$
$$\leq \sum_{i=1}^N\left(\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1})\right) + \frac{N(1+8L\eta)}{L}\epsilon_1. \qquad (94)$$

Taking the expectation over the randomness in p, conditioning on the information before the communication

the successive difference of $\mathcal{L}_i$,

$$\mathbb{E}_{r+1} \frac{1}{N} \sum_{i=1}^{N} [\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1})] = \frac{1}{N} \sum_{i=1}^{N} [\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1})]$$

$$+ \mathbb{E}_{r+1} \frac{1}{N} \sum_{i=1}^{N} [\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1})]$$

$$\overset{(a)}{=} \frac{1}{N} \sum_{i=1}^{N} [\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1})]$$

$$+ \frac{1}{N} \sum_{i=1}^{N} [p\mathcal{L}_i(\mathbf{x}_i^{r+1}, \overline{\mathbf{x}}_0^{r+1}, \lambda_i^{r+1}) + (1-p)\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1})] \qquad (95)$$

$$= \frac{1}{N} \sum_{i=1}^{N} [\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1})]$$

$$+ p\frac{1}{N} \sum_{i=1}^{N} [\mathcal{L}_i(\mathbf{x}_i^{r+1}, \overline{\mathbf{x}}_0^{r+1}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1})]$$

$$\overset{(b)}{\leq} \frac{1}{N} \sum_{i=1}^{N} [\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1})] + p\frac{N-1}{2\eta N}(\triangle \mathbf{x}_0^{r+1})^2,$$

where (a) expands the expectation on p, and use the fact that with probability p, $\mathbf{x}_{0,i}^{r+1} = \mathbf{x}_{0,i}^{r+}$, and with probability $(1-p)$ $\mathbf{x}_0^{r+1}$ will be updated; in (b) we apply Lemma 8 to the last term.

Combine (94) and (95), we have

$$\min\{\frac{1-2L\eta-4L^2\eta^2}{2\eta}, \frac{1}{2\eta}, \frac{1+8L\eta}{2L}\}\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{r+1} \left[\left\| \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\|^2 + \left\| \mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^r \right\|^2 + \epsilon_1 \right]$$

$$\leq \frac{1}{N} \sum_{i=1}^{N} \left[\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1}) \right] + \frac{1+8L\eta}{L}\epsilon_1 + p\frac{(N-1)}{\eta N}(\triangle \mathbf{x}_0^{r+1})^2. \qquad (96)$$

Combining (91), (94) and (96), define $C_7 = 2C_6/\min\{\frac{1-2L\eta-4L^2\eta^2}{2\eta}, \frac{1}{2\eta}, \frac{1+8L\eta}{2L}\}$ and sum up the iterations, we have

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{r=0}^{T} \mathbb{E} \left\| \nabla \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2$$

$$\overset{(91)(94)}{\leq} \frac{2C_6}{N} \sum_{i=1}^{N} \sum_{r=0}^{T} \mathbb{E} \left[\left\| \mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r+} \right\|^2 + \left\| \mathbf{x}_i^r - \mathbf{x}_i^{r+1} \right\|^2 + \frac{(N-1)}{\eta N}(\triangle \mathbf{x}_0^{r+1})^2 + \epsilon_1 \right]$$

$$\overset{(96)}{\leq} C_7 \sum_{r=0}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} (\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1})) + \frac{1+8L\eta}{L}\epsilon_1 \right) \qquad (97)$$

$$+ pC_7 \sum_{r=0}^{T} \frac{N-1}{N\eta} \mathbb{E}(\triangle \mathbf{x}_0^{r+1})^2.$$

Next we bound the last term. By iteratively applying Lemma 7 from $\tau = 0$ to r and use the fact that $\Delta^0 = 0$, we have

$$\mathbb{E} \triangle \mathbf{x}_0^{r+1} \leq [1,0] \sum_{\tau=0}^{r} (\frac{A}{1-L\eta})^\tau \eta \frac{[p(3+L\eta),2]^T}{1-L\eta} (\delta + 2\sqrt{\epsilon_1}). \tag{98}$$

From Lemma 7 we have:

$$\lambda_{\max} \left(\frac{1}{1-L\eta} A \right) = \frac{p(1+L\eta)+L\eta}{1-L\eta} \triangleq C_8.$$

So by taking norm square on both side of (98), we have

$$\begin{aligned}
\mathbb{E}(\triangle \mathbf{x}_0^{r+1})^2 &\leq \left\| [1,0] \sum_{\tau=1}^{r} (\frac{A}{1-L\eta})^\tau \eta \frac{[p(3+L\eta),2]^T}{1-L\eta} (\delta + 2\sqrt{\epsilon_1}) \right\|^2 \\
&\leq \left(\sum_{\tau=1}^{r} C_8^\tau \right)^2 \frac{8\eta^2(p^2(3+L\eta)^2+4)}{(1-L\eta)^2} (\delta^2 + \epsilon_1) \\
&\leq \frac{(1-C_8^{r+1})^2 \times 8\eta^2(p^2(3+L\eta)^2+4)}{(1-C_8)^2(1-L\eta)^2} (\delta^2 + \epsilon_1).
\end{aligned} \tag{99}$$

Substitute (99) into (97) and divide both side by T we have

$$\begin{aligned}
&\frac{1}{N} \sum_{i=1}^{N} \frac{1}{T} \sum_{r=0}^{T} \mathbb{E} \left\| \nabla \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2 \\
&\leq \frac{C_7}{T} \left(\mathcal{L}(\mathbf{x}_0^0, \mathbf{x}_i^0, \lambda_i^0) - \mathcal{L}(\mathbf{x}_i^T, \mathbf{x}_{0,i}^T, \lambda_i^T) \right) + \frac{C_7(1+8L\eta)}{L} \epsilon_1 \\
&\quad + \frac{8p\eta C_7(N-1)(1-C_8^{1/(1-p)})^2(p^2(3+L\eta)^2+4)}{N(1-C_8)^2(1-L\eta)^2} (\delta^2 + \epsilon_1).
\end{aligned}$$

From the initial conditions we have $\mathcal{L}(\mathbf{x}_0^0, \mathbf{x}_i^0, \lambda_i^0) = f(\mathbf{x}_0^0)$ and apply Lemma 9 we obtain

$$\begin{aligned}
&\frac{1}{NT} \sum_{i=1}^{N} \sum_{r=0}^{T} \mathbb{E} \left\| \nabla \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2 \\
&\leq \frac{C_7(f(\mathbf{x}_0^0) - f(\mathbf{x}_0^T))}{T} + \frac{C_7(1+8L\eta)}{L} \epsilon_1 \\
&\quad + \frac{8p\eta C_7(N-1)(1-C_8^{1/(1-p)})^2(p^2(3+L\eta)^2+4)}{N(1-C_8)^2(1-L\eta)^2} (\delta^2 + \epsilon_1).
\end{aligned}$$

Finally we bound $\|\nabla f(\mathbf{x}_0^r)\|^2$ by

$$\|\nabla f(\mathbf{x}_0^r)\|^2 \leq 2 \left\| \nabla f(\mathbf{x}_0^r) - \frac{1}{N}\sum_{i=1}^N \nabla_{\mathbf{x}_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2 + \frac{2}{N}\sum_{i=1}^N \left\| \nabla_{\mathbf{x}_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2$$

$$\leq \frac{4}{N}\sum_{i=1}^N \|\nabla f_i(\mathbf{x}_0^r) - \nabla f_i(\mathbf{x}_i^r)\|^2 + 4 \left\| \frac{1}{N\eta}\sum_{i=1}^N (\eta\lambda_i^r + \mathbf{x}_i^r - \mathbf{x}_{0,i}^r) \right\|^2$$

$$+ \frac{2}{N}\sum_{i=1}^N \left\| \nabla_{\mathbf{x}_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2$$

$$\overset{(a)}{\leq} \frac{4L^2}{N}\sum_{i=1}^N \|\mathbf{x}_0^r - \mathbf{x}_i^r\|^2 + \frac{2}{N}\sum_{i=1}^N \left\| \nabla_{\mathbf{x}_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2 \qquad (100)$$

$$= \frac{4L^2}{N}\sum_{i=1}^N \left\| \nabla_{\lambda_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2 + \frac{2}{N}\sum_{i=1}^N \left\| \nabla_{\mathbf{x}_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2$$

$$\leq \frac{4L^2}{N}\sum_{i=1}^N \left\| \nabla\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2,$$

where in (a) we use the same argument in (87) and (88).

Therefore Theorem 1 is proved. During the proof, we need all $C_2, \ldots, C_7, C_8 > 0$, therefore, $0 < \eta < \frac{\sqrt{5}-1}{4L}$.

Finally, let us note that if the local problems are solved with SGD, then the local problem needs to be solved such that the condition (68) holds true. As no other information of the local solvers except error term $\mathbf{e}_i^r$ is used in the proof, the proofs and results of FedPD with SGD as local solver will not change much, except that all the results hold in expectation. Therefore we skip the proof for the SGD version.

F.1.6 Constants used in the proofs

In this subsection we list all the constants $C_2, \ldots, C_8$ used in the proof of Theorem 1.

$$C_2 \geq 4L^2 C_7, \qquad C_3 = C_8, \qquad C_4 \geq \frac{C_2(1 + 8L\eta)}{L},$$

$$C_5 = 8C_2, \qquad C_6 \geq \max\{(\frac{1 + L\eta}{\eta})^2, (1 + 2\eta)^2, L^2\eta^2\},$$

$$C_7 = 2C_6 / \min\{\frac{1 - 2L\eta - 4L^2\eta^2}{2\eta}, \frac{1}{2\eta}, \frac{1 + 8L\eta}{2L}\},$$

$$C_8 = \frac{p(1 + L\eta) + L\eta}{1 - L\eta},$$

we can see that when $0 < \eta < \frac{\sqrt{5}-1}{4L}$, all the terms are positive.

G FedPD with Variance Reduction

In this section we provided an alternative oracle for FedPD which has a lower sample complexity.

G.1 Algorithm description

Alternatively, when instantiating the local oracle using Algorithm 4, the original local problems are not required to solve to ϵ_1 accuracy. Instead, we successively optimize a linearized AL function:

$$\tilde{\mathcal{L}}_i^r(\mathbf{x}_i) \triangleq \tilde{f}_i(\mathbf{x}_i; \mathbf{x}_i^{r,q}) + \langle \lambda_r^i, \mathbf{x}_i - \mathbf{x}_{0,i}^r \rangle + \frac{1}{2\eta} \left\| \mathbf{x}_i - \mathbf{x}_{0,i}^r \right\|^2.$$

In the above expression, we linearize $f_i(\mathbf{x}_i)$ at inner iteration $\mathbf{x}_i^{r,q}$ as (γ is a constant)

$$\tilde{f}_i^r(\mathbf{x}_i; \mathbf{x}_i^{r,q}) \triangleq f(\mathbf{x}_i^{r,q}) + \langle g_i^{r,q}, \mathbf{x}_i - \mathbf{x}_i^r \rangle + \frac{1}{2\gamma} \left\| \mathbf{x}_i - \mathbf{x}_i^{r,q} \right\|^2,$$

where $g_i^{r,q}$ is an approximation of $\nabla f_i(\mathbf{x}_i^{r,q})$. The optimizer has a closed-form expression:

$$\mathbf{x}_i^{r,q+1} = \frac{\eta}{\eta + \gamma} \mathbf{x}_i^{r,q} + \frac{\gamma}{\eta + \gamma} \mathbf{x}_{0,i}^r - \frac{\eta\gamma}{\eta + \gamma}(g_i^{r,q} + \lambda_i^r).$$

Algorithm 4 Oracle Choice II

> **Input:** $\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r), Q, I, B$
> Initialize: $\mathbf{x}_i^{r,0} = \mathbf{x}_i^r,$
> **if** $r \mod I = 0$ **then** $g_i^{r,0} = \nabla f_i(\mathbf{x}_i^{r,0})$
> **else** $g_i^{r,0} = g_i^{r-1,Q}$
> **end if**
> **for** $q = 0, \dots, Q-1$ **do**
> $\mathbf{x}_i^{r,q+1} = \arg\min_{\mathbf{x}_i} \tilde{\mathcal{L}}_i(\mathbf{x}_i, \mathbf{x}_{0,i}^r, \lambda_i^r; \mathbf{x}_i^{r,q}, g_i^{r,q})$
> $g_i^{r,q+1} = g_i^{r,q} + \frac{1}{B} \sum_{b=1}^{B} (h_i(\mathbf{x}_i^{r,q+1}; \xi_{i,b}^{r,q}) - h_i(\mathbf{x}_i^{r,q}; \xi_{i,b}^{r,q}))$
> **end for**
> **Output:** $\mathbf{x}_i^{r+1} \triangleq \mathbf{x}_i^{r,Q}, g_i^{r,Q}$

In Oracle II, an agent i first decides whether to compute the full gradient $\nabla f_i(\mathbf{x}_i^{r,0})$, or to keep using the previous estimate $g_i^{r-1,Q}$. Then Q local steps are performed, each requires B local data samples. In this scheme, Q can be chosen as *any* positive integer.

It is important to note that this oracle does not simply apply the VR technique (such as F-SVRG) to solve the subproblem of optimizing $\mathcal{L}_i(\mathbf{x}_i, \mathbf{x}_{0,i}^r, \lambda_i^r)$. That is, it is *not* a variation of Oracle I. Instead, the VR technique is applied to the entire primal-dual iteration, and the full gradient evaluation $\nabla f_i(\mathbf{x}_i^{r,0})$ is only needed every I iteration r. Later we will see that if I is large enough, then there is an $\mathcal{O}(\sqrt{M})$ reduction of sample complexity.

G.2 Algorithm Convergence and Complexity

The convergence result of FedPD with Oracle II is given as follows:

Theorem 3. *Suppose A1–A2 hold, and consider FedPD with Oracle II. Choose $p = 0, \eta \in \left(0, \frac{1}{3(Q+\sqrt{QI/B})L}\right)$, and $\gamma > \frac{5\eta}{B\sqrt{L}}$. Then, the following holds (where $C_9 > 0$ is a constant):*

$$\frac{1}{T} \sum_{r=0}^{T} \mathbb{E} \left\| \nabla f(\mathbf{x}_0^r) \right\|^2 \leq \frac{C_9}{T} (f(\mathbf{x}_0^0) - f(\mathbf{x}^\star)). \tag{101}$$

Remark 5. (**Communication complexity**): As $p = 0$, the communication round to achieve ϵ accuracy is $T = \mathcal{O}(1/\epsilon)$ which is independent of Q.

Remark 6. (**Computation complexity**): Note that the total number full gradient evaluation is $T/I + 1$, each uses M samples. Meanwhile, the total number of mini-batch stochastic gradient evaluation is TQ, each uses $2B$ samples per node. So the total sample complexity is $\mathcal{O}(M + MT/I + 2TQBN)$. In order to keep the same convergence speed, we need stepsize η to be unchanged. Therefore, we choose $I = \sqrt{M}, B = I/QN = \sqrt{M}/QN$, then the SC of Algorithm 4 is $\mathcal{O}(M + \frac{\sqrt{M}}{\epsilon})$.

G.3 Proof of Theorem 3

Following the similar proof of Theorem 1, we first analyze the descent between each outer iteration. Notice throughout the proof, we assume that $p = 0$, that is, there is no delayed communication. It follows that the following holds:

$$\mathbf{x}_{0,i}^{r+1} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_{0,j}^{r+}, \quad \forall i = 1, \ldots, N.$$

We also recall that r is the (outer) stage index, and q is the local update index. First we provide a series of lemmas.

Lemma 10. *Under Assumption 1, consider FedPD with Algorithm 4 (Oracle II) as the update rule. The difference of the local AL is bounded by:*

$$\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)$$

$$\leq -\frac{1}{2\eta} \left\| \mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^r \right\|^2 - \left(\frac{1}{2\eta} + \frac{1}{\gamma} - L - \frac{3\eta}{\gamma^2} \right) \left\| \mathbf{x}_i^{r,Q} - \mathbf{x}_i^{r,Q-1} \right\|^2$$

$$- \left(\frac{1}{2\eta} + \frac{1}{\gamma} - L - 9Q^2 L^2 \eta \right) \sum_{q=1}^{Q-1} \left\| \mathbf{x}_i^{r,q} - \mathbf{x}_i^{r,q-1} \right\|^2$$

$$+ \left(9Q^2 L^2 \eta + \frac{3\eta}{\gamma^2} \right) \left\| \mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1} \right\|^2 + \frac{1}{2L} \sum_{q=0}^{Q-2} \| \nabla f_i(\mathbf{x}_i^{r,q}) - g_i^{r,q} \|^2$$

$$+ \left(\frac{1}{2L} + 9\eta \right) \left\| g_i^{r,Q-1} - \nabla f_i(\mathbf{x}_i^{r,Q-1}) \right\|^2 + 9\eta \left\| g_i^{r-1,Q-1} - \nabla f_i(\mathbf{x}_i^{r-1,Q-1}) \right\|^2$$

$$+ \left\langle \lambda_i^{r+1} + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+1}), \mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^r \right\rangle.$$

Then we deal with the variance of the stochastic gradient estimations.

Lemma 11. *Suppose A1 holds true and the samples are randomly sampled according to (13), consider FedPD with Algorithm 4 (Oracle II) as the update rule. The expected norm square of the difference between $g_i^{r,q+1}$ and $\nabla f_i(\mathbf{x}_i^{r,q+1})$ is bounded by*

$$\mathbb{E} \left\| g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1}) \right\|^2 \leq \frac{L^2}{B} \sum_{\tau=\{r_0,1\}}^{\{r,q+1\}} \mathbb{E} \left\| \mathbf{x}_i^\tau - \mathbf{x}_i^{\tau-1} \right\|^2. \tag{102}$$

Lastly we upper bound the original loss function.

Lemma 12. *Under A1 and A2, the difference between the original loss and the AL is bounded as below:*

$$\mathbb{E} f(\mathbf{x}_0^r) \leq \mathbb{E} \mathcal{L}(\mathbf{x}_0^r, \mathbf{x}_1^r, \ldots, \mathbf{x}_N^r, \lambda_1^r, \ldots, \lambda_N^r) - \frac{1 - 3L\eta}{2N\eta} \sum_{i=1}^N \mathbb{E} \left\| \mathbf{x}_i^r - \mathbf{x}_0^r \right\|^2$$

$$+ \frac{(1 + L\gamma)^2 + L^2\gamma^2}{4L\gamma^2} \left[\frac{1}{B} \sum_{\tau=\{r_0,1\}}^{\{r-1,Q-1\}} \mathbb{E} \left\| \mathbf{x}_i^\tau - \mathbf{x}_i^{\tau-1} \right\|^2 + \mathbb{E} \left\| \mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1} \right\|^2 \right]. \tag{103}$$

G.3.1   Proof of Lemma 10

Let us first express the difference of the local AL as following:

$$\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \tag{104}$$

$$= \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) + \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r)$$

$$+ \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1}),$$

where the above three differences respectively correspond to the three steps in the algorithm's update steps.

Let us bound the above three differences one by one. First, note that we have the following decomposition (by using the fact that $\mathbf{x}_i^{r,Q+1} = \mathbf{x}_i^{r+1}$ and $\mathbf{x}_i^{r,1} = \mathbf{x}_i^r$):

$$\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) = \sum_{q=1}^Q \left(\mathcal{L}_i(\mathbf{x}_i^{r,q+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r,q}, \mathbf{x}_{0,i}^r, \lambda_i^r) \right). \tag{105}$$

Each term on the right hand side (RHS) of the above equality can be bounded by (see a similar arguments in (76)):

$$\mathcal{L}_i(\mathbf{x}_i^{r,q+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r,q}, \mathbf{x}_{0,i}^r, \lambda_i^r) \leq \left\langle \nabla f_i(\mathbf{x}_i^{r,q}) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}^{r,q+1} - \mathbf{x}_{0,i}^r), \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\rangle$$

$$- \frac{1 - L\eta}{2\eta} \left\| \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\|^2$$

$$\overset{(a)}{=} \left\langle \nabla f_i(\mathbf{x}_i^{r,q}) - g_i^{r,q} - \frac{1}{\gamma}(\mathbf{x}^{r,q+1} - \mathbf{x}_i^{r,q}), \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\rangle$$

$$- (\frac{1}{2\eta} - \frac{L}{2}) \left\| \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\|^2 \tag{106}$$

$$= \left\langle \nabla f_i(\mathbf{x}_i^{r,q}) - g_i^{r,q}, \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\rangle - (\frac{1}{2\eta} + \frac{1}{\gamma} - \frac{L}{2}) \left\| \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\|^2$$

$$\overset{(b)}{\leq} \frac{1}{2L} \left\| \nabla f_i(\mathbf{x}_i^{r,q}) - g_i^{r,q} \right\|^2 - (\frac{1}{2\eta} + \frac{1}{\gamma} - L) \left\| \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\|^2,$$

where in (a) we use the optimal condition that $\nabla_{\mathbf{x}_i} \tilde{\mathcal{L}}_i(\mathbf{x}_i^{r,q+1}, \mathbf{x}_{0,i}^r, \lambda_i^r; \mathbf{x}_i^{r,q}, g_i^{r,q}) = 0$ which gives us the following relation

$$\lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r,q+1} - \mathbf{x}_{0,i}^r) + g_i^{r,q} + \frac{1}{\gamma}(\mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q}) = 0; \tag{107}$$

in (b) we use the fact that $2\langle a, b\rangle \leq L \|a\|^2 + \frac{1}{L}\|b\|^2$. Therefore, the first difference in the RHS of (104) is given by

$$\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)$$

$$\leq \frac{1}{2L}\sum_{q=1}^{Q}\|\nabla f_i(\mathbf{x}_i^{r,q}) - g_i^{r,q}\|^2 - (\frac{1}{2\eta} + \frac{1}{\gamma} - L)\sum_{q=1}^{Q}\left\|\mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q}\right\|^2. \tag{108}$$

The other two differences in (104) can be explicitly expressed as:

$$\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) = \eta\left\|\lambda_i^{r+1} - \lambda_i^r\right\|^2, \tag{109}$$

$$\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1})$$

$$= -\frac{1}{2\eta}\left\|\mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^r\right\|^2 + \left\langle\lambda_i^{r+1} + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+1}), \mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^r\right\rangle. \tag{110}$$

Next we bound $\left\|\lambda_i^{r+1} - \lambda_i^r\right\|^2$. Notice that the from the update rule the following holds:

$$\lambda_i^{r+1} = \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r,Q} - \mathbf{x}_{0,i}^r) \overset{(107)}{=} -\frac{1}{\gamma}(\mathbf{x}_i^{r,Q} - \mathbf{x}_i^{r,Q-1}) - g^{r,Q-1}. \tag{111}$$

Using the above property, we have

$$\left\|\lambda_i^{r+1} - \lambda_i^r\right\|^2 = \left\|\frac{1}{\gamma}(\mathbf{x}_i^{r,Q} - \mathbf{x}_i^{r,Q-1}) + g_i^{r,Q-1} - \frac{1}{\gamma}(\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}) - g_i^{r-1,Q-1}\right\|^2$$

$$\overset{(a)}{\leq} 3\left\|g_i^{r,Q-1} - g_i^{r-1,Q-1}\right\|^2 + \frac{3}{\gamma^2}\left\|\mathbf{x}_i^{r,Q} - \mathbf{x}_i^{r,Q-1}\right\|^2 + \frac{3}{\gamma^2}\left\|\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}\right\|^2. \tag{112}$$

where in (a) we apply Cauchy-Schwarz inequality. Next we bound $\left\|g_i^{r,Q-1} - g_i^{r-1,Q-1}\right\|^2$ by

$$\left\|g_i^{r,Q-1} - g_i^{r-1,Q-1}\right\|^2$$

$$= \left\|g_i^{r,Q-1} - \nabla f_i(\mathbf{x}_i^{r,Q-1}) + \nabla f_i(\mathbf{x}_i^{r,Q-1}) - \nabla f_i(\mathbf{x}_i^{r-1,Q-1}) + \nabla f_i(\mathbf{x}_i^{r-1,Q-1}) - g_i^{r-1,Q-1}\right\|^2$$

$$\overset{(a)}{\leq} 3\left\|g_i^{r,Q-1} - \nabla f_i(\mathbf{x}_i^{r,Q-1})\right\|^2 + 3\left\|g_i^{r-1,Q-1} - \nabla f_i(\mathbf{x}_i^{r-1,Q-1})\right\|^2 + 3L^2\left\|\mathbf{x}_i^{r,Q-1} - \mathbf{x}_i^{r-1,Q-1}\right\|^2$$

$$\overset{(b)}{\leq} 3\left\|g_i^{r,Q-1} - \nabla f_i(\mathbf{x}_i^{r,Q-1})\right\|^2 + 3\left\|g_i^{r-1,Q-1} - \nabla f_i(\mathbf{x}_i^{r-1,Q-1})\right\|^2$$

$$+ 3Q^2 L^2\sum_{q=1}^{Q-1}\left\|\mathbf{x}_i^{r,q} - \mathbf{x}_i^{r,q-1}\right\|^2 + 3Q^2 L^2\left\|\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}\right\|^2,$$

where in (a) and (b) we both apply Cauchy-Schwarz inequality, in (a) we use A1 to the last term and in (b) we notice $\mathbf{x}_i^{r-1,Q} = \mathbf{x}_i^{r,0}$.

Substitute (113) to (112) and sum the three parts, we have

43

$$\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)$$

$$\leq -\frac{1}{2\eta} \left\| \mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^r \right\|^2 - (\frac{1}{2\eta} + \frac{1}{\gamma} - L - \frac{3\eta}{\gamma^2}) \left\| \mathbf{x}_i^{r,Q} - \mathbf{x}_i^{r,Q-1} \right\|^2$$

$$- (\frac{1}{2\eta} + \frac{1}{\gamma} - L - 9Q^2L^2\eta) \sum_{q=1}^{Q-1} \left\| \mathbf{x}_i^{r,q} - \mathbf{x}_i^{r,q-1} \right\|^2$$

$$+ (9Q^2L^2\eta + \frac{3\eta}{\gamma^2}) \left\| \mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1} \right\|^2 + \frac{1}{2L} \sum_{q=0}^{Q-2} \|\nabla f_i(\mathbf{x}_i^{r,q}) - g_i^{r,q}\|^2$$

$$+ (\frac{1}{2L} + 9\eta) \left\| g_i^{r,Q-1} - \nabla f_i(\mathbf{x}_i^{r,Q-1}) \right\|^2 + 9\eta \left\| g_i^{r-1,Q-1} - \nabla f_i(\mathbf{x}_i^{r-1,Q-1}) \right\|^2$$

$$+ \left\langle \lambda_i^{r+1} + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+1}), \mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^r \right\rangle,$$

which complete the proof of Lemma 10.

G.3.2   Proof of Lemma 11

To study $\mathbb{E} \|g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q})\|^2$, we denote the latest iteration before r that computes full gradients as r_0. That is, in r_0 we have $g_i^{r_0,0} = \nabla f_i(\mathbf{x}_i^{r_0,0})$. By the description of the algorithm we know

$$r_0 = kI, \quad k \in \mathbb{N}, \quad rQ + q - r_0Q \leq IQ.$$

That is, r_0 is a multiple of I and there is no more than IQ local update steps between step $\{r_0, 0\}$ and step $\{r, q\}$. By the update rule of $g_i^{r,q}$, we have

$$g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1}) = g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q+1}) + \frac{1}{B} \sum_{b=1}^{B}(h_i(\mathbf{x}_i^{r,q+1}; \xi_{i,b}^{r,q}) - h_i(\mathbf{x}_i^{r,q}; \xi_{i,b}^{r,q})). \tag{113}$$

Take expectation on both sides, we have

$$\mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^B}[g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1})]$$

$$= g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q+1}) + \mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^B}[\frac{1}{B} \sum_{b=1}^{B}(h_i(\mathbf{x}_i^{r,q+1}; \xi_{i,b}^{r,q}) - h_i(\mathbf{x}_i^{r,q}; \xi_{i,b}^{r,q}))]$$

$$= g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q+1}) + \nabla f_i(\mathbf{x}_i^{r,q+1}) - \nabla f_i(\mathbf{x}_i^{r,q})$$
$$= g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q}).$$

By using the fact that $\mathbb{E}[X^2] = [\mathbb{E}\,X]^2 + \mathbb{E}[[X - \mathbb{E}\,X]^2]$ and substitute (114) we have

$$
\mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^B} \left\| g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1}) \right\|^2
$$

$$
= \left\| \mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^B}[g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1})] \right\|^2
$$

$$
+ \mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^B} \left\| g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1}) - \mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^B}[g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1})] \right\|^2
$$

$$
\overset{(114)}{=} \left\| g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2
$$

$$
+ \mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^B} \left\| \frac{1}{B} \sum_{b=1}^B (h_i(\mathbf{x}_i^{r,q+1}; \xi_{i,b}^{r,q} - h_i(\mathbf{x}_i^{r,q}; \xi_{i,b}^{r,q})) - \nabla f_i(\mathbf{x}_i^{r,q+1}) + \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2
$$

$$
\overset{(a)}{\leq} \left\| g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2 + \frac{1}{B^2} \sum_{b=1}^B \mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^B} \left\| h_i(\mathbf{x}_i^{r,q+1}; \xi_{i,b}^{r,q}) - h_i(\mathbf{x}_i^{r,q}; \xi_{i,b}^{r,q}) \right\|^2
$$

$$
\overset{(b)}{\leq} \left\| g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2 + \frac{L^2}{B} \left\| \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\|^2,
$$

where (a) comes form the fact that we view $h_i(\mathbf{x}_i^{r,q+1}; \xi_{i,b}^{r,q}) - h_i(\mathbf{x}_i^{r,q}; \xi_{i,b}^{r,q})$ as X and by identically random sampling strategy we have $\mathbb{E}\,X = \nabla f_i(\mathbf{x}_i^{r,q+1}) - \nabla f_i(\mathbf{x}_i^{r,q})$ and $\mathbb{E}[[X - \mathbb{E}\,X]^2 \leq \mathbb{E}[X]^2$, in (b) we use A1.

Iteratively taking expectation until $\{r, q\} = \{r_0, 0\}$, we have

$$
\mathbb{E} \left\| g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1}) \right\|^2 \leq \frac{L^2}{B} \sum_{\tau=\{r_0,1\}}^{\{r,q+1\}} \mathbb{E} \left\| \mathbf{x}_i^\tau - \mathbf{x}_i^{\tau-1} \right\|^2, \tag{114}
$$

which completes the proof.

G.3.3   Proof of Lemma 12

Applying A1, we have

$$
f_i(\mathbf{x}_0^r) \leq f_i(\mathbf{x}_i^r) + \langle \nabla f_i(\mathbf{x}_i^r), \mathbf{x}_0^r - \mathbf{x}_i^r \rangle + \frac{L}{2} \left\| \mathbf{x}_0^r - \mathbf{x}_i^r \right\|^2
$$

$$
= \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_0^r, \lambda_i^r) - \langle \nabla f_i(\mathbf{x}_i^r) + \lambda_i^r, \mathbf{x}_0^r - \mathbf{x}_i^r \rangle - \frac{1 - L\eta}{2\eta} \left\| \mathbf{x}_0^r - \mathbf{x}_i^r \right\|^2 \tag{115}
$$

$$
\leq \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_0^r, \lambda_i^r) + \frac{1}{4L} \left\| \nabla f_i(\mathbf{x}_i^r) + \lambda_i^r \right\|^2 - \frac{1 - 3L\eta}{2\eta} \left\| \mathbf{x}_0^r - \mathbf{x}_i^r \right\|^2.
$$

45

Then notice $\mathbf{x}_i^r = \mathbf{x}_i^{r-1,Q}$ and apply (111), we can bound $\mathbb{E}\|\nabla f_i(\mathbf{x}_i^r) + \lambda_i^r\|^2$ by the following:

$$
\begin{aligned}
\mathbb{E}\|\nabla f_i(\mathbf{x}_i^r) + \lambda_i^r\|^2 &\overset{(111)}{=} \mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{r-1,Q}) - g_i^{r-1,Q-1} - \frac{1}{\gamma}(\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1})\right\|^2 \\
&\overset{(a)}{\leq} (1 + \frac{(1+L\gamma)^2}{L^2\gamma^2}) \mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{r-1,Q-1}) - g_i^{r-1,Q-1}\right\|^2 \\
&\quad + (1 + \frac{L^2\gamma^2}{(1+L\gamma)^2})(1 + \frac{1}{L\gamma}) \mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{r-1,Q}) - \nabla f_i(\mathbf{x}_i^{r-1,Q-1})\right\|^2 \\
&\quad + \frac{(1 + \frac{L^2\gamma^2}{(1+L\gamma)^2})(1+L\gamma)}{\gamma^2} \mathbb{E}\left\|\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}\right\|^2 \\
&\overset{(b)}{\leq} \frac{(1+L\gamma)^2 + L^2\gamma^2}{B\gamma^2} \sum_{\tau=\{r_0,1\}}^{\{r-1,Q-1\}} \mathbb{E}\left\|\mathbf{x}_i^\tau - \mathbf{x}_i^{\tau-1}\right\|^2 \\
&\quad + (1 + \frac{L^2\gamma^2}{(1+L\gamma)^2})\left((1 + \frac{1}{L\gamma})L^2 + \frac{1+L\gamma}{\gamma^2}\right) \mathbb{E}\left\|\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}\right\|^2 \\
&= \frac{(1+L\gamma)^2 + L^2\gamma^2}{B\gamma^2} \sum_{\tau=\{r_0,1\}}^{\{r-1,Q-1\}} \mathbb{E}\left\|\mathbf{x}_i^\tau - \mathbf{x}_i^{\tau-1}\right\|^2 \\
&\quad + \frac{(1+L\gamma)^2 + L^2\gamma^2}{\gamma^2} \mathbb{E}\left\|\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}\right\|^2,
\end{aligned}
\tag{116}
$$

where in (a) we apply Cauchy-Schwarz inequality twice, that is

$$
\|x + y + z\|^2 \leq (1 + \frac{1}{a})\|x\|^2 + (1+a)\|y+z\|^2 \leq (1 + \frac{1}{a})\|x\|^2 + (1+a)(1+b)\|y\|^2 + (1+a)(1 + \frac{1}{b})\|z\|^2;
$$

in (b) we apply Lemma 11 to the first term and apply A1 to the second term.

Substitute (116) to (115) and average over the agents, Lemma 12 is proved.

G.3.4 Proof of Theorem 3

By the update step of $\mathbf{x}_0^r$, following (87) we have

$$
\left\|\frac{1}{N}\sum_{i=1}^{N} \nabla_{\mathbf{x}_{0,i}}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\| = \left\|\frac{1}{N}\sum_{i=1}^{N}(\frac{1}{\eta}(\mathbf{x}_i^r - \mathbf{x}_{0,i}^r) + \lambda_i^r)\right\| = 0.
$$

We also have

$$
\left\| \nabla \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2
$$

$$
= \left\| \nabla_{\mathbf{x}_i} \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2 + \left\| \nabla_{\lambda_i} \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2
$$

$$
= \left\| \nabla f_i(\mathbf{x}_i^r) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^r - \mathbf{x}_{0,i}^r) \right\|^2 + \left\| \mathbf{x}_i^r - \mathbf{x}_{0,i}^r \right\|^2
$$

$$
\overset{(a)}{=} \left\| \nabla f_i(\mathbf{x}_i^r) - g_i^{r,0} - \frac{\eta + \gamma}{\eta \gamma}(\mathbf{x}_i^{r,1} - \mathbf{x}_i^r) \right\|^2 + \left\| \mathbf{x}_i^r - \mathbf{x}_{0,i}^r + \mathbf{x}_{0,i}^{r-1} - \mathbf{x}_{0,i}^{r-1} \right\|^2
$$

$$
\leq \left\| \nabla f_i(\mathbf{x}_i^r) - g_i^{r,0} - \frac{\eta + \gamma}{\eta \gamma}(\mathbf{x}_i^{r,1} - \mathbf{x}_i^r) \right\|^2 + 2\left\| \mathbf{x}_i^r - \mathbf{x}_{0,i}^{r-1} \right\|^2 + 2\left\| \mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r-1} \right\|^2
$$

$$
\leq 2\left\| \nabla f_i(\mathbf{x}_i^r) - g_i^{r,0} \right\|^2 + 2(\frac{\eta + \gamma}{\eta \gamma})^2 \left\| \mathbf{x}_i^{r,1} - \mathbf{x}_i^r \right\|^2 + 2\eta^2 \left\| \lambda_i^r - \lambda_i^{r-1} \right\|^2 + 2\left\| \mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r-1} \right\|^2.
$$

where in (a), the first term is obtained by plugging in (111) given below

$$
\lambda_i^r = -g_i^{r,0} - \frac{1}{\gamma}(\mathbf{x}_i^{r,1} - \mathbf{x}_i^r) - \frac{1}{\eta}(\mathbf{x}_i^{r,1} - \mathbf{x}_{0,i}^r).
$$

Next we take expectation and substitute (112), (113),

$$
\mathbb{E}\left\| \nabla \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \right\|^2
$$

$$
\leq 2\,\mathbb{E}\left\| \nabla f_i(\mathbf{x}_i^r) - g_i^{r,0} \right\|^2 + 2(\frac{\eta + \gamma}{\eta \gamma})^2 \mathbb{E}\left\| \mathbf{x}_i^{r,1} - \mathbf{x}_i^r \right\|^2 + 2\,\mathbb{E}\left\| \mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r-1} \right\|^2
$$

$$
+ \frac{6\eta^2}{\gamma^2}(\gamma^2\,\mathbb{E}\left\| g_i^{r-1,Q-1} - g_i^{r-2,Q-1} \right\|^2 + \mathbb{E}\left\| \mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1} \right\|^2 + E\left\| \mathbf{x}_i^{r-2,Q} - \mathbf{x}_i^{r-2,Q-1} \right\|^2)
$$

$$
\overset{(a)}{\leq} \frac{2L^2}{B} \sum_{\tau=\{r_0,1\}}^{\{r,0\}} \mathbb{E}\left\| \mathbf{x}_i^\tau - \mathbf{x}_i^{\tau-1} \right\|^2 + 2(\frac{\eta + \gamma}{\eta \gamma})^2 \mathbb{E}\left\| \mathbf{x}_i^{r,1} - \mathbf{x}_i^r \right\|^2 + 2\,\mathbb{E}\left\| \mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r-1} \right\|^2
$$

$$
+ \frac{6\eta^2}{\gamma^2}(\mathbb{E}\left\| \mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1} \right\|^2 + \mathbb{E}\left\| \mathbf{x}_i^{r-2,Q} - \mathbf{x}_i^{r-2,Q-1} \right\|^2)
$$

$$
+ 18\eta^2 \left(\mathbb{E}\left\| g_i^{r-1,Q-1} - \nabla f_i(\mathbf{x}_i^{r-1,Q-1}) \right\|^2 + \mathbb{E}\left\| g_i^{r-2,Q-1} - \nabla f_i(\mathbf{x}_i^{r-2,Q-1}) \right\|^2 \right)
$$

$$
+ 18\eta^2 Q^2 L^2 \left(\sum_{q=1}^{Q-1} \mathbb{E}\left\| \mathbf{x}_i^{r-1,q} - \mathbf{x}_i^{r-1,q-1} \right\|^2 + \mathbb{E}\left\| \mathbf{x}_i^{r-2,Q} - \mathbf{x}_i^{r-2,Q-1} \right\|^2 \right),
$$

where we substitute Lemma 11 and (113) in (a).

Taking expectation of (102), summing over $r = 0$ to $r = T - 1$ and average over the agents, we have

the following

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathcal{L}_i(\mathbf{x}_i^T,\mathbf{x}_{0,i}^T,\lambda_i^T)-\mathcal{L}_i(\mathbf{x}_i^0,\mathbf{x}_{0,i}^0,\lambda_i^0)] \leq -\frac{1}{2\eta}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_0^{r+1}-\mathbf{x}_0^r\right\|^2$$

$$-(\frac{1}{2\eta}+\frac{1}{\gamma}-L-\frac{6\eta}{\gamma^2}-9Q^2L^2\eta)\frac{1}{N}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_i^{r,q+1}-\mathbf{x}_i^{r,q-1}\right\|^2$$

$$+(\frac{1}{2L}+18\eta)\frac{1}{N}\sum_{i=1}^{N}\sum_{r=0}^{T-1}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{r,q})-g_i^{r,q}\right\|^2$$

$$+\sum_{r=0}^{T-1}\frac{1}{N}\mathbb{E}\left\langle\sum_{i=1}^{N}(\lambda_i^{r+1}+\frac{1}{\eta}(\mathbf{x}_i^{r+1}-\mathbf{x}_{0,i}^{r+1})),\mathbf{x}_{0,i}^{r+1}-\mathbf{x}_{0,i}^r\right\rangle$$

$$\overset{(a)}{\leq} -(\frac{1}{2\eta}+\frac{1}{\gamma}-L-\frac{6\eta}{\gamma^2}-9Q^2L^2\eta)\frac{1}{N}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_i^{r,q+1}-\mathbf{x}_i^{r,q-1}\right\|^2$$

$$-\frac{1}{2\eta}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_0^{r+1}-\mathbf{x}_0^r\right\|^2$$

$$+\frac{(1+18L\eta)LIQ}{2B}\frac{1}{N}\sum_{i=1}^{N}\sum_{r=0}^{T-1}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\mathbf{x}_i^{r,q+1}-\mathbf{x}_i^{r,q-1}\right\|^2$$

$$= -\frac{C_{10}}{N}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_i^{r,q+1}-\mathbf{x}_i^{r,q-1}\right\|^2-\frac{1}{2\eta}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_0^{r+1}-\mathbf{x}_0^r\right\|^2, \tag{117}$$

where in (a) we apply Lemma 11 and (87).

Finally, in the last equation of (117), we have defined the constant C_{10} as

$$C_{10} := \frac{1}{2\eta}+\frac{1}{\gamma}-L-\frac{6\eta}{\gamma^2}-9Q^2L^2\eta-\frac{(1+18L\eta)LIQ}{2B}.$$

Then by taking expectation and applying Lemma 12, we obtain

$$\mathbb{E}[f(\mathbf{x}_0^T)-f(\mathbf{x}_0^0)]$$
$$\leq -\frac{C_{10}-\frac{(1+L\gamma)^2+L^2\gamma^2}{4BL\gamma^2}}{N}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_i^{r,q+1}-\mathbf{x}_i^{r,q-1}\right\|^2-\frac{1}{2\eta}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_0^{r+1}-\mathbf{x}_0^r\right\|^2,$$

where by the initialization that $\mathbf{x}_i^0=\mathbf{x}_0^0$ we have $f(\mathbf{x}_0^0)=\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i(\mathbf{x}_i^0,\mathbf{x}_{0,i}^0,\lambda_i^0)$.

Combine (117) and (118), we can find a positive constant C_{11} satisfying

$$C_{11} \leq \min\left\{C_{12}/C_{13}, 1/(4\eta)\right\}, \tag{118}$$

where we have defined

$$C_{12} \triangleq C_{10}-\frac{(1+L\gamma)^2+L^2\gamma^2}{4BL\gamma^2},$$

$$C_{13} \triangleq Q\left(2(\frac{\eta+\gamma}{\eta\gamma})^2+\frac{2I(1+18\eta^2)L^2}{B}+\frac{3L(1+9L\eta)\eta^2}{2B\gamma^2}+18Q^2L^2\eta^2\right) \tag{119}$$

48

so that the following holds

$$\frac{C_{11}}{NT}\sum_{r=0}^{T}\sum_{i=1}^{N}\mathbb{E}\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^r,\mathbf{x}_{0,i}^r,\lambda_i^r)\right\|^2 \leq \frac{C_{10}-\frac{(1+L\gamma)^2+L^2\gamma^2}{4BL\gamma^2}}{NT}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_i^{r,q+1}-\mathbf{x}_i^{r,q-1}\right\|^2$$

$$+\frac{1}{2\eta T}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_0^{r+1}-\mathbf{x}_0^r\right\|^2 \tag{120}$$

$$\leq \frac{1}{T}(f(\mathbf{x}_0^0)-\mathbb{E}\,f(\mathbf{x}_0^T)) \leq \frac{1}{T}(f(\mathbf{x}_0^0)-f(\mathbf{x}^\star)).$$

Similar to the proof of Theorem 1, we can bound $\|\nabla f(\mathbf{x}_0^r)\|^2$ by $\frac{1}{N}\sum_{i=1}^{N}\|\nabla\mathcal{L}_i(\mathbf{x}_i^r,\mathbf{x}_0^r,\lambda_i^r)\|^2$, therefore Theorem 3 is proved.

During the prove we need

$$C_9 = 4L^2/C_{11}, \quad C_{10} = \frac{1}{2\eta}+\frac{1}{\gamma}-L-\frac{6\eta}{\gamma^2}-9Q^2L^2\eta-\frac{(1+18L\eta)LIQ}{2B},$$

$$C_{11} \leq \min\left\{\frac{\left(C_{10}-\frac{(1+L\gamma)^2+L^2\gamma^2}{4BL\gamma^2}\right)}{Q\left(2(\frac{\eta+\gamma}{\eta\gamma})^2+\frac{2I(1+18\eta^2)L^2}{B}+\frac{3L(1+9L\eta)\eta^2}{2B\gamma^2}+18Q^2L^2\eta^2\right)}, \frac{1}{4\eta}\right\}$$

to be positive constant. By selecting $\gamma > \frac{5}{B\sqrt{L}}\eta$, and $0 < \eta < \frac{1}{3(Q+\sqrt{QI/B})L}$, this is guaranteed.

H   Additional Numerical Results

H.1   Penalized Logistic Regression

In this experiment, we consider the penalized regression problem [30], whose loss function evaluated on a single sample $(\mathbf{a},b)=\xi$ is given by:

$$F(\mathbf{x};(\mathbf{a},b)) = \log(1+\exp(-b\mathbf{x}^T\mathbf{a}))+\sum_{d=1}^{D}\frac{\beta\alpha(\mathbf{x}[d])^2}{1+\alpha(\mathbf{x}[d])^2}. \tag{121}$$

Here $\mathbf{x}[d]$ denotes the dth component of $\mathbf{x}$. The feature vector and model parameter $\mathbf{a},\mathbf{x}\in\mathbb{R}^D$ have dimension D and $b\in\{-1,1\}$ is the label corresponding to the feature. During the simulation, we set the constants to be $\alpha=1$ and $\beta=0.1$.

In the experiment, we use two ways to generate the data. In the first case (referred to as the "weakly non-i.i.d" case), the features and the labels on the agents are randomly generated, so the local data sets are not very non-i.i.d. In the second case (referred to as the "strong non-i.i.d." case), we first generate the feature vector $\mathbf{a}$'s following the standard Normal distribution, then we generate the local model $\mathbf{x}_i$ on the ith agent by using uniform distribution in the range of $[-10,10]$ for each component. Then we compute the label b's according to the local models and the features and add some uniform noise. In this case, the data distribution on the agents are more non-i.i.d. compared to the first case. In both cases, there are 400 samples on each agent with total 100 agents.

The total number of iterations T is set as 600 for all algorithms. We choose the stepsize to be $\eta = 4$ for FedAvg-GD with local update number $Q = 8$ and for FedAvg-SGD we use diminishing stepsize $\eta = 4/\sqrt{Qr + q + 1}$ with $Q = 600$. For FedProx we use VR algorithm as the local solver and set $Q = 8$, $\rho = 1$ and stepsize $\eta = 4$. For FedPD, we also use the same stepsize $\eta = 4$ with $Q = 8$ with local GD. For FedPD-SGD, we also set $\eta = 4$ and uses local step size $\eta_1 = \frac{1}{Q}$ with inner iteration number $Q = 600$. Lastly for FedPD with VR, we set the parameters to be $\eta = 4$, $\gamma = 4$, $I = 100$, $Q = 2$ and $B = 1$. The choice of the stepsize is the same among all the algorithms. We used grid search on stepsizes $\eta \in \{5, 2, 1, 0.1, 0.01\}$ and the relative performance of the algorithms are similar to what we will show shortly.
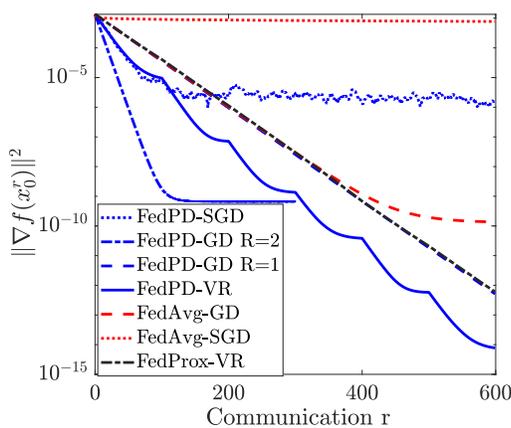
(a) The stationary gap of FedAvg, FedProx and FedPD with respect to the number of communication rounds.
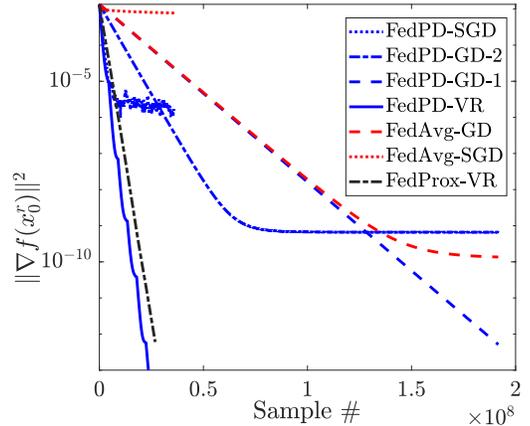
(b) The stationary gap of of FedAvg, FedProx and FedPD with respect to the number of samples.

Figure 4: The convergence result of the algorithms on penalized logistic regression with weakly non-i.i.d. data.

Fig. 4 shows the convergence results of the penalized logistic regression problem with the first data set. In Fig. 4(a), we compare the convergence of the tested algorithms w.r.t the communication rounds. It is clear that FedProx and FedPD with $R = 1$ (i.e., no communication skipping) are comparable. Meanwhile, FedAvg with local GD will not converge to the stationary point with a constant stepsize when local update step $Q > 1$. By skipping half of the communication, FedPD with local GD can still achieve a similar error as FedAvg, but using fewer communication rounds. In Fig. 4(b), we compare the sample complexity of different algorithms. It can be shown that when using the same number of samples for computation, FedPD with Oracle II (FedPD-VR) converges the fastest among all the algorithms. FedProx uses VR to solve the inner problem and converges the second fastest. Fig 5 shows the convergence results with the strongly non-i.i.d. data set. We can see that the algorithms using stochastic solvers become less stable compared with the case when the data sets are weakly non-i.i.d. Further, FedPD-VR and FedPD-GD with $R = 1$ are able still to converge to the global stationary point while FedProx will achieve a similar error as the FedAvg with local GD.
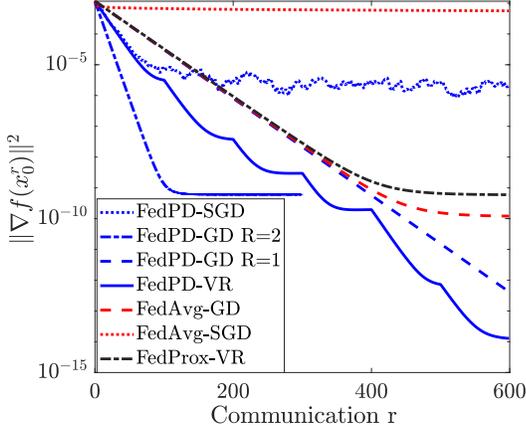
(a) The stationary gap of FedAvg, FedProx and FedPD with respect to the number of communication rounds.
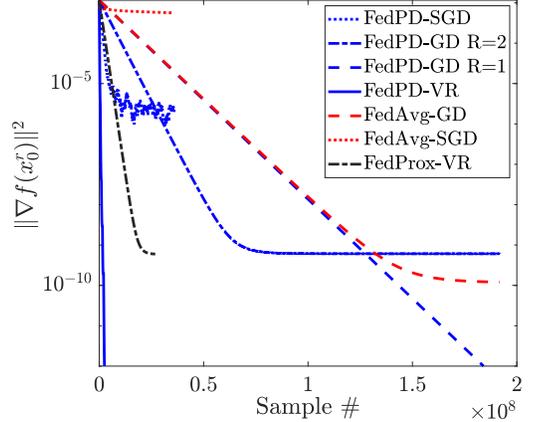
(b) The stationary gap of of FedAvg, FedProx and FedPD with respect to the number of samples.

Figure 5: The convergence result of the algorithms on penalized logistic regression with strongly non-i.i.d. data.

H.2   Handwritten Character Classification

In the second experiment, we compare FedPD with FedAvg and FedProx on the FEMNIST data set [31]. The FEMNIST data set collects the handwritten characters, including numbers 1–10 and the upper- and lower-case letters A–Z and a–z, from different writers and is separated by the writers, therefore the data set naturally preserves non-i.i.d.-ness.

The entire data set contains 805,000 samples collected from 3,550 writers. In our experiments, we use the data collected from 100 writers with an average of 300 samples per writer and the size of the whole data set is 29,214. We set the number of agent $N = 90$, the first ten agents are assigned with data from two writers, and the rest of the agents are assigned with data form one writer. Therefore, the data distribution is neither i.i.d. nor balanced. We use the neural network given in [31] as the training model, which consists of 2 convolutional layers and two fully connected layers. The output layer has 62 neurons that matches the number of classes in the FEMNIST data set.

The numerical results shown in Fig. 6 in the main text were generated by running MATLAB codes on Amazon Web Services (AWS), with Intel Xeon E5-2686 v4 CPUs. In the training phase, we train the CNN model with FedAvg, FedProx and FedPD. In Fig. 6(a), for FedAvg, we use gradient descent for $Q = 8$ local update steps between each communication rounds; to solve the local problem for FedProx, we use SARAH with $Q = 20$ local steps; we use FedPD with Oracle II, computing full gradient every $I = 20$ communication rounds and perform $Q = 2$ local steps between two communication rounds. The hyper-parameters we use for FedAvg is $\eta = 0.005$; for FedProx we use $\rho = 1$ and stepsize $\eta = 0.01$; for FedPD we use $\eta = 100$ and $\gamma = 400$. In Fig. 6(b), we use FedPD with Oracle I, with $Q = 20$, $\eta = 100$ and $\gamma = 400$ and the mini-batch size 2. We set the communication saving to $p = 0$ and $p = 0.5$.

The results shown in Fig. 7 were generated by running Python codes (using the the PyTorch package [1]) with AMD EPYC 7702 CPUs and an NVIDIA V100 GPU.

---

[1]PyTorch: An Imperative Style, High-Performance Deep Learning Library, `https://pytorch.org/`
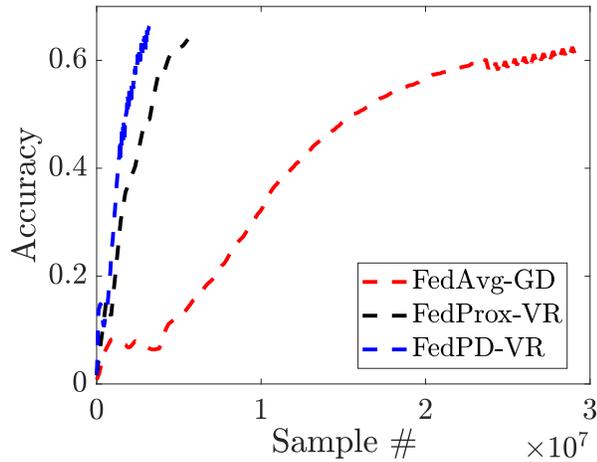
(a) The testing accuracy of FedAvg-GD, FedProx-VR and FedPD-VR with respect to the number of samples.
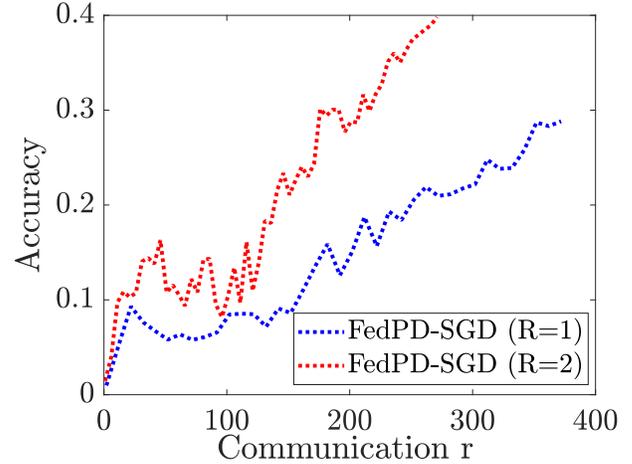
(b) The testing accuracy of FedPD-SGD with $R = 1$ and $R = 2$ with respect to the number of communications.

Figure 6: The convergence result of the algorithms on training neural network for handwriting character classification.

In the training phase, we train with FedProx, FedAvg and FedPD with a total $T = 1000$ outer iterations. The local problems are solved with SGD for $Q = 300$ local iterations and the mini-batch size in evaluating the stochastic gradient is 2. The stepsize choice for FedAvg, FedProx and FedPD are 0.001, 0.01 and 0.01, the hyper-parameter of FedProx is $\rho = 1$ and for FedPD $\eta = 1$. In the experiment, we set the communication saving for FedPD to be $p = 0$, $p = 0.5$ and $p = 0.25$. Note that we also tested FedAvg with larger stepsize 0.01, but the algorithm becomes unstable, and its performance degrages significantly. As shown in Fig. 7 and 8, FedAvg is slower than FedPD and FedProx, while FedProx has similar performance as FedPD when $R = 1$. Further, we can see that as the frequency of communication of FedPD decreases, the final accuracy decreases and the final loss increases. However, the drop of accuracy is not significant, so FedPD is able to achieve a better performance with the same number of communication rounds.
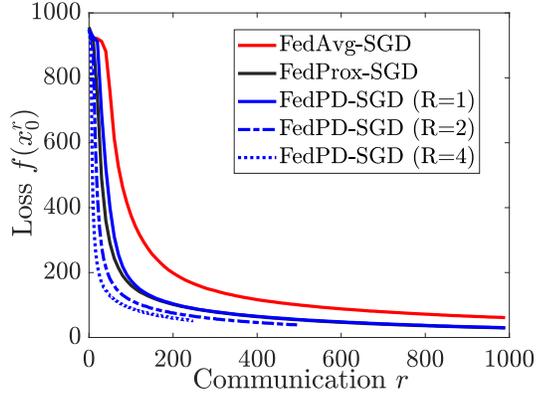
(a) The loss value of FedAvg-SGD, FedProx-SGD and FedPD-SGD with respect to the number of communication rounds.
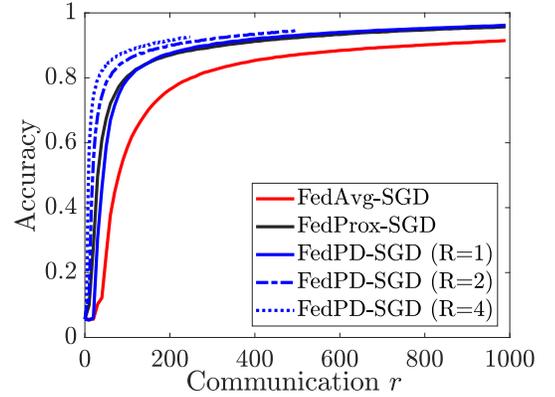
(b) The training accuracy of of FedAvg-SGD, FedProx-SGD and FedPD-SGD with respect to the number of communication rounds.

Figure 7: The convergence results of the algorithms on training neural networks on the federated handwritten characters classification problem.
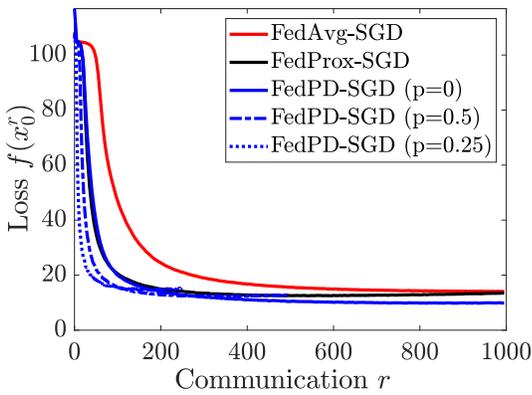
(a) The testing loss value of FedAvg-SGD, FedProx-SGD and FedPD-SGD with respect to the number of communication rounds.
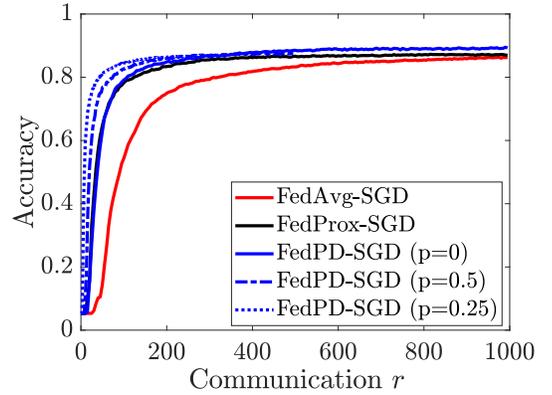
(b) The testing accuracy of of FedAvg-SGD, FedProx-SGD and FedPD-SGD with respect to the number of communication rounds.

Figure 8: The convergence results of the algorithms on training neural networks on the federated handwritten characters classification problem with test data set.