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 k^{th} sample $\xi_{i,k}$ in data set $\mathcal{D}_i$ consists 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

$$f_i(\mathbf{x}) = \frac{1}{|\mathcal{D}_i|} \sum_{(\mathbf{a}_k,b_k)\in\mathcal{D}_i} \frac{1}{1+\exp(b_k - \mathbf{a}_k^T\mathbf{x})}. \tag{19}$$

The gradient of this loss function is

$$\nabla f_i(\mathbf{x}) = \frac{1}{|\mathcal{D}_i|} \sum_{(\mathbf{a}_k,b_k)\in\mathcal{D}_i} \frac{\mathbf{a}_k \exp(b_k - \mathbf{a}_k^T\mathbf{x})}{(1+\exp(b_k - \mathbf{a}_k^T\mathbf{x}))^2}. \tag{20}$$

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

$$\mathbf{v}(\mathcal{D}_i, \mathbf{x}) = [v(\mathbf{a}_1, b_1, \mathbf{x}); \dots; v(\mathbf{a}_{|\mathcal{D}_i|}, b_{|\mathcal{D}_i|}, \mathbf{x})],$$

Further we define A_i as the stacked matrix of all $\mathbf{a}_k \in \mathcal{D}_i$ (i.e., $A_i = [\mathbf{a}_1, \dots, \mathbf{a}_{|\mathcal{D}_i|}]$), then we can express $\nabla f_i(\mathbf{x})$ as

$$\nabla f_i(\mathbf{x}) = \frac{1}{|\mathcal{D}_i|} A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x}). \tag{21}$$

The difference between the gradients of f_i and f_j is

$$\begin{aligned} \|\nabla f_i(\mathbf{x}) - \nabla f_j(\mathbf{x})\|^2 &= \left\| \frac{1}{|\mathcal{D}_i|} A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x}) - \frac{1}{|\mathcal{D}_j|} A_j \mathbf{v}(\mathcal{D}_j, \mathbf{x}) \right\|^2 \\ &\le \frac{1}{|\mathcal{D}_i|^2} \|A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x})\|^2 + \frac{1}{|\mathcal{D}_j|^2} \|A_j \mathbf{v}(\mathcal{D}_j, \mathbf{x})\|^2 \\ &\le \frac{1}{|\mathcal{D}_i|^2} \|A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x})\|^2 + \frac{1}{|\mathcal{D}_j|^2} \|A_j \mathbf{v}(\mathcal{D}_j, \mathbf{x})\|^2. \end{aligned} \tag{22}$$

As $v(\mathbf{a}, b, \mathbf{x}) \in (0,1)$, we know $\|\mathbf{v}(\mathcal{D}_i, \mathbf{x})\|^2 \le \|[1, \dots, 1]\|^2 = |\mathcal{D}_i|$, which implies:

$$\|A_i\|^2 \ge \frac{\|A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x})\|^2}{\|\mathbf{v}(\mathcal{D}_i, \mathbf{x})\|^2} \ge \frac{\|A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x})\|^2}{|\mathcal{D}_i|}.$$

Plug in the above inequality into (22), we obtain:

$$\begin{aligned} \|\nabla f_i(\mathbf{x}) - \nabla f_j(\mathbf{x})\|^2 &\le \frac{1}{|\mathcal{D}_i|^2} \|A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x})\|^2 + \frac{1}{|\mathcal{D}_j|^2} \|A_j \mathbf{v}(\mathcal{D}_j, \mathbf{x})\|^2 \\ &\le \frac{1}{|\mathcal{D}_i|} \|A_i\|^2 + \frac{1}{|\mathcal{D}_j|} \|A_j\|^2. \end{aligned} \tag{23}$$

So we can define $\delta = \max_{i,j} \left\{ \sqrt{\frac{1}{|\mathcal{D}_i|} \|A_i\|^2 + \frac{1}{|\mathcal{D}_j|} \|A_j\|^2} \right\}$ 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 expression for δ, which can be made to zero when A_i's are all the same.

Using similar analysis steps, we can also show that A5 holds for other loss functions such as the hyperbolic tangent function which is commonly used in neural network models.

B. Proof of Claim 2.1

Proof. The proof leverages on techniques developed from the classical and recent works that characterize lower bounds for first-order methods, in both centralized (Nesterov, 2004; Carmon et al., 2019) and decentralized (Scaman et al., 2017) settings. The major 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. Instead, one main point we would like to make is that there exist constructions of *local* functions f_i's such that *no matter* how local 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 mainly 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 $\nabla F(x_i^k;\xi_i)$. That is, we consider the following slightly simplified model for now:

$$x^t = V^t(\{x_i^{t-1,Q}\}_{i=1}^N),\ x_i^{t,0} = x^t,\ \ \forall\, i \in [N] \tag{24a}$$

$$x_i^{t,q} \in W_i^t\left(\{x_i^{r,k}, \left\{\nabla f_i(x_i^{r,k})\}\right\}_{r=0:t}^{k=0:q-1}\right), q\in[Q],\ \forall\, i \tag{24b}$$

We first introduce the main notations used in this section.

B.1. Notations.

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 with, we construct the following two non-convex functions

$$g(x) := \frac{1}{N}\sum_{i=1}^N g_i(x), \quad f(x) := \frac{1}{N}\sum_{i=1}^N f_i(x). \tag{25}$$

Here we have $x \in \mathbb{R}^{T+1}$. Note here that we assume N is considered as a constant, and T is the total number of stages, which is a large number and potentially can 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 (25) as follows.

$$g_i(x) = \Theta(x,1) + \sum_{j=1}^{T/N} \Theta(x,(j-1)N+i+1), \tag{26}$$

where we have defined the following functions

$$\begin{aligned}&\Theta(x,j) := \Psi(-x[j-1])\Phi(-x[j]) - \Psi(x[j-1])\Phi(x[j]),\ \forall\, j = 2,\cdots,T+1\\ &\Theta(x,1) := -\Psi(1)\Phi(x[1]).\end{aligned} \tag{27a}$$

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

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

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

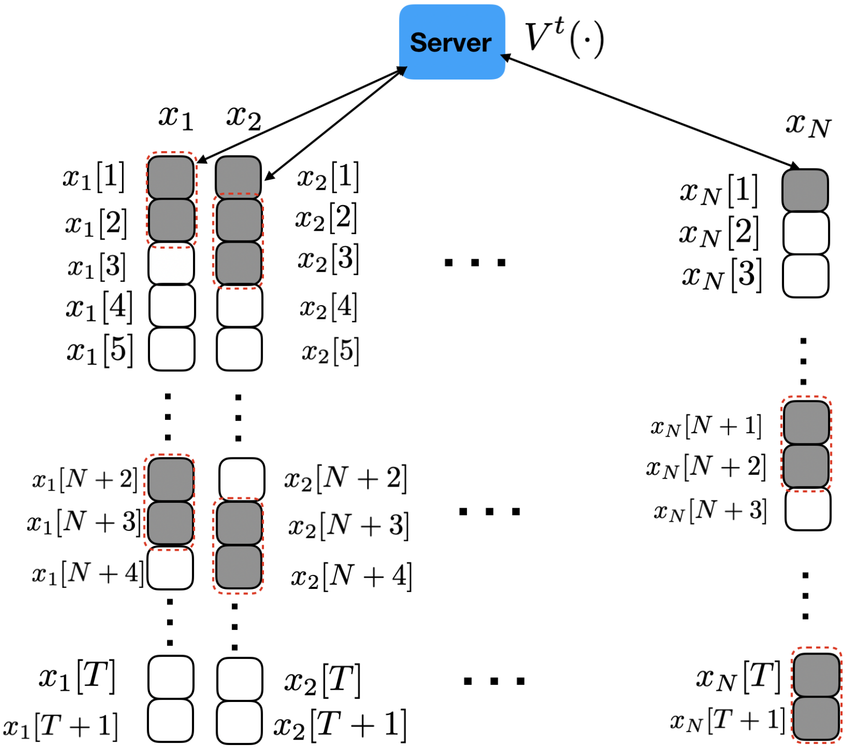


Figure 3: The example constructed for proving Claim 2.1. Here each agent has a local length $T+1$ vector x_i; each block in the figure represents one dimension of the local vector. If for agent i, its jth block is white it means that f_i is not a function of $x_i[j]$, while if jth block is shaded means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

By the above definition, the average function becomes:

$$\begin{aligned} g(x) &:= \frac{1}{M}\sum_{j=1}^{M} g_i(x) = \Theta(x,1) + \sum_{j=2}^{T+1}\Theta(x,j) \qquad (28)\\ &= -\Psi(1)\Phi\left(x[1]\right) + \sum_{j=2}^{T+1}\left[\Psi\left(-x[j-1]\right)\Phi\left(-x[j]\right) - \Psi\left(x[j-1]\right)\Phi\left(x[j]\right)\right]. \end{aligned}$$

See Fig. 3 for such a construction.

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

$$f_i(x) := \frac{2\pi\epsilon}{L} g_i\left(\frac{xL}{\pi\sqrt{2\epsilon}}\right). \qquad (29)$$

Therefore we also have

$$f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x) = \frac{2\pi\epsilon}{L} g\left(\frac{xL}{\pi\sqrt{2\epsilon}}\right). \qquad (30)$$

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 $w \le 0$, $\Psi(w) = 0$, $\Psi'(w) = 0$.*

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

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

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

3. *The following key property holds:*

$$\Psi(w)\Phi'(v) > 1, \quad \forall\, w \ge 1,\ |v| < 1. \tag{31}$$

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 (resp. g_i) is Lipschitz continuous with constant $\ell = 27\pi$ (resp. $\ell_i = 27\pi$, $\forall\, i$).*

Proof. Property 1) is easy to check.

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

$$\Psi(w) = 1 - e^{-w^2},\ \ \Psi'(w) = 2e^{-w^2}w,\ \ \Psi''(w) = 2e^{-w^2} - 4e^{-w^2}w^2,\ \forall\, w > 0. \tag{32}$$

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.,

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

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

$$\Phi(w) = 4\arctan w + 2\pi,\ \ \Phi'(w) = \frac{4}{w^2+1},\ \ \Phi''(w) = -\frac{8w}{(w^2+1)^2}. \tag{33}$$

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

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

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 (26)

$$\inf_x g_i(x) \geq -\Psi(1)\Phi(x[1]) - \sum_{j=1}^{T/N} \sup_{w,v} \Psi(w)\Phi(v) \tag{34}$$

$$\geq -4\pi - 4(T/N)\pi \geq -5\pi T/N, \tag{35}$$

where the first inequality follows $\Psi(w)\Phi(v) > 0$, the second follows $\Psi(w)\Phi(v) < 4\pi$, 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.

To proceed, let us note a few properties of the construction in (28) (also see Fig. 3). First, for a given node q, its local function h_q is only related to the following $x[j]$'s

$$\begin{aligned} &j = 1 + q + \ell \times N \geq 1,\ \ell = 0, \cdots, (N-1) \\ &j = q + \ell \times N \geq 1,\ \ell = 0, \cdots, (N-1), \end{aligned}$$

or equivalently

$$\begin{aligned} &q = j - 1 - \ell \times N \geq 1,\ \ell = 0, \cdots, (N-1) \\ &q = j - \ell \times N \geq 1,\ \ell = 0, \cdots, (N-1). \end{aligned}$$

Then the first-order partial derivative of $g_q(y)$ can be expressed below.

Case I) If $j \neq 1$ we have

$$\frac{\partial g_q}{\partial x[j]} = \begin{cases} \left(-\Psi\left(-x[j-1]\right)\Phi'\left(-x[j]\right) - \Psi\left(x[j-1]\right)\Phi'\left(x[j]\right)\right), \\ \qquad q = j - 1 - N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N} - 1, j = 2, 3, \cdots, T+1 \\ \left(-\Psi'\left(-x[j]\right)\Phi\left(-x[j+1]\right) - \Psi'\left(x[j]\right)\Phi\left(x[j+1]\right)\right), \\ \qquad q = j - N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N} - 1, j = 3, 4, \cdots T \\ 0 \qquad\qquad \text{otherwise}. \end{cases} . \tag{36}$$

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

$$\frac{\partial g_q}{\partial x[1]} = \begin{cases} -\Psi(1)\Phi'(x[1]) + \left(-\Psi'\left(-x[1]\right)\Phi\left(-x[2]\right) - \Psi'\left(x[1]\right)\Phi\left(x[2]\right)\right), & q = 1 \\ -\Psi(1)\Phi'(x[1]), & q \neq 1 \end{cases} . \tag{37}$$

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

$$\frac{\partial^2 g_q}{\partial x[j] \partial x[j]} = \begin{cases} \left(\Psi\left(-x[j-1]\right)\Phi''\left(-x[j]\right) - \Psi\left(x[j-1]\right)\Phi''\left(x[j]\right)\right), & \\ \qquad q = j-1-N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N}-1, j = 2, 3, \cdots, T+1 & \\ \left(\Psi''\left(-x[j]\right)\Phi\left(-x[j+1]\right) - \Psi''\left(x[j]\right)\Phi\left(x[j+1]\right)\right), & \\ \qquad q = j-N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N}-1, j = 3, 4, \cdots, T & \\ 0, & \text{otherwise} \end{cases} \tag{38}$$

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

$$\frac{\partial^2 g_q}{\partial x[j] \partial x[j-1]} = \begin{cases} \left(\Psi'\left(-x[j-1]\right)\Phi'\left(-x[j]\right) - \Psi'\left(x[j-1]\right)\Phi'\left(x[j]\right)\right), & \\ \qquad q = j-N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N}-1, j = 2, 3, \cdots, T+1 & \\ 0, & \text{otherwise} \end{cases}. \tag{40}$$

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 $\Psi\left(-x[j-1]\right)\Phi''\left(-x[j]\right)$ or $-\Psi\left(x[j-1]\right)\Phi''\left(x[j]\right)$ is zero. It follows that

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

Therefore, take the maximum over equations (38) to (40) and plug in the above inequalities, we obtain

$$\begin{aligned} \left|\frac{\partial^2 g_q}{\partial x[j_1] \partial x[j_2]}\right| &\leq \max\{\sup_w |\Psi''(w)| \sup_v |\Phi(v)|, \sup_w |\Psi(w)| \sup_v |\Phi''(v)|, \sup_w |\Psi'(w)| \sup_v |\Phi'(v)|\} \\ &= \max\left\{8\pi, \frac{3\sqrt{3}}{2}, 4\sqrt{\frac{2}{e}}\right\} < 8\pi, \quad \forall\ j_1 \neq 1 \end{aligned}$$

where the equality comes from Lemma 1 – ii).

When $j = 1$, by using (37), we have the following:

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

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

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

$$\begin{aligned} \left|\frac{\partial^2 g_q(x)}{\partial x[1] \partial x[j_2]}\right| &\leq \max\{\sup_w |\Psi(1)\Phi''(w)| + \sup_w |\Psi''(w)| \sup_v |\Phi(v)|, \sup_w |\Psi'(w)| \sup_v |\Phi'(v)|\} \\ &= \max\left\{\frac{3\sqrt{3}}{2}(1-e^{-1}) + 8\pi, 4\sqrt{\frac{2}{e}}\right\} < 9\pi,\ \forall\ j_2. \end{aligned}$$

Summarizing the above results, we obtain:

$$\begin{aligned}|\ell_q''(0;x,v)| &= |\sum_{\delta\in\{0,1,-1\}}\sum_{j=1}^{T}\frac{\partial^2 g_q(y)}{\partial x[j]\partial x[j+\delta]}v[j]v[j+\delta]| \\ &\le 9\pi\sum_{\delta\in\{0,1,-1\}}|\sum_{j=1}^{T}v[j]v[j+\delta]| \\ &\le 9\pi\left(|\sum_{j=1}^{T}v[j]^2|+2|\sum_{j=1}^{T}v[j]v[j+1]|\right) \\ &\le 27\pi\sum_{j=1}^{T}|v[j]^2| = 27\pi.\end{aligned}$$

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* (30) *and* (29).

1. *We have* $\forall\, x\in\mathbb{R}^{T+1}$
$$f(0)-\inf_x f(x)\le\frac{10\pi^2\epsilon}{LN}T.$$
2. *We have*
$$\|\nabla f(x)\| = \sqrt{2\epsilon}\left\|\nabla g\left(\frac{xL}{\pi\sqrt{2\epsilon}}\right)\right\|,\ \forall\, x\in\mathbb{R}^{T+1}. \tag{41}$$
3. *The first-order derivatives of f and that for each $f_i, i\in[N]$ are Lipschitz continuous, with the same constant $U>0$.*

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 $T\ge 1$, the following holds

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

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:

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

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 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 (37), we have the following inequalities:

$$\frac{\partial g_i(x)}{\partial x[1]} \overset{\text{(i)}}{\leq} -\Psi(1)\Phi'(x[1]) \overset{\text{(ii)}}{<} -1, \forall i \tag{42}$$

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:

$$\frac{\partial g_i(x)}{\partial x[j]} = \begin{cases} \left(-\Psi\left(-x[j-1]\right)\Phi'\left(-x[j]\right) - \Psi\left(x[j-1]\right)\Phi'\left(x[j]\right)\right), & \\ \qquad i = j-1-N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N}-1, j = 2, 3, \cdots, T+1 & \\ \left(-\Psi'\left(-x[j]\right)\Phi\left(-x[j+1]\right) - \Psi'\left(x[j]\right)\Phi\left(x[j+1]\right)\right), & \\ \qquad i = j-1-N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N}-1, j = 3, 4, \cdots, T & \\ 0 & \text{otherwise} \end{cases} . \tag{43}$$

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 equals to:

$$-\Psi(|x[j-1]|)\Phi'\left(|x[j]|\right)] \overset{(31)}{<} -1. \tag{44}$$

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 in the form of* (24) *to solve the following problem:*

$$\min_{x \in \mathbb{R}^{T+1}} \ g(x) = \frac{1}{N} \sum_{i=1}^{N} g_i(x). \tag{45}$$

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 (24b) *is performed, it needs at least T communication step* (24a)*, to have $\bar{x}[T] \neq 0$.*

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 (36))

$$\frac{\partial g_i(x_i)}{\partial x_i[j]} = -\left(\Psi\left(-x_i[j-1]\right)\Phi'\left(-x_i[j]\right)\right) + \left(\Psi\left(x_i[j-1]\right)\Phi'\left(x_i[j]\right)\right), \tag{46}$$

$$i = j - 1 - N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N} - 1, j = 2, 3, \cdots, T+1. \tag{47}$$

Clearly, if $x_i[j-1] = 0$, then by the definition of $\Psi(\cdot)$, the above partial gradient is also zero. In another word, 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 liner 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 happen:

- 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 (37). It is also important to observe that, for all the nodes $i \neq 1$, if they were to perform subsequent local updates (24b), 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

$$\frac{\partial g_i(x_i)}{\partial x_i[2]} = \left(-\Psi'\left(-x[2]\right)\Phi\left(-x[3]\right) - \Psi'\left(x[2]\right)\Phi\left(x[3]\right)\right) = 0, \tag{48}$$

since $x[2] = 0$ implies $\Psi'(-x[2]) = 0$. 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 (46), we have

$$\frac{\partial g_1(x_1)}{\partial x_1[2]} = -\left(\Psi\left(-x_1[1]\right)\Phi'\left(-x_1[2]\right)\right) + \left(\Psi\left(x_1[1]\right)\Phi'\left(x_1[2]\right)\right). \tag{49}$$

Since $x_1[1]$ is possible to be non-zero, then this partial gradient is also possible to be non-zero. Further, by the 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 (24b)), 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 do 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 3. *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* (13)*, 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* (24a)*)*

$$t \geq \frac{\left(f(0) - \inf_x f(x)\right) LN}{10\pi^2} \epsilon^{-1} \tag{50}$$

to achieve the following error

$$h_t^* = \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x^t) \right\|^2 < \epsilon. \tag{51}$$

Proof of Claim 2.1. First, let us show that the algorithm obeying the rules given in (24) 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

$$\|\nabla f(\bar{x}[T])\| = \sqrt{2\epsilon} \left\| \nabla h \left(\frac{\bar{x}[T]U}{\pi\sqrt{2\epsilon}} \right) \right\| > \sqrt{2\epsilon}/N, \tag{52}$$

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

$$T \geq \frac{\left(f(0) - \inf_x f(x)\right) LN}{10\pi^2} \epsilon^{-1}. \tag{53}$$

This completes the proof.

Second, consider the algorithm obeying the rules give in (13), 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

$$\hat{f}_i(x) = \sum_{\xi_i \in D_i} F(x; \xi_i) \tag{54}$$

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

$$F(\mathbf{x}; \xi_i) = \delta(\xi_i) f_i(x), \quad \text{where } f_i(x) \text{ is defined in (29).} \tag{55}$$

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

Proof. First let us consider FedAvg with local-GD update (12). We consider the following problem with $N = 2$, which satisfies both Assumptions 1 and 2, with $f(\mathbf{x}) = 0, \ \forall\, \mathbf{x}$

$$f_1(\mathbf{x}) = \mathbf{x}^2, \quad f_2(\mathbf{x}) = -\mathbf{x}^2. \tag{56}$$

Each local iteration of the FedAvg is given by

$$\mathbf{x}_1^{r+1} = (1-\eta^{r+1})\mathbf{x}_1^r, \quad \mathbf{x}_2^{r+1} = (1+\eta^{r+1})\mathbf{x}_2^r. \tag{57}$$

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

$$\mathbf{y}^{(k+1)Q} = \mathbf{D}^{Q-1}\mathbf{y}^{kQ+1}, \quad \mathbf{y}^{kQ+1} = \frac{1}{2}\mathbf{1}\mathbf{1}^T\mathbf{D}\mathbf{y}^{kQ}. \tag{58}$$

Therefore overall we have

$$\mathbf{y}^{(k+1)Q} = \frac{1}{2}\mathbf{D}^{Q-1}\mathbf{1}\mathbf{1}^T\mathbf{D}\mathbf{y}^{kQ}. \tag{59}$$

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. Proof of Claim 2.3

Before we prove Claim 2.3, the following lemma is needed.

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

$$\begin{aligned} f(\mathbf{x}^{r+1}) - f(\mathbf{x}^r) \leq & -(\eta^{r,0}(1-L\eta^{r,0}) + \sum_{q=1}^{Q-1}\frac{\eta^{r,q}}{2})\left\|\nabla f(\mathbf{x}^r)\right\|^2 \\ & - \sum_{q=1}^{Q-1}(\frac{\eta^{r,q}}{2} - 2L(Q-1)(\eta^{r,q})^2)\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q})\right\|^2 \\ & + \frac{QG^2}{2}((\eta^{r,0})^2 + \sum_{q=1}^{Q-1}(\eta^{r,q})^2)\sum_{q=1}^{Q-1}\eta^{r,q}, \end{aligned} \tag{60}$$

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

Proof: By using A1 we have:

$$\begin{aligned} & f(\mathbf{x}^{r+1}) - f(\mathbf{x}^r) \\ & \leq \left\langle \nabla f(\mathbf{x}^r), \mathbf{x}^{r+1} - \mathbf{x}^r \right\rangle + \frac{L}{2}\left\|\mathbf{x}^{r+1} - \mathbf{x}^r\right\|^2 \\ & \overset{(a)}{=} -\left\langle \nabla f(\mathbf{x}^r), \frac{1}{N}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\eta^{r,q}\nabla f_i(\mathbf{x}_i^{r,q}) \right\rangle + \frac{L}{2}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\eta^{r,q}\nabla f_i(\mathbf{x}_i^{r,q})\right\|^2 \\ & \overset{(b)}{\leq} -\sum_{q=1}^{Q-1}\eta^{r,q}\left\langle \nabla f(\mathbf{x}^r), \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q}) \right\rangle + L(\eta^{r,0})^2\left\|\nabla f(\mathbf{x}^r)\right\|^2 \\ & \quad + (Q-1)L\sum_{q=1}^{Q-1}(\eta^{r,q})^2\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q})\right\|^2 \\ & \overset{(c)}{=} -\eta^{r,0}\left\|\nabla f(\mathbf{x}^r)\right\|^2 - \sum_{q=1}^{Q-1}\eta^{r,q}\left\langle \nabla f(\mathbf{x}^r), \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q}) \right\rangle \\ & \quad + L(\eta^{r,0})^2\left\|\nabla f(\mathbf{x}^r)\right\|^2 + (Q-1)L\sum_{q=1}^{Q-1}(\eta^{r,q})^2\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q})\right\|^2, \end{aligned} \tag{61}$$

where (a) comes form the update rule in Algorithm 1, in (b) we use Cauchy-Schwarz inequality and notice $\mathbf{x}_i^{r,0} = \mathbf{x}^r$ so in (c) we extract the terms with index $(r, 0)$ form 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

$$\begin{aligned}
&-\eta^{r,q}\left\langle \nabla f(\mathbf{x}^r), \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q})\right\rangle + (Q-1)L(\eta^{r,q})^2\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q})\right\|^2\\
&= -\frac{\eta^{r,q}}{2}\|\nabla f(\mathbf{x}^r)\|^2 - \frac{\eta^{r,q}}{2}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q})\right\|^2 + \frac{\eta^{r,q}}{2}\left\|\nabla f(\mathbf{x}^r) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q})\right\|^2 + (Q-1)L(\eta^{r,q})^2\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q})\right\|^2\\
&\overset{(a)}{\leq} -\frac{\eta^{r,q}}{2}\|\nabla f(\mathbf{x}^r)\|^2 + \frac{\eta^{r,q}}{2N}\sum_{i=1}^{N}\|\nabla f_i(\mathbf{x}^r) - \nabla f_i(\mathbf{x}_i^{r,q})\|^2 - \frac{\eta^{r,q}}{2}((1-2(Q-1)L\eta^{r,q}))\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q})\right\|^2\\
&\overset{(b)}{\leq} -\frac{\eta^{r,q}}{2}\|\nabla f(\mathbf{x}^r)\|^2 + \frac{L^2\eta^{r,q}}{2N}\sum_{i=1}^{N}\|\mathbf{x}^r - \mathbf{x}_i^{r,q}\|^2 - \frac{\eta^{r,q}}{2}((1-2(Q-1)L\eta^{r,q}))\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{r,q})\right\|^2,
\end{aligned} \tag{62}$$

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

Further note that

$$\begin{aligned}
\|\mathbf{x}^r - \mathbf{x}_i^{r,q}\|^2 &= \left\|\mathbf{x}^r - \mathbf{x}^r + \sum_{\tau=0}^{q-1}\eta^{r,\tau}\nabla f_i(\mathbf{x}_i^{r,\tau})\right\|^2\\
&= \left\|\sum_{\tau=0}^{q-1}\eta^{r,\tau}\nabla f_i(\mathbf{x}_i^{r,\tau})\right\|^2\\
&\overset{(a)}{\leq} 2(q-1)\sum_{\tau=1}^{q-1}(\eta^{r,\tau})^2\|\nabla f_i(\mathbf{x}_i^{r,\tau})\|^2 + 2(\eta^{r,0})^2\left\|\nabla f_i(\mathbf{x}_i^{r,0})\right\|^2\\
&\overset{(b)}{\leq} 2\left((q-1)\sum_{\tau=1}^{q-1}(\eta^{r,\tau})^2 + (\eta^{r,0})^2\right)G^2.
\end{aligned} \tag{63}$$

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

Substitute (63) to (62) and then to (61), rearrange the terms we obtain (60), which ends the proof of the lemma. ■

D.1. Proof of Claim 2.3

Next we prove Claim 2.3

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

$$\begin{aligned}
f(\mathbf{x}^{r+1}) - f(\mathbf{x}^r) &\leq -(C_1 + \sum_{q=1}^{Q-1}\frac{\eta^{r,q}}{2})\|\nabla f(\mathbf{x}^r)\|^2\\
&+ \frac{QG^2}{2}((\eta_1)^2 + \sum_{q=1}^{Q-1}(\eta^{r,q})^2)\sum_{q=1}^{Q-1}\eta^{r,q},
\end{aligned} \tag{64}$$

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

$$\begin{aligned}
f(\mathbf{x}^T) - f(\mathbf{x}^0) &\leq -\sum_{r=0}^{T-1}(C_1 + \sum_{q=1}^{Q-1}\frac{\eta^{r,q}}{2})\|\nabla f(\mathbf{x}^r)\|^2\\
&+ \frac{QG^2}{2}\sum_{r=0}^{T-1}((\eta_1)^2 + \sum_{q=1}^{Q-1}(\eta^{r,q})^2)\sum_{q=1}^{Q-1}\eta^{r,q}.
\end{aligned} \tag{65}$$

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

$$(\frac{1}{T}+\frac{\sum_{r=0}^{T-1}\sum_{q=1}^{Q-1}\eta^{r,q}}{TC_1})\sum_{r=0}^{T}\|\nabla f(\mathbf{x}^r)\|^2 \le \frac{2(f(\mathbf{x}^0)-f(\mathbf{x}^\star))}{C_1T}+\frac{QG^2}{C_1T}\sum_{r=0}^{T-1}((\eta_1)^2+\sum_{q=1}^{Q-1}(\eta^{r,q})^2)\sum_{q=1}^{Q-1}\eta^{r,q}. \quad (66)$$

Choose $\eta^{r,q} \le \eta_1/Q$, then $(\eta_1)^2+\sum_{q=1}^{Q-1}(\eta^{r,q})^2 \le 2(\eta_1)^2$. Choose $\{\eta^{r,q}\}$ as a sequence that diminishes to 0, then for all $q \neq 0$, as $T \to \infty$, $\frac{2\eta_1 Q^2G^2}{C_1}\frac{1}{QT}\sum_{r=0}^{T-1}\sum_{q=1}^{Q-1}\eta^{r,q} \to 0$. Therefore the right hand side converges to 0, Claim 2.3 is proved.

E. Proof of Claim 2.4

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

$$f_1(\mathbf{x})=\mathbf{x}^2, \quad f_2(\mathbf{x})=-\mathbf{x}^2. \quad (67)$$

Each local iteration of the FedAvg is given by

$$\mathbf{x}_1^{r+1}=(1-\eta^r)\mathbf{x}_1^r, \quad \mathbf{x}_2^{r+1}=(1+\eta^r)\mathbf{x}_2^r. \quad (68)$$

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 $k\ge 0$, can be expressed as

$$\mathbf{y}^{(k+1)Q}=\prod_{r=kQ+1}^{(k+1)Q-1}\mathbf{D}^r\mathbf{y}^{kQ+1}, \quad \mathbf{y}^{kQ+1}=\frac{1}{2}\mathbf{1}\mathbf{1}^T\mathbf{D}^{kQ}\mathbf{y}^{kQ}. \quad (69)$$

Therefore overall we have

$$\mathbf{y}^{(k+1)Q}=\frac{1}{2}\prod_{r=kQ+1}^{(k+1)Q-1}\mathbf{D}^r\mathbf{1}\mathbf{1}^T\mathbf{D}^{kQ}\mathbf{y}^{kQ}. \quad (70)$$

In specific, 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 as the following holds true:

$$\left\|\nabla_{\mathbf{x}_i}\mathcal{L}(\mathbf{x}_i^{r+1},\mathbf{x}_0^r,\lambda_i^r)\right\|^2 \le \epsilon_1. \quad (71)$$

Note that if SGD is applied in Oracle I to solve the local problem, then this condition (71) is replaced with the following

$$\mathbb{E}[\left\|\nabla_{\mathbf{x}_i}\mathcal{L}(\mathbf{x}_i^{r+1},\mathbf{x}_0^r,\lambda_i^r)\right\|^2] \le \epsilon_1. \quad (72)$$

The difference does not significantly change the proofs and the results. So throughout the proof of Theorem 1, we use (71) as the condition.

Then we define the error between different nodes as

$$\triangle^r \triangleq [\triangle \mathbf{x}_0^r; \triangle \mathbf{x}^r], \text{ with } \triangle \mathbf{x}_0^r \triangleq \max_{i,j} \left\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,j}^r\right\|, \; \triangle \mathbf{x}^r \triangleq \max_{i,j} \left\|\mathbf{x}_i^r - \mathbf{x}_j^r\right\|. \tag{73}$$

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 (71)), we have

$$\nabla f_i(\mathbf{x}_i^{r+1}) + \lambda_i^{r+1} = \nabla f_i(\mathbf{x}_i^{r+1}) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r) = \mathbf{e}_i^{r+1}, \text{ where } \left\|\mathbf{e}_i^{r+1}\right\|^2 \leq \epsilon_1. \tag{74}$$

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* (73) *is satisfied, the difference of the local augmented Lagrangian is bounded by*

$$\begin{aligned}&\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\\ &\leq -\frac{1-2L\eta}{2\eta}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 - \frac{1}{2\eta}\left\|\mathbf{x}_{0,i}^{r+} - \mathbf{x}_{0,i}^r\right\|^2 + \eta\left\|\lambda_i^{r+1} - \lambda_i^r\right\|^2 + \frac{\epsilon_1}{2L}.\end{aligned} \tag{75}$$

Then we derive a key lemma about how the error propagate 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* (73) *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*

$$\triangle^{r+1} \leq \frac{1}{1-L\eta}(A\triangle^r + \eta B[\delta, \sqrt{\epsilon_1}]^T). \tag{76}$$

where $A = [1+L\eta, 1]^T[1, L\eta]$ *and* $B = [2, 3+L\eta]^T[1, 2]$ *constant matrices.*

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.

Lemma 8. *Suppose A1 holds. Consider FedPD with Algorithm 4 (Oracle I) as the update rule. When the local problem is solved such that* (73) *is satisfied, the difference between local AL and the global AL is bounded as below:*

$$\frac{1}{N}\sum_{i=1}^N \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) - \mathcal{L}(\overline{\mathbf{x}}_0^{r+1}, \mathbf{x}_1^{r+1}, \ldots, \mathbf{x}_N^{r+1}, \lambda_1^{r+1}, \ldots, \lambda_N^{r+1}) \geq -\frac{N-1}{2N\eta}(\triangle \mathbf{x}_0^{r+1})^2. \tag{77}$$

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.*

$$f(\mathbf{x}_0^r) \leq \mathcal{L}(\mathbf{x}_0^r, \mathbf{x}_1^r, \ldots, \mathbf{x}_N^r, \lambda_1^r, \ldots, \lambda_N^r) - \frac{1-2L\eta}{N\eta}\sum_{i=1}^N \left\|\mathbf{x}_i^r - \mathbf{x}_0^r\right\|^2 + \frac{\epsilon_1}{2L}. \tag{78}$$

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 (75), 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:

$$\begin{aligned}\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) &= \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r \lambda_i^r) \\ &+ \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) \\ &+ \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1}),\end{aligned} \tag{79}$$

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

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

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

and obtain the following series of inequalities:

$$\begin{aligned}\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) &\le \left\langle \nabla f_i(\mathbf{x}_i^{r+1}), \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\rangle + \frac{L}{2}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 + \left\langle \lambda_i^r, \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\rangle \\ &+ \frac{1}{2\eta}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r\right\|^2 - \frac{1}{2\eta}\left\|\mathbf{x}_i^r - \mathbf{x}_{0,i}^r\right\|^2 \\ &\overset{(a)}{=} \left\langle \nabla f_i(\mathbf{x}_i^{r+1}) + \lambda_i^r, \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\rangle + \frac{L}{2}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 \\ &+ \frac{1}{2\eta}\left\langle \mathbf{x}_i^{r+1} + \mathbf{x}_i^r - 2\mathbf{x}_{0,i}^r, \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\rangle \\ &\overset{(b)}{=} \left\langle \nabla f_i(\mathbf{x}_i^{r+1}) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r), \mathbf{x}_i^{r+1} - \mathbf{x}_i^r \right\rangle + \frac{L}{2}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 \\ &- \frac{1}{2\eta}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 \\ &\overset{(c)}{\le} \frac{1}{2L}\left\|\nabla f_i(\mathbf{x}_i^{r+1}) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r)\right\|^2 + \frac{L}{2}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 \\ &- \frac{1 - L\eta}{2\eta}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 \\ &\overset{(d)}{\le} -\frac{1 - 2L\eta}{2\eta}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 + \frac{\epsilon_1}{2L}.\end{aligned} \tag{80}$$

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 $\langle a, b\rangle \le \frac{L}{2}\|a\|^2 + \frac{1}{2L}\|b\|^2)$; in (d) we apply the fact that $\mathbf{x}_i^{r+1}$ is the inexact solution; see (74).

Then we bound the second difference in (79) by the following:

$$\begin{aligned}\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) &= \left\langle \lambda_i^{r+1} - \lambda_i^r, \mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r \right\rangle \\ &\overset{(a)}{=} \left\langle \lambda_i^{r+1} - \lambda_i^r, \eta(\lambda_i^{r+1} - \lambda_i^r) \right\rangle \\ &= \eta\left\|\lambda_i^{r+1} - \lambda_i^r\right\|^2,\end{aligned} \tag{81}$$

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

Further we bound the third difference in (79) by the following:

$$\begin{aligned}
&\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r}, \lambda_i^{r+1}) \\
&= \left\langle \lambda_i^{r+1}, \mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+}\right\rangle - \left\langle \lambda_i^{r+1}, \mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r}\right\rangle + \frac{1}{2\eta}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+}\right\|^2 - \frac{1}{2\eta}\left\|\mathbf{x}_i^{r+1}-\mathbf{x}_{0,i}^r\right\|^2 \\
&\overset{(a)}{=} \left\langle \lambda_i^{r+1}, \mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r+}\right\rangle + \frac{1}{2\eta}\left\langle 2\mathbf{x}_i^{r+1} - 2\mathbf{x}_{0,i}^{r+} + \mathbf{x}_{0,i}^{r+} - \mathbf{x}_{0,i}^r, \mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r+}\right\rangle \\
&= \left\langle \frac{1}{\eta}(\eta\lambda_i^{r+1} + \mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+}), \mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r+}\right\rangle - \frac{1}{2\eta}\left\|\mathbf{x}_{0,i}^{r+} - \mathbf{x}_{0,i}^r\right\|^2 \\
&\overset{(b)}{=} -\frac{1}{2\eta}\left\|\mathbf{x}_{0,i}^{r+}-\mathbf{x}_{0,i}^r\right\|^2,
\end{aligned} \tag{82}$$

where (a) we use the same reasoning as in (80) (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 (80), (81), (82) 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 inexact minimization:

$$\begin{aligned}
\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_j^{r+1}\right\| &\overset{(74)}{=} \left\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,j}^r - \eta(\nabla f_i(\mathbf{x}_i^{r+1}) + \lambda_i^r - \mathbf{e}_i^{r+1} - \nabla f_j(\mathbf{x}_j^{r+1}) - \lambda_j^r + \mathbf{e}_j^{r+1})\right\| \\
&\le \left\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,j}^r\right\| + \eta\left\|\nabla f_i(\mathbf{x}_i^{r+1}) - \nabla f_j(\mathbf{x}_j^{r+1})\right\| + \eta\left\|\lambda_i^r - \lambda_j^r\right\| + \eta(\left\|\mathbf{e}_i^{r+1}\right\| + \left\|\mathbf{e}_j^{r+1}\right\|) \\
&\overset{(a)}{\le} \triangle\mathbf{x}_0^r + \eta\left\|\nabla f_i(\mathbf{x}_i^{r+1}) - \nabla f_i(\mathbf{x}_j^{r+1}) + \nabla f_i(\mathbf{x}_j^{r+1}) - \nabla f_j(\mathbf{x}_j^{r+1})\right\| + \eta\left\|\lambda_i^r - \lambda_j^r\right\| + 2\eta\sqrt{\epsilon_1} \\
&\overset{(b)}{\le} \triangle\mathbf{x}_0^r + L\eta\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_j^{r+1}\right\| + \eta\left\|\nabla f_i(\mathbf{x}_j^{r+1}) - \nabla f_j(\mathbf{x}_j^{r+1})\right\| + \eta\left\|\lambda_i^r - \lambda_j^r\right\| + 2\eta\sqrt{\epsilon_1} \\
&\overset{(c)}{\le} \triangle\mathbf{x}_0^r + L\eta\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_j^{r+1}\right\| + \eta\delta + \eta\left\|\lambda_i^r - \lambda_j^r\right\| + 2\eta\sqrt{\epsilon_1} \\
&\overset{(d)}{=} \frac{1}{1-L\eta}\triangle\mathbf{x}_0^r + \frac{\eta}{1-L\eta}\delta + \frac{\eta}{1-L\eta}\left\|\lambda_i^r - \lambda_j^r\right\| + \frac{2\eta}{1-L\eta}\sqrt{\epsilon_1}
\end{aligned} \tag{83}$$

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 (74), and note that $\lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r) = \lambda_i^{r+1}$:

$$\begin{aligned}
\left\|\lambda_i^r - \lambda_j^r\right\| &= \left\|-\nabla f_i(\mathbf{x}_i^r) + \mathbf{e}_i^r + \nabla f_j(\mathbf{x}_j^r) - \mathbf{e}_j^r\right\| \\
&\overset{(a)}{\le} \left\|\nabla f_i(\mathbf{x}_i^r) - \nabla f_i(\mathbf{x}_j^r)\right\| + \left\|\nabla f_i(\mathbf{x}_j^r) - \nabla f_j(\mathbf{x}_j^r)\right\| + 2\sqrt{\epsilon_1} \\
&\overset{(b)}{\le} L\left\|\mathbf{x}_i^r - \mathbf{x}_j^r\right\| + \delta + 2\sqrt{\epsilon_1} \\
&\overset{(c)}{\le} L\triangle\mathbf{x}^r + \delta + 2\sqrt{\epsilon_1},
\end{aligned} \tag{84}$$

where (a) and (b) follow the same argument in (a), (b) and (c) of (83) ; in (c) we plug in the definition of $\triangle\mathbf{x}^r$.

Next we bound the difference $\left\|\mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,j}^{r+1}\right\|$. When $r + 1 \mod R = 0$ (when the aggregation step has just been done at iteration r), $\mathbf{x}_{0,i}^{r+1} = \mathbf{x}_{0,j}^{r+1}$. Otherwise, we have

$$\begin{aligned}
\left\|\mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,j}^{r+1}\right\| &= \left\|\mathbf{x}_i^{r+1} - \mathbf{x}_j^{r+1} + \eta(\lambda_i^{r+1} - \lambda_j^{r+1})\right\| \\
&\le \left\|\mathbf{x}_i^{r+1} - \mathbf{x}_j^{r+1}\right\| + \eta\left\|\lambda_i^{r+1} - \lambda_j^{r+1}\right\| \\
&\overset{(a)}{\le} (1 + L\eta)\triangle\mathbf{x}^{r+1} + \eta\delta + 2\eta\sqrt{\epsilon_1}
\end{aligned} \tag{85}$$

where in (a) we plug in the definition of $\triangle\mathbf{x}^{r+1}$ and (84). As these relations hold true for arbitrary (i,j) pairs, they are also true for the maximum of $\|\mathbf{x}_i^{r+1}-\mathbf{x}_j^{r+1}\|$ and $\|\mathbf{x}_{0,i}^{r+1}-\mathbf{x}_{0,j}^{r+1}\|$.

Therefore stacking (83) and (85) and plug in (84), we have

$$\begin{aligned}\triangle\mathbf{x}^{r+1} \leq& \frac{1}{1-L\eta}(L\eta\triangle\mathbf{x}^r+\triangle\mathbf{x}_0^r)+\frac{2\eta}{1-L\eta}(\delta+2\sqrt{\epsilon_1}),\\ \triangle\mathbf{x}_0^{r+1} \leq& \frac{1+L\eta}{1-L\eta}(L\eta\triangle\mathbf{x}^r+\triangle\mathbf{x}_0^r)+\frac{\eta(3+L\eta)}{1-L\eta}(\delta+2\sqrt{\epsilon_1}).\end{aligned} \tag{86}$$

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:

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

Similar to (82), we have

$$\begin{aligned}\mathcal{L}_i(\mathbf{x}_i^{r+1},\mathbf{x}_{0,i}^{r+},\lambda_i^{r+1})-\mathcal{L}_i(\mathbf{x}_i^{r+1},\overline{\mathbf{x}}_0^{r+1},\lambda_i^{r+1}) &= \left\langle\lambda_i^{r+1},\mathbf{x}_i^{r+1}-\mathbf{x}_{0,i}^{r+}\right\rangle-\left\langle\lambda_i^{r+1},\mathbf{x}_i^{r+1}-\overline{\mathbf{x}}_0^{r+1}\right\rangle\\ &\quad+\frac{1}{2\eta}\left\|\mathbf{x}_i^{r+1}-\mathbf{x}_{0,i}^{r+}\right\|^2-\frac{1}{2\eta}\left\|\mathbf{x}_i^{r+1}-\overline{\mathbf{x}}_0^{r+1}\right\|^2\\ &\overset{(a)}{=}-\frac{1}{2\eta}\left\|\mathbf{x}_{0,i}^{r+}-\overline{\mathbf{x}}_0^{r+1}\right\|^2\\ &\overset{(b)}{=}-\frac{1}{2\eta}\left\|\mathbf{x}_{0,i}^{r+}-\frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_{0,j}^{r+}\right\|^2\\ &=-\frac{1}{2\eta}\left\|\frac{1}{N}\sum_{j=1}^{N}(\mathbf{x}_{0,i}^{r+}-\mathbf{x}_{0,j}^{r+})\right\|^2\\ &\overset{(c)}{\geq}-\frac{1}{2\eta N}\sum_{j\neq i}\left\|\mathbf{x}_{0,i}^{r+}-\mathbf{x}_{0,j}^{r+}\right\|^2\\ &\overset{(d)}{\geq}-\frac{N-1}{2\eta N}(\triangle\mathbf{x}_0^{r+1})^2,\end{aligned} \tag{87}$$

where (a) follows the same argument in (82); in (b),we plug in the definition of $\overline{\mathbf{x}}_0^{r+1}$; in (c) we use Jensen's inequality and we bound the term with $\triangle\mathbf{x}_0^{r+1}$. Then the lemma is proved.

F.1.4. Proof of Lemma 9

Applying A1, we have

$$\begin{aligned}f_i(\mathbf{x}_0^r) &\leq f_i(\mathbf{x}_i^r)+\langle\nabla f_i(\mathbf{x}_i^r),\mathbf{x}_0^r-\mathbf{x}_i^r\rangle+\frac{L}{2}\left\|\mathbf{x}_0^r-\mathbf{x}_i^r\right\|^2\\ &\overset{(74)}{=}\mathcal{L}_i(\mathbf{x}_i^r,\mathbf{x}_0^r,\lambda_i^r)-\langle\mathbf{e}_i^r,\mathbf{x}_0^r-\mathbf{x}_i^r\rangle-\frac{1-L\eta}{2\eta}\left\|\mathbf{x}_0^r-\mathbf{x}_i^r\right\|^2\\ &\leq\mathcal{L}_i(\mathbf{x}_i^r,\mathbf{x}_0^r,\lambda_i^r)+\frac{\epsilon_1}{2L}-\frac{1-2L\eta}{2\eta}\left\|\mathbf{x}_0^r-\mathbf{x}_i^r\right\|^2.\end{aligned} \tag{88}$$

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 (74), the following holds:

$$\left\|\lambda_i^r - \lambda_i^{r-1}\right\|^2 \le 2L^2 \left\|\mathbf{x}_i^r - \mathbf{x}_i^{r-1}\right\|^2 + 4\epsilon_1. \tag{89}$$

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

$$\begin{aligned}
\left\|\nabla_{\mathbf{x}_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\| &= \left\|\nabla f_i(\mathbf{x}_i^r) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^r - \mathbf{x}_{0,i}^r)\right\| \\
&\overset{(74)}{=} \left\|\nabla f_i(\mathbf{x}_i^r) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^r - \mathbf{x}_{0,i}^r) - \nabla f_i(\mathbf{x}_i^{r+1}) - \lambda_i^r - \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r) + \mathbf{e}_i^{r+1}\right\| \\
&\le \frac{1+L\eta}{\eta}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\| + \sqrt{\epsilon_1},
\end{aligned} \tag{90}$$

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

$$\left\|\nabla_{\mathbf{x}_0}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\| = \left\|\lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^r - \mathbf{x}_{0,i}^r)\right\| = 0. \tag{91}$$

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

$$\left\|\nabla_{\mathbf{x}_0}\mathcal{L}(\mathbf{x}_0^r, \mathbf{x}_1^r, \ldots, \mathbf{x}_N^r, \lambda_1^r, \ldots, \lambda_N^r)\right\| = \left\|\frac{1}{N}\sum_{i=1}^N(\lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^r - \mathbf{x}_{0,i}^r))\right\| = 0. \tag{92}$$

Further we have:

$$\begin{aligned}
\left\|\nabla_{\lambda_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\| &= \left\|\mathbf{x}_i^r - \mathbf{x}_{0,i}^r\right\| \\
&\le \left\|\mathbf{x}_i^r - \mathbf{x}_0^{r-1}\right\| + \left\|\mathbf{x}_{0,i}^{r-1} - \mathbf{x}_{0,i}^r\right\| \\
&\le \eta\left\|\lambda_i^r - \lambda_i^{r-1}\right\| + \left\|\mathbf{x}_{0,i}^{r-1} - \mathbf{x}_{0,i}^r\right\| \\
&\le \eta(L\left\|\mathbf{x}_i^r - \mathbf{x}_i^{r-1}\right\| + 2\sqrt{\epsilon_1}) + \left\|\mathbf{x}_{0,i}^{r-1} - \mathbf{x}_{0,i}^r\right\|.
\end{aligned} \tag{93}$$

Summing (90) and (93), 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

$$\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\| \le \left\|\mathbf{x}_{0,i}^{r-1} - \mathbf{x}_{0,i}^r\right\| + \frac{1+L\eta}{\eta}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\| + L\eta\left\|\mathbf{x}_i^r - \mathbf{x}_i^{r-1}\right\| + (1+2\eta)\sqrt{\epsilon_1}. \tag{94}$$

Squaring both sides of the above inequality, we obtain:

$$\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 \le C_6\left(\left\|\mathbf{x}_{0,i}^{r-1} - \mathbf{x}_{0,i}^r\right\|^2 + \left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 + \left\|\mathbf{x}_i^r - \mathbf{x}_i^{r-1}\right\|^2 + \epsilon_1\right), \tag{95}$$

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

Apply (89) to Lemma 6 we have

$$\begin{aligned}
&\frac{1-2L\eta-4L^2\eta^2}{2\eta}\left\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\right\|^2 + \frac{1}{2\eta}\left\|\mathbf{x}_{0,i}^{r+} - \mathbf{x}_{0,i}^r\right\|^2 + \frac{1+8L\eta}{2L}\epsilon_1 \\
&\qquad \le \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+}, \lambda_i^{r+1}) + \frac{1+8L\eta}{L}\epsilon_1.
\end{aligned} \tag{96}$$

Define $C_7 = C_6 / \min\{\frac{1-2L\eta-4L^2\eta^2}{2\eta}, \frac{1}{2\eta}, \frac{1+8L\eta}{2L}\}$, apply (95) with Lemma 6 and Lemma 8 and sum up the iterations, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\sum_{r=0}^{T}\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 \leq C_7\sum_{r=0}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1})) + \frac{1+8L\eta}{L}\epsilon_1\right) \\ + C_7\sum_{r+1 \mod R=0}\frac{N-1}{N\eta}(\triangle\mathbf{x}_0^{r+1})^2.\end{aligned} \tag{97}$$

Next we bound the last term. Since $\triangle\mathbf{x}_0$ is a component of $\triangle$, then to bound $(\triangle\mathbf{x}_0^{r+1})^2$ it is sufficient to bound $(\triangle^{r+1})^2$. By iteratively applying Lemma 7 from $r = 0$ to $R-1$, we have

$$\triangle\mathbf{x}^{r+1} \leq \sum_{r=0}^{R-2}(\frac{A}{1-L\eta})^r\eta\frac{B}{1-L\eta}(\delta + \sqrt{\epsilon_1}) \tag{98}$$

From the definition of A in Lemma 7 we have:

$$\lambda_{\max}\left(\frac{1}{1-L\eta}A\right) = \frac{1}{1-L\eta}\sqrt{1+L^2\eta^2}\sqrt{2+L^2\eta^2+2L\eta} \triangleq C_8.$$

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

$$\begin{aligned}(\triangle\mathbf{x}_0^{r+1})^2 \leq \left\|\triangle^{r+1}\right\|^2 &\leq \left\|\sum_{r=0}^{R-2}(\frac{A}{1-L\eta})^r\eta\frac{B}{1-L\eta}(\delta + \sqrt{\epsilon_1})\right\|^2 \\ &\leq \left(\sum_{r=0}^{R-2}C_8^r\right)^2\eta^2\frac{\|B\|^2}{(1-L\eta)^2}(\delta^2 + \epsilon_1) \\ &\leq \frac{(C_8^{(R-1)}-1)^2 \times 5\eta^2(13+6L\eta+L^2\eta^2)}{(C_8-1)^2(1-L\eta)^2}(\delta^2+\epsilon_1).\end{aligned} \tag{99}$$

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

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{r=0}^{T}\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 \leq \frac{C_7}{T}\left(\mathcal{L}(\mathbf{x}_0^0, \mathbf{x}_i^0, \lambda_i^0) - \mathcal{L}(\mathbf{x}_i^T, \mathbf{x}_{0,i}^T, \lambda_i^T)\right) + \frac{C_7(1+8L\eta)}{L}\epsilon_1 \\ + \frac{5\eta C_7(13+6L\eta+L^2\eta^2)(N-1)(C_8^{(R-1)}-1)^2}{NR(C_8-1)^2(1-L\eta)^2}(\delta^2+\epsilon_1).\end{aligned} \tag{100}$$

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

$$\begin{aligned}\frac{1}{NT}\sum_{i=1}^{N}\sum_{r=0}^{T}\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 \leq \frac{C_7(f(\mathbf{x}_0^0) - f(\mathbf{x}_0^T))}{T} + \frac{C_7(1+8L\eta)}{L}\epsilon_1 \\ + \frac{5\eta C_7(13+6L\eta+L^2\eta^2)(N-1)(C_8^{(R-1)}-1)^2}{NR(C_8-1)^2(1-L\eta)^2}(\delta^2+\epsilon_1).\end{aligned} \tag{101}$$

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

$$
\begin{aligned}
\|\nabla f(\mathbf{x}_0^r)\|^2 &\leq 2\left\|\nabla f(\mathbf{x}_0^r) - \frac{1}{N}\sum_{i=1}^{N}\nabla_{\mathbf{x}_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 + \frac{2}{N}\sum_{i=1}^{N}\left\|\nabla_{\mathbf{x}_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 \\
&\leq \frac{4}{N}\sum_{i=1}^{N}\|\nabla f_i(\mathbf{x}_0^r) - \nabla f_i(\mathbf{x}_i^r)\|^2 + 4\left\|\frac{1}{N\eta}\sum_{i=1}^{N}(\eta\lambda_i^r + \mathbf{x}_i^r - \mathbf{x}_{0,i}^r)\right\|^2 \\
&\quad + \frac{2}{N}\sum_{i=1}^{N}\left\|\nabla_{\mathbf{x}_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 \\
&\overset{(a)}{\leq} \frac{4L^2}{N}\sum_{i=1}^{N}\|\mathbf{x}_0^r - \mathbf{x}_i^r\|^2 + \frac{2}{N}\sum_{i=1}^{N}\left\|\nabla_{\mathbf{x}_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 \\
&= \frac{4L^2}{N}\sum_{i=1}^{N}\left\|\nabla_{\lambda_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 + \frac{2}{N}\sum_{i=1}^{N}\left\|\nabla_{\mathbf{x}_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 \\
&\leq \frac{4L^2}{N}\sum_{i=1}^{N}\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2,
\end{aligned}
\tag{102}
$$

where in (a) we use the same argument in (91) and (92).

Therefore Theorem 1 is proved. During the proof, we need all $C_2, \ldots, C_8 > 0$, 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 (72) 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.

$$
\begin{aligned}
&C_2 \geq 4L^2C_7, \qquad C_3 = C_8, \qquad C_4 \geq \frac{C_2(1+8L\eta)}{L} \\
&C_5 = \frac{C_2(13+6L\eta+L^2\eta^2)}{(C_8-1)^2(1-L\eta)^2}, \qquad C_6 \geq \max\{(\frac{1+L\eta}{\eta})^2, (1+2\eta)^2, L^2\eta^2\} \\
&C_7 = C_6/\min\{\frac{1-2L\eta-4L^2\eta^2}{2\eta}, \frac{1}{2\eta}, \frac{1+8L\eta}{2L}\} \\
&C_8 = \frac{1}{1-L\eta}\sqrt{1+L^2\eta^2}\sqrt{2+L^2\eta^2+2L\eta},
\end{aligned}
$$

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

F.2. Proof of Theorem 2

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

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

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:*

$$
\begin{aligned}
\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r}, \mathbf{x}_{0,i}, \lambda_i^{r}) \le & -\frac{1}{2\eta}\left\|\mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^{r}\right\|^2 - \left(\frac{1}{2\eta} + \frac{1}{\gamma} - L - \frac{3\eta}{\gamma^2}\right)\left\|\mathbf{x}_i^{r,Q} - \mathbf{x}_i^{r,Q-1}\right\|^2 \\
& - (\frac{1}{2\eta} + \frac{1}{\gamma} - L - 9Q^2L^2\eta)\sum_{q=1}^{Q-1}\left\|\mathbf{x}_i^{r,q} - \mathbf{x}_i^{r,q-1}\right\|^2 \\
& + \left(9Q^2L^2\eta + \frac{3\eta}{\gamma^2}\right)\left\|\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}\right\|^2 + \frac{1}{2L}\sum_{q=0}^{Q-2}\|\nabla f_i(\mathbf{x}_i^{r,q}) - g_i^{r,q}\|^2 \\
& + (\frac{1}{2L} + 9\eta)\left\|g_i^{r,Q-1} - \nabla f_i(\mathbf{x}_i^{r,Q-1})\right\|^2 + 9\eta\left\|g_i^{r-1,Q-1} - \nabla f_i(\mathbf{x}_i^{r-1,Q-1})\right\|^2 \\
& + \left\langle \lambda_i^{r+1} + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+1}), \mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^{r}\right\rangle .
\end{aligned}
\tag{103}
$$

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* (16)*, 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*

$$
\mathbb{E}\left\|g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1})\right\|^2 \le \frac{L^2}{B}\sum_{\tau=\{r_0,1\}}^{\{r,q+1\}}\mathbb{E}\left\|\mathbf{x}_i^{\tau} - \mathbf{x}_i^{\tau-1}\right\|^2 . \tag{104}
$$

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:*

$$
\begin{aligned}
\mathbb{E}\, f(\mathbf{x}_0^r) \le\ & \mathbb{E}\,\mathcal{L}(\mathbf{x}_0^r, \mathbf{x}_1^r, \ldots, \mathbf{x}_N^r, \lambda_1^r, \ldots, \lambda_N^r) - \frac{1-3L\eta}{2N\eta}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^r - \mathbf{x}_0^r\right\|^2 \\
& + \frac{(1+L\gamma)^2 + L^2\gamma^2}{4L\gamma^2}\left[\frac{1}{B}\sum_{\tau=\{r_0,1\}}^{\{r-1,Q-1\}}\mathbb{E}\left\|\mathbf{x}_i^{\tau} - \mathbf{x}_i^{\tau-1}\right\|^2 + \mathbb{E}\left\|\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}\right\|^2\right].
\end{aligned}
\tag{105}
$$

F.2.1. Proof of Lemma 10

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

$$
\begin{aligned}
& \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r}, \mathbf{x}_{0,i}^{r}, \lambda_i^{r}) \\
& = \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r}, \lambda_i^{r}) - \mathcal{L}_i(\mathbf{x}_i^{r}, \mathbf{x}_{0,i}^{r}, \lambda_i^{r}) + \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r}, \lambda_i^{r}) \\
& \quad + \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r}, \lambda_i^{r+1}).
\end{aligned}
\tag{106}
$$

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 $\mathbf{x}_i^{r,1} = \mathbf{x}_i^{r}$):

$$
\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r}, \lambda_i^{r}) - \mathcal{L}_i(\mathbf{x}_i^{r}, \mathbf{x}_{0,i}^{r}, \lambda_i^{r}) = \sum_{q=1}^{Q}\left(\mathcal{L}_i(\mathbf{x}_i^{r,q+1}, \mathbf{x}_{0,i}^{r}, \lambda_i^{r}) - \mathcal{L}_i(\mathbf{x}_i^{r,q}, \mathbf{x}_{0,i}^{r}, \lambda_i^{r})\right). \tag{107}
$$

Each term on the right hand side (RHS) of the above equality can be bounded by (see a similar arguments in (80)):

$$
\begin{aligned}
\mathcal{L}_i(\mathbf{x}_i^{r,q+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^{r,q}, \mathbf{x}_{0,i}^r, \lambda_i^r) &\leq \left\langle \nabla f_i(\mathbf{x}_i^{r,q}) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}^{r,q+1} - \mathbf{x}_{0,i}^r), \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\rangle \\
&\quad - \frac{1 - L\eta}{2\eta} \left\| \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\|^2 \\
&\overset{(a)}{=} \left\langle \nabla f_i(\mathbf{x}_i^{r,q}) - g_i^{r,q} - \frac{1}{\gamma}(\mathbf{x}^{r,q+1} - \mathbf{x}_i^{r,q}), \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\rangle \\
&\quad - (\frac{1}{2\eta} - \frac{L}{2}) \left\| \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\|^2 \\
&= \left\langle \nabla f_i(\mathbf{x}_i^{r,q}) - g_i^{r,q}, \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\rangle - (\frac{1}{2\eta} + \frac{1}{\gamma} - \frac{L}{2}) \left\| \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\|^2 \\
&\overset{(b)}{\leq} \frac{1}{2L} \left\| \nabla f_i(\mathbf{x}_i^{r,q}) - g_i^{r,q} \right\|^2 - (\frac{1}{2\eta} + \frac{1}{\gamma} - L) \left\| \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\|^2
\end{aligned}
\tag{108}
$$

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

$$
\lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r,q+1} - \mathbf{x}_{0,i}^r) + g_i^{r,q} + \frac{1}{\gamma}(\mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q}) = 0; \tag{109}
$$

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 (106) is given by

$$
\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) - \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r) \leq \frac{1}{2L} \sum_{q=1}^{Q} \left\| \nabla f_i(\mathbf{x}_i^{r,q}) - g_i^{r,q} \right\|^2 - (\frac{1}{2\eta} + \frac{1}{\gamma} - L) \sum_{q=1}^{Q} \left\| \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\|^2. \tag{110}
$$

The other two differences in (106) can be explicitly expressed as:

$$
\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^r) = \eta \left\| \lambda_i^{r+1} - \lambda_i^r \right\|^2, \tag{111}
$$

$$
\begin{aligned}
&\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^r, \lambda_i^{r+1}) \\
&= -\frac{1}{2\eta} \left\| \mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^r \right\|^2 + \left\langle \lambda_i^{r+1} + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+1}), \mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^r \right\rangle.
\end{aligned}
\tag{112}
$$

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

$$
\lambda_i^{r+1} = \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^{r,Q} - \mathbf{x}_{0,i}^r) \overset{(109)}{=} -\frac{1}{\gamma}(\mathbf{x}_i^{r,Q} - \mathbf{x}_i^{r,Q-1}) - g^{r,Q-1}. \tag{113}
$$

Using the above property, we have

$$
\begin{aligned}
\left\| \lambda_i^{r+1} - \lambda_i^r \right\|^2 &= \left\| \frac{1}{\gamma}(\mathbf{x}_i^{r,Q} - \mathbf{x}_i^{r,Q-1}) + g_i^{r,Q-1} - \frac{1}{\gamma}(\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}) - g_i^{r-1,Q-1} \right\|^2 \\
&\overset{(a)}{\leq} 3 \left\| g_i^{r,Q-1} - g_i^{r-1,Q-1} \right\|^2 + \frac{3}{\gamma^2} \left\| \mathbf{x}_i^{r,Q} - \mathbf{x}_i^{r,Q-1} \right\|^2 + \frac{3}{\gamma^2} \left\| \mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1} \right\|^2.
\end{aligned}
\tag{114}
$$

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

$$\begin{aligned}\left\|g_i^{r,Q-1} - g_i^{r-1,Q-1}\right\|^2 &= \left\|g_i^{r,Q-1} - \nabla f_i(\mathbf{x}_i^{r,Q-1}) + \nabla f_i(\mathbf{x}_i^{r,Q-1}) - \nabla f_i(\mathbf{x}_i^{r-1,Q-1}) + \nabla f_i(\mathbf{x}_i^{r-1,Q-1}) - g_i^{r-1,Q-1}\right\|^2 \\ &\overset{(a)}{\leq} 3\left\|g_i^{r,Q-1} - \nabla f_i(\mathbf{x}_i^{r,Q-1})\right\|^2 + 3\left\|g_i^{r-1,Q-1} - \nabla f_i(\mathbf{x}_i^{r-1,Q-1})\right\|^2 + 3L^2\left\|\mathbf{x}_i^{r,Q-1} - \mathbf{x}_i^{r-1,Q-1}\right\|^2 \\ &\overset{(b)}{\leq} 3\left\|g_i^{r,Q-1} - \nabla f_i(\mathbf{x}_i^{r,Q-1})\right\|^2 + 3\left\|g_i^{r-1,Q-1} - \nabla f_i(\mathbf{x}_i^{r-1,Q-1})\right\|^2 \\ &\quad + 3Q^2L^2\sum_{q=1}^{Q-1}\left\|\mathbf{x}_i^{r,q} - \mathbf{x}_i^{r,q-1}\right\|^2 + 3Q^2L^2\left\|\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}\right\|^2,\end{aligned} \tag{115}$$

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 (115) to (114) and sum the three parts, we have

$$\begin{aligned}\mathcal{L}_i(\mathbf{x}_i^{r+1}, \mathbf{x}_{0,i}^{r+1}, \lambda_i^{r+1}) - \mathcal{L}_i(\mathbf{x}_i^{r}, \mathbf{x}_{0,i}^{r}, \lambda_i^{r}) &\leq -\frac{1}{2\eta}\left\|\mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^{r}\right\|^2 - (\frac{1}{2\eta} + \frac{1}{\gamma} - L - \frac{3\eta}{\gamma^2})\left\|\mathbf{x}_i^{r,Q} - \mathbf{x}_i^{r,Q-1}\right\|^2 \\ &\quad - (\frac{1}{2\eta} + \frac{1}{\gamma} - L - 9Q^2L^2\eta)\sum_{q=1}^{Q-1}\left\|\mathbf{x}_i^{r,q} - \mathbf{x}_i^{r,q-1}\right\|^2 \\ &\quad + (9Q^2L^2\eta + \frac{3\eta}{\gamma^2})\left\|\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}\right\|^2 + \frac{1}{2L}\sum_{q=0}^{Q-2}\left\|\nabla f_i(\mathbf{x}_i^{r,q}) - g_i^{r,q}\right\|^2 \\ &\quad + (\frac{1}{2L} + 9\eta)\left\|g_i^{r,Q-1} - \nabla f_i(\mathbf{x}_i^{r,Q-1})\right\|^2 + 9\eta\left\|g_i^{r-1,Q-1} - \nabla f_i(\mathbf{x}_i^{r-1,Q-1})\right\|^2 \\ &\quad + \left\langle \lambda_i^{r+1} + \frac{1}{\eta}(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^{r+1}), \mathbf{x}_{0,i}^{r+1} - \mathbf{x}_{0,i}^{r} \right\rangle,\end{aligned} \tag{116}$$

which complete the proof of Lemma 10.

F.2.2. Proof of Lemma 11

To study $\mathbb{E}\left\|g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q})\right\|^2$, 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

$$g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1}) = g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q+1}) + \frac{1}{B}\sum_{b=1}^{B}(h_i(\mathbf{x}_i^{r,q+1}; \xi_{i,b}^{r,q}) - h_i(\mathbf{x}_i^{r,q}; \xi_{i,b}^{r,q})). \tag{117}$$

Take expectation on both sides, we have

$$\begin{aligned}\mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^{B}}[g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1})] &= g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q+1}) + \mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^{B}}[\frac{1}{B}\sum_{b=1}^{B}(h_i(\mathbf{x}_i^{r,q+1}; \xi_{i,b}^{r,q}) - h_i(\mathbf{x}_i^{r,q}; \xi_{i,b}^{r,q}))] \\ &= g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q+1}) + \nabla f_i(\mathbf{x}_i^{r,q+1}) - \nabla f_i(\mathbf{x}_i^{r,q}) \\ &= g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q})\end{aligned} \tag{118}$$

By using the fact that $\mathbb{E}[X^2] = [\mathbb{E}\, X]^2 + \mathbb{E}[[X - \mathbb{E}\, X]^2]$ and substitute (118) we have

$$\begin{aligned}
&\mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^{B}} \left\| g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1}) \right\|^2 \\
&= \left\| \mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^{B}} [g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1})] \right\|^2 + \mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^{B}} \left\| g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1}) - \mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^{B}} [g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1})] \right\|^2 \\
&\overset{(118)}{=} \|g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q})\|^2 + \mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^{B}} \left\| \frac{1}{B} \sum_{b=1}^{B} (h_i(\mathbf{x}_i^{r,q+1}; \xi_{i,b}^{r,q} - h_i(\mathbf{x}_i^{r,q}; \xi_{i,b}^{r,q})) - \nabla f_i(\mathbf{x}_i^{r,q+1}) + \nabla f_i(\mathbf{x}_i^{r,q}) \right\|^2 \\
&\overset{(a)}{\leq} \|g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q})\|^2 + \frac{1}{B^2} \sum_{b=1}^{B} \mathbb{E}_{\{\xi_{i,b}^{r,q}\}_{b=1}^{B}} \left\| h_i(\mathbf{x}_i^{r,q+1}; \xi_{i,b}^{r,q}) - h_i(\mathbf{x}_i^{r,q}; \xi_{i,b}^{r,q})) \right\|^2 \\
&\overset{(b)}{\leq} \|g_i^{r,q} - \nabla f_i(\mathbf{x}_i^{r,q})\|^2 + \frac{L^2}{B} \left\| \mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q} \right\|^2,
\end{aligned}$$

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

$$\mathbb{E} \left\| g_i^{r,q+1} - \nabla f_i(\mathbf{x}_i^{r,q+1}) \right\|^2 \leq \frac{L^2}{B} \sum_{\tau=\{r_0,1\}}^{\{r,q+1\}} \mathbb{E} \left\| \mathbf{x}_i^{\tau} - \mathbf{x}_i^{\tau-1} \right\|^2, \tag{119}$$

which completes the proof.

F.2.3. Proof of Lemma 12

Applying A1, we have

$$\begin{aligned}
f_i(\mathbf{x}_0^r) &\leq f_i(\mathbf{x}_i^r) + \langle \nabla f_i(\mathbf{x}_i^r), \mathbf{x}_0^r - \mathbf{x}_i^r \rangle + \frac{L}{2} \|\mathbf{x}_0^r - \mathbf{x}_i^r\|^2 \\
&= \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_0^r, \lambda_i^r) - \langle \nabla f_i(\mathbf{x}_i^r) + \lambda_i^r, \mathbf{x}_0^r - \mathbf{x}_i^r \rangle - \frac{1 - L\eta}{2\eta} \|\mathbf{x}_0^r - \mathbf{x}_i^r\|^2 \\
&\leq \mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_0^r, \lambda_i^r) + \frac{1}{4L} \|\nabla f_i(\mathbf{x}_i^r) + \lambda_i^r\|^2 - \frac{1 - 3L\eta}{2\eta} \|\mathbf{x}_0^r - \mathbf{x}_i^r\|^2.
\end{aligned} \tag{120}$$

Then notice $\mathbf{x}_i^r = \mathbf{x}_i^{r-1,Q}$ and apply (113), we can bound $\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^r) + \lambda_i^r\right\|^2$ by the following:

$$
\begin{aligned}
\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^r) + \lambda_i^r\right\|^2 &\overset{(113)}{=} \mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{r-1,Q}) - g_i^{r-1,Q-1} - \frac{1}{\gamma}(\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1})\right\|^2 \\
&\overset{(a)}{\le} (1 + \frac{(1+L\gamma)^2}{L^2\gamma^2})\,\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{r-1,Q-1}) - g_i^{r-1,Q-1}\right\|^2 \\
&\quad + (1 + \frac{L^2\gamma^2}{(1+L\gamma)^2})(1 + \frac{1}{L\gamma})\,\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{r-1,Q}) - \nabla f_i(\mathbf{x}_i^{r-1,Q-1})\right\|^2 \\
&\quad + \frac{(1 + \frac{L^2\gamma^2}{(1+L\gamma)^2})(1+L\gamma)}{\gamma^2}\,\mathbb{E}\left\|\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}\right\|^2 \\
&\overset{(b)}{\le} \frac{(1+L\gamma)^2 + L^2\gamma^2}{B\gamma^2}\sum_{\tau=\{r_0,1\}}^{\{r-1,Q-1\}}\mathbb{E}\left\|\mathbf{x}_i^{\tau} - \mathbf{x}_i^{\tau-1}\right\|^2 \\
&\quad + (1 + \frac{L^2\gamma^2}{(1+L\gamma)^2})\left((1+\frac{1}{L\gamma})L^2 + \frac{1+L\gamma}{\gamma^2}\right)\mathbb{E}\left\|\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}\right\|^2 \\
&= \frac{(1+L\gamma)^2 + L^2\gamma^2}{B\gamma^2}\sum_{\tau=\{r_0,1\}}^{\{r-1,Q-1\}}\mathbb{E}\left\|\mathbf{x}_i^{\tau} - \mathbf{x}_i^{\tau-1}\right\|^2 \\
&\quad + \frac{(1+L\gamma)^2 + L^2\gamma^2}{\gamma^2}\,\mathbb{E}\left\|\mathbf{x}_i^{r-1,Q} - \mathbf{x}_i^{r-1,Q-1}\right\|^2,
\end{aligned}
\tag{121}
$$

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

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

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

Substitute (121) to (120) and average over the agents, Lemma 12 is proved.

F.2.4. PROOF OF THEOREM 2

By the update step of $\mathbf{x}_0^r$, following (91) 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

$$
\begin{aligned}
\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 &= \left\|\nabla_{\mathbf{x}_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 + \left\|\nabla_{\lambda_i}\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 \\
&= \left\|\nabla f_i(\mathbf{x}_i^r) + \lambda_i^r + \frac{1}{\eta}(\mathbf{x}_i^r - \mathbf{x}_{0,i}^r)\right\|^2 + \left\|\mathbf{x}_i^r - \mathbf{x}_{0,i}^r\right\|^2 \\
&\overset{(a)}{=} \left\|\nabla f_i(\mathbf{x}_i^r) - g_i^{r,0} - \frac{\eta+\gamma}{\eta\gamma}(\mathbf{x}_i^{r,1} - \mathbf{x}_i^r)\right\|^2 + \left\|\mathbf{x}_i^r - \mathbf{x}_{0,i}^r + \mathbf{x}_{0,i}^{r-1} - \mathbf{x}_{0,i}^{r-1}\right\|^2 \\
&\le \left\|\nabla f_i(\mathbf{x}_i^r) - g_i^{r,0} - \frac{\eta+\gamma}{\eta\gamma}(\mathbf{x}_i^{r,1} - \mathbf{x}_i^r)\right\|^2 + 2\left\|\mathbf{x}_i^r - \mathbf{x}_{0,i}^{r-1}\right\|^2 + 2\left\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r-1}\right\|^2 \\
&\le 2\left\|\nabla f_i(\mathbf{x}_i^r) - g_i^{r,0}\right\|^2 + 2(\frac{\eta+\gamma}{\eta\gamma})^2\left\|\mathbf{x}_i^{r,1} - \mathbf{x}_i^r\right\|^2 + 2\eta^2\left\|\lambda_i^r - \lambda_i^{r-1}\right\|^2 + 2\left\|\mathbf{x}_{0,i}^r - \mathbf{x}_{0,i}^{r-1}\right\|^2.
\end{aligned}
\tag{122}
$$

where in (a), the first term is obtained by plugging in (113) 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 (114), (115),

$$\begin{aligned}
\mathbb{E}\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^r,\mathbf{x}_{0,i}^r,\lambda_i^r)\right\|^2 \leq{}& 2\,\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^r)-g_i^{r,0}\right\|^2+2(\frac{\eta+\gamma}{\eta\gamma})^2\,\mathbb{E}\left\|\mathbf{x}_i^{r,1}-\mathbf{x}_i^r\right\|^2+2\,\mathbb{E}\left\|\mathbf{x}_{0,i}^r-\mathbf{x}_{0,i}^{r-1}\right\|^2\\
&+\frac{6\eta^2}{\gamma^2}(\gamma^2\,\mathbb{E}\left\|g_i^{r-1,Q-1}-g_i^{r-2,Q-1}\right\|^2+\mathbb{E}\left\|\mathbf{x}_i^{r-1,Q}-\mathbf{x}_i^{r-1,Q-1}\right\|^2+E\left\|\mathbf{x}_i^{r-2,Q}-\mathbf{x}_i^{r-2,Q-1}\right\|^2)\\
&\overset{(a)}{\leq}\frac{2L^2}{B}\sum_{\tau=\{r_0,1\}}^{\{r,0\}}\mathbb{E}\left\|\mathbf{x}_i^\tau-\mathbf{x}_i^{\tau-1}\right\|^2+2(\frac{\eta+\gamma}{\eta\gamma})^2\,\mathbb{E}\left\|\mathbf{x}_i^{r,1}-\mathbf{x}_i^r\right\|^2+2\,\mathbb{E}\left\|\mathbf{x}_{0,i}^r-\mathbf{x}_{0,i}^{r-1}\right\|^2\\
&+\frac{6\eta^2}{\gamma^2}(\mathbb{E}\left\|\mathbf{x}_i^{r-1,Q}-\mathbf{x}_i^{r-1,Q-1}\right\|^2+\mathbb{E}\left\|\mathbf{x}_i^{r-2,Q}-\mathbf{x}_i^{r-2,Q-1}\right\|^2)\\
&+18\eta^2\left(\mathbb{E}\left\|g_i^{r-1,Q-1}-\nabla f_i(\mathbf{x}_i^{r-1,Q-1})\right\|^2+\mathbb{E}\left\|g_i^{r-2,Q-1}-\nabla f_i(\mathbf{x}_i^{r-2,Q-1})\right\|^2\right)\\
&+18\eta^2Q^2L^2\left(\sum_{q=1}^{Q-1}\mathbb{E}\left\|\mathbf{x}_i^{r-1,q}-\mathbf{x}_i^{r-1,q-1}\right\|^2+\mathbb{E}\left\|\mathbf{x}_i^{r-2,Q}-\mathbf{x}_i^{r-2,Q-1}\right\|^2\right),
\end{aligned}\tag{123}$$

where we substitute Lemma 11 and (115) in (a).

Taking expectation of (103), summing over $r=0$ to $r=T-1$ and average over the agents, we have the following

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N\mathbb{E}[\mathcal{L}_i(\mathbf{x}_i^T,\mathbf{x}_{0,i}^T,\lambda_i^T)-\mathcal{L}_i(\mathbf{x}_i^0,\mathbf{x}_{0,i}^0,\lambda_i^0)]\leq{}&-\frac{1}{2\eta}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_0^{r+1}-\mathbf{x}_0^r\right\|^2\\
&-(\frac{1}{2\eta}+\frac{1}{\gamma}-L-\frac{6\eta}{\gamma^2}-9Q^2L^2\eta)\frac{1}{N}\sum_{i=1}^N\sum_{q=0}^{Q-1}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_i^{r,q+1}-\mathbf{x}_i^{r,q-1}\right\|^2\\
&+(\frac{1}{2L}+18\eta)\frac{1}{N}\sum_{i=1}^N\sum_{r=0}^{T-1}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{r,q})-g_i^{r,q}\right\|^2\\
&+\sum_{r=0}^{T-1}\frac{1}{N}\mathbb{E}\left\langle\sum_{i=1}^N(\lambda_i^{r+1}+\frac{1}{\eta}(\mathbf{x}_i^{r+1}-\mathbf{x}_{0,i}^{r+1})),\mathbf{x}_{0,i}^{r+1}-\mathbf{x}_{0,i}^r\right\rangle\\
\overset{(a)}{\leq}{}&-(\frac{1}{2\eta}+\frac{1}{\gamma}-L-\frac{6\eta}{\gamma^2}-9Q^2L^2\eta)\frac{1}{N}\sum_{i=1}^N\sum_{q=0}^{Q-1}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_i^{r,q+1}-\mathbf{x}_i^{r,q-1}\right\|^2\\
&-\frac{1}{2\eta}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_0^{r+1}-\mathbf{x}_0^r\right\|^2\\
&+\frac{(1+18L\eta)LIQ}{2B}\frac{1}{N}\sum_{i=1}^N\sum_{r=0}^{T-1}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\mathbf{x}_i^{r,q+1}-\mathbf{x}_i^{r,q-1}\right\|^2\\
={}&-\frac{C_{10}}{N}\sum_{i=1}^N\sum_{q=0}^{Q-1}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_i^{r,q+1}-\mathbf{x}_i^{r,q-1}\right\|^2-\frac{1}{2\eta}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_0^{r+1}-\mathbf{x}_0^r\right\|^2,
\end{aligned}\tag{124}$$

where in (a) we apply Lemma 11 and (91).

Finally, in the last equation of (124), 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

$$\mathbb{E}[f(\mathbf{x}_0^T) - f(\mathbf{x}_0^0)] \leq \, - \frac{C_{10} - \frac{(1+L\gamma)^2 + L^2\gamma^2}{4BL\gamma^2}}{N} \sum_{i=1}^{N} \sum_{q=0}^{Q-1} \sum_{r=0}^{T-1} \mathbb{E}\left\|\mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q-1}\right\|^2 - \frac{1}{2\eta}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_0^{r+1} - \mathbf{x}_0^{r}\right\|^2, \tag{125}$$

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 (123) and (125), we can find a positive constant C_{11} satisfying

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

so that the following holds

$$\begin{aligned} \frac{C_{11}}{NT}\sum_{r=0}^{T}\sum_{i=1}^{N}\mathbb{E}\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^r, \mathbf{x}_{0,i}^r, \lambda_i^r)\right\|^2 &\leq \frac{C_{10} - \frac{(1+L\gamma)^2 + L^2\gamma^2}{4BL\gamma^2}}{NT}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_i^{r,q+1} - \mathbf{x}_i^{r,q-1}\right\|^2 \\ &\quad + \frac{1}{2\eta T}\sum_{r=0}^{T-1}\mathbb{E}\left\|\mathbf{x}_0^{r+1} - \mathbf{x}_0^{r}\right\|^2 \\ &\leq \frac{1}{T}(f(\mathbf{x}_0^0) - \mathbb{E}\, f(\mathbf{x}_0^T)) \leq \frac{1}{T}(f(\mathbf{x}_0^0) - f(\mathbf{x}^\star)). \end{aligned} \tag{126}$$

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 2 is proved.

During the prove we need

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

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

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.

G. Numerical Experiments

G.1. Penalized Logistic Regression

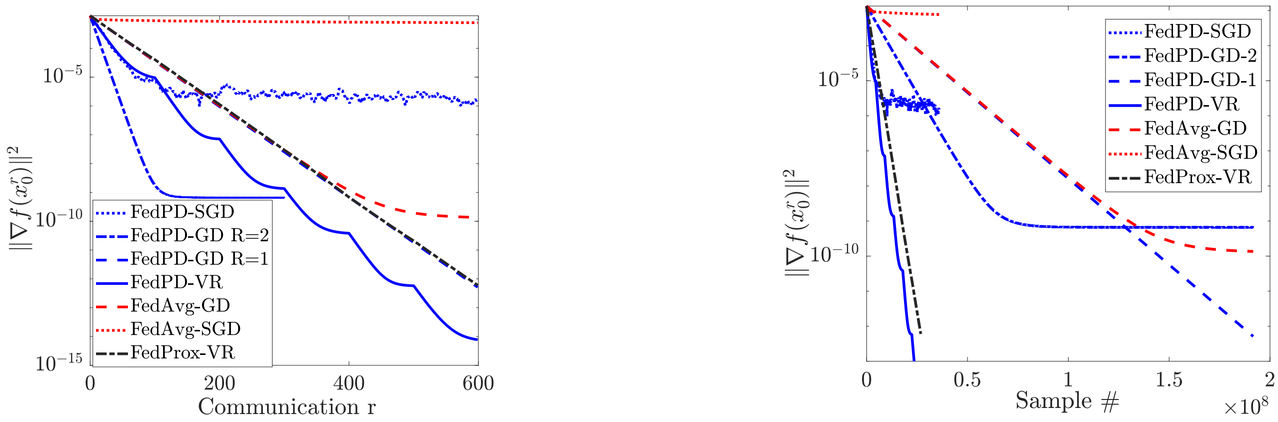
In this experiment, we consider the penalized regression problem (Antoniadis et al., 2011), whose loss function evaluated on a single sample $(\mathbf{a}, b) = \xi$ is given by:

$$F(\mathbf{x}; (\mathbf{a}, b)) = \log(1 + \exp(-b\mathbf{x}^T\mathbf{a})) + \sum_{d=1}^{D}\frac{\beta\alpha(\mathbf{x}[d])^2}{1 + \alpha(\mathbf{x}[d])^2}. \tag{127}$$

Here $\mathbf{x}[d]$ denotes the d^{th} 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 i^{th} 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.

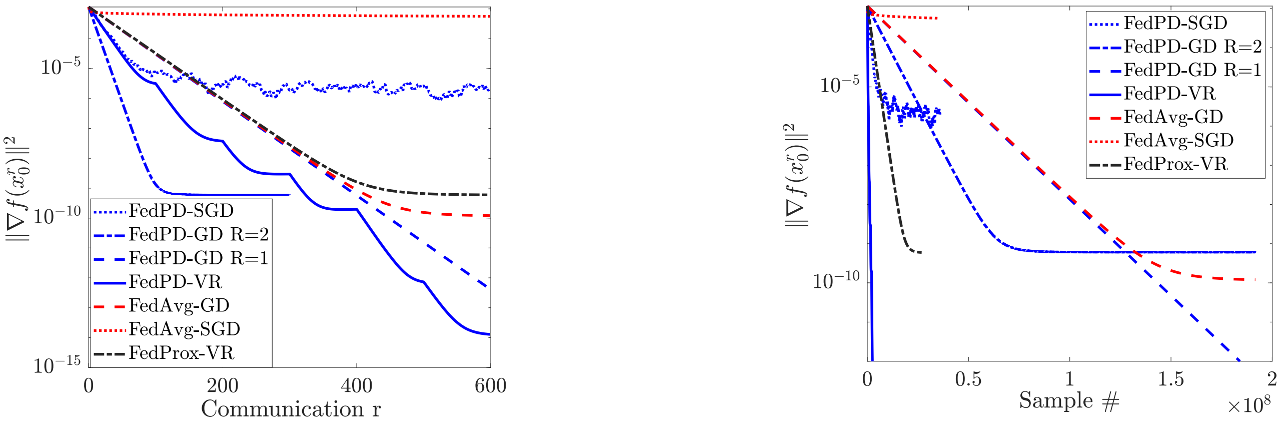
The total number of iterations T is set as 600 for all algorithms. We choose the stepsize to be $\eta = 4$ for FedAvg-GD with local update number $Q = 8$ and for FedAvg-SGD we use diminishing stepsize $\eta = 4/\sqrt{Qr + q + 1}$ with $Q = 600$. For FedProx we use VR algorithm as the local solver and 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 also tried other 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.



(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.

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.

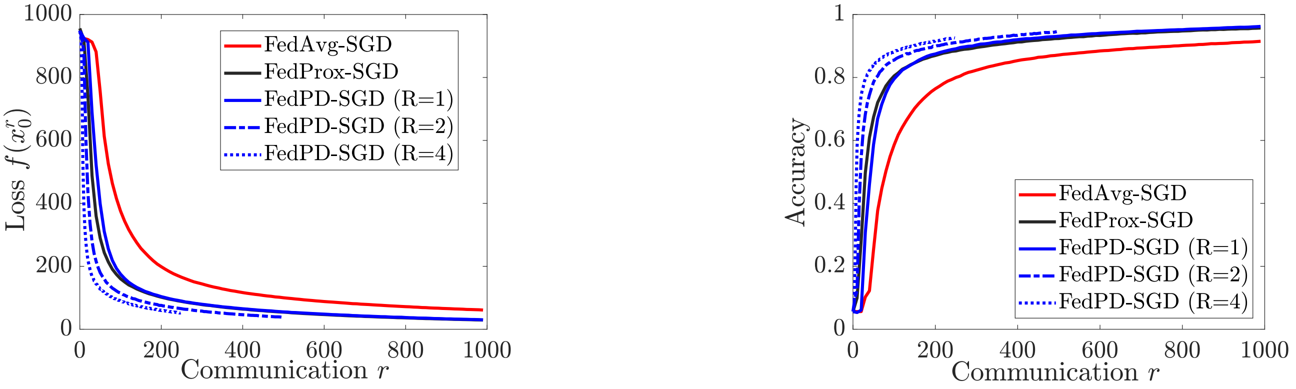
G.2. Handwritten Character Classification

In this subsection, we state the procedure in generating the figures for the handwritten character classification problem. As stated in Section 4, we train a neural network that classifies 62 classes of handwritten characters, including numbers 1–10 and the upper- and lower-case letters A–Z and a–z. 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. We assign the data to 90 agents, where 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.

The numerical results shown in Fig. 2 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. 2(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. 2(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 frequency to $R = 1$ and $R = 2$.

The results shown in Fig. 6 were generated by running Python codes (using the the PyTorch package [1]) with AMD EPYC 7702 CPUs and an NVIDIA V100 GPU.

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 frequency for FedPD to be $R = 1$, $R = 2$ and $R = 4$. 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. 6, 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 6: The convergence results of the algorithms on training neural networks on the federated handwritten characters classification problem.

[1] PyTorch: An Imperative Style, High-Performance Deep Learning Library, `https://pytorch.org/`

References

Antoniadis, A., Gijbels, I., and Nikolova, M. Penalized likelihood regression for generalized linear models with non-quadratic penalties. *Annals of the Institute of Statistical Mathematics*, 63(3):585–615, 2011.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konecny, J., Mazzocchi, S., McMahan, H. B., et al. Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046*, 2019.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Caldas, S., Wu, P., Li, T., Konečnỳ, J., McMahan, H. B., Smith, V., and Talwalkar, A. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points i. *Mathematical Programming*, Jun 2019. ISSN 1436-4646. doi: 10.1007/s10107-019-01406-y. URL `https://doi.org/10.1007/s10107-019-01406-y`.

Cen, S., Zhang, H., Chi, Y., Chen, W., and Liu, T.-Y. Convergence of distributed stochastic variance reduced methods without sampling extra data. *arXiv preprint arXiv:1905.12648*, 2019.

Chen, T., Giannakis, G., Sun, T., and Yin, W. LAG: Lazily aggregated gradient for communication-efficient distributed learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 5050–5060. Curran Associates, Inc., 2018.

Haddadpour, F., Kamani, M. M., Mahdavi, M., and Cadambe, V. Trading redundancy for communication: Speeding up distributed SGD for non-convex optimization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2545–2554, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Khaled, A., Mishchenko, K., and Richtárik, P. First analysis of local GD on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*, 2019.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. FedDANE: A federated newton-type method. *arXiv preprint arXiv:2001.01920*, 2020.

Li, W., Liu, Y., Tian, Z., and Ling, Q. COLA: Communication-censored linearized admm for decentralized consensus optimization. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5237–5241, May 2019. doi: 10.1109/ICASSP.2019.8682575.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2017.

Liang, X., Shen, S., Liu, J., Pan, Z., Chen, E., and Cheng, Y. Variance reduced local SGD with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

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

Sahu, A. K., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., and Smith, V. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

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

Sharma, P., Khanduri, P., Bulusu, S., Rajawat, K., and Varshney, P. K. Parallel restarted spider–communication efficient distributed nonconvex optimization with optimal computation complexity. *arXiv preprint arXiv:1912.06036*, 2019.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. Federated multi-task learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4424–4434. Curran Associates, Inc., 2017.

Stich, S. U. Local sgd converges fast and communicates little. *ICLR 2019 - International Conference on Learning Representations*, pp. 17, 2019.

Wang, J. and Joshi, G. Cooperative SGD: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Wang, S., Tuor, T., Salonidis, T., Leung, K. K., Makaya, C., He, T., and Chan, K. When edge meets learning: Adaptive control for resource-constrained distributed machine learning. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pp. 63–71. IEEE, 2018.

Wang, X., Han, Y., Wang, C., Zhao, Q., Chen, X., and Chen, M. In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning. *IEEE Network*, 33(5):156–165, Sep. 2019. doi: 10.1109/MNET.2019.1800286.

Yu, H., Jin, R., and Yang, S. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7184–7193, Long Beach, California, USA, 09–15 Jun 2019a. PMLR.

Yu, H., Yang, S., and Zhu, S. 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*, volume 33, pp. 5693–5700, 2019b.

Yuan, K., Ying, B., Vlaski, S., and Sayed, A. H. Stochastic gradient descent with finite samples sizes. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, Sep. 2016. doi: 10.1109/MLSP.2016.7738878.

Zhao, B., Mopuri, K. R., and Bilen, H. iDLG: Improved deep leakage from gradients, 2020.