Online Supplementary Materials

APPENDIX E
EXAMPLES OF COST FUNCTIONS SATISFY A4

In this part, we provide a commonly used function that satisfies A4.

Logistic Regression

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

$$f_i(\mathbf{x}) = \frac{1}{|\mathcal{D}_i|} \sum_{(\mathbf{a}_k, b_k) \in \mathcal{D}_i} \frac{1}{1 + \exp(b_k - \mathbf{a}_k^T \mathbf{x})}. \tag{55}$$

The gradient of this loss function is

$$\nabla f_i(\mathbf{x}) = \frac{1}{|\mathcal{D}_i|} \sum_{(\mathbf{a}_k, b_k) \in \mathcal{D}_i} \frac{\mathbf{a}_k \exp(b_k - \mathbf{a}_k^T \mathbf{x})}{(1 + \exp(b_k - \mathbf{a}_k^T \mathbf{x}))^2}. \tag{56}$$

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

$$\nabla f_i(\mathbf{x}) = \frac{1}{|\mathcal{D}_i|} A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x}). \tag{57}$$

The difference between the gradients of f_i and f_j is

$$\begin{aligned} \|\nabla f_i(\mathbf{x}) - \nabla f_j(\mathbf{x})\| &= \left\| \frac{1}{|\mathcal{D}_i|} A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x}) - \frac{1}{|\mathcal{D}_j|} A_j \mathbf{v}(\mathcal{D}_j, \mathbf{x}) \right\| \\ &\leq \frac{1}{|\mathcal{D}_i|} \|A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x})\| + \frac{1}{|\mathcal{D}_j|} \|A_j \mathbf{v}(\mathcal{D}_j, \mathbf{x})\|. \end{aligned} \tag{58}$$

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

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

Utilizing the above inequality in (58), we obtain:

$$\begin{aligned} \|\nabla f_i(\mathbf{x}) - \nabla f_j(\mathbf{x})\| &\leq \frac{1}{|\mathcal{D}_i|} \|A_i \mathbf{v}(\mathcal{D}_i, \mathbf{x})\| + \frac{1}{|\mathcal{D}_j|} \|A_j \mathbf{v}(\mathcal{D}_j, \mathbf{x})\| \\ &\leq \frac{1}{\sqrt{|\mathcal{D}_i|}} \|A_i\| + \frac{1}{\sqrt{|\mathcal{D}_j|}} \|A_j\|. \end{aligned} \tag{59}$$

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

Hyperbolic Tangent

Similar to logistic regression, we can also show that A4 holds for hyperbolic tangent function which is commonly used in neural network models. First, notice that the hyperbolic tangent is a rescaled version of logistic regression:

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

Therefore we have

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

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

Special Case in Linear Regression

Consider the linear regression problem

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

We have

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

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

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

APPENDIX F
PROOF OF CLAIM II.1

The proof is related to techniques developed in classical and recent works that characterize lower bounds for first-order methods in centralized [18], [19] and decentralized [20], [21] settings. Technically, our computational/communication model is *different* compared to the aforementioned works, since we allow arbitrary number of local processing iterations, and we have a central aggregator. The difference here is that our goal is *not* to show the lower bounds on the number of total (centralized) gradient access, nor to show the optimal graph dependency. The main point we would like to make is that there exist constructions of *local* functions f_i's such that *no matter* how many times that local first-order processing is performed, without *communication* and *aggregation*, no significant progress can be made in reducing the stationarity gap of the original problem.

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

$$x^t = V^t(\{x_i^{t-1,Q}\}_{i=1}^N),\ x_i^{t,0} = x^t,\ \ \forall\ i \in [N], \tag{60a}$$

$$x_i^{t,q} \in W_i^t\left(\{x_i^{r,k}, \left\{\nabla f_i(x_i^{r,k})\right\}_{r=0:t}^{k=0:q-1}\right), q \in [Q],\ \forall\ i. \tag{60b}$$

A. Notation.

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

B. Main Constructions.

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

$$g(x) := \frac{1}{N}\sum_{i=1}^N g_i(x), \quad f(x) := \frac{1}{N}\sum_{i=1}^N f_i(x). \tag{61}$$

Here we have $x \in \mathbb{R}^{T+1}$. We assume N is constant, and T is the total number of stages (a large number and one that can potentially increase). For notational simplicity, and without loss of generality, we assume that $T \geq N$, and it is divisible by N. Let us define the component functions g_i's in (61) as follows.

$$g_i(x) = \Theta(x, 1) + \sum_{j=1}^{T/N} \Theta(x, (j-1)N + i + 1), \tag{62}$$

where we have defined the following functions

$$\begin{aligned}&\Theta(x, j) := \Psi(-x[j-1])\Phi(-x[j]) - \Psi(x[j-1])\Phi(x[j]),\ \forall\ j = 2, \cdots, T+1,\\ &\Theta(x, 1) := -\Psi(1)\Phi(x[1]).\end{aligned} \tag{63a}$$

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

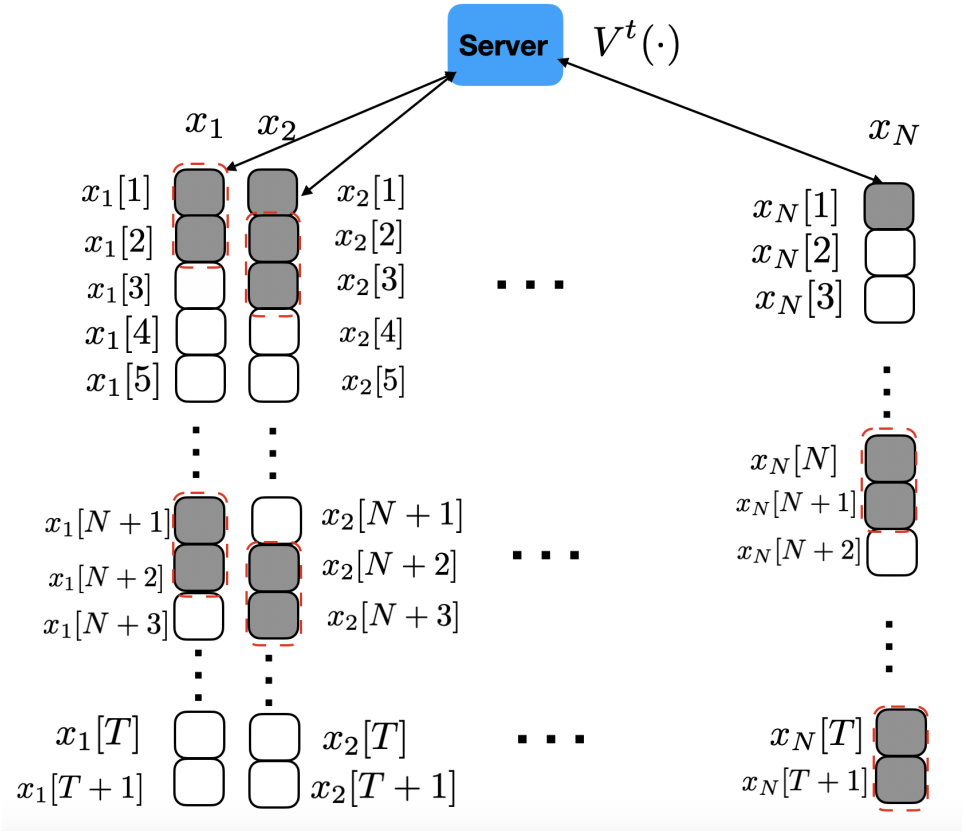

Fig. 6: The example constructed for proving Claim 2.1. Here each agent has a local length $T+1$ vector x_i; each block in the figure represents one dimension of the local vector. If for agent i, its jth block is white it means that f_i is not a function of $x_i[j]$, while if jth block is shaded means f_i is a function of $x_i[j]$. Each dashed red box contains two variables that are coupled together by a function $\Theta(\cdot)$.

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

$$\Psi(w) := \begin{cases} 0 & w \le 0 \\ 1 - e^{-w^2} & w > 0, \end{cases}$$
$$\Phi(w) := 4 \arctan w + 2\pi.$$

By the above definition, the average function becomes:

$$g(x) := \frac{1}{M}\sum_{j=1}^{M} g_i(x) = \Theta(x,1) + \sum_{j=2}^{T+1} \Theta(x,j) \tag{64}$$
$$= -\Psi(1)\Phi\left(x[1]\right) + \sum_{j=2}^{T+1} \left[\Psi\left(-x[j-1]\right)\Phi\left(-x[j]\right) - \Psi\left(x[j-1]\right)\Phi\left(x[j]\right)\right].$$

See Fig. 6 for an illustration of the construction discussed above.

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

$$f_i(x) := \frac{2\pi\epsilon}{L} g_i\left(\frac{xL}{\pi\sqrt{2\epsilon}}\right). \tag{65}$$

Therefore, we also have

$$f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x) = \frac{2\pi\epsilon}{L} g\left(\frac{xL}{\pi\sqrt{2\epsilon}}\right). \tag{66}$$

C. Properties.

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

Lemma 5. *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:*

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

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

3) *The following key property holds:*

$$\Psi(w)\Phi'(v) > 1, \quad \forall\ w \ge 1,\ |v| < 1. \tag{67}$$

4) *The function* h *is lower bounded as follows:*

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

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

Proof. Property 1) is easy to check.

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

$$\Psi(w) = 1 - e^{-w^2}, \quad \Psi'(w) = 2e^{-w^2}w, \quad \Psi''(w) = 2e^{-w^2} - 4e^{-w^2}w^2, \ \forall\ w > 0. \tag{68}$$

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

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

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

$$\Phi(w) = 4\arctan w + 2\pi, \quad \Phi'(w) = \frac{4}{w^2 + 1}, \quad \Phi''(w) = -\frac{8w}{(w^2 + 1)^2}. \tag{69}$$

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

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

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

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

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

Next we show Property 4). Note that $0 \le \Psi(w) < 1$ and $0 < \Phi(w) < 4\pi$. Therefore we have $g(0) = -\Psi(1)\Phi(0) < 0$ and using the construction in (62)

$$\inf_x g_i(x) \ge -\Psi(1)\Phi(x[1]) - \sum_{j=1}^{T/N} \sup_{w,v} \Psi(w)\Phi(v) \tag{70}$$
$$\ge -4\pi - 4(T/N)\pi \ge -5\pi T/N, \tag{71}$$

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

Finally, we show Property 5), using the fact that a function is Lipschitz if it is piecewise smooth with bounded derivative. Before proceeding, let us note a few properties of the construction in (64) (also see Fig. 6). 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

$$\frac{\partial g_q}{\partial x[j]} = \begin{cases} \left(-\Psi\left(-x[j-1]\right)\Phi'\left(-x[j]\right) - \Psi\left(x[j-1]\right)\Phi'\left(x[j]\right)\right), \\ \qquad\qquad q = j - 1 - N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N} - 1, j = 2, 3, \cdots, T+1 \\ \left(-\Psi'\left(-x[j]\right)\Phi\left(-x[j+1]\right) - \Psi'\left(x[j]\right)\Phi\left(x[j+1]\right)\right), \\ \qquad\qquad q = j - N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N} - 1, j = 3, 4, \cdots T \\ 0 \\ \qquad\qquad \text{otherwise}. \end{cases}. \tag{72}$$

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

$$\frac{\partial g_q}{\partial x[1]} = \begin{cases} -\Psi(1)\Phi'(x[1]) + \left(-\Psi'\left(-x[1]\right)\Phi\left(-x[2]\right) - \Psi'\left(x[1]\right)\Phi\left(x[2]\right)\right), & q = 1 \\ -\Psi(1)\Phi'(x[1]), & q \neq 1 \end{cases}. \tag{73}$$

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:

$$\begin{aligned} \ell_q^{''}(0; x, v) &= \sum_{j_1, j_2 = 1}^{T} \frac{\partial^2 g_q(x)}{\partial x[j_1]\partial x[j_2]} v[j_1] v[i_2] \\ &= \sum_{G \in \{0, 1, -1\}} \sum_{j=1}^{T} \frac{\partial^2 g_q(x)}{\partial x[j]\partial x[j+G]} v[j] v[j+G], \end{aligned}$$

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

By using (72) and the above facts, the second-order partial derivative of $g_q(x)$ ($\forall x \neq 0$) is given as follows when $j \neq 1$:

$$\frac{\partial^2 g_q}{\partial x[j]\partial x[j]} = \begin{cases} \left(\Psi\left(-x[j-1]\right)\Phi''\left(-x[j]\right) - \Psi\left(x[j-1]\right)\Phi''\left(x[j]\right)\right), \\ \qquad\qquad q = j - 1 - N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N} - 1, j = 2, 3, \cdots, T+1 \\ \left(\Psi''\left(-x[j]\right)\Phi\left(-x[j+1]\right) - \Psi''\left(x[j]\right)\Phi\left(x[j+1]\right)\right), \\ \qquad\qquad q = j - N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N} - 1, j = 3, 4, \cdots, T \\ 0, \\ \qquad\qquad \text{otherwise} \end{cases} \tag{74}$$

$$\frac{\partial^2 g_q}{\partial x[j]\partial x[j+1]} = \begin{cases} \left(\Psi'\left(-x[j]\right)\Phi'\left(-x[j+1]\right) - \Psi'\left(x[j]\right)\Phi'\left(x[j+1]\right)\right), \\ \qquad\qquad q = j - N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N} - 1, j = 3, 4, \cdots, T \\ 0, \\ \qquad\qquad \text{otherwise} \end{cases} \tag{75}$$

$$\frac{\partial^2 g_q}{\partial x[j]\partial x[j-1]} = \begin{cases} \left(\Psi'\left(-x[j-1]\right)\Phi'\left(-x[j]\right) - \Psi'\left(x[j-1]\right)\Phi'\left(x[j]\right)\right), \\ \qquad\qquad q = j - N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N} - 1, j = 2, 3, \cdots, T+1 \\ 0, \\ \qquad\qquad \text{otherwise} \end{cases}. \tag{76}$$

By applying Lemma 5 – 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

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

Taking the maximum over equations (74) to (76) 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 5 – ii).

When $j = 1$, by using (73), 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 5 – 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:

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

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 6. *We have the following properties for the functions f defined in* (66) *and* (65)*:*

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

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

2) *We have*

$$\|\nabla f(x)\| = \sqrt{2\epsilon}\left\|\nabla g\left(\frac{xL}{\pi\sqrt{2\epsilon}}\right)\right\|,\ \forall\ x \in \mathbb{R}^{T+1}. \tag{77}$$

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 5 we have that for any $T \geq 1$, the following holds

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

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

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

Property 3) is true because the following:

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

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

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

Lemma 7. *If there exists $k \in [T]$ such that $|x[k]| < 1$, then*

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

Proof. The first inequality holds for all $k \in [T]$, since $\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial y[k]}g_i(x)$ is one element of $\frac{1}{N}\sum_{i=1}^{N}\nabla g_i(x)$. We divide the proof for the second inequality into two cases.

Case 1. Suppose $|x[j-1]| < 1$ for all $2 \leq j \leq k$. Therefore, we have $|x[1]| < 1$. Using (73), we have the following inequalities:

$$\frac{\partial g_i(x)}{\partial x[1]} \overset{\text{(i)}}{\leq} -\Psi(1)\Phi'(x[1]) \overset{\text{(ii)}}{<} -1, \forall i \tag{78}$$

where (i) is true because $\Psi'(w), \Phi(w)$ are all non-negative from Lemma 5 -(2); (ii) is true due to Lemma 5 – (3). Therefore, we have the following

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

Case 2) Suppose there exists $2 \leq j \leq k$ such that $|x[j-1]| \geq 1$.

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

$$\frac{\partial g_i(x)}{\partial x[j]} = \begin{cases} \left(-\Psi\left(-x[j-1]\right)\Phi'\left(-x[j]\right) - \Psi\left(x[j-1]\right)\Phi'\left(x[j]\right)\right), \\ \qquad i = j-1-N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N}-1, j = 2, 3, \cdots, T+1 \\ \left(-\Psi'\left(-x[j]\right)\Phi\left(-x[j+1]\right) - \Psi'\left(x[j]\right)\Phi\left(x[j+1]\right)\right), \\ \qquad i = j-1-N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N}-1, j = 3, 4, \cdots, T \\ 0 \\ \qquad \text{otherwise} \end{cases}. \tag{79}$$

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 5 – (3), and the fact that the negative part is zero for Ψ, and Φ' is even function, the expression further simplifies to:

$$-\Psi(|x[j-1]|)\Phi'\left(|x[j]|\right)] \overset{(67)}{<} -1. \tag{80}$$

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 8. *Consider using an algorithm of the form* (60) *to solve the following problem:*

$$\min_{x\in\mathbb{R}^{T+1}}\ g(x) = \frac{1}{N}\sum_{i=1}^{N} g_i(x). \tag{81}$$

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 (60b) *are performed, at least T communication steps* (60a) *are needed to ensure $\bar{x}[T] \neq 0$.*

Proof. For a given $j \geq 2$, suppose that $x_i[j], x_i[j+1], ..., x_i[T] = 0$, $\forall i$, that is, support$\{x_i\} \subseteq \{1, 2, 3, ..., j-1\}$ for all i. Then $\Psi'(x_i[j]) = \Psi'(-x_i[j]) = 0$ for all i, and g_i has the following partial derivative (see (72))

$$\frac{\partial g_i(x_i)}{\partial x_i[j]} = -\left(\Psi\left(-x_i[j-1]\right)\Phi'\left(-x_i[j]\right)\right) + \left(\Psi\left(x_i[j-1]\right)\Phi'\left(x_i[j]\right)\right), \tag{82}$$

$$i = j - 1 - N(\ell) \geq 1,\ \ell = 0, \cdots, \frac{T}{N} - 1, j = 2, 3, \cdots, T+1. \tag{83}$$

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

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

- If before the aggregation step (i.e., at stage $t-1$), some other node q has $x_q[j]$ being nonzero.
- If $\frac{\partial g_i(x_i)}{\partial x_i[j]}$ is nonzero at stage t.

Now suppose that the initial solution is $x_i[j] = 0$ for all (i, j). Then at the first iteration only $\frac{\partial g_i(x_i)}{\partial x_i[1]}$ is non-zero for all i, due to the fact that $\frac{\partial g_i(x_i)}{\partial x_i[1]} = \Psi(1)\Phi'(0) = 4(1 - e^{-1})$ for all i from (73). It is also important to observe that, if all nodes $i \neq 1$ were to perform subsequent local updates (60b), 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

$$\frac{\partial g_i(x_i)}{\partial x_i[2]} = \left(-\Psi'\left(-x[2]\right)\Phi\left(-x[3]\right) - \Psi'\left(x[2]\right)\Phi\left(x[3]\right)\right) = 0, \tag{84}$$

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 (82), we have

$$\frac{\partial g_1(x_1)}{\partial x_1[2]} = -\left(\Psi\left(-x_1[1]\right)\Phi'\left(-x_1[2]\right)\right) + \left(\Psi\left(x_1[1]\right)\Phi'\left(x_1[2]\right)\right). \tag{85}$$

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

So for the first stage $t = 1$, we conclude that, no matter how many local computation that the nodes perform (in the form of the computation step given in (60b)), only x_1 can have two non-zero entries, while the rest of the local variables only have one non-zero entries.

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

D. Main Result for Claim 2.1.

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

Theorem 4. *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* (5)*, 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* (60a)*)*

$$t \geq \frac{\left(f(0) - \inf_x f(x)\right)LN}{10\pi^2}\epsilon^{-1} \tag{86}$$

to achieve the following error

$$h_t^* = \left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x^t)\right\|^2 < \epsilon. \tag{87}$$

Proof of Claim 2.1. First, let us show that the algorithm obeying the rules given in (60) 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 8 we have $\bar{x}[T] = 0$ for all $t < T$. Then by applying Lemma 6 – (2) and Lemma 7, we can conclude that the following holds

$$\|\nabla f(\bar{x}[T])\| = \sqrt{2\epsilon} \left\| \nabla h \left(\frac{\bar{x}[T]U}{\pi\sqrt{2\epsilon}} \right) \right\| > \sqrt{2\epsilon}/N, \tag{88}$$

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 7.

The third part of Lemma 6 ensures that f_i's are L-Lipschitz continuous gradient, and the first part shows

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

Therefore we obtain

$$T \geq \frac{(f(0) - \inf_x f(x))\, LN}{10\pi^2} \epsilon^{-1}. \tag{89}$$

This completes the proof.

Second, consider the algorithm obeying the rules give in (5), 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

$$\hat{f}_i(x) = \sum_{\xi_i \in D_i} F(x; \xi_i) \tag{90}$$

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

$$F(\mathbf{x}; \xi_i) = G(\xi_i) f_i(x), \quad \text{where } f_i(x) \text{ is defined in (65).} \tag{91}$$

where $G(\xi_i)$'s satisfy $G(\xi_i) > 0$ and $\sum_{\xi_i \in D_i} G(\xi_i) = 1$. 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. 6). 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.