Layerwise and Dimensionwise Locally Adaptive Optimization Method (Supplementary Material)

Plan of Supplementary Material: The supplementary material of this paper is composed of two main parts. Section A contains detailed proofs of our results and Section B where additional runs are provided. In particular, Theorem 1 is proved in subsection A.2.

A Theoretical Analysis

We first recall in Table 1 some important notations that will be used in our following analysis.

R, T	$\triangleq$	Number of communications rounds and local iterations (resp.)
n, D, i	$\triangleq$	Total number of clients, portion sampled uniformly and client index
h, ℓ	$\triangleq$	Total number of layers in the DNN and its index
$\phi(\cdot)$	$\triangleq$	Scaling factor in FED-LAMBupdate
$\overline{\theta}$	$\triangleq$	Global model (after periodic averaging)
$p_{r,i}^t$	$\triangleq$	ratio computed at round r, local iteration t and for device i. $p_{r,i}^{\ell,t}$ denotes its component at layer ℓ

Table 1: Summary of notations used in the paper.

We now provide the proofs for the theoretical results of the main paper, including the intermediary Lemmas and the main convergence result, Theorem 1.

A.1 Intermediary Lemmas

Lemma. *Consider $\{\overline{\theta_r}\}_{r>0}$, the sequence of parameters obtained running Algorithm 1. Then for $i \in [\![n]\!]$:*

$$\|\overline{\theta_r} - \theta_{r,i}\|^2 \leq \alpha^2 M^2 \phi_M^2 \frac{(1-\beta_2)p}{v_0} \,,$$

where ϕ_M is defined in H4 and p is the total number of dimensions $p = \sum_{\ell=1}^{\mathsf{h}} p_\ell$.

Proof. Assuming the simplest case when $T = 1$, i.e. one local iteration, then by construction of Algorithm 1, we have for all $\ell \in [\![\mathsf{h}]\!]$, $i \in [\![n]\!]$ and $r > 0$:

$$\theta_{r,i}^\ell = \overline{\theta_r}^\ell - \alpha \phi(\|\theta_{r,i}^{\ell,t-1}\|) p_{r,i}^j / \|p_{r,i}^\ell\| = \overline{\theta_r}^\ell - \alpha \phi(\|\theta_{r,i}^{\ell,t-1}\|) \frac{m_{r,i}^t}{\sqrt{v_r^t}} \frac{1}{\|p_{r,i}^\ell\|}$$

leading to

$$\begin{aligned}
\|\overline{\theta_r} - \theta_{r,i}\|^2 &= \sum_{\ell=1}^{\mathsf{h}} \left\langle \overline{\theta_r}^\ell - \theta_{r,i}^\ell \,|\, \overline{\theta_r}^\ell - \theta_{r,i}^\ell \right\rangle \\
&\leq \alpha^2 M^2 \phi_M^2 \frac{(1-\beta_2)p}{v_0} \,,
\end{aligned}$$

which concludes the proof. ∎

Lemma. *Consider $\{\overline{\theta_r}\}_{r>0}$, the sequence of parameters obtained running Algorithm 1. Then for $r > 0$:*

$$\left\| \frac{\overline{\nabla} f(\theta_r)}{\sqrt{v_r^t}} \right\|^2 \geq \frac{1}{2} \left\| \frac{\nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\|^2 - \overline{L} \alpha^2 M^2 \phi_M^2 \frac{(1-\beta_2)p}{v_0}$$

where M is defined in H2, p is the total number of dimensions $p = \sum_{\ell=1}^{\mathsf{h}} p_\ell$ and ϕ_M is defined in H4.

Proof. Consider the following sequence:

$$\left\| \frac{\overline{\nabla} f(\theta_r)}{\sqrt{v_r^t}} \right\|^2 \geq \frac{1}{2} \left\| \frac{\nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\|^2 - \left\| \frac{\overline{\nabla} f(\theta_r) - \nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\|^2 ,$$

where the inequality is due to the Cauchy-Schwartz inequality.

Under the smoothness assumption H1 and using Lemma 1, we have

$$\begin{aligned}
\left\| \frac{\overline{\nabla} f(\theta_r)}{\sqrt{v_r^t}} \right\|^2 &\geq \frac{1}{2} \left\| \frac{\nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\| - \left\| \frac{\overline{\nabla} f(\theta_r) - \nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\|^2 \\
&\geq \frac{1}{2} \left\| \frac{\nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\|^2 - \overline{L} \alpha^2 M^2 \phi_M^2 \frac{(1-\beta_2)p}{v_0} ,
\end{aligned}$$

which concludes the proof. □

A.2 Proof of Theorem 1

We now develop a proof for the two intermediary lemmas, Lemma 1 and Lemma 2, in the case when each local model is obtained after more than one local update. Then the two quantities, either the gap between the periodically averaged parameter and each local update, *i.e.*, $\|\overline{\theta_r} - \theta_{r,i}\|^2$, and the ratio of the average gradient, more particularly its relation to the gradient of the average global model (*i.e.*, $\left\|\frac{\overline{\nabla} f(\theta_r)}{\sqrt{v_r^t}}\right\|$ and $\left\|\frac{\nabla f(\overline{\theta_r})}{\sqrt{v_r^t}}\right\|$), are impacted.

Theorem. *Assume **H1-H4**. Consider $\{\overline{\theta_r}\}_{r>0}$, the sequence of parameters obtained running Algorithm 1 with a decreasing learning rate α. Let the number of local epochs be $T \geq 1$ and $\lambda = 0$. Then, at iteration τ, we have:*

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{\theta_t})}{\hat{v}_t^{1/4}}\right\|^2\right] \leq \sqrt{\frac{M^2 p}{n}}\frac{\mathbb{E}[f(\bar{\theta}_1)] - \min_{\theta\in\Theta} f(\theta)}{\mathsf{h}\alpha_r \tau} + \frac{\phi_M \sigma^2}{\tau n}\sqrt{\frac{1-\beta_2}{M^2 p}}$$
$$+ 4\alpha\left[\frac{\alpha^2 L_\ell}{\sqrt{v_0}}M^2(T-1)^2\phi_M^2(1-\beta_2)p + \frac{M^2}{\sqrt{v_0}} + \phi_M^2\sqrt{M^2+p\sigma^2} + \phi_M\frac{\mathsf{h}\sigma^2}{\sqrt{n}}\right] + cst.$$

If one considers a deceasing stepsize as $\alpha_\tau = \mathcal{O}(\frac{1}{L\sqrt{\tau}})$, then:

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{\theta_t})}{\hat{v}_t^{1/4}}\right\|^2\right] \leq \mathcal{O}\left(\sqrt{\frac{M^2 p}{n}}\frac{1}{\sqrt{\mathsf{h}\tau}} + \frac{\sigma^2}{\tau n\sqrt{p}} + \frac{(T-1)^2 p}{\tau^{3/2}L^3}\right)$$

Discussion on the bound: Obviously, the last term containing the number of local updates T is small as long as $T \leq \mathcal{O}(\frac{\tau^{1/2}L^{5/4}}{(np)^{1/4}})$. Treating $p^{1/4}/L = \mathcal{O}(1)$ which is usually small, the result implies that we can get the same rate of convergence as the algorithm using one local update, with $O(\tau^{1/2}/n^{1/4})$ rounds of communication. When the number of workers n increases, then a constraint on the number of local updates T is occurring, meaning that we would need more rounds of communication to achieve the same convergence rate, for a identical $\epsilon-$stationary point. We recall that a $\epsilon-$stationary point is defined by the number of communication rounds $\mathcal{R}$ such that $\frac{1}{\tau}\sum_{t=1}^{\mathcal{R}}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{\theta_t})}{\hat{v}_t^{1/4}}\right\|^2\right] \leq \epsilon$.

We now provide the proof for Theorem 1.

Proof. Using H1, we have:

$$f(\bar{\vartheta}_{r+1}) \leq f(\bar{\vartheta}_r) + \left\langle \nabla f(\bar{\vartheta}_r) \,|\, \bar{\vartheta}_{r+1} - \bar{\vartheta}_r \right\rangle + \sum_{\ell=1}^{L}\frac{L_\ell}{2}\|\bar{\vartheta}_{r+1}^\ell - \bar{\vartheta}_r^\ell\|^2$$
$$\leq f(\bar{\vartheta}_r) + \sum_{\ell=1}^{\mathsf{h}}\sum_{j=1}^{p_\ell}\nabla_\ell f(\bar{\vartheta}_r)^j(\bar{\vartheta}_{r+1}^{\ell,j} - \bar{\vartheta}_r^{\ell,j}) + \sum_{\ell=1}^{L}\frac{L_\ell}{2}\|\bar{\vartheta}_{r+1}^\ell - \bar{\vartheta}_r^\ell\|^2 \,.$$

Taking expectations on both sides leads to:

$$-\mathbb{E}[\left\langle \nabla f(\bar{\vartheta}_r) \,|\, \bar{\vartheta}_{r+1} - \bar{\vartheta}_r \right\rangle] \leq \mathbb{E}[f(\bar{\vartheta}_r) - f(\bar{\vartheta}_{r+1})] + \sum_{\ell=1}^{L}\frac{L_\ell}{2}\mathbb{E}[\|\bar{\vartheta}_{r+1}^\ell - \bar{\vartheta}_r^\ell\|^2] \,. \tag{7}$$

Yet, we observe that, using the classical intermediate quantity, used for proving convergence results of adaptive optimization methods, see for instance [27], we have:

$$\bar{\vartheta}_r = \bar{\theta}_r + \frac{\beta_1}{1-\beta_1}(\bar{\theta}_r - \bar{\theta}_{r-1}) \,, \tag{8}$$

where $\bar{\theta_r}$ denotes the average of the local models at round r. Then for each layer ℓ,

$$\bar{\vartheta}^{\ell}_{r+1} - \bar{\vartheta}^{\ell}_{r} = \frac{1}{1-\beta_1}(\bar{\theta}^{\ell}_{r+1} - \bar{\theta}^{\ell}_{r}) - \frac{\beta_1}{1-\beta_1}(\bar{\theta}^{\ell}_{r} - \bar{\theta}^{\ell}_{r-1}) \tag{9}$$

$$= \frac{\alpha_r}{1-\beta_1}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta^{\ell}_{r,i}\|)}{\|p^{\ell}_{r,i}\|}p^{\ell}_{r,i} - \frac{\alpha_{r-1}}{1-\beta_1}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta^{\ell}_{r-1,i}\|)}{\|p^{\ell}_{r-1,i}\|}p^{\ell}_{r-1,i} \tag{10}$$

$$= \frac{\alpha\beta_1}{1-\beta_1}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\phi(\|\theta^{\ell}_{r,i}\|)}{\sqrt{v^t_r}\|p^{\ell}_{r,i}\|} - \frac{\phi(\|\theta^{\ell}_{r-1,i}\|)}{\sqrt{v^t_{r-1}}\|p^{\ell}_{r-1,i}\|}\right)m^t_{r-1} + \frac{\alpha}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta^{\ell}_{r,i}\|)}{\sqrt{v^t_r}\|p^{\ell}_{r,i}\|}g_{r,i}\ , \tag{11}$$

where we have assumed a constant learning rate α.

We note for all $\theta \in \Theta$, the majorant $G > 0$ such that $\phi(\|\theta\|) \leq G$. Then, following (7), we obtain:

$$-\mathbb{E}[\langle \nabla f(\bar{\vartheta}_r)\,|\,\bar{\vartheta}_{r+1} - \bar{\vartheta}_r\rangle] \leq \mathbb{E}[f(\bar{\vartheta}_r) - f(\bar{\vartheta}_{r+1})] + \sum_{\ell=1}^{L}\frac{L_\ell}{2}\mathbb{E}[\|\bar{\vartheta}_{r+1} - \bar{\vartheta}_r\|^2]\ . \tag{12}$$

Developing the LHS of (12) using (9) leads to

$$\langle \nabla f(\bar{\vartheta}_r)\,|\,\bar{\vartheta}_{r+1} - \bar{\vartheta}_r\rangle = \sum_{\ell=1}^{\mathsf{h}}\sum_{j=1}^{p_\ell}\nabla_\ell f(\bar{\vartheta}_r)^j(\bar{\vartheta}^{\ell,j}_{r+1} - \bar{\vartheta}^{\ell,j}_{r}) \tag{13}$$

$$= \frac{\alpha\beta_1}{1-\beta_1}\frac{1}{n}\sum_{\ell=1}^{\mathsf{h}}\sum_{j=1}^{p_\ell}\nabla_\ell f(\bar{\vartheta}_r)^j\left[\sum_{i=1}^{n}\left(\frac{\phi(\|\theta^{\ell}_{r,i}\|)}{\sqrt{v^t_r}\|p^{\ell}_{r,i}\|} - \frac{\phi(\|\theta^{\ell}_{r-1,i}\|)}{\sqrt{v^t_{r-1}}\|p^{\ell}_{r-1,i}\|}\right)m^t_{r-1}\right] \tag{14}$$

$$\underbrace{-\frac{\alpha}{n}\sum_{\ell=1}^{\mathsf{h}}\sum_{j=1}^{p_\ell}\nabla_\ell f(\bar{\vartheta}_r)^j\sum_{i=1}^{n}\frac{\phi(\|\theta^{\ell}_{r,i}\|)}{\sqrt{v^t_r}\|p^{\ell}_{r,i}\|}g^{l,j}_{r,i}\ .}_{=A_1} \tag{15}$$

We change all index r to iteration t. Suppose T is the number of local iterations. We can write (15) as

$$A_1 = -\alpha_t\langle \nabla f(\bar{\vartheta}_t), \frac{\bar{g}_t}{\sqrt{\hat{v}_t}}\rangle,$$

where $\bar{g}_t = \frac{1}{n}\sum_{i=1}^{n}\bar{g}_{t,i}$, with $\bar{g}_{t,i} = \left[\frac{\phi(\|\theta^1_{t,i}\|)}{\|p^1_{t,i}\|}g^1_{t,i}, ..., \frac{\phi(\|\theta^L_{t,i}\|)}{\|p^L_{t,i}\|}g^L_{t,i}\right]$ representing the normalized gradient (concatenated by layers) of the i-th device. It holds that

$$\langle \nabla f(\bar{\vartheta}_t), \frac{\bar{g}_t}{\sqrt{\hat{v}_t}}\rangle = \frac{1}{2}\|\frac{\nabla f(\bar{\vartheta}_t)}{\hat{v}_t^{1/4}}\|^2 + \frac{1}{2}\|\frac{\bar{g}_t}{\hat{v}_t^{1/4}}\|^2 - \|\frac{\nabla f(\bar{\vartheta}_t) - \bar{g}_t}{\hat{v}_t^{1/4}}\|^2. \tag{16}$$

To bound the last term on the RHS, we have

$$\begin{aligned}\|\frac{\nabla f(\bar{\vartheta}_t) - \bar{g}_t}{\hat{v}_t^{1/4}}\|^2 &= \|\frac{\frac{1}{n}\sum_{i=1}^{n}(\nabla f(\bar{\vartheta}_t) - \bar{g}_{t,i})}{\hat{v}_t^{1/4}}\|^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\|\frac{\nabla f(\bar{\vartheta}_t) - \bar{g}_{t,i}}{\hat{v}_t^{1/4}}\|^2 \\ &\leq \frac{2}{n}\sum_{i=1}^{n}\Big(\|\frac{\nabla f(\bar{\vartheta}_t) - \nabla f(\bar{\theta}_t)}{\hat{v}_t^{1/4}}\|^2 + \|\frac{\nabla f(\bar{\theta}_t) - \bar{g}_{t,i}}{\hat{v}_t^{1/4}}\|^2\Big).\end{aligned}$$

By Lipschitz smoothness of the loss function, the first term admits

$$\begin{aligned}\frac{2}{n}\sum_{i=1}^{n}\|\frac{\nabla f_i(\bar{\vartheta}_t)-\nabla f_i(\bar{\theta}_t)}{\hat{v}_t^{1/4}}\|^2 &\leq \frac{2}{n\sqrt{v_0}}\sum_{i=1}^{n}L_\ell\|\bar{\vartheta}_t-\bar{\theta}_t\|^2\\ &= \frac{2L_\ell}{n\sqrt{v_0}}\frac{\beta_1^2}{(1-\beta_1)^2}\sum_{i=1}^{n}\|\bar{\theta}_t-\bar{\theta}_{t-1}\|^2\\ &\leq \frac{2\alpha_r^2 L_\ell}{n\sqrt{v_0}}\frac{\beta_1^2}{(1-\beta_1)^2}\sum_{l=1}^{L}\sum_{i=1}^{n}\|\frac{\phi(\|\theta_{t,i}^l\|)}{\|p_{t,i}^l\|}p_{t,i}^l\|^2\\ &\leq \frac{2\alpha_r^2 L_\ell p\phi_M^2}{\sqrt{v_0}}\frac{\beta_1^2}{(1-\beta_1)^2}.\end{aligned}$$

For the second term,

$$\frac{2}{n}\sum_{i=1}^{n}\|\frac{\nabla f(\bar{\theta}_t)-\bar{g}_{t,i}}{\hat{v}_t^{1/4}}\|^2 \leq \frac{4}{n}\Big(\underbrace{\sum_{i=1}^{n}\|\frac{\nabla f(\bar{\theta}_t)-\nabla f(\theta_{t,i})}{\hat{v}_t^{1/4}}\|^2}_{B_1}+\underbrace{\sum_{i=1}^{n}\|\frac{\nabla f(\theta_{t,i})-\bar{g}_{t,i}}{\hat{v}_t^{1/4}}\|^2}_{B_2}\Big). \tag{17}$$

Using the smoothness of f_i we can transform B_1 into consensus error by

$$\begin{aligned}B_1 &\leq \frac{L}{\sqrt{v_0}}\sum_{i=1}^{n}\|\bar{\theta}_t-\theta_{t,i}\|^2\\ &= \frac{\alpha_r^2 L}{\sqrt{v_0}}\sum_{i=1}^{n}\sum_{l=1}^{L}\|\sum_{j=\lfloor t\rfloor_T+1}^{t}\Big(\frac{\phi(\|\theta_{j,i}^l\|)}{\|p_{j,i}^l\|}p_{j,i}^l-\frac{1}{n}\sum_{k=1}^{n}\frac{\phi(\|\theta_{j,k}^l\|)}{\|p_{j,k}^l\|}p_{j,k}^l\Big)\|^2 \qquad (18)\\ &\leq n\frac{\alpha_t^2 L}{\sqrt{v_0}}M^2(T-1)^2\phi_M^2(1-\beta_2)p\end{aligned}$$

where the last inequality stems from Lemma 1 in the particular case where $\theta_{t,i}$ are averaged every $ct+1$ local iterations for any integer c, since $(t-1)-(\lfloor t\rfloor_T+1)+1\leq T-1$.

We now develop the expectation of B_2 under the simplification that $\beta_1=0$:

$$\begin{aligned}\mathbb{E}[B_2] &= \mathbb{E}[\sum_{i=1}^{n}\|\frac{\nabla f(\theta_{t,i})-\bar{g}_{t,i}}{\hat{v}_t^{1/4}}\|^2]\\ &\leq \frac{nM^2}{\sqrt{v_0}}+n\phi_M^2\sqrt{M^2+p\sigma^2}-2\sum_{i=1}^{n}\mathbb{E}[\langle\nabla f(\theta_{t,i}),\bar{g}_{t,i}\rangle/\sqrt{\hat{v}_t}]\\ &= \frac{nM^2}{\sqrt{v_0}}+n\phi_M^2\sqrt{M^2+p\sigma^2}-2\sum_{i=1}^{n}\sum_{\ell=1}^{L}\mathbb{E}[\langle\nabla_\ell f(\theta_{t,i}),\frac{\phi(\|\theta_{t,i}^l\|)}{\|p_{t,i}^l\|}g_{t,i}^l\rangle/\sqrt{\hat{v}_t^l}]\\ &= \frac{nM^2}{\sqrt{v_0}}+n\phi_M^2\sqrt{M^2+p\sigma^2}-2\sum_{i=1}^{n}\sum_{l=1}^{L}\sum_{i=1}^{p_l}\mathbb{E}[\nabla_l f(\theta_{t,i})^j\frac{\phi(\|\theta_{t,i}^{l,j}\|)}{\sqrt{\hat{v}_t^{l,j}}\|p_{t,i}^{l,j}\|}g_{t,i}^{l,j}]\\ &\leq \frac{nM^2}{\sqrt{v_0}}+n\phi_M^2\sqrt{M^2+p\sigma^2}-2\sum_{i=1}^{n}\sum_{l=1}^{L}\sum_{i=1}^{p_l}\mathbb{E}\left[\sqrt{\frac{1-\beta_2}{M^2p_\ell}}\phi(\|\theta_{r,i}^{l,j}\|)\nabla_l f(\theta_{t,i})^j g_{t,i}^{l,j}\right]\\ &\quad -2\sum_{i=1}^{n}\sum_{l=1}^{L}\sum_{j=1}^{p_l}E\left[\left(\phi(\|\theta_{r,i}^{l,j}\|)\nabla_l f(\theta_{t,i})^j\frac{g_{r,i}^{l,j}}{\|p_{r,i}^{l,j}\|}\right)\mathbb{1}\left(\text{sign}(\nabla_l f(\theta_{t,i})^j\neq\text{sign}(g_{r,i}^{l,j})\right)\right]\end{aligned}$$

where we use assumption H2, H3 and H4. Yet,

$$-\mathbb{E}\left[\left(\phi(\|\theta_{r,i}^{l,j}\|)\nabla_l f(\theta_{t,i})^j\frac{g_{r,i}^{l,j}}{\|p_{r,i}^{l,j}\|}\right)\mathbb{1}\left(\text{sign}(\nabla_l f(\theta_{t,i})^j\neq\text{sign}(g_{r,i}^{l,j})\right)\right]\leq\phi_M\nabla_l f(\theta_{t,i})^j\mathbb{P}\left[\text{sign}(\nabla_l f(\theta_{t,i})^j\neq\text{sign}(g_{r,i}^{l,j})\right]$$

Then we have:

$$\mathbb{E}[B_2] \leq \frac{nM^2}{\sqrt{v_0}} + n\phi_M^2\sqrt{M^2+p\sigma^2} - 2\phi_m\sqrt{\frac{1-\beta_2}{M^2p}}\sum_{i=1}^{n}\mathbb{E}[\|[\nabla f(\theta_{t,i})\|^2] + \phi_M\frac{\mathsf{h}\sigma^2}{\sqrt{n}}$$

Thus, (17) becomes:

$$\frac{2}{n}\sum_{i=1}^{n}\|\frac{\nabla f_i(\bar{\theta}_t)-\bar{g}_{t,i}}{\hat{v}_t^{1/4}}\|^2 \leq 4\left[\frac{\alpha_t^2 Ll}{\sqrt{v_0}}\alpha_r^2M^2(T-1)^2\phi_M^2(1-\beta_2)p + \frac{M^2}{\sqrt{v_0}} + \phi_M^2\sqrt{M^2+p\sigma^2} + \phi_M\frac{\mathsf{h}\sigma^2}{\sqrt{n}}\right]$$

Substituting all ingredients into (16), we obtain

$$\begin{aligned}-\alpha_t\mathbb{E}[\langle\nabla f(\vartheta_t),\frac{\bar{g}_t}{\sqrt{\hat{v}_t}}\rangle] \leq -\frac{\alpha_t}{2}\mathbb{E}\big[\|\frac{\nabla f(\bar{\vartheta}_t)}{\hat{v}_t^{1/4}}\|^2\big] - \frac{\alpha_t}{2}\mathbb{E}\big[\|\frac{\bar{g}_t}{\hat{v}_t^{1/4}}\|^2\big] + \frac{2\alpha_t^3 L_\ell p\phi_M^2}{\sqrt{v_0}}\frac{\beta_1^2}{(1-\beta_1)^2}&\\ + 4\left[\frac{\alpha_t^2 L}{\sqrt{v_0}}M^2(T-1)^2\phi_M^2(1-\beta_2)p + \frac{M^2}{\sqrt{v_0}} + \phi_M^2\sqrt{M^2+p\sigma^2} + \phi_M\frac{\mathsf{h}\sigma^2}{\sqrt{n}}\right].&\end{aligned}$$

At the same time, we have

$$\begin{aligned}\mathbb{E}\big[\|\frac{\bar{g}_t}{\hat{v}_t^{1/4}}\|^2\big] &= \frac{1}{n^2}\mathbb{E}\big[\|\frac{\sum_{i=1}^{n}\bar{g}_{t,i}}{\hat{v}_t^{1/4}}\|^2\big]\\ &= \frac{1}{n^2}\mathbb{E}\big[\sum_{l=1}^{L}\sum_{i=1}^{n}\|\frac{\phi(\|\theta_{t,i}^l\|)}{\hat{v}^{1/4}\|p_{t,i}^l\|}g_{t,i}^l\|^2\big]\\ &\geq \phi_m^2(1-\beta_2)\mathbb{E}\left[\|\frac{1}{n}\sum_{i=1}^{n}\frac{\nabla f(\theta_{t,i})}{\hat{v}^{1/4}}\|^2\right]\\ &= \phi_m^2(1-\beta_2)\mathbb{E}\left[\|\frac{\overline{\nabla}f(\theta_t)}{\hat{v}^{1/4}}\|^2\right]\end{aligned}$$

Regarding $\left\|\frac{\overline{\nabla}f(\theta_t)}{\hat{v}_t^{1/4}}\right\|^2$, we have

$$\begin{aligned}\left\|\frac{\overline{\nabla}f(\theta_t)}{\hat{v}_t^{1/4}}\right\|^2 &\geq \frac{1}{2}\left\|\frac{\nabla f(\overline{\theta}_t)}{\hat{v}_t^{1/4}}\right\|^2 - \left\|\frac{\overline{\nabla}f(\theta_t)-\nabla f(\overline{\theta}_t)}{\hat{v}_t^{1/4}}\right\|^2\\ &\geq \frac{1}{2}\left\|\frac{\nabla f(\overline{\theta}_t)}{\hat{v}_t^{1/4}}\right\|^2 - \left\|\frac{\frac{1}{n}\sum_{i=1}^{n}(\nabla f(\theta_{t,i})-\nabla f(\bar{\theta}_i))}{\hat{v}_t^{1/4}}\right\|^2\\ &\geq \frac{1}{2}\left\|\frac{\nabla f(\overline{\theta}_t)}{\hat{v}_t^{1/4}}\right\|^2 - \frac{\alpha_t^2 L_\ell}{\sqrt{v_0}}M^2(T-1)^2\phi_M^2(1-\beta_2)p,\end{aligned}$$

where the last line is due to (18). Therefore, we have obtained

$$\begin{aligned}A_1 \leq -\frac{\phi_m^2(1-\beta_2)}{2}\left\|\frac{\nabla f(\overline{\theta}_t)}{\hat{v}_t^{1/4}}\right\|^2 + \frac{\alpha_r^2 L_\ell}{\sqrt{v_0}}M^2(T-1)^2\phi_m^2\phi_M^2(1-\beta_2)^2p + \frac{2\alpha^3 L_\ell p\phi_M^2}{\sqrt{v_0}}\frac{\beta_1^2}{(1-\beta_1)^2}&\\ + 4\alpha_t\left[\frac{\alpha_t^2 L}{\sqrt{v_0}}M^2(T-1)^2\phi_M^2(1-\beta_2)p + \frac{M^2}{\sqrt{v_0}} + \phi_M^2\sqrt{M^2+p\sigma^2} + \phi_M\frac{\mathsf{h}\sigma^2}{\sqrt{n}}\right].&\end{aligned}$$

Substitute back into (15), and leave other derivations unchanged. Assuming $M \leq 1$, we have the following:

$$
\begin{aligned}
&\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{\theta_t})}{\hat{v}_t^{1/4}}\right\|^2\right] \\
&\lesssim \sqrt{\frac{M^2 p}{n}}\frac{f(\bar{\vartheta}_1)-\mathbb{E}[f(\bar{\vartheta}_{\tau+1})]}{\mathsf{h}\alpha_t\tau} + \frac{\alpha_t}{n^2}\sum_{r=1}^{\tau}\sum_{i=1}^{n}\sigma_i^2\mathbb{E}\left[\left\|\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_t}\|p_{r,i}^{\ell}\|}\right\|^2\right] + \frac{2\alpha^3 L_\ell p\phi_M^2}{\sqrt{v_0}}\frac{\beta_1^2}{(1-\beta_1)^2} \\
&+ 4\alpha_t\left[\frac{\alpha_t^2 L_\ell}{\sqrt{v_0}}M^2(T-1)^2\phi_M^2(1-\beta_2)p + \frac{M^2}{\sqrt{v_0}} + \phi_M^2\sqrt{M^2+p\sigma^2} + \phi_M\frac{\mathsf{h}\sigma^2}{\sqrt{n}}\right] + \frac{\overline{L}\beta_1^2\mathsf{h}(1-\beta_2)M^2\phi_M^2 n}{2(1-\beta_1)^2 v_0} \\
&+ \frac{\alpha_t\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{h}M^2}{\sqrt{v_0}} + \overline{L}\alpha_t^2 M^2\phi_M^2\frac{(1-\beta_2)p}{Tv_0} \\
&\leq \sqrt{\frac{M^2 p}{n}}\frac{\mathbb{E}[f(\bar{\theta}_1)] - \min_{\theta\in\Theta} f(\theta)}{\mathsf{h}\alpha_t\tau} + \frac{\phi_M\sigma^2}{\tau n}\sqrt{\frac{1-\beta_2}{M^2 p}} \\
&+ 4\alpha_t\left[\frac{\alpha_t^2 L_\ell}{\sqrt{v_0}}M^2(T-1)^2\phi_M^2(1-\beta_2)p + \frac{M^2}{\sqrt{v_0}} + \phi_M^2\sqrt{M^2+p\sigma^2} + \phi_M\frac{\mathsf{h}\sigma^2}{\sqrt{n}}\right] \\
&+ \frac{\alpha_t\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{h}M^2}{\sqrt{v_0}} + \overline{L}\alpha_t^2 M^2\phi_M^2\frac{(1-\beta_2)p}{Tv_0} + \frac{\overline{L}\beta_1^2\mathsf{h}(1-\beta_2)M^2\phi_M^2 n}{2(1-\beta_1)^2 v_0} + \frac{2\alpha^3 L_\ell p\phi_M^2}{\sqrt{v_0}}\frac{\beta_1^2}{(1-\beta_1)^2}.
\end{aligned}
$$

And if we set the learning rate to be of order $\mathcal{O}(\frac{1}{L\sqrt{\tau}})$ then:

$$
\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{\theta_t})}{\hat{v}_t^{1/4}}\right\|^2\right] \leq \mathcal{O}\left(\sqrt{\frac{M^2 p}{n}}\frac{1}{\sqrt{\mathsf{h}\tau}} + \frac{\sigma^2}{\tau n\sqrt{p}} + \frac{(T-1)^2 p}{\tau^{3/2}L^3}\right),
$$

concluding our proof.

□

A.3 Proof Corollary 1

H5. *For $t > 0$ and $r > 0$, there exists some constant such that $\|\sqrt{v_r^t}\| \leq V^2$.*

Corollary. *Assume **H1**-**H4**. Consider $\{\overline{\theta_r}\}_{r>0}$, the sequence of parameters obtained running Algorithm 1. Then, if the number of local epochs is set to $T = 1$, $\epsilon = \lambda = 0$ we have, under **H5**:*

$$
\frac{1}{R}\mathbb{E}\left[\left\|\nabla f(\overline{\theta_R})\right\|^2\right] \leq \mathcal{O}\left(\sqrt{\frac{p}{n}}\frac{1}{\mathsf{h}\sqrt{R}}\right)
$$

Proof. From the bound in Theorem 1 and with assumption **H5**. □

B Additional Numerical Experiments

B.1 CIFAR-10 with Residual Neural Network

In Figure 4, we report the test accuracies of a ResNet trained on CIFAR-10 dataset, where the data is iid allocated among clients. We run 1 and 3 local epochs for 10 clients and see great performances of our method.

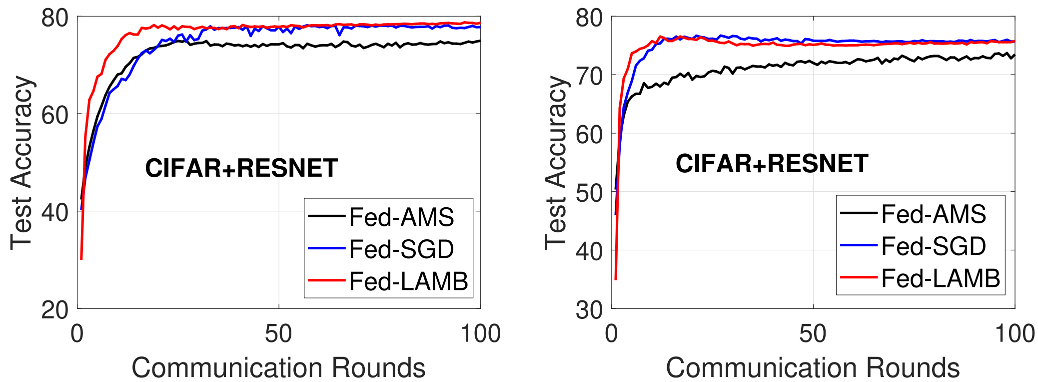

Figure 4: **From Left to Right**: Test accuracy on CIFAR+ResNet, with iid data distribution. 10 clients and (Left) 1 local epoch, (Right) 3 local epoch