

Layerwise and Dimensionwise Locally Adaptive Optimization Method (Supplementary Material)

Plan of the 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	$:=$	Number of communications rounds and local iterations (resp.)
n, D, i	$:=$	Total number of clients, portion sampled uniformly and client index
h, ℓ	$:=$	Total number of layers in the DNN and its index
$\phi(\cdot)$	$:=$	Scaling factor in FED-LAMBUpdate
$\bar{\theta}$	$:=$	Global model (after periodic averaging)
$p_{r,i}^t$	$:=$	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 $\{\bar{\theta}_r\}_{r>0}$, the sequence of parameters obtained running Algorithm 1. Then for $i \in \llbracket n \rrbracket$:

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

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

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

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

leading to

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

which concludes the proof. \square

Lemma. Consider $\{\bar{\theta}_r\}_{r>0}$, the sequence of parameters obtained running Algorithm 1. Then for $r > 0$:

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

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

486 *Proof.* Consider the following sequence:

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

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

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

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

489 which concludes the proof. \square

490 A.2 Proof of Theorem 1

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

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

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

499 *Proof.* Using H1, we have:

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

500 Taking expectations on both sides leads to:

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

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

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

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

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

$$= \frac{\alpha_r}{1 - \beta_1} \frac{1}{n} \sum_{i=1}^n \frac{\phi(\|\theta_{r,i}^\ell\|)}{\|p_{r,i}^\ell\|} p_{r,i}^\ell - \frac{\alpha_{r-1}}{1 - \beta_1} \frac{1}{n} \sum_{i=1}^n \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\|p_{r-1,i}^\ell\|} p_{r-1,i}^\ell \quad (10)$$

$$= \frac{\alpha\beta_1}{1 - \beta_1} \frac{1}{n} \sum_{i=1}^n \left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t} \|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t} \|p_{r-1,i}^\ell\|} \right) m_{r-1}^t + \frac{\alpha}{n} \sum_{i=1}^n \frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t} \|p_{r,i}^\ell\|} g_{r,i}^\ell, \quad (11)$$

504 where we have assumed a constant learning rate α .

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

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

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

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

$$= \frac{\alpha\beta_1}{1 - \beta_1} \frac{1}{n} \sum_{\ell=1}^h \sum_{j=1}^{p_\ell} \nabla_\ell f(\bar{\vartheta}_r)^j \left[\sum_{i=1}^n \left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t} \|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t} \|p_{r-1,i}^\ell\|} \right) m_{r-1}^t \right] \quad (14)$$

$$- \underbrace{\frac{\alpha}{n} \sum_{\ell=1}^h \sum_{j=1}^{p_\ell} \nabla_\ell f(\bar{\vartheta}_r)^j \sum_{i=1}^n \frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t} \|p_{r,i}^\ell\|} g_{r,i}^{\ell,j}}_{=A_1}. \quad (15)$$

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

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

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

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

510 To bound the last term on the RHS, we have

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

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

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

512 For the second term,

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

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

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

514 where the last inequality stems from Lemma 1 in the particular case where $\theta_{t,i}$ are averaged every
515 $ct + 1$ local iterations for any integer c , since $(t-1) - (\lfloor t \rfloor_T + 1) + 1 \leq T-1$.

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

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

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

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

518 Then we have:

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

519 Thus, (17) becomes:

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

520 Substituting all ingredients into (16), we obtain

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

521 At the same time, we have

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

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

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

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

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

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

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

This concludes the proof.

□

B Additional Numerical Experiments

In below we provide two more results on CIFAR dataset trained on a Linux server with four Nvidia Tesla V100 cards, which is the hardware setting for all the experiments conducted in this paper. In Figure 4, we report the test accuracies of a ResNet-9 [8] trained on CIFAR-10 dataset, where the data is iid allocated among clients. We run 1 and 3 local epochs for 10 clients. From the figures, we observe similar advantage as the set of experiments presented in the main paper: faster convergence than local AMS and local SGD. In both cases, FED-LAMB also generalizes better than local SGD.

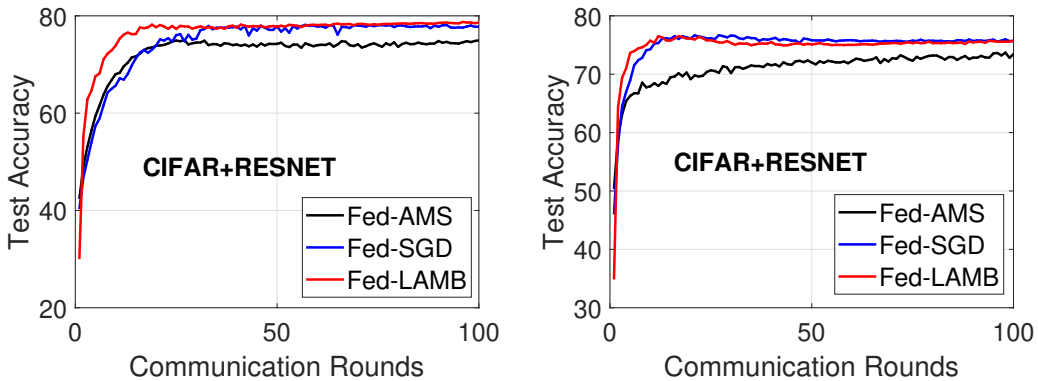


Figure 4: **From Left to Right:** Test accuracy on CIFAR+ResNet, with iid data distribution. 10 clients and (Left) 1 local epoch, (Right) 3 local epochs.