# Fed-LAMB: Layer-wise and Dimension-wise Locally Adaptive Federated Learning (Supplemental Material)

**Belhal Karimi, Ping Li, Xiaoyun Li**

Cognitive Computing Lab
Baidu Research
10900 NE 8th St, Bellevue, WA 98004, USA
{belhal.karimi, pingli98, lixiaoyun996}@gmail.com

## A EXPERIMENT DETAILS AND RESULTS

### A.1 THE ADP-FED ALGORITHM

The Adp-Fed (Adaptive Federated Optimization) is one of the baseline methods compared with Fed-LAMB in our paper. The algorithm is given in Algorithm 1. The key difference between Adp-Fed and Fed-AMS [Chen et al., 2020] is that, in Adp-Fed, each client runs local SGD (Line 8), and an Adam optimizer is maintained for the global adaptive optimization (Line 15). In the Fed-AMS framework (as well as our Fed-LAMB), each clients runs local (adaptive) AMSGrad method, and the global model is simply obtained by averaging the local models. [Li and Li, 2023] proposed a variant of Adp-Fed algorithm with communication compression.

**Algorithm 1** Adp-Fed: Adaptive Federated Optimization [Reddi et al., 2021]

```
 1: **Input**: parameter $0 < \beta_1, \beta_2 < 1$, and learning rate $\alpha_t$, weight decaying parameter $\lambda \in [0, 1]$.
 2: **Initialize**: $\theta_{0,i} \in \Theta \subseteq \mathbb{R}^d$, $m_0 = 0$, $v_0 = \epsilon$, $\forall i \in [\![n]\!]$, and $\theta_0 = \frac{1}{n}\sum_{i=1}^n \theta_{0,i}$.
 3: **for** $r = 1, \dots, R$ **do**
 4:     **parallel for device** $i$ **do**:
 5:         Set $\theta_{r,i}^0 = \theta_{r-1}$.
 6:         **for** $t = 1, \dots, T$ **do**
 7:             Compute stochastic gradient $g_{r,i}^t$ at $\theta_{r,i}^0$.
 8:             $\theta_{r,i}^t = \theta_{r,i}^{t-1} - \eta_l g_{r,i}^t$
 9:         **end for**
10:         Devices send $\triangle_{r,i} = \theta_{r,i}^T - \theta_{r,i}^0$ to server.
11:     **end for**
12:     Server computes $\bar{\triangle}_r = \frac{1}{n}\sum_{i=1}^n \triangle_{r,i}$
13:     $m_r = \beta_1 m_{r-1} + (1-\beta_1)\bar{\triangle}_r$
14:     $v_r = \beta_2 v_{r-1} + (1-\beta_2)\bar{\triangle}_r^2$
15:     $\theta_r = \theta_{r-1} + \eta_g \frac{m_r}{\sqrt{v_r}}$
16: **end for**
17: **Output**: Global model parameter $\theta_R$.
```

*Accepted for the $39^{th}$ Conference on Uncertainty in Artificial Intelligence* (UAI 2023).

## A.2 HYPER-PARAMETER TUNING

In our empirical study, we tune the learning rate of each algorithm carefully such that the best performance is achieved. The search grids in all our experiments are provided in Table 1.

Table 1: Search grids of the learning rate.

| | Learning rate range |
|---|---|
| Fed-SGD | $[0.001, 0.003, 0.005, 0.01, 0.03, 0.05, 0.1, 0.3, 0.5]$ |
| Fed-AMS | $[0.0001, 0.0003, 0.0005, 0.001, 0.003, 0.005, 0.01, 0.03, 0.05, 0.1]$ |
| Fed-LAMB | $[0.001, 0.003, 0.005, 0.01, 0.03, 0.05, 0.1, 0.3, 0.5]$ |
| Adp-Fed | Local $\eta_l$: $[0.0001, 0.0003, 0.0005, 0.001, 0.003, 0.005, 0.01, 0.03, 0.05, 0.1, 0.3, 0.5]$<br>Global $\eta_g$: $[0.0001, 0.0003, 0.0005, 0.001, 0.003, 0.005, 0.01, 0.03, 0.05, 0.1]$ |
| Mime | $[0.0001, 0.0003, 0.0005, 0.001, 0.003, 0.005, 0.01, 0.03, 0.05, 0.1]$ |
| Mime-LAMB | $[0.001, 0.003, 0.005, 0.01, 0.03, 0.05, 0.1, 0.3, 0.5]$ |

## A.3 MORE WORKERS

In Figure 1, we provide additional figures with larger number of workers $n = 200$, on MNIST and FMNIST with non-IID data. The conclusions stay the same: we see that the proposed Fed-LAMB and Mime-LAMB perform much better than the baseline algorithms, with faster convergence and better accuracy at the end of 100 FL training rounds.

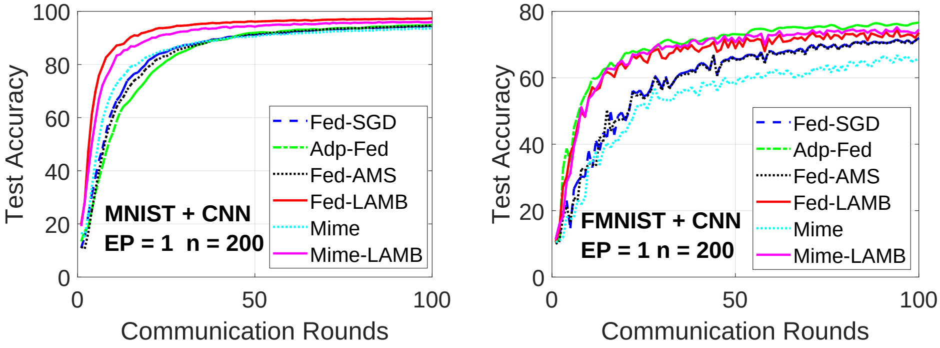

Figure 1: Test accuracy on MNIST and FMNIST with $n = 200$ workers, full participation, local batch size 64. Data are non-IID distributed among clients.

## B THEORETICAL ANALYSIS

We first recall in Table 2 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 |
| $\mathsf{h}, \ell$ | $:=$ | Total number of layers in the DNN and its index |
| $\phi(\cdot)$ | $:=$ | Scaling factor in Fed-LAMB update |
| $\bar{\theta}$ | $:=$ | Global model (after periodic averaging) |
| $\psi_{r,i}^{t}$ | $:=$ | ratio computed at round $r$, local iteration $t$ and for device $i$. $\psi_{r,i}^{\ell,t}$ denotes its component at layer $\ell$ |

Table 2: 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 5.

### B.1 INTERMEDIARY LEMMA

We now develop the proof of the convergence rate of Fed-LAMB. We need a supporting Lemma 1 for this.

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

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

*where $\phi_M$ is defined in Assumption 4 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 $r > 0$:

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

leading to

$$\|\overline{\theta_r} - \theta_{r,i}\|^2 = \sum_{\ell=1}^{\mathsf{h}} \|\overline{\theta_r}^{\ell} - \theta_{r,i}^{\ell}\|^2 \leq \alpha^2 M^2 T^2 \phi_M^2 \frac{(1-\beta_2)p}{\epsilon} ,$$

which concludes the proof. □

### B.2 PROOF OF THEOREM 5

**Theorem.** Suppose **Assumption 1 - Assumption 4** holds. Consider $\{\overline{\theta_r}\}_{r>0}$, the sequence of parameters obtained running Algorithm 1 with a constant learning rate $\alpha$. Let the number of local epochs be $T \geq 1$ and $\lambda = 0$. Then, for any round $R > 0$, we have

$$\frac{1}{R}\sum_{r=1}^{R} \mathbb{E}\left[\left\|\frac{\nabla f(\overline{\theta_r})}{\hat{v}_r^{1/4}}\right\|^2\right] \leq \sqrt{\frac{M^2 p}{n}}\frac{\triangle}{\mathsf{h}\alpha R} + \frac{4\alpha^2 L M^2 T^2 \phi_M^2 (1-\beta_2)p}{\sqrt{\epsilon}} \tag{1}$$

$$+ 4\alpha^2 \frac{M^2}{\sqrt{\epsilon}} + \frac{\phi_M \sigma^2}{Rn}\sqrt{\frac{1-\beta_2}{M^2 p}} + 4\alpha\left[\phi_M \frac{\mathsf{h}\sigma^2}{\sqrt{n}}\right] + 4\alpha^2\left[\phi_M^2 \sqrt{M^2 + p\sigma^2}\right] ,$$

where $\triangle = \mathbb{E}[f(\bar{\theta}_1)] - \min_{\theta \in \Theta} f(\theta)$.

*Proof.* Our proof will make use of an intermediary virtual sequence defined as

$$\bar{\vartheta}_r = \bar{\theta}_r + \frac{\beta_1}{1-\beta_1}(\bar{\theta}_r - \bar{\theta}_{r-1}) \ , \tag{2}$$

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

$$\begin{aligned}
\bar{\vartheta}^{\ell}_{r+1} - \bar{\vartheta}^{\ell}_{r} &= \frac{1}{1-\beta_1}(\bar{\theta}^{\ell}_{r+1} - \bar{\theta}^{\ell}_{r}) - \frac{\beta_1}{1-\beta_1}(\bar{\theta}^{\ell}_{r} - \bar{\theta}^{\ell}_{r-1}) \\
&= \frac{\alpha_r}{1-\beta_1}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta^{\ell}_{r,i}\|)}{\|\psi^{\ell}_{r,i}\|}\psi^{\ell}_{r,i} - \frac{\alpha_{r-1}}{1-\beta_1}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta^{\ell}_{r-1,i}\|)}{\|\psi^{\ell}_{r-1,i}\|}\psi^{\ell}_{r-1,i} \\
&= \frac{\alpha\beta_1}{1-\beta_1}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\phi(\|\theta^{\ell}_{r,i}\|)}{\sqrt{v^t_r}\|\psi^{\ell}_{r,i}\|} - \frac{\phi(\|\theta^{\ell}_{r-1,i}\|)}{\sqrt{v^t_{r-1}}\|\psi^{\ell}_{r-1,i}\|}\right)m^t_{r-1} + \frac{\alpha}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta^{\ell}_{r,i}\|)}{\sqrt{v^t_r}\|\psi^{\ell}_{r,i}\|}g^t_{r,i} \ ,
\end{aligned} \tag{3}$$

where we have assumed a constant learning rate $\alpha$.

Using Assumption 1, we have

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

Taking expectations on both sides leads to

$$-\mathbb{E}[\big\langle \nabla f(\bar{\vartheta}_r) \,|\, \bar{\vartheta}_{r+1} - \bar{\vartheta}_r \big\rangle] \leq \mathbb{E}[f(\bar{\vartheta}_r) - f(\bar{\vartheta}_{r+1})] + \sum_{\ell=1}^{L}\frac{L_\ell}{2}\mathbb{E}[\|\bar{\vartheta}^{\ell}_{r+1} - \bar{\vartheta}^{\ell}_{r}\|^2] \ . \tag{4}$$

We note for all $\theta \in \Theta$, the majorant $G > 0$ such that $\phi(\|\theta\|) \leq G$. Then, following (4), we obtain

$$-\mathbb{E}[\big\langle \nabla f(\bar{\vartheta}_r) \,|\, \bar{\vartheta}_{r+1} - \bar{\vartheta}_r \big\rangle] \leq \mathbb{E}[f(\bar{\vartheta}_r) - f(\bar{\vartheta}_{r+1})] + \sum_{\ell=1}^{L}\frac{L_\ell}{2}\mathbb{E}[\|\bar{\vartheta}_{r+1} - \bar{\vartheta}_{r}\|^2] \ . \tag{5}$$

Developing the LHS of (5) using (3) leads to

$$\begin{aligned}
\big\langle \nabla f(\bar{\vartheta}_r) \,|\, \bar{\vartheta}_{r+1} - \bar{\vartheta}_r \big\rangle &= \sum_{\ell=1}^{\mathsf{h}}\sum_{j=1}^{p_\ell}\nabla_\ell f(\bar{\vartheta}_r)^j(\bar{\vartheta}^{\ell,j}_{r+1} - \bar{\vartheta}^{\ell,j}_{r}) \\
&= \frac{\alpha\beta_1}{1-\beta_1}\frac{1}{n}\sum_{\ell=1}^{\mathsf{h}}\sum_{j=1}^{p_\ell}\nabla_\ell f(\bar{\vartheta}_r)^j\left[\sum_{i=1}^{n}\left(\frac{\phi(\|\theta^{\ell}_{r,i}\|)}{\sqrt{v^t_r}\|\psi^{\ell}_{r,i}\|} - \frac{\phi(\|\theta^{\ell}_{r-1,i}\|)}{\sqrt{v^t_{r-1}}\|\psi^{\ell}_{r-1,i}\|}\right)m^t_{r-1}\right] \\
&\quad \underbrace{-\frac{\alpha}{n}\sum_{\ell=1}^{\mathsf{h}}\sum_{j=1}^{p_\ell}\nabla_\ell f(\bar{\vartheta}_r)^j\sum_{i=1}^{n}\frac{\phi(\|\theta^{\ell}_{r,i}\|)}{\sqrt{v^t_r}\|\psi^{\ell}_{r,i}\|}g^{t,l,j}_{r,i}}_{=A_1} \ .
\end{aligned} \tag{6}$$

Suppose $T$ is the total number of local iterations and $R$ is the number of rounds. We can write (6) as

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

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

$$\langle \nabla f(\bar{\vartheta}_r), \frac{\bar{g}_r}{\sqrt{\hat{v}_r}}\rangle = \frac{1}{2}\|\frac{\nabla f(\bar{\vartheta}_r)}{\hat{v}_r^{1/4}}\|^2 + \frac{1}{2}\|\frac{\bar{g}_r}{\hat{v}_r^{1/4}}\|^2 - \|\frac{\nabla f(\bar{\vartheta}_r) - \bar{g}_r}{\hat{v}_r^{1/4}}\|^2. \tag{7}$$

To bound the last term on the RHS, we have

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

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

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

For the second term,

$$\frac{2}{n}\sum_{i=1}^n\|\frac{\nabla f(\bar{\theta}_r)-\bar{g}_{t,i}}{\hat{v}_r^{1/4}}\|^2 \leq \frac{4}{n}\Big(\underbrace{\sum_{i=1}^n\|\frac{\nabla f(\bar{\theta}_r)-\nabla f(\theta_{t,i})}{\hat{v}_r^{1/4}}\|^2}_{B_1}+\underbrace{\sum_{i=1}^n\|\frac{\nabla f(\theta_{t,i})-\bar{g}_{t,i}}{\hat{v}_r^{1/4}}\|^2}_{B_2}\Big). \tag{8}$$

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

$$\begin{aligned}B_1 \leq \frac{L}{\sqrt{\epsilon}}\sum_{i=1}^n\|\bar{\theta}_r-\theta_{t,i}\|^2 &= \frac{\alpha^2L}{\sqrt{\epsilon}}\sum_{i=1}^n\sum_{l=1}^L\|\sum_{j=\lfloor t\rfloor_r+1}^t\Big(\frac{\phi(\|\theta_{j,i}^l\|)}{\|\psi_{j,i}^l\|}\psi_{j,i}^l-\frac{1}{n}\sum_{k=1}^n\frac{\phi(\|\theta_{j,k}^l\|)}{\|\psi_{j,k}^l\|}\psi_{j,k}^l\Big)\|^2\\ &\leq n\frac{\alpha^2L}{\sqrt{\epsilon}}M^2T^2\phi_M^2(1-\beta_2)p,\end{aligned} \tag{9}$$

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 $(t-1)-(\lfloor t\rfloor_r+1)+1\leq T-1$.

We now bound $B_2$ (under the simplification that $\beta_1=0$):

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

where we use Assumption 2, Assumption 3 and Assumption 4. Yet,

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

Then we have

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

Thus, (8) becomes

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

Substituting all ingredients into (7), we obtain

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

To bound the second term on the RHS in above, we notice that

$$
\begin{aligned}
\mathbb{E}[\|\frac{\bar{g}_r}{\hat{v}_r^{1/4}}\|^2] = \frac{1}{n^2}\mathbb{E}[\|\frac{\sum_{i=1}^n \bar{g}_{r,i}}{\hat{v}_r^{1/4}}\|^2] &= \frac{1}{n^2}\mathbb{E}[\sum_{l=1}^L\sum_{i=1}^n \|\frac{\phi(\|\theta_{r,i}^l\|)}{\hat{v}^{1/4}\|\psi_{r,i}^l\|} g_{r,i}^l\|^2] \\
&\geq \phi_m^2(1-\beta_2)\mathbb{E}\left[\|\frac{1}{n}\sum_{i=1}^n \frac{\nabla f(\theta_{r,i})}{\hat{v}_r^{1/4}}\|^2\right] \\
&= \phi_m^2(1-\beta_2)\mathbb{E}\left[\|\frac{\overline{\nabla} f(\theta_r)}{\hat{v}_r^{1/4}}\|^2\right].
\end{aligned}
\tag{10}
$$

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

$$
\begin{aligned}
\left\|\frac{\overline{\nabla} f(\theta_r)}{\hat{v}_r^{1/4}}\right\|^2 &\geq \frac{1}{2}\left\|\frac{\nabla f(\overline{\theta_r})}{\hat{v}_r^{1/4}}\right\|^2 - \left\|\frac{\overline{\nabla} f(\theta_r) - \nabla f(\overline{\theta_r})}{\hat{v}_r^{1/4}}\right\|^2 \\
&\geq \frac{1}{2}\left\|\frac{\nabla f(\overline{\theta_r})}{\hat{v}_r^{1/4}}\right\|^2 - \left\|\frac{\frac{1}{n}\sum_{i=1}^n(\nabla f_i(\theta_r) - \nabla f(\bar{\theta}_r))}{\hat{v}_r^{1/4}}\right\|^2 \\
&\geq \frac{1}{2}\left\|\frac{\nabla f(\overline{\theta_r})}{\hat{v}_r^{1/4}}\right\|^2 - \frac{\alpha^2 L_\ell}{\sqrt{\epsilon}} M^2 T^2 (\sigma^2 + G^2)(1-\beta_2)p,
\end{aligned}
$$

where the last line is due to (9) and Assumption 3. Therefore, we have obtained

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

Substitute back into (6), assuming $M \leq 1$, we have the following by taking the telescope sum

$$\begin{aligned}
& \frac{1}{R}\sum_{t=1}^{R}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{\theta_r})}{\hat{v}_r^{1/4}}\right\|^2\right] \\
& \lesssim \sqrt{\frac{M^2 p}{n}}\frac{f(\bar{\vartheta}_1)-\mathbb{E}[f(\bar{\vartheta}_{R+1})]}{\mathsf{h}\alpha R} + \frac{\alpha}{n^2}\sum_{r=1}^{R}\sum_{i=1}^{n}\sigma_i^2\mathbb{E}\left[\left\|\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r}\|\psi_{r,i}^\ell\|}\right\|^2\right] + \frac{2\alpha^3\overline{L}p\phi_M^2}{\sqrt{\epsilon}}\frac{\beta_1^2}{(1-\beta_1)^2} \\
& + 4\Big[\frac{\alpha^2\overline{L}}{\sqrt{\epsilon}}M^2T^2G^2(1-\beta_2)p + \frac{\alpha M^2}{\sqrt{\epsilon}} + \alpha\phi_M^2\sqrt{M^2+p\sigma^2} \\
& + \sigma^2\left(\frac{\alpha^2\overline{L}}{\sqrt{\epsilon}}M^2T^2(1-\beta_2)p + \phi_M\alpha\frac{\mathsf{h}}{\sqrt{n}}\right)\Big] + \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{h}M^2}{\sqrt{\epsilon}} + \overline{L}\alpha^2M^2\phi_M^2\frac{(1-\beta_2)p}{T\epsilon} \\
& \leq \sqrt{\frac{M^2 p}{n}}\frac{\mathbb{E}[f(\bar{\theta}_1)] - \min_{\theta\in\Theta} f(\theta)}{\mathsf{h}\alpha R} + \frac{\phi_M\sigma^2}{Rn}\sqrt{\frac{1-\beta_2}{M^2 p}} \\
& + 4\Big[\frac{\alpha^2\overline{L}}{\sqrt{\epsilon}}M^2T^2G^2(1-\beta_2)p + \frac{M^2\alpha}{\sqrt{\epsilon}} + \phi_M^2\alpha\sqrt{M^2+p\sigma^2} \\
& + \sigma^2\Big(\frac{\alpha^2\overline{L}}{\sqrt{\epsilon}}M^2T^2(1-\beta_2)p + \phi_M\frac{\mathsf{h}}{\sqrt{n}}\Big)\Big] + \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{h}M^2}{\sqrt{\epsilon}} \\
& \qquad + \overline{L}\alpha^2M^2\phi_M^2\frac{(1-\beta_2)p}{T\epsilon} + \frac{2\alpha^3\overline{L}p\phi_M^2}{\sqrt{\epsilon}}\frac{\beta_1^2}{(1-\beta_1)^2}.
\end{aligned}$$

Organizing terms, we conclude the proof. □

## References

Xiangyi Chen, Xiaoyun Li, and Ping Li. Toward communication efficient adaptive gradient method. In *Proceedings of the ACM-IMS Foundations of Data Science Conference (FODS)*, pages 119–128, Virtual Event, USA, 2020.

Xiaoyun Li and Ping Li. Analysis of error feedback in federated non-convex optimization with biased compression: Fast convergence and partial participation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Honolulu, HI, 2023.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021.