Appendix for "Layerwise and Dimensionwise Local Adaptive Method for Federated Learning "

A. Theoretical Analysis

We first provide the proofs for the theoretical results of the main paper, including the intermediary Lemmas and the main convergence result, Theorem 1. Then, we provide the extension of this result *to the setting of multiple local updates* in Section A.4.

A.1. Intermediary Lemmas

Lemma. *Consider $\{\overline{\theta_r}\}_{r>0}$, the sequence of parameters obtained running Algorithm 2. 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 $p = \sum_{\ell=1}^{\mathsf{L}} p_\ell$.

Proof. Assuming the simplest case when $T = 1$, i.e. one local iteration, then by construction of Algorithm 2, we have for all $\ell \in [\![\mathsf{L}]\!]$, $i \in [\![n]\!]$ 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

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

which concludes the proof. □

Lemma. *Consider $\{\overline{\theta_r}\}_{r>0}$, the sequence of parameters obtained running Algorithm 2. 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 $p = \sum_{\ell=1}^{\mathsf{L}} p_\ell$ 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

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

concluding the proof. □

A.2. Proof of Theorem 1

Theorem. *Consider $\{\overline{\theta_r}\}_{r>0}$, the sequence of parameters obtained running Algorithm 2. Then, if the number of local epochs is set to $T=1$ and $\epsilon = \lambda = 0$, we have:*

$$\frac{1}{R}\sum_{r=1}^{R} \mathbb{E}[\|\nabla f(\overline{\theta_r})\|^2 \leq dd$$

Case with $T=1$, $\epsilon=0$ and $\lambda=0$: Using H1, we have:

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

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 (Reddi et al., 2018), 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,

$$\bar{\vartheta}_{r+1}^\ell - \bar{\vartheta}_r^\ell = \frac{1}{1-\beta_1}(\bar{\theta}_{r+1}^\ell - \bar{\theta}_r^\ell) - \frac{\beta_1}{1-\beta_1}(\bar{\theta}_r^\ell - \bar{\theta}_{r-1}^\ell) \tag{9}$$
$$= \frac{\alpha_r}{1-\beta_1}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^\ell\|)}{\|p_{r,i}^\ell\|}p_{r,i}^\ell - \frac{\alpha_{r-1}}{1-\beta_1}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\|p_{r-1,i}^\ell\|}p_{r-1,i}^\ell \tag{10}$$
$$= \frac{\alpha\beta_1}{1-\beta_1}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right) m_{r-1}^t + \frac{\alpha}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}g_{r,i} \tag{11}$$

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:

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

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

$$
\begin{aligned}
\left\langle \nabla f(\bar{\vartheta}_r) \,|\, \bar{\vartheta}_{r+1} - \bar{\vartheta}_r \right\rangle &= \sum_{\ell=1}^{\mathsf{L}} \sum_{j=1}^{p_\ell} \nabla_\ell f(\bar{\vartheta}_r)^j (\bar{\vartheta}_{r+1}^{\ell,j} - \bar{\vartheta}_r^{\ell,j}) && (13) \\
&= \frac{\alpha\beta_1}{1-\beta_1} \frac{1}{n} \sum_{\ell=1}^{\mathsf{L}} \sum_{j=1}^{p_\ell} \nabla_\ell f(\bar{\vartheta}_r)^j \left[\sum_{i=1}^{n} \left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|} \right) m_{r-1}^t \right] && (14) \\
&\underbrace{- \frac{\alpha}{n} \sum_{\ell=1}^{\mathsf{L}} \sum_{j=1}^{p_\ell} \nabla_\ell f(\bar{\vartheta}_r)^j \sum_{i=1}^{n} \frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} g_{r,i}^{l,j}}_{=A_1} && (15)
\end{aligned}
$$

Term A_1: Since we have that $\|p_{r,i}^\ell\| \leq \sqrt{\frac{p_\ell}{1-\beta_2}}$ and $1/\sqrt{v_r^t} \leq 1/\sqrt{v_0}$, using H2, we develop the term A_1 as follows:

$$
\begin{aligned}
A_1 &\leq -\frac{\alpha}{n} \sum_{\ell=1}^{\mathsf{L}} \sum_{j=1}^{p_\ell} \nabla_\ell f(\bar{\vartheta}_r)^j \sum_{i=1}^{n} \frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} g_{r,i}^{l,j} \\
&\leq -\frac{\alpha}{n} \sum_{\ell=1}^{\mathsf{L}} \sqrt{\frac{1-\beta_2}{M^2 p_\ell}} \sum_{i=1}^{n} \sum_{j=1}^{p_\ell} \phi(\|\theta_{r,i}^\ell\|) \nabla_\ell f(\bar{\vartheta}_r)^j g_{r,i}^{\ell,j} \\
&\quad - \frac{\alpha}{n} \sum_{\ell=1}^{\mathsf{L}} \sum_{i=1}^{n} \sum_{j=1}^{p_\ell} \left(\phi(\|\theta_{r,i}^\ell\|) \nabla_\ell f(\bar{\vartheta}_r)^j \frac{p_{r,i}^\ell}{\|p_{r,i}^\ell\|} \right) \mathbb{1} \left(\text{sign}(\nabla_\ell f(\bar{\vartheta}_r)^j) \neq \text{sign}(g_{r,i}^{\ell,j}) \right)
\end{aligned}
$$

Taking the expectations on both sides yields and using H4:

$$
\begin{aligned}
\mathbb{E}[A_1] &\leq -\alpha \sum_{\ell=1}^{\mathsf{L}} \sqrt{\frac{1-\beta_2}{M^2 p_\ell}} \sum_{i=1}^{n} \sum_{j=1}^{p_\ell} \mathbb{E}\left[\phi(\|\theta_{r,i}^\ell\|) \nabla_\ell f(\bar{\vartheta}_r)^j g_{r,i}^{\ell,j} \right] \\
&\quad - \frac{\alpha}{n} \sum_{\ell=1}^{\mathsf{L}} \sum_{i=1}^{n} \sum_{j=1}^{p_\ell} \mathbb{E}\left[\phi(\|\theta_{r,i}^\ell\|) \nabla_\ell f(\bar{\vartheta}_r)^j \frac{p_{r,i}^\ell}{\|p_{r,i}^\ell\|} \mathbb{1} \left(\text{sign}(\nabla_\ell f(\bar{\vartheta}_r)^j) \neq \text{sign}(g_{r,i}^{\ell,j}) \right) \right] \\
&\leq -\frac{\alpha}{n} \sum_{\ell=1}^{\mathsf{L}} \phi_m \sqrt{\frac{1-\beta_2}{M^2 p_\ell}} \sum_{i=1}^{n} \sum_{j=1}^{p_\ell} (\nabla_\ell f(\bar{\vartheta}_r)^j)^2 \\
&\quad - \frac{\alpha}{n} \sum_{\ell=1}^{\mathsf{L}} \sum_{i=1}^{n} \sum_{j=1}^{p_\ell} \phi_M \mathbb{E}\left[\left| \nabla_\ell f(\bar{\vartheta}_r)^j \frac{p_{r,i}^\ell}{\|p_{r,i}^\ell\|} \right| \mathbb{1} \left(\text{sign}(\nabla_\ell f(\bar{\vartheta}_r)^j) \neq \text{sign}(g_{r,i}^{\ell,j}) \right) \right]
\end{aligned}
$$

where we have used assumption H4.

Since for any ℓ, i, j, we have

$$
\mathbb{E}\left[\left| \nabla_\ell f(\bar{\vartheta}_r)^j \frac{p_{r,i}^\ell}{\|p_{r,i}^\ell\|} \right| \mathbb{1} \left(\text{sign}(\nabla_\ell f(\bar{\vartheta}_r)^j) \neq \text{sign}(g_{r,i}^{\ell,j}) \right) \right] \leq \left| \nabla_\ell f(\bar{\vartheta}_r)^j \right| \mathbb{P}\left(\text{sign}(\nabla_\ell f(\bar{\vartheta}_r)^j) \neq \text{sign}(g_{r,i}^{\ell,j}) \right)
$$

Then, we obtain

$$
\mathbb{E}[A_1] \leq -\alpha\phi_m \sqrt{\frac{\mathsf{L}(1-\beta_2)}{M^2 p}} \mathbb{E}[\|\overline{\nabla f}(\bar{\vartheta}_r)\|^2] - \alpha\phi_M \sum_{\ell=1}^{\mathsf{L}} \sum_{i=1}^{n} \sum_{j=1}^{p_\ell} \frac{\sigma_i^{\ell,j}}{\sqrt{n}} \tag{16}
$$

where for any $\theta \in \Theta$ we define $\overline{\nabla f}(\cdot) = \sum_{i=1}^n \nabla f_i(\cdot)$.

We now need to bound the following terms:

$$
\begin{aligned}
&A_r^2 := \mathbb{E}[\|\bar{\vartheta}_{r+1} - \bar{\vartheta}_r\|^2] \\
&A_r^3 := \frac{\alpha\beta_1}{1-\beta_1}\frac{1}{n}\sum_{\ell=1}^{\mathsf{L}}\sum_{j=1}^{p_\ell} \nabla_\ell f(\bar{\vartheta}_r)^j \left[\sum_{i=1}^n \left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right) m_{r-1}^t\right]
\end{aligned}
$$

Term A_r^2: According to definition (8), for each layer $\ell \in [\![\mathsf{L}]\!]$, we have, using the Cauchy-Schwartz inequality, that:

$$
\begin{aligned}
\|\bar{\vartheta}_{r+1}^\ell - \bar{\vartheta}_r^\ell\|^2 &= \left\|\frac{\alpha\beta_1}{1-\beta_1}\frac{1}{n}\sum_{i=1}^n \left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right) m_{r-1}^t + \frac{\alpha}{n}\sum_{i=1}^n \frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} g_{r,i}\right\|^2 \\
&\leq 2\frac{\alpha^2}{n^2}\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^n \left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right) m_{r-1}^t\right\|^2 + \frac{\alpha^2}{n^2}\left\|\sum_{i=1}^n \frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} g_{r,i}\right\|^2
\end{aligned}
$$

Taking the expectation on both sides leads to:

$$
\begin{aligned}
\mathbb{E}[\|\bar{\vartheta}_{r+1}^\ell - \bar{\vartheta}_r^\ell\|^2] &\leq 2\alpha^2\mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^n \left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right) m_{r-1}^t\right\|^2\right] + \frac{\alpha^2}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n \frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} g_{r,i}\right\|^2\right] \\
&\leq 2\frac{\alpha^2}{n^2}\mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^n \left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right) m_{r-1}^t\right\|^2\right] \\
&+ \frac{\alpha^2}{n^2}\mathbb{E}\left[\left|\sum_{i=1}^n\sum_{j=1}^p \left\langle \Gamma_{r,i}^j(\nabla f_i(\theta_r)^j + g_{r,i}^j - \nabla f_i(\theta_r)^j) \,|\, \Gamma_{r,i}^j(\nabla f_i(\theta_r)^j + g_{r,i}^j - \nabla f_i(\theta_r)^j)\right\rangle\right|\right] \\
&\leq 2\frac{\alpha^2}{n^2}\mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^n \left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right) m_{r-1}^t\right\|^2\right] \\
&+ \frac{\alpha^2}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n \frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\nabla f_i(\theta_r)\right\|^2\right] + \frac{\alpha^2}{n^2}\sum_{i=1}^n \sigma_i^2\mathbb{E}\left[\left\|\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\right\|^2\right]
\end{aligned}
\tag{17}
$$

where the last line uses assumptions H2 and H3 (unbiased gradient and bounded variance of the stochastic gradient) and $\Gamma_{r,i}^\ell := \frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}$.

On the other hand, using the bound on the gradient H2,

$$
\begin{aligned}
&\sum_{r=1}^{R}\mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^{n}\left(\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}-\frac{\phi(\|\theta_{r-1,i}^{\ell}\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^{\ell}\|}\right)m_{r-1}^t\right\|^2\right]\\
&\leq \frac{\beta_1^2}{(1-\beta_1)^2}M^2\phi_M^2\sum_{r=1}^{R}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\left(\frac{1}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}-\frac{1}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^{\ell}\|}\right)\right\|^2\right]\\
&\leq \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\sum_{r=1}^{R}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\left(\frac{1}{\sqrt{v_r^t}}-\frac{1}{\sqrt{v_{r-1}^t}}\right)\right\|^2\right]\\
&\leq \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\sum_{r=1}^{R}\mathbb{E}\left[\left|\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\frac{1}{\sqrt{v_r^{t,j}}}-\frac{1}{\sqrt{v_{r-1}^{t,j}}}\right)\right|\right]\\
&\leq \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\frac{np}{v_0}
\end{aligned}
\tag{18}
$$

where, in the telescopic sum, we have used the initial value v_0 of the non decreasing sequence $\{v_r^t\}_{r>0}$ by construction (max operator).

Combining (18) into (17) and summing over the total number of rounds R yields

$$
\begin{aligned}
\sum_{r=1}^{R}A_r^2 := \sum_{r=1}^{R}\mathbb{E}[\|\bar{\vartheta}_{r+1}^{\ell}-\bar{\vartheta}_r^{\ell}\|^2] \leq \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\frac{np}{v_0}\\
+\sum_{r=1}^{R}\left[\frac{\alpha^2}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}\nabla f_i(\theta_r)\right\|^2\right]+\frac{\alpha^2}{n^2}\sum_{i=1}^{n}\sigma_i^2\mathbb{E}\left[\left\|\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}\right\|^2\right]\right]
\end{aligned}
\tag{19}
$$

Term A_r^3: According to similar arguments on the non decreasing sequence involved in the algorithm as in the previous series of calculations, observe that

$$
\sum_{r=1}^{R}A_r^3 \leq \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{L}M^2}{\sqrt{v_0}}
\tag{20}
$$

Plugging (16) into (12) combined with (19) and (20) injected into the original smoothness definition (7) summed over the total number of rounds:

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

gives:

$$
\begin{aligned}
&\sum_{r=1}^{R}\alpha\phi_m\sqrt{\frac{\mathsf{L}(1-\beta_2)}{M^2p}}\mathbb{E}[\|\overline{\nabla}f(\bar{\vartheta}_r)\|^2]-\alpha\phi_M\sum_{\ell=1}^{\mathsf{L}}\sum_{i=1}^{n}\sum_{j=1}^{p_\ell}\frac{\sigma_i^{\ell,j}}{\sqrt{n}}+\frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{L}M^2}{\sqrt{v_0}}\\
&\leq \sum_{r=1}^{R}\mathbb{E}[f(\bar{\vartheta}_r)-f(\bar{\vartheta}_{r+1})]+\sum_{\ell=1}^{L}\frac{L_\ell}{2}\frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\frac{np}{v_0}\\
&\quad-\sum_{r=1}^{R}\left[\frac{\alpha^2}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}\nabla f_i(\theta_r)\right\|^2\right]+\frac{\alpha^2}{n^2}\sum_{i=1}^{n}\sigma_i^2\mathbb{E}\left[\left\|\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}\right\|^2\right]\right]
\end{aligned}
$$

Noting that $\sum_{r=1}^{R} \mathbb{E}[f(\bar{\vartheta}_r) - f(\bar{\vartheta}_{r+1})] = f(\bar{\vartheta}_1) - \mathbb{E}[f(\bar{\vartheta}_{R+1})]$, we obtain

$$
\begin{aligned}
&\sum_{r=1}^{R} \alpha\phi_m \sqrt{\frac{\mathsf{L}(1-\beta_2)}{M^2 p}} \mathbb{E}[\|\overline{\nabla} f(\bar{\vartheta}_r)\|^2] + \frac{1}{n^2}\sum_{r=1}^{R} \mathbb{E}\left[\left\|\sum_{i=1}^{n} \frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}\nabla f_i(\theta_r)\right\|^2\right] \\
&\leq f(\bar{\vartheta}_1) - \mathbb{E}[f(\bar{\vartheta}_{R+1})] + \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 \mathbb{E}\left[\left\|\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}\right\|^2\right] + \alpha\phi_M \sum_{\ell=1}^{\mathsf{L}}\sum_{i=1}^{n}\sum_{j=1}^{p_\ell} \frac{\sigma_i^{\ell,j}}{\sqrt{n}} + \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{L}M^2}{\sqrt{v_0}} \\
&+ \sum_{\ell=1}^{L} \frac{L_\ell}{2} \frac{\beta_1^2}{(1-\beta_1)^2} \frac{\mathsf{L}(1-\beta_2)}{p} M^2 \phi_M^2 \frac{np}{v_0}
\end{aligned}
$$

leading to

$$
\begin{aligned}
\sum_{r=1}^{R} \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}\nabla f_i(\theta_r)\right\|^2\right] &\leq \frac{1}{\alpha}\left[f(\bar{\vartheta}_1) - \mathbb{E}[f(\bar{\vartheta}_{R+1})]\right] + \sum_{r=1}^{R}\frac{\alpha}{n^2}\sum_{i=1}^{n}\sigma_i^2 \mathbb{E}\left[\left\|\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}\right\|^2\right] \\
&+ \alpha\phi_M \sigma \mathsf{L} p\sqrt{n} + \frac{\overline{L}\beta_1^2 \mathsf{L}(1-\beta_2)M^2\phi_M^2 n}{2(1-\beta_1)^2 v_0} + \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{L}M^2}{\sqrt{v_0}}
\end{aligned}
\tag{21}
$$

where $\overline{L} = \sum_{\ell=1}^{\mathsf{L}} L_\ell$ is the sum of all smoothness constants.

Consider the following inequality:

$$
\frac{1}{n}\sum_{i=1}^{n} \frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}\nabla f_i(\theta_r) \leq \phi_M(1-\beta_2)\frac{\overline{\nabla} f(\theta_r)}{\sqrt{v_r^t}}
$$

where $\overline{\nabla} f(\theta_r) := \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\theta_r)$. And using the Cauchy-Schwartz inequality we have

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

Using Lemma 1 and the smoothness assumption H1, we have

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

which is due to Lemma 2. Plugging the above inequality into (21) and dividing both sides by R yields:

$$
\begin{aligned}
\frac{1}{R}\sum_{r=1}^{R} \mathbb{E}\left[\left\|\frac{\nabla f(\overline{\theta_r})}{\sqrt{v_r^t}}\right\|^2\right] &\leq \sqrt{\frac{M^2 p}{n}}\frac{f(\bar{\vartheta}_1) - \mathbb{E}[f(\bar{\vartheta}_{R+1})]}{\mathsf{L}\alpha R} + \frac{\alpha}{n^2}\sum_{r=1}^{R}\sum_{i=1}^{n}\sigma_i^2 \mathbb{E}\left[\left\|\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}\right\|^2\right] \\
&+ \alpha\phi_M \sigma \mathsf{L} p\sqrt{n} + \frac{\overline{L}\beta_1^2 \mathsf{L}(1-\beta_2)M^2\phi_M^2 n}{2(1-\beta_1)^2 v_0} + \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{L}M^2}{\sqrt{v_0}} + \overline{L}\alpha^2 M^2\phi_M^2 \frac{(1-\beta_2)p}{R v_0}
\end{aligned}
$$

Then using the fact that $\vartheta_1 = \theta_1$ and that for all $\vartheta \in \Theta$, $-f(\vartheta) \leq -\min_{\theta\in\Theta} f(\theta)$ concludes the proof of our main convergence result.

A.3. Proof Corollary 1

Corollary. *Assume **H1-H4**. Consider $\{\overline{\theta_r}\}_{r>0}$, the sequence of parameters obtained running Algorithm 2. Then, if the number of local epochs is set to $T = 1$, $\epsilon = \lambda = 0$, and $R \geq$ we have, under **H5**:*

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

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

A.4. Extension of Theorem 1 to multiple local updates

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 3. *Assume **H1-H4**. Consider $\{\overline{\theta_r}\}_{r>0}$, the sequence of parameters obtained running Algorithm 2 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:*

$$\begin{aligned}\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{\theta_t})}{\hat{v}_t^{1/4}}\right\|^2\right] \leq \sqrt{\frac{M^2 p}{n}}\frac{\mathbb{E}[f(\bar{\theta}_1)] - \min_{\theta\in\Theta} f(\theta)}{\mathsf{L}\alpha_r\tau} + \frac{\phi_M\sigma^2}{\tau n}\sqrt{\frac{1-\beta_2}{M^2 p}} \\ + 4\alpha\left[\frac{\alpha^2 L_\ell}{\sqrt{v_0}}M^2(T-1)^2\phi_M^2(1-\beta_2)p + \frac{M^2}{\sqrt{v_0}} + \phi_M^2\sqrt{M^2+p\sigma^2} + \phi_M\frac{\mathsf{L}\sigma^2}{\sqrt{n}}\right] + cst.\end{aligned} \tag{22}$$

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

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

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 3.

Proof. The structure of proof generally follows previous analysis. 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 $\bar{g}_{t,i} = \left[\frac{\phi(\|\theta_{t,i}^1\|)}{\|p_{t,i}^1\|}g_{t,i}^1, \ldots, \frac{\phi(\|\theta_{t,i}^L\|)}{\|p_{t,i}^L\|}g_{t,i}^L\right]$ representing the normalized gradient (concatenated by layers) of the i-th device. It holds that

$$\langle\nabla f(\bar{\vartheta}_t), \frac{\bar{g}_t}{\sqrt{\hat{v}_t}}\rangle = \frac{1}{2}\|\frac{\nabla f(\bar{\vartheta}_t)}{\hat{v}_t^{1/4}}\|^2 + \frac{1}{2}\|\frac{\bar{g}_t}{\hat{v}_t^{1/4}}\|^2 - \|\frac{\nabla f(\bar{\vartheta}_t) - \bar{g}_t}{\hat{v}_t^{1/4}}\|^2. \tag{23}$$

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,

$$
\frac{2}{n}\sum_{i=1}^n \|\frac{\nabla f(\bar{\theta}_t) - \bar{g}_{t,i}}{\hat{v}_t^{1/4}}\|^2 \leq \frac{4}{n}\Big(\underbrace{\sum_{i=1}^n \|\frac{\nabla f(\bar{\theta}_t) - \nabla f(\theta_{t,i})}{\hat{v}_t^{1/4}}\|^2}_{B_1} + \underbrace{\sum_{i=1}^n \|\frac{\nabla f(\theta_{t,i}) - \bar{g}_{t,i}}{\hat{v}_t^{1/4}}\|^2}_{B_2}\Big). \tag{24}
$$

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

$$
\begin{aligned}
B_1 &\leq \frac{L}{\sqrt{v_0}}\sum_{i=1}^n \|\bar{\theta}_t - \theta_{t,i}\|^2 \\
&= \frac{\alpha_r^2 L}{\sqrt{v_0}}\sum_{i=1}^n\sum_{l=1}^L \|\sum_{j=\lfloor t\rfloor_T+1}^t \Big(\frac{\phi(\|\theta_{j,i}^l\|)}{\|p_{j,i}^l\|}p_{j,i}^l - \frac{1}{n}\sum_{k=1}^n \frac{\phi(\|\theta_{j,k}^l\|)}{\|p_{j,k}^l\|}p_{j,k}^l\Big)\|^2 \\
&\leq n\frac{\alpha_t^2 L}{\sqrt{v_0}}M^2(T-1)^2\phi_M^2(1-\beta_2)p
\end{aligned} \tag{25}
$$

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 $\beta_1 = 0$:

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

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

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

Then we have:

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

Thus, (24) becomes:

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

Substituting all ingredients into (23), we obtain

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

At the same time, we have

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

Regarding (22), 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 (25). Therefore, we have obtained

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

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

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

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

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

concluding our proof.

□