Supplemental Material for "Faster Distributed Deep Net Training: Computation and Communication Decoupled Stochastic Gradient Descent"

Appendix: proofs

At first, we bound the partially accumulated local gradients.

Lemma 1 *Under Assumption 1, we have the following inequality*

$$\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma G_\tau^i\right\|^2 \leq 4\gamma^2(t-t')\left(\frac{N\sigma^2}{\sum_{l=1}^{N}M_l}+(t-t')\,\zeta^2+L^2\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|x_\tau^i-\hat{x}_\tau\right\|^2+\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|\nabla f(\hat{x}_\tau)\right\|^2\right). \tag{1}$$

Proof. By the definition of G_τ^i, we have

$$\begin{aligned}
&\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma G_\tau^i\right\|^2\\
=\;&\gamma^2\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\frac{1}{M_i}\sum_{j=1}^{M_i}\nabla f_i(x_\tau^i,\xi_\tau^{i,j})\right\|^2\\
\leq\;&4\gamma^2\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\left(\underbrace{\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\left(\frac{1}{M_i}\sum_{j=1}^{M_i}\nabla f_i\left(x_\tau^i,\xi_\tau^{i,j}\right)-\nabla f_i\left(x_\tau^i\right)\right)\right\|^2}_{T_1}+\underbrace{\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\left(\nabla f_i\left(x_\tau^i\right)-\nabla f_i\left(\hat{x}_\tau\right)\right)\right\|^2}_{T_2}\right.\\
&\left.+\underbrace{\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\left(\nabla f_i\left(\hat{x}_\tau\right)-\nabla f\left(\hat{x}_\tau\right)\right)\right\|^2}_{T_3}+\underbrace{\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\nabla f\left(\hat{x}_\tau\right)\right\|^2}_{T_4}\right),
\end{aligned} \tag{2}$$

where the inequality follows from Cauchy's inequality. We next bound T_1

$$\begin{aligned}
T_1 =\;&\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|\frac{1}{M_i}\sum_{j=1}^{M_i}\nabla f_i\left(x_\tau^i,\xi_\tau^{i,j}\right)-\nabla f_i\left(x_\tau^i\right)\right\|^2\\
&+2\sum_{t'\leq\tau_1<\tau_2\leq t-1}\mathbb{E}\left\langle\frac{1}{M_i}\sum_{j=1}^{M_i}\nabla f_i\left(x_{\tau_1}^i,\xi_{\tau_1}^{i,j}\right)-\nabla f_i\left(x_{\tau_1}^i\right),\frac{1}{M_i}\sum_{j=1}^{M_i}\nabla f_i\left(x_{\tau_2}^i,\xi_{\tau_2}^{i,j}\right)-\nabla f_i\left(x_{\tau_2}^i\right)\right\rangle\\
=\;&\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|\frac{1}{M_i}\sum_{j=1}^{M_i}\nabla f_i\left(x_\tau^i,\xi_\tau^{i,j}\right)-\nabla f_i\left(x_\tau^i\right)\right\|^2\\
=\;&\sum_{\tau=t'}^{t-1}\frac{1}{M_i^2}\left(\sum_{j=1}^{M_i}\mathbb{E}\left\|\nabla f_i(x_\tau^i,\xi_\tau^{i,j})-\nabla f_i(x_\tau^i)\right\|^2\right.\\
&\left.+2\sum_{1\leq j_1<j_2\leq M_i}\mathbb{E}\left\langle\nabla f_i(x_\tau^i,\xi_\tau^{i,j_1})-\nabla f_i(x_\tau^i),\nabla f_i(x_\tau^i,\xi_\tau^{i,j_2})-\nabla f_i(x_\tau^i)\right\rangle\right)\\
=\;&\sum_{\tau=t'}^{t-1}\frac{1}{M_i^2}\sum_{j=1}^{M_i}\mathbb{E}\left\|\nabla f_i(x_\tau^i,\xi_\tau^{i,j})-\nabla f_i(x_\tau^i)\right\|^2\\
\leq\;&\frac{(t-t')\,\sigma^2}{M_i},
\end{aligned} \tag{3}$$

where the second and the fourth equalities hold because $\mathbb{E}_{\xi_\tau^{i,j}\in\mathcal{D}_i}\nabla f_i(x_\tau^i,\xi_\tau^{i,j})=\nabla f_i(x_\tau^i)$ and $\xi_\tau^{i,j}$'s are independent, and the inequality follows from Assumption 1 (3). According to Cauchy's inequality, we can bound T_2, T_3 and T_4 as

$$T_2 \leq (t-t')\sum_{\tau=t'}^{t-1}\left\|\nabla f_i\left(x_\tau^i\right)-\nabla f_i\left(\hat{x}_\tau\right)\right\|^2 \leq (t-t')\,L^2\sum_{\tau=t'}^{t-1}\left\|x_\tau^i-\hat{x}_\tau\right\|^2, \tag{4}$$

$$\sum_{i=1}^N\frac{M_i}{\sum_{l=1}^N M_l}T_3 \leq (t-t')\sum_{\tau=t'}^{t-1}\mathbb{E}\sum_{i=1}^N\frac{M_i}{\sum_{l=1}^N M_l}\left\|\nabla f_i\left(\hat{x}_\tau\right)-\nabla f\left(\hat{x}_\tau\right)\right\|^2 \leq (t-t')^2\zeta^2, \tag{5}$$

$$T_4 \leq (t-t')\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|\nabla f(\hat{x}_\tau)\right\|^2, \tag{6}$$

where the second inequality in (4) and the second inequality in (5) follow Assumption 1 (1) and (3), respectively. Substituting (3), (4), (5) and (6) into (2), we obtain

$$\begin{aligned}
&\sum_{i=1}^N\frac{M_i}{\sum_{l=1}^N M_l}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma G_\tau^i\right\|^2\\
\leq\ & 4\gamma^2\left(\frac{N(t-t')\sigma^2}{\sum_{l=1}^N M_l}+(t-t')L^2\sum_{i=1}^N\frac{M_i}{\sum_{l=1}^N M_l}\sum_{\tau=t'}^{t-1}\left\|x_\tau^i-\hat{x}_\tau\right\|^2+(t-t')^2\zeta^2+(t-t')\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|\nabla f(\hat{x}_\tau)\right\|^2\right)\\
=\ & 4\gamma^2(t-t')\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l}+(t-t')\zeta^2+L^2\sum_{i=1}^N\frac{M_i}{\sum_{l=1}^N M_l}\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|x_\tau^i-\hat{x}_\tau\right\|^2+\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|\nabla f(\hat{x}_\tau)\right\|^2\right),
\end{aligned} \tag{7}$$

which completes the proof.

Next, we bound the difference between the local models and the global average model.

Lemma 2 *Under Assumption 1, the difference of $\hat{x}_t$ and x_t^i's can be bounded as*

$$\sum_{i=1}^N\frac{M_i}{\sum_{l=1}^N M_l}\sum_{t=0}^{T-1}\mathbb{E}\left\|\hat{x}_t-x_t^i\right\|^2 \leq \frac{8\gamma^2k}{1-16\gamma^2k^2L^2}\left(\frac{TN\sigma^2}{\sum_{l=1}^N M_l}+2kT\zeta^2+2k\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{x}_t)\right\|^2\right). \tag{8}$$

Proof. According to the updating scheme in Algorithms 1, x_t^i can be represented as

$$x_t^i=\hat{x}_{(\lfloor\frac{t}{k}\rfloor-1)k}-\sum_{\tau=(\lfloor\frac{t}{k}\rfloor-1)k}^{t-1}\gamma G_\tau^i, \tag{9}$$

since the result of the last complete communication is the average of the models at step $(\lfloor\frac{t}{k}\rfloor-1)k$. On the other hand, by the definition of $\hat{x}_t$, we can represent it as

$$\hat{x}_t=\hat{x}_{(\lfloor\frac{t}{k}\rfloor-1)k}-\sum_{\tau=(\lfloor\frac{t}{k}\rfloor-1)k}^{t-1}\gamma\sum_{j=1}^N\frac{M_j}{\sum_{l=1}^N M_l}G_\tau^j. \tag{10}$$

Substituting (9) and (10) into the left hand side of (8) , we have

$$
\begin{aligned}
&\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\mathbb{E}\left\|\hat{x}_t-x_t^i\right\|^2\\
=\;&\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\mathbb{E}\left\|\left(\hat{x}_{t'}-\sum_{\tau=t'}^{t-1}\gamma\sum_{j=1}^{N}\frac{M_j}{\sum_{l=1}^{N}M_l}G_\tau^j\right)-\left(\hat{x}_{t'}-\sum_{\tau=t'}^{t-1}\gamma G_\tau^i\right)\right\|^2\\
=\;&\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma G_\tau^i-\sum_{\tau=t'}^{t-1}\gamma\sum_{j=1}^{N}\frac{M_j}{\sum_{l=1}^{N}M_l}G_\tau^j\right\|^2\\
=\;&\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma G_\tau^i\right\|^2+\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma\sum_{j=1}^{N}\frac{M_j}{\sum_{l=1}^{N}M_l}G_\tau^j\right\|^2\\
&-2\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\mathbb{E}\left\langle\sum_{\tau=t'}^{t-1}\gamma G_\tau^i,\sum_{\tau=t'}^{t-1}\gamma\sum_{j=1}^{N}\frac{M_j}{\sum_{l=1}^{N}M_l}G_\tau^j\right\rangle\\
=\;&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\|\sum_{\tau=t'}^{t-1}\gamma G_\tau^i\|^2+\mathbb{E}\|\sum_{\tau=t'}^{t-1}\gamma\sum_{j=1}^{N}\frac{M_j}{\sum_{l=1}^{N}M_l}G_\tau^j\|^2-2\mathbb{E}\|\sum_{\tau=t'}^{t-1}\gamma\sum_{j=1}^{N}\frac{M_j}{\sum_{l=1}^{N}M_l}G_\tau^j\|^2\\
=\;&\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma G_\tau^i\right\|^2-\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma\sum_{j=1}^{N}\frac{M_j}{\sum_{l=1}^{N}M_l}G_\tau^j\right\|^2\\
\le\;&\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma G_\tau^i\right\|^2\\
\le\;&4\gamma^2(t-t')\left(\frac{N\sigma^2}{\sum_{l=1}^{N}M_l}+(t-t')\,\zeta^2+L^2\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|x_\tau^i-\hat{x}_\tau\right\|^2+\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|\nabla f(\hat{x}_\tau)\right\|^2\right),
\end{aligned}
\tag{11}
$$

where the last inequality follows from Lemma 1. Since $t'=(\lfloor\frac{t}{k}\rfloor-1)k$, we have $t'\ge t-2k$ and can further obtain

$$
\begin{aligned}
&\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\mathbb{E}\|\hat{x}_t-x_t^i\|^2\\
\le\;&8\gamma^2k\left(\frac{N\sigma^2}{\sum_{l=1}^{N}M_l}+2k\zeta^2+L^2\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\sum_{\tau=t-2k}^{t-1}\mathbb{E}\|x_\tau^i-\hat{x}_\tau\|^2+\sum_{\tau=t-2k}^{t-1}\mathbb{E}\|\nabla f(\hat{x}_\tau)\|^2\right).
\end{aligned}
\tag{12}
$$

Summing up this inequality from $t=0$ to $T-1$, we have

$$
\begin{aligned}
&\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\sum_{t=0}^{T-1}\mathbb{E}\|\hat{x}_t-x_t^i\|^2\\
\le\;&8\gamma^2k\left(\frac{TN\sigma^2}{\sum_{l=1}^{N}M_l}+2kT\zeta^2+L^2\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\sum_{t=0}^{T-1}\sum_{\tau=t-2k}^{t-1}\mathbb{E}\left\|x_\tau^i-\hat{x}_\tau\right\|^2+\sum_{t=0}^{T-1}\sum_{\tau=t-2k}^{t-1}\mathbb{E}\left\|\nabla f\left(\hat{x}_\tau\right)\right\|^2\right)\\
\le\;&8\gamma^2k\left(\frac{TN\sigma^2}{\sum_{l=1}^{N}M_l}+2kT\zeta^2+2kL^2\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\sum_{t=0}^{T-1}\mathbb{E}\left\|x_t^i-\hat{x}_t\right\|^2+2k\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{x}_t)\right\|^2\right),
\end{aligned}
\tag{13}
$$

where the last inequality can be obtained by using a simple counting argument $\sum_{t=0}^{T-1}\sum_{\tau=t-2k}^{t-1}A_\tau\le 2k\sum_{t=0}^{T-1}A_t$. Rearranging the inequality, we obtain

$$
\left(1-16\gamma^2k^2L^2\right)\sum_{i=1}^{N}\frac{M_i}{\sum_{l=1}^{N}M_l}\sum_{t=0}^{T-1}\mathbb{E}\left\|\hat{x}_t-x_t^i\right\|^2\le 8\gamma^2k\left(\frac{TN\sigma^2}{\sum_{l=1}^{N}M_l}+2kT\zeta^2+2k\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{x}_t)\right\|^2\right).
\tag{14}
$$

Dividing $\left(1-16\gamma^2k^2L^2\right)$ on both sides yields the result.

Theorem 1 *Under Assumption 1, if the learning rate satisfies $\gamma \leq \frac{1}{L}$, we have the following convergence result for Algorithm 1:*

$$\frac{1}{T}\sum_{t=0}^{T-1} D_1 \mathbb{E}\|\nabla f(\hat{x}_t)\|^2 \leq \frac{2(f(\hat{x}_0)-f^*)}{T\gamma} + D_2\left(\frac{N\sigma^2}{\sum_{i=1}^N M_i} + 2k\zeta^2\right) + \frac{\gamma L \sigma^2}{\sum_{i=1}^N M_i}, \tag{15}$$

where

$$D_1 = 1 - 2kD_2, \quad D_2 = \frac{8\gamma^2 L^2 k}{1 - 16\gamma^2 k^2 L^2}. \tag{16}$$

Proof. Since $f_i(\cdot), i = 1, 2, \cdots, N$ are L-smooth, it is easy to verify that $f(\cdot)$ is L-smooth. We have

$$\begin{aligned} f(\hat{x}_{t+1}) &\leq f(\hat{x}_t) + \langle \nabla f(\hat{x}_t), \hat{x}_{t+1} - \hat{x}_t \rangle + \frac{L}{2}\|\hat{x}_{t+1} - \hat{x}_t\|^2 \\ &= f(\hat{x}_t) - \gamma \left\langle \nabla f(\hat{x}_t), \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} G_i^t \right\rangle + \frac{L\gamma^2}{2}\left\| \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} G_t^i \right\|^2. \end{aligned} \tag{17}$$

By applying expectation with respect to all the random variables at step t and conditional on the past (denote by $\mathbb{E}_{t|\cdot}$), we have

$$\begin{aligned} & \mathbb{E}_{t|\cdot} f(\hat{x}_{t+1}) \\ \leq\ & f(\hat{x}_t) - \gamma \left\langle \nabla f(\hat{x}_t), \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} \nabla f_i(x_t^i) \right\rangle + \frac{L\gamma^2}{2}\mathbb{E}_{t|\cdot}\left\| \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} G_t^i \right\|^2 \\ =\ & f(\hat{x}_t) - \frac{\gamma}{2}\left(\|\nabla f(\hat{x}_t)\|^2 + \left\| \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} \nabla f_i(x_t^i) \right\|^2 - \left\| \nabla f(\hat{x}_t) - \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} \nabla f_i(x_t^i) \right\|^2 \right) \\ & + \frac{L\gamma^2}{2}\mathbb{E}_{t|\cdot}\left\| \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} G_t^i \right\|^2. \end{aligned} \tag{18}$$

Note that

$$\begin{aligned} & \mathbb{E}_{t|\cdot}\left\| \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} G_t^i \right\|^2 \\ =\ & \mathbb{E}_{t|\cdot}\left\| \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} G_t^i - \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} \nabla f_i(x_t^i) + \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} \nabla f_i(x_t^i) \right\|^2 \\ =\ & \mathbb{E}_{t|\cdot}\left\| \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} G_t^i - \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} \nabla f_i(x_t^i) \right\|^2 + \left\| \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} \nabla f_i(x_t^i) \right\|^2 \\ & + 2\mathbb{E}_{t|\cdot}\left\langle \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} G_t^i - \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} \nabla f_i(x_t^i), \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} \nabla f_i(x_t^i) \right\rangle \\ =\ & \mathbb{E}_{t|\cdot}\left\| \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} G_t^i - \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} \nabla f_i(x_t^i) \right\|^2 + \left\| \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} \nabla f_i(x_t^i) \right\|^2, \end{aligned} \tag{19}$$

where the last equality holds because $\mathbb{E}_{t|\cdot}\left(\frac{1}{N}\sum_{i=1}^N G_t^i - \frac{1}{N}\sum_{i=1}^N \nabla f_i(x_i^t)\right) = 0$, and

$$
\begin{aligned}
& \mathbb{E}_{t|\cdot}\left\|\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l} G_t^i - \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\nabla f_i(x_t^i)\right\|^2 \\
= \;& \mathbb{E}_{t|\cdot}\sum_{i=1}^N \frac{M_i^2}{(\sum_{l=1}^N M_l)^2}\left\|G_t^i - \nabla f_i(x_t^i)\right\|^2 \\
& +2\sum_{1\le i_1<i_2\le N}\mathbb{E}_{t|\cdot}\left\langle \frac{M_{i_1}}{\sum_{l=1}^N M_l}G_t^{i_1} - \frac{M_{i_1}}{\sum_{l=1}^N M_l}\nabla f_{i_1}(x_t^{i_1}), \frac{M_{i_2}}{\sum_{l=1}^N M_l}G_t^{i_2} - \frac{M_{i_2}}{\sum_{l=1}^N M_l}\nabla f_{i_2}(x_t^{i_2})\right\rangle \\
= \;& \mathbb{E}_{t|\cdot}\sum_{i=1}^N \frac{M_i^2}{(\sum_{l=1}^N M_l)^2}\left\|G_t^i - \nabla f_i(x_t^i)\right\|^2 \\
= \;& \mathbb{E}_{t|\cdot}\sum_{i=1}^N \frac{M_i^2}{(\sum_{l=1}^N M_l)^2}\left\|\frac{1}{M_i}\sum_{j=1}^{M_i}\nabla f_i(x_t^i,\xi_t^{i,j}) - \nabla f_i(x_t^i)\right\|^2 \\
= \;& \mathbb{E}_{t|\cdot}\sum_{i=1}^N \frac{1}{(\sum_{l=1}^N M_l)^2}\left(\sum_{j=1}^{M_i}\left\|\nabla f_i(x_t^i,\xi_t^{i,j}) - \nabla f_i(x_t^i)\right\|^2\right. \\
& \left. +2\sum_{1\le j_1<j_2\le M_i}\left\langle \nabla f_i(x_t^i,\xi_t^{i,j_1}) - \nabla f_i(x_t^i), \nabla f_i(x_t^i,\xi_t^{i,j_2}) - \nabla f_i(x_t^i)\right\rangle\right) \\
= \;& \sum_{i=1}^N \frac{1}{(\sum_{l=1}^N M_l)^2}\sum_{j=1}^{M_i}\mathbb{E}_{t|\cdot}\|\nabla f_i(x_t^i,\xi_t^{i,j}) - \nabla f_i(x_t^i)\|^2 \\
\le \;& \sum_{i=1}^N \frac{1}{(\sum_{l=1}^N M_l)^2} M_i\sigma^2 = \frac{\sigma^2}{\sum_{l=1}^N M_l},
\end{aligned} \tag{20}
$$

where the second equality and the fifth equality hold because the random variables on different workers and the random variables in one mini-batch are independent, and the last inequality follows from Assumption 1 (3). We have

$$
\mathbb{E}_{t|\cdot}\left\|\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}G_i^t\right\|^2 \le \frac{\sigma^2}{\sum_{l=1}^N M_l} + \mathbb{E}_{t|\cdot}\left\|\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\nabla f_i(x_i^t)\right\|^2. \tag{21}
$$

Substituting (21) into (18) and applying expectation with respect to all the random variables, we obtain

$$
\begin{aligned}
\mathbb{E}f(\hat{x}_{t+1}) \;\le\; & \mathbb{E}f(\hat{x}_t) - \frac{\gamma}{2}\mathbb{E}\|\nabla f(\hat{x}_t)\|^2 - \frac{\gamma}{2}(1-L\gamma)\mathbb{E}\left\|\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\nabla f_i(x_t^i)\right\|^2 \\
& +\frac{\gamma}{2}\mathbb{E}\left\|\nabla f(\hat{x}_t) - \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\nabla f_i(x_t^i)\right\|^2 + \frac{\gamma^2 L\sigma^2}{2\sum_{l=1}^N M_l}.
\end{aligned} \tag{22}
$$

We then bound the difference of $\nabla f(\hat{x}^t)$ and $\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\nabla f_i(x_i^t)$ as

$$\begin{aligned}
\mathbb{E}\left\|\nabla f(\hat{x}^t) - \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\nabla f_i(x_i^t)\right\|^2 &= \mathbb{E}\left\|\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\left(\nabla f_i(\hat{x}^t) - \nabla f_i(x_i^t)\right)\right\|^2 \\
&\leq \mathbb{E}\left(\sum_{i=1}^N \left(\frac{\sqrt{M_i}}{\sum_{l=1}^N M_l}\right)^2\right)\left(\sum_{i=1}^N \left\|\sqrt{M_i}\left(\nabla f_i(\hat{x}^t) - \nabla f_i(x_i^t)\right)\right\|^2\right) \\
&= \sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\mathbb{E}\left\|\nabla f_i(\hat{x}^t) - \nabla f_i(x_i^t)\right\|^2 \\
&\leq L^2\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\mathbb{E}\left\|\hat{x}^t - x_i^t\right\|^2, \qquad (23)
\end{aligned}$$

where the two inequalities follow from Cauchy's inequality and L-smooth assumption, respectively. Substituting (23) into (22) yields

$$\begin{aligned}
\mathbb{E}f(\hat{x}_{t+1}) \leq \mathbb{E}f(\hat{x}_t) - \frac{\gamma}{2}\mathbb{E}\|\nabla f(\hat{x}_t)\|^2 - \frac{\gamma}{2}(1-L\gamma)\mathbb{E}\left\|\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\nabla f_i(x_t^i)\right\|^2 \\
+\frac{\gamma L^2}{2}\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\mathbb{E}\|\tilde{x}^t - x_i^t\|^2 + \frac{\gamma^2 L\sigma^2}{2\sum_{l=1}^N M_l}. \qquad (24)
\end{aligned}$$

Rearranging the inequality and summing up both sides from $t=0$ to $T-1$, we have

$$\begin{aligned}
&\sum_{t=0}^{T-1}\left(\frac{\gamma}{2}\mathbb{E}\|\nabla f(\hat{x}_t)\|^2 + \frac{\gamma}{2}(1-L\gamma)\mathbb{E}\left\|\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\nabla f_i(x_t^i)\right\|^2\right) \\
\leq\ & f(\hat{x}_0) - f^* + \frac{\gamma L^2}{2}\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\sum_{t=0}^{T-1}\mathbb{E}\|\tilde{x}^t - x_i^t\|^2 + \frac{T\gamma^2 L\sigma^2}{2\sum_{l=1}^N M_l}. \qquad (25)
\end{aligned}$$

Substituting Lemma 2 into (25), we obtain

$$\begin{aligned}
&\sum_{t=0}^{T-1}\left(\frac{\gamma}{2}\mathbb{E}\|\nabla f(\hat{x}_t)\|^2 + \frac{\gamma}{2}(1-L\gamma)\mathbb{E}\left\|\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\nabla f_i(x_t^i)\right\|^2\right) \\
\leq\ & f(\hat{x}_0) - f^* + \frac{4\gamma^3 L^2 k}{1-16\gamma^2k^2L^2}\left(T\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right) + 2k\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\hat{x}_t)\|^2\right) + \frac{T\gamma^2 L\sigma^2}{2\sum_{l=1}^N M_l}. \qquad (26)
\end{aligned}$$

Rearranging this inequality and dividing both sides by $\frac{T\gamma}{2}$, we get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(\left(1 - \frac{16\gamma^2L^2k^2}{1-16\gamma^2k^2L^2}\right)\mathbb{E}\|\nabla f(\hat{x}_t)\|^2 + (1-L\gamma)\mathbb{E}\left\|\sum_{i=1}^N \frac{M_i}{\sum_{l=1}^N M_l}\nabla f_i(x_t^i)\right\|^2\right) \\
\leq\ & \frac{2(f(\hat{x}_0) - f(\hat{x}_T))}{T\gamma} + \frac{8\gamma^2L^2k}{1-16\gamma^2k^2L^2}\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right) + \frac{\gamma L\sigma^2}{\sum_{l=1}^N M_l} \\
\leq\ & \frac{2(f(\hat{x}_0) - f^*)}{T\gamma} + \frac{8\gamma^2L^2k}{1-16\gamma^2k^2L^2}\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right) + \frac{\gamma L\sigma^2}{\sum_{l=1}^N M_l}. \qquad (27)
\end{aligned}$$

If the learning rate satisfies $\gamma \leq \frac{1}{L}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(1 - \frac{16\gamma^2L^2k^2}{1-16\gamma^2k^2L^2}\right)\mathbb{E}\|\nabla f(\hat{x}_t)\|^2 \leq \frac{2(f(\hat{x}_0) - f^*)}{T\gamma} + \frac{8\gamma^2L^2k}{1-16\gamma^2k^2L^2}\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right) + \frac{\gamma L\sigma^2}{\sum_{l=1}^N M_l}, \qquad (28)$$

which completes the proof.

Corollary 1 *Under Assumption 1, when the learning rate is set as* $\gamma = \frac{1}{\sigma\sqrt{\frac{T}{\sum_{i=1}^N M_i}}}$ *and the total number of iterations satisfies*

$$T \geq \max\left\{\frac{L^2(\sum_{i=1}^N M_i)}{\sigma^2}, \frac{48(\sum_{i=1}^N M_i)L^2k^2}{\sigma^2}, \frac{144(\sum_{i=1}^N M_i)^3}{\sigma^6}L^2k^2\left(\frac{N\sigma^2}{\sum_{i=1}^N M_i} + 2k\zeta^2\right)^2\right\}, \tag{29}$$

we have the following convergence result for Algorithm 1:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\hat{x}_t)\|^2 \leq \frac{4\sigma(f(\hat{x}_0) - f^* + L)}{\sqrt{T\sum_{i=1}^N M_i}}. \tag{30}$$

Proof. Since $\gamma = \frac{1}{\sigma\sqrt{\frac{T}{\sum_{l=1}^N M_l}}}$ and $T \geq \frac{L^2\sum_{l=1}^N M_l}{\sigma^2}$, we immediately have $\gamma \leq \frac{1}{L}$, then we have the result in (15) and get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\hat{x}_t)\|^2 \leq \frac{2(f(\hat{x}_0) - f^*)}{T\gamma D_1} + \frac{D_2}{D_1}\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right) + \frac{\gamma L\sigma^2}{(\sum_{l=1}^N M_l)D_1}. \tag{31}$$

By setting $\gamma = \frac{1}{\sigma\sqrt{\frac{T}{\sum_{l=1}^N M_l}}}$ and $T \geq \frac{48(\sum_{l=1}^N M_l)L^2k^2}{\sigma^2}$, we have

$$16\gamma^2L^2k^2 = \frac{16\sum_{l=1}^N M_l}{\sigma^2 T}L^2k^2 \leq \frac{1}{3}. \tag{32}$$

Now we can bound D_1 as

$$D_1 = 1 - 2kD_2 = 1 - \frac{16\gamma^2L^2k^2}{1 - 16\gamma^2L^2k^2} \geq \frac{1}{2}. \tag{33}$$

Combining (32) with $T \geq \frac{144(\sum_{l=1}^N M_l)^3}{\sigma^6}L^2k^2\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right)^2$, $D_2\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right)$ can be bounded as

$$\begin{aligned}
D_2\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right) &= \frac{8\gamma^2L^2k}{1 - 16\gamma^2L^2k^2}\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right) \\
&\leq 12\gamma^2L^2k\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right) \\
&= \frac{12\sum_{l=1}^N M_l}{\sigma^2 T}L^2k\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right) \\
&\leq \frac{12\sum_{l=1}^N M_l}{\sigma^2\sqrt{T}}L^2k\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right)\cdot\frac{1}{\frac{12(\sum_{l=1}^N M_l)^{\frac{3}{2}}}{\sigma^3}Lk\left(\frac{N\sigma^2}{\sum_{l=1}^N M_l} + 2k\zeta^2\right)} \\
&= \frac{\sigma L}{\sqrt{T\sum_{l=1}^N M_l}}.
\end{aligned} \tag{34}$$

Substituting $\gamma = \frac{1}{\sigma\sqrt{\frac{T}{\sum_{l=1}^N M_l}}}$, (33) and (34) into (31), we can get the final result:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\hat{x}_t)\|^2 \leq \frac{4\sigma\left(f(\hat{x}_0) - f^* + L\right)}{\sqrt{T\sum_{l=1}^N M_l}}, \tag{35}$$

which completes the proof.