A Wait-free (continuous) training and communication

The theoretical guarantee of AD-PSGD relies on the the doubly stochastic property of matrix W. The implication is the averaging of the weights between two workers should be atomic. This brings a special challenge for current distributed deep learning frameworks where the computation (gradients calculation and weights update) runs on GPU devices and the communication runs on CPU (or its peripherals such as infiniband or RDMA), because when there is averaging happening on a worker, the GPU is not allowed to update gradients into the weights. This can be solve by using CPU to update weights while GPUs only calculate gradients. Every worker (including active and passive workers) runs two threads in parallel with a shared buffer g, one thread for computation and the other for communication. Algorithm 2, Algorithm 3, and Algorithm 4 illustrate the task on each thread. The communication thread is run by CPUs, while the computation thread is run by GPUs. In this way GPUs can continuously calculate new gradients by putting the results in CPUs' buffer regardless of whether there is averaging happening. Recall in D-PSGD, communication only occurs once in each iteration. In contrast, AD-PSGD can exchange weights at any time by using this implementation.

Algorithm 2 Computation thread on active or passive worker (worker index is i)

Require: Batch size M

while not terminated **do**
  Pull model x^i from the communication thread.
  Update locally in the thread $x^i \leftarrow x^i - \gamma g$.[a]
  Randomly sample a batch $\xi^i := (\xi^i_1, \xi^i_2, \ldots, \xi^i_M)$ from local data of the i-th worker and compute the stochastic gradient $g^i(x^i; \xi^i) := \sum_{m=1}^{M} \nabla F(x^i; \xi^i_m)$ locally.
  wait until $g = 0$ **then**
    Local buffer $g \leftarrow g^i(x^i; \xi^i)$.[b]
  end wait until
end while

[a] At this time the communication thread may have not update g into x^i so the computation thread pulls an old model. We compensate this by doing local update in computation thread. We observe this helps the scaling.

[b] We can also make a queue of gradients here to avoid the waiting. Note that doing this will make the effective batch-size different from M.

Algorithm 3 Communication thread on active worker (worker index is i)

Require: Initialize local model x^i, learning rate γ.

while not terminated **do**
  if $g \neq 0$ **then**
    $x^i \leftarrow x^i - \gamma g, \quad g \leftarrow 0.$
  end if
  Randomly select a neighbor (namely worker j). Send x^i to worker j and fetch x^j from it.
  $x^i \leftarrow \frac{1}{2}(x^i + x^j)$.
end while

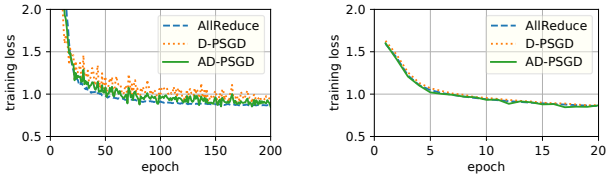
Algorithm 4 Communication thread on passive worker (worker index is j)

Require: Initialize local model x^j, learning rate γ.

while not terminated **do**
  if $g \neq 0$ **then**
    $x^j \leftarrow x^j - \gamma g, \quad g \leftarrow 0.$
  end if
  if receive the request of reading local model (say from worker i) **then**
    Send x^j to worker i.
    $x^j \leftarrow \frac{1}{2}(x^i + x^j)$.
  end if
end while

B NLC experiments

In this section, we use IBM proprietary natural language processing dataset and model to evaluate AD-PSGD against other algorithms.



(a) NLC-Joule (b) NLC-Yelp

Figure 7. Training loss comparison for IBM NLC model on Joule and Yelp datasets. AllReduce-SGD, D-PSGD and AD-PSGD converge alike w.r.t. epochs.

The IBM NLC task is to classify input sentences into a target category in a predefined label set. The NLC model is a CNN model that has a word-embedding lookup table layer, a convolutional layer and a fully connected layer with a softmax output layer. We use two datasets in our evaluation. The first dataset Joule is an in-house customer dataset that has 2.5K training samples, 1K test samples, and 311 different classes. The second dataset Yelp, which is a public dataset, has 500K training samples, 2K test samples and 5 different classes. Figure 7 shows that AD-PSGD converges (w.r.t epochs) similarly to AllReduce-SGD and D-PSGD on NLC tasks.

Above results show AD-PSGD converges similarly (w.r.t) to AllReduce-SGD and D-PSGD for IBM NLC workload, which is an example of proprietary workloads.

C Appendix: proofs

In the following analysis we define

$$M_k := \sum_{i=1}^{n} p_i \left\| \frac{X_k \mathbf{1}_n}{n} - X_k e_i \right\|^2, \tag{9}$$

and

$$\hat{M}_k := M_{k-\tau_k}. \tag{10}$$

We also define

$$\begin{aligned}
\partial f(X_k) :=& n \left[\begin{array}{cccc} p_1 \nabla f_1(x_k^1) & p_2 \nabla f_2(x_k^2) & \cdots & p_n \nabla f_n(x_k^n) \end{array} \right] \in \mathbb{R}^{N \times n}, \\
\partial f(X_k, i) :=& \left[\begin{array}{ccccc} 0 & \cdots & \nabla f_i(x_k^i) & \cdots & 0 \end{array} \right] \in \mathbb{R}^{N \times n}, \\
\partial g(\hat{X}_k, \xi_k) :=& n \left[\begin{array}{cccc} p_1 \sum_{j=1}^{M} \nabla F(\hat{x}_k^1, \xi_{k,j}^1) & \cdots & p_n \sum_{j=1}^{M} \nabla F(\hat{x}_k^n, \xi_{k,j}^n) \end{array} \right] \in \mathbb{R}^{N \times n}.
\end{aligned}$$

$$\begin{aligned}
\bar{\rho} :=& \frac{n-1}{n} \left(\frac{1}{1-\rho} + \frac{2\sqrt{\rho}}{\left(1-\sqrt{\rho}\right)^2} \right), \\
C_1 :=& 1 - 24M^2L^2\gamma^2 \left(T\frac{n-1}{n} + \bar{\rho} \right), \\
C_2 :=& \frac{\gamma M}{2n} - \frac{\gamma^2 LM^2}{n^2} - \frac{2M^3L^2T^2\gamma^3}{n^3} - \left(\frac{6\gamma^2L^3M^2}{n^2} + \frac{\gamma M}{n}L^2 + \frac{12M^3L^4T^2\gamma^3}{n^3} \right) \frac{4M^2\gamma^2(T\frac{n-1}{n} + \bar{\rho})}{C_1}, \\
C_3 :=& \frac{1}{2} + \frac{2\left(6\gamma^2L^2M^2 + \gamma nML + \frac{12M^3L^3T^2\gamma^3}{n}\right)\bar{\rho}}{C_1} + \frac{LT^2\gamma M}{n}.
\end{aligned}$$

Proof to Theorem 1. We start from

$$\begin{aligned}
& \mathbb{E} f\left(\frac{X_{k+1}\mathbf{1}_n}{n} \right) \\
=& \mathbb{E} f\left(\frac{X_k W_k \mathbf{1}_n}{n} - \gamma \frac{\partial g(\hat{X}_k; \xi_k^{i_k}, i_k)\mathbf{1}_n}{n} \right) = \mathbb{E} f\left(\frac{X_k \mathbf{1}_n}{n} - \gamma \frac{\partial g(\hat{X}_k; \xi_k^{i_k}, i_k)\mathbf{1}_n}{n} \right) \\
\leqslant& \mathbb{E} f\left(\frac{X_k \mathbf{1}_n}{n} \right) - \gamma \mathbb{E} \left\langle \nabla f\left(\frac{X_k \mathbf{1}_n}{n} \right), \frac{\partial g(\hat{X}_k; \xi_k^{i_k}, i_k)\mathbf{1}_n}{n} \right\rangle + \frac{\gamma^2 L}{2} \mathbb{E} \left\| \frac{\partial g(\hat{X}_k; \xi_k^{i_k}, i_k)\mathbf{1}_n}{n} \right\|^2 \\
\overset{(3),\ \text{Lemma 4}}{\leqslant}& \mathbb{E} f\left(\frac{X_k \mathbf{1}_n}{n} \right) - \frac{\gamma M}{n} \mathbb{E} \left\langle \nabla f\left(\frac{X_k \mathbf{1}_n}{n} \right), \frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n} \right\rangle + \frac{\gamma^2 L\sigma^2 M}{2n^2} + \frac{\gamma^2 LM^2}{2n^2} \sum_{i=1}^{n} p_i \mathbb{E} \|\nabla f_i(\hat{x}_k^i)\|^2 \\
=& \mathbb{E} f\left(\frac{X_k \mathbf{1}_n}{n} \right) + \frac{\gamma M}{2n} \left\| \nabla f\left(\frac{X_k \mathbf{1}_n}{n} \right) - \frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n} \right\|^2 - \frac{\gamma M}{2n} \mathbb{E} \left\| \nabla f\left(\frac{X_k \mathbf{1}_n}{n} \right) \right\|^2 - \frac{\gamma M}{2n} \mathbb{E} \left\| \frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n} \right\|^2 \\
& + \frac{\gamma^2 LM^2}{2n^2} \sum_{i=1}^{n} p_i \mathbb{E} \|\nabla f_i(\hat{x}_k^i)\|^2 + \frac{\gamma^2 L\sigma^2 M}{2n^2}.
\end{aligned}$$

Using the upper bound of $\sum_{i=1}^{n} p_i \mathbb{E} \|\nabla f_i(\hat{x}_k^i)\|^2$ in Lemma 5:

$$\begin{aligned}
& \mathbb{E} f\left(\frac{X_{k+1}\mathbf{1}_n}{n} \right) \\
\leqslant& \mathbb{E} f\left(\frac{X_k \mathbf{1}_n}{n} \right) + \frac{\gamma M}{2n} \mathbb{E} \left\| \nabla f\left(\frac{X_k \mathbf{1}_n}{n} \right) - \frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n} \right\|^2 - \frac{\gamma M}{2n} \mathbb{E} \left\| \nabla f\left(\frac{X_k \mathbf{1}_n}{n} \right) \right\|^2 - \frac{\gamma M}{2n} \mathbb{E} \left\| \frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n} \right\|^2 \\
& + \frac{\gamma^2 LM^2}{2n^2} \left(12L^2 \hat{M}_k + 6\varsigma^2 + 2 \sum_{i=1}^{n} p_i \mathbb{E} \left\| \sum_{j=1}^{n} p_j \nabla f_j(\hat{x}_k^j) \right\|^2 \right) + \frac{\gamma^2 L\sigma^2 M}{2n^2}
\end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}f\left(\frac{X_k\mathbf{1}_n}{n}\right)+\frac{\gamma M}{2n}\underbrace{\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)-\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2}_{T_1}-\frac{\gamma M}{2n}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2\\
&-\left(\frac{\gamma M}{2n}-\frac{\gamma^2LM^2}{n^2}\right)\mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2+\frac{\gamma^2L(\sigma^2M+6\varsigma^2M^2)}{2n^2}+\frac{6\gamma^2L^3M^2}{n^2}\hat{M}_k. \qquad (11)
\end{aligned}
$$

For T_1 we have

$$
\begin{aligned}
T_1=&\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)-\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2\\
\leqslant&2\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)-\nabla f\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)\right\|^2+2\mathbb{E}\left\|\nabla f\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)-\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2\\
=&2\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)-\nabla f\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)\right\|^2+2\mathbb{E}\left\|\sum_i p_i\left(\nabla f_i\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)-\nabla f_i(\hat{x}_k^i)\right)\right\|^2\\
\leqslant&2\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)-\nabla f\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)\right\|^2+2\mathbb{E}\sum_i p_i\left\|\nabla f_i\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)-\nabla f_i(\hat{x}_k^i)\right\|^2\\
\overset{Assumption\ 1:1}{\leqslant}&2L^2\mathbb{E}\left\|\frac{(X_k-\hat{X}_k)\mathbf{1}_n}{n}\right\|^2+2L^2\mathbb{E}\hat{M}_k. \qquad (12)
\end{aligned}
$$

From (11) and (12) we obtain

$$
\begin{aligned}
\mathbb{E}f\left(\frac{X_{k+1}\mathbf{1}_n}{n}\right)\leqslant&\mathbb{E}f\left(\frac{X_k\mathbf{1}_n}{n}\right)+\frac{\gamma M}{2n}\mathbb{E}\left(2L^2\left\|\frac{(X_k-\hat{X}_k)\mathbf{1}_n}{n}\right\|^2+2L^2\hat{M}_k\right)\\
&-\frac{\gamma M}{2n}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2-\left(\frac{\gamma M}{2n}-\frac{\gamma^2LM^2}{n^2}\right)\mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2\\
&+\frac{6\gamma^2L^3M^2}{n^2}\mathbb{E}\hat{M}_k+\frac{\gamma^2L(\sigma^2M+6\varsigma^2M^2)}{2n^2}\\
=&\mathbb{E}f\left(\frac{X_k\mathbf{1}_n}{n}\right)-\frac{\gamma M}{2n}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2-\left(\frac{\gamma M}{2n}-\frac{\gamma^2LM^2}{n^2}\right)\mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2\\
&+\left(\frac{6\gamma^2L^3M^2}{n^2}+\frac{\gamma M}{n}L^2\right)\mathbb{E}\hat{M}_k+\frac{\gamma M}{n}L^2\mathbb{E}\left\|\frac{(X_k-\hat{X}_k)\mathbf{1}_n}{n}\right\|^2+\frac{\gamma^2L(\sigma^2M+6\varsigma^2M^2)}{2n^2}\\
\overset{\text{Lemma 8}}{\leqslant}&\mathbb{E}f\left(\frac{X_k\mathbf{1}_n}{n}\right)-\frac{\gamma M}{2n}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2-\left(\frac{\gamma M}{2n}-\frac{\gamma^2LM^2}{n^2}\right)\mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2\\
&+\left(\frac{6\gamma^2L^3M^2}{n^2}+\frac{\gamma M}{n}L^2\right)\mathbb{E}\hat{M}_k+\frac{\gamma M}{n}L^2\left(\frac{\tau_k^2\gamma^2\sigma^2M}{n^2}+\tau_k\gamma^2\sum_{t=1}^{\tau_k}\left(\frac{M^2}{n^2}\sum_{i=1}^n p_i\mathbb{E}\|\nabla f_i(\hat{x}_{k-t}^i)\|^2\right)\right)\\
&+\frac{\gamma^2L(\sigma^2M+6\varsigma^2M^2)}{2n^2}\\
\leqslant&\mathbb{E}f\left(\frac{X_k\mathbf{1}_n}{n}\right)-\frac{\gamma M}{2n}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2-\left(\frac{\gamma M}{2n}-\frac{\gamma^2LM^2}{n^2}\right)\mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2\\
&+\left(\frac{6\gamma^2L^3M^2}{n^2}+\frac{\gamma M}{n}L^2\right)\mathbb{E}\hat{M}_k+\frac{\gamma^2L(\sigma^2M+6\varsigma^2M^2)}{2n^2}+\frac{L^2T^2\gamma^3\sigma^2M^2}{n^3}
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{M^3L^2\tau_k\gamma^3}{n^3}\sum_{t=1}^{\tau_k}\left(\sum_{i=1}^{n}p_i\mathbb{E}\|\nabla f_i(\hat{x}^i_{k-t})\|^2\right)\\
\overset{\text{Lemma 5}}{\leqslant}&\mathbb{E}f\left(\frac{X_k\mathbf{1}_n}{n}\right)-\frac{\gamma M}{2n}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2-\left(\frac{\gamma M}{2n}-\frac{\gamma^2LM^2}{n^2}\right)\mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2\\
&+\left(\frac{6\gamma^2L^3M^2}{n^2}+\frac{\gamma M}{n}L^2\right)\mathbb{E}\hat{M}_k+\frac{\gamma^2L(\sigma^2M+6\varsigma^2M^2)}{2n^2}+\frac{L^2T^2\gamma^3\sigma^2M^2}{n^3}\\
&+\frac{M^3L^2\tau_k\gamma^3}{n^3}\sum_{t=1}^{\tau_k}\left(12L^2\hat{M}_{k-t}+6\varsigma^2+2\mathbb{E}\left\|\sum_{j=1}^{n}p_j\nabla f_j(\hat{x}^j_{k-t})\right\|^2\right)\\
=&\mathbb{E}f\left(\frac{X_k\mathbf{1}_n}{n}\right)-\frac{\gamma M}{2n}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2-\left(\frac{\gamma M}{2n}-\frac{\gamma^2LM^2}{n^2}\right)\mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2\\
&+\left(\frac{6\gamma^2L^3M^2}{n^2}+\frac{\gamma M}{n}L^2\right)\mathbb{E}\hat{M}_k+\frac{\gamma^2L(\sigma^2M+6\varsigma^2M^2)}{2n^2}+\frac{L^2T^2\gamma^3M(\sigma^2M+6\varsigma^2M^2)}{n^3}\\
&+\frac{2M^3L^2T\gamma^3}{n^3}\sum_{t=1}^{\tau_k}\left(6L^2\mathbb{E}\hat{M}_{k-t}+\mathbb{E}\left\|\frac{\partial f(\hat{X}_{k-t})\mathbf{1}_n}{n}\right\|^2\right).
\end{aligned}
$$

Summing from $k=0$ to $k=K-1$ we obtain

$$
\begin{aligned}
\mathbb{E}f\left(\frac{X_K\mathbf{1}_n}{n}\right)\leqslant&\mathbb{E}f\left(\frac{X_0\mathbf{1}_n}{n}\right)-\frac{\gamma M}{2n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2-\left(\frac{\gamma M}{2n}-\frac{\gamma^2LM^2}{n^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2\\
&+\left(\frac{6\gamma^2L^3M^2}{n^2}+\frac{\gamma M}{n}L^2\right)\sum_{k=0}^{K-1}\mathbb{E}\hat{M}_k+\frac{\gamma^2L(\sigma^2M+6\varsigma^2M^2)K}{2n^2}+\frac{L^2T^2\gamma^3M(\sigma^2M+6\varsigma^2M^2)K}{n^3}\\
&+\frac{2M^3L^2T\gamma^3}{n^3}\sum_{k=0}^{K-1}\sum_{t=1}^{\tau_k}\left(6L^2\mathbb{E}\hat{M}_{k-t}+\mathbb{E}\left\|\frac{\partial f(\hat{X}_{k-t})\mathbf{1}_n}{n}\right\|^2\right)\\
\leqslant&\mathbb{E}f\left(\frac{X_0\mathbf{1}_n}{n}\right)-\frac{\gamma M}{2n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2-\left(\frac{\gamma M}{2n}-\frac{\gamma^2LM^2}{n^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\|^2\\
&+\left(\frac{6\gamma^2L^3M^2}{n^2}+\frac{\gamma M}{n}L^2\right)\sum_{k=0}^{K-1}\mathbb{E}\hat{M}_k+\frac{\gamma^2L(\sigma^2M+6\varsigma^2M^2)K}{2n^2}+\frac{L^2T^2\gamma^3M(\sigma^2M+6\varsigma^2M^2)K}{n^3}\\
&+\frac{2M^3L^2T^2\gamma^3}{n^3}\sum_{k=0}^{K-1}\left(6L^2\mathbb{E}\hat{M}_k+\mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2\right)\\
=&\mathbb{E}f\left(\frac{X_0\mathbf{1}_n}{n}\right)-\frac{\gamma M}{2n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2\\
&-\left(\frac{\gamma M}{2n}-\frac{\gamma^2LM^2}{n^2}-\frac{2M^3L^2T^2\gamma^3}{n^3}\right)\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2\\
&+\left(\frac{6\gamma^2L^3M^2}{n^2}+\frac{\gamma M}{n}L^2+\frac{12M^3L^4T^2\gamma^3}{n^3}\right)\sum_{k=0}^{K-1}\mathbb{E}\hat{M}_k\\
&+\frac{\gamma^2L(\sigma^2M+6\varsigma^2M^2)K}{2n^2}+\frac{L^2T^2\gamma^3M(\sigma^2M+6\varsigma^2M^2)K}{n^3}.\\
\leqslant&\mathbb{E}f\left(\frac{X_0\mathbf{1}_n}{n}\right)-\frac{\gamma M}{2n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&- \left(\frac{\gamma M}{2n} - \frac{\gamma^2 L M^2}{n^2} - \frac{2M^3L^2T^2\gamma^3}{n^3}\right) \sum_{k=0}^{K-1} \mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2 \\
&+ \left(\frac{6\gamma^2L^3M^2}{n^2} + \frac{\gamma M}{n}L^2 + \frac{12M^3L^4T^2\gamma^3}{n^3}\right) \sum_{k=0}^{K-1} \mathbb{E}\hat{M}_k \\
&+ \frac{\gamma^2 L(\sigma^2 M + 6\varsigma^2 M^2)K}{2n^2} + \frac{L^2T^2\gamma^3 M(\sigma^2 M + 6\varsigma^2 M^2)K}{n^3} \\
\overset{C_1>0,\text{Lemma 7}}{\leqslant}\ & \mathbb{E}f\left(\frac{X_0\mathbf{1}_n}{n}\right) - \frac{\gamma M}{2n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2 \\
&- \left(\frac{\gamma M}{2n} - \frac{\gamma^2 L M^2}{n^2} - \frac{2M^3L^2T^2\gamma^3}{n^3}\right) \sum_{k=0}^{K-1} \mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2 \\
&+ \left(\frac{6\gamma^2L^3M^2}{n^2} + \frac{\gamma M}{n}L^2 + \frac{12M^3L^4T^2\gamma^3}{n^3}\right) K\frac{2\gamma^2(M\sigma^2 + 6M^2\varsigma^2)\bar{\rho}}{C_1} \\
&+ \left(\frac{6\gamma^2L^3M^2}{n^2} + \frac{\gamma M}{n}L^2 + \frac{12M^3L^4T^2\gamma^3}{n^3}\right) \frac{4M^2\gamma^2\left(T\frac{n-1}{n} + \bar{\rho}\right)\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^n p_i \nabla f_i(\hat{x}_k^i)\right\|^2}{C_1} \\
&+ \frac{\gamma^2 L(\sigma^2 M + 6\varsigma^2 M^2)K}{2n^2} + \frac{L^2T^2\gamma^3 M(\sigma^2 M + 6\varsigma^2 M^2)K}{n^3} \\
=& \mathbb{E}f\left(\frac{X_0\mathbf{1}_n}{n}\right) - \frac{\gamma M}{2n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2 \\
&- C_2\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{\partial f(\hat{X}_k)\mathbf{1}_n}{n}\right\|^2 + C_3\frac{\gamma^2 LK}{n^2}(M\sigma^2 + 6M^2\varsigma^2).
\end{aligned}
$$

Thus while $C_3 \leqslant 1$ and $C_2 \geqslant 0$ we have

$$
\begin{aligned}
\frac{\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2}{K} \leqslant& \frac{2\left(\mathbb{E}f\left(\frac{X_0\mathbf{1}_n}{n}\right) - \mathbb{E}f\left(\frac{X_K\mathbf{1}_n}{n}\right)\right)}{\gamma KM/n} + \frac{2\gamma L}{Mn}(M\sigma^2 + 6M^2\varsigma^2) \\
\leqslant& \frac{2(\mathbb{E}f(x_0) - \mathbb{E}f^*)}{\gamma KM/n} + \frac{2\gamma L}{n}(\sigma^2 + 6M\varsigma^2).
\end{aligned}
$$

It completes the proof. □

Lemma 3. *Define $\prod_{k=1}^0 W_k = I$, where I is the identity matrix. Then*

$$
\mathbb{E}\left\|\frac{\mathbf{1}_n}{n} - \prod_{k=1}^K W_k e_i\right\|^2 \leq \frac{n-1}{n}\rho^K, \quad \forall K \geq 0.
$$

Proof. Let $y_K = \frac{\mathbf{1}_n}{n} - \prod_{k=1}^K W_k e_i$. Then noting that $y_{K+1} = W_{K+1}y_K$ we have

$$
\begin{aligned}
&\mathbb{E}\|y_{K+1}\|^2 \\
=&\mathbb{E}\|W_{k+1}y_K\|^2 \\
=&\mathbb{E}\langle W_{K+1}y_K, W_{K+1}y_K\rangle \\
=&\mathbb{E}\langle y_K, W_{K+1}^\top W_{K+1}y_K\rangle \\
=&\mathbb{E}\langle y_K, \mathbb{E}\mathbb{E}_{i_{K+1}}(W_{K+1}^\top W_{K+1})y_K\rangle \\
=&\mathbb{E}\langle y_K, \mathbb{E}(W_{K+1}^\top W_{K+1})y_K\rangle.
\end{aligned}
$$

Note that $\mathbb{E}(W_{K+1}^\top W_{K+1})$ is symmetric and doubly stochastic and $\mathbf{1}_n$ is an eigenvector of $\mathbb{E}(W_{K+1}^\top W_{K+1})$ with eigenvalue 1. Starting from $\mathbf{1}_n$ we construct a basis of $\mathbb{R}^n$ composed by the eigenvectors of $\mathbb{E}(W_{K+1}^\top W_{K+1})$, which is guaranteed to exist by the spectral theorem of Hermitian matrices. From (2) the magnitude of all other eigenvectors' associated eigenvalues should be smaller or equal to ρ. Noting y_K is orthogonal to $\mathbf{1}_n$, we decompose y_K using this constructed basis

and it follows that

$$\mathbb{E}\|y_{K+1}\|^2 \leqslant \rho \mathbb{E}\|y_K\|^2.$$

Noting that $\|y_0\|^2 = \|\mathbf{1}_n/n - e_i\|^2 = \frac{(n-1)^2}{n^2} + \sum_{i=1}^{n-1}\frac{1}{n^2} = \frac{n^2-2n+1+n-1}{n^2} = \frac{n-1}{n}$, by induction, we complete the proof. □

Lemma 4.

$$\mathbb{E}\left\|\frac{\partial g(\hat{X}_k;\xi_k^{i_k},i_k)\mathbf{1}_n}{n}\right\|^2 \leqslant \frac{\sigma^2 M}{n^2} + \frac{M^2}{n^2}\sum_{i=1}^n p_i \mathbb{E}\|\nabla f_i(\hat{x}_k^i)\|^2, \quad \forall k \geq 0.$$

Proof. The LHS can be bounded by

$$\begin{aligned}
\mathbb{E}\left\|\frac{\partial g(\hat{X}_k;\xi_k^{i_k},i_k)\mathbf{1}_n}{n}\right\|^2 &\overset{(1)}{=} \sum_{i=1}^n p_i \mathbb{E}\left\|\frac{\sum_{j=1}^M \nabla F(\hat{x}_k^i,\xi_{k,j}^i)}{n}\right\|^2 \\
&= \sum_{i=1}^n p_i \mathbb{E}\left\|\frac{\sum_{j=1}^M (\nabla F(\hat{x}_k^i,\xi_{k,j}^i) - \nabla f_i(\hat{x}_k^i))}{n}\right\|^2 + \sum_{i=1}^n p_i \mathbb{E}\left\|\frac{M\nabla f_i(\hat{x}_k^i)}{n}\right\|^2 \\
&\overset{(5)}{\leqslant} \frac{\sigma^2 M}{n^2} + \frac{M^2}{n^2}\sum_{i=1}^n p_i \mathbb{E}\|\nabla f_i(\hat{x}_k^i)\|^2.
\end{aligned}$$

□

Lemma 5.

$$\sum_{i=1}^n p_i \mathbb{E}\|\nabla f_i(\hat{x}_k^i)\|^2 \leqslant 12L^2\mathbb{E}\hat{M}_k + 6\varsigma^2 + 2\mathbb{E}\left\|\sum_{j=1}^n p_j \nabla f_j(\hat{x}_k^j)\right\|^2, \quad \forall k \geq 0.$$

Proof. The LHS can be bounded by

$$\begin{aligned}
\sum_{i=1}^n p_i \mathbb{E}\|\nabla f_i(\hat{x}_k^i)\|^2 &\leqslant \sum_{i=1}^n p_i \mathbb{E}\|\nabla f_i(\hat{x}_k^i) - \sum_{j=1}^n p_j \nabla f_j(\hat{x}_k^j) + \sum_{j=1}^n p_j \nabla f_j(\hat{x}_k^j)\|^2 \\
&\leqslant 2\sum_{i=1}^n p_i \mathbb{E}\left\|\nabla f_i(\hat{x}_k^i) - \sum_{j=1}^n p_j \nabla f_j(\hat{x}_k^j)\right\|^2 + 2\sum_{i=1}^n p_i \mathbb{E}\left\|\sum_{j=1}^n p_j \nabla f_j(\hat{x}_k^j)\right\|^2 \\
&= 2\sum_{i=1}^n p_i \mathbb{E}\left\|\nabla f_i(\hat{x}_k^i) - \sum_{j=1}^n p_j \nabla f_j(\hat{x}_k^j)\right\|^2 + 2\mathbb{E}\left\|\sum_{j=1}^n p_j \nabla f_j(\hat{x}_k^j)\right\|^2. \qquad (13)
\end{aligned}$$

For the first term on the RHS we have

$$\begin{aligned}
&\sum_{i=1}^n p_i \mathbb{E}\left\|\nabla f_i(\hat{x}_k^i) - \sum_{j=1}^n p_j \nabla f_j(\hat{x}_k^j)\right\|^2 \\
\leqslant& 3\sum_{i=1}^n p_i \mathbb{E}\left\|\nabla f_i(\hat{x}_k^i) - \nabla f_i\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)\right\|^2 + 3\sum_{i=1}^n p_i \mathbb{E}\left\|\nabla f_i\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right) - \sum_{j=1}^n p_j \nabla f_j\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)\right\|^2 \\
&+ 3\sum_{i=1}^n p_i \mathbb{E}\left\|\sum_{j=1}^n p_j \nabla f_j(\hat{x}_k^j) - \sum_{j=1}^n p_j \nabla f_j\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)\right\|^2 \\
\leqslant& 3L^2\sum_{i=1}^n p_i \mathbb{E}\left\|\hat{x}_k^i - \frac{\hat{X}_k\mathbf{1}_n}{n}\right\|^2 + 3\sum_{i=1}^n p_i \mathbb{E}\left\|\nabla f_i\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right) - \sum_{j=1}^n p_j \nabla f_j\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&+3\mathbb{E}\left\|\sum_{j=1}^{n}p_j\nabla f_j(\hat{x}_k^j)-\sum_{j=1}^{n}p_j\nabla f_j\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)\right\|^2\\
\leqslant&3L^2\mathbb{E}\hat{M}_k+3\sum_{i=1}^{n}p_i\mathbb{E}\left\|\nabla f_i\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)-\nabla f\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)\right\|^2+3\sum_{j=1}^{n}p_j\mathbb{E}\left\|\nabla f_j(\hat{x}_k^j)-\nabla f_j\left(\frac{\hat{X}_k\mathbf{1}_n}{n}\right)\right\|^2\\
\leqslant&6L^2\mathbb{E}\hat{M}_k+3\varsigma^2.
\end{aligned}
$$

Plugging this upper bound into (13) we complete the proof. □

Lemma 6. *For any $k\geq -1$ we have*

$$
\begin{aligned}
&\mathbb{E}\left\|\frac{X_{k+1}\mathbf{1}_n}{n}-X_{k+1}e_i\right\|^2\\
\leqslant&2\gamma^2(M\sigma^2+6M^2\varsigma^2)\bar{\rho}\\
&+2\frac{n-1}{n}M^2\gamma^2\mathbb{E}\sum_{j=0}^{k}\left(12L^2\hat{M}_j+2\mathbb{E}\left\|\sum_{i=1}^{n}p_i\nabla f_i(\hat{x}_j^i)\right\|^2\right)\left(\rho^{k-j}+2(k-j)\rho^{\frac{k-j}{2}}\right).
\end{aligned}
$$

Proof. Note that for $k=-1$, we have

$$
\mathbb{E}\left\|\frac{X_{k+1}\mathbf{1}_n}{n}-X_{k+1}e_i\right\|^2=0.
$$

Also note that the columns of X_0 are the same (all workers start with the same model), we have $X_0W_k=X_0$ for all k and $X_0\mathbf{1}_n/n-X_0e_i=0,\forall i$. It follows that

$$
\begin{aligned}
&\mathbb{E}\left\|\frac{X_{k+1}\mathbf{1}_n}{n}-X_{k+1}e_i\right\|^2\\
=&\mathbb{E}\left\|\frac{X_k\mathbf{1}_n-\gamma\partial g(\hat{X}_k;\xi_k^{i_k},i_k)\mathbf{1}_n}{n}-(X_kW_ke_i-\gamma\partial g(\hat{X}_k,\xi_k^{i_k},i_k)e_i)\right\|^2\\
=&\mathbb{E}\left\|\frac{X_0\mathbf{1}_n-\sum_{j=0}^{k}\gamma\partial g(\hat{X}_j;\xi_j^{i_j},i_j)\mathbf{1}_n}{n}-\left(X_0\prod_{j=0}^{k}W_je_i-\sum_{j=0}^{k}\gamma\partial g(\hat{X}_j;\xi_j^{i_j},i_j)\prod_{q=j+1}^{k}W_qe_i\right)\right\|^2\\
=&\mathbb{E}\left\|-\sum_{j=0}^{k}\gamma\partial g(\hat{X}_j;\xi_j^{i_j},i_j)\frac{\mathbf{1}_n}{n}+\sum_{j=0}^{k}\gamma\partial g(\hat{X}_j;\xi_j^{i_j},i_j)\prod_{q=j+1}^{k}W_qe_i\right\|^2\\
=&\gamma^2\mathbb{E}\left\|\sum_{j=0}^{k}\partial g(\hat{X}_j,\xi_j^{i_j},i_j)\left(\frac{\mathbf{1}_n}{n}-\prod_{q=j+1}^{k}W_qe_i\right)\right\|^2\\
\leqslant&2\gamma^2\underbrace{\mathbb{E}\left\|\sum_{j=0}^{k}(\partial g(\hat{X}_j,\xi_j^{i_j},i_j)-M\partial f(\hat{X}_j,i_j))\left(\frac{\mathbf{1}_n}{n}-\prod_{q=j+1}^{k}W_qe_i\right)\right\|^2}_{A_1}\\
&+2M^2\gamma^2\underbrace{\mathbb{E}\left\|\sum_{j=0}^{k}\partial f(\hat{X}_j,i_j)\left(\frac{\mathbf{1}_n}{n}-\prod_{q=j+1}^{k}W_qe_i\right)\right\|^2}_{A_2}.
\end{aligned}
\tag{14}
$$

For A_1,

$$
A_1=\mathbb{E}\left\|\sum_{j=0}^{k}(\partial g(\hat{X}_j,\xi_j^{i_j},i_j)-M\partial f(\hat{X}_j,i_j))\left(\frac{\mathbf{1}_n}{n}-\prod_{q=j+1}^{k}W_qe_i\right)\right\|^2
$$

$$= \underbrace{\sum_{j=0}^{k} \mathbb{E} \left\| (\partial g(\hat{X}_j, \xi_j^{i_j}, i_j) - M\partial f(\hat{X}_j, i_j)) \left(\frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right) \right\|^2}_{A_3}$$

$$+ \underbrace{2\,\mathbb{E} \sum_{k \geqslant j > j' \geqslant 0} \begin{matrix} \left\langle (\partial g(\hat{X}_j, \xi_j^{i_j}, i_j) - M\partial f(\hat{X}_j, i_j)) \left(\frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right), \right. \\ \left. (\partial g(\hat{X}_{j'} \xi_{j'}^{i_{j'}}, i_{j'}) - M\partial f(\hat{X}_{j'}, i_{j'})) \left(\frac{\mathbf{1}_n}{n} - \prod_{q=j'+1}^{k} W_q e_i \right) \right\rangle \end{matrix}}_{A_4}.$$

A_3 can be bounded by a constant:

$$\begin{aligned} A_3 =& \sum_{j=0}^{k} \mathbb{E} \left\| (\partial g(\hat{X}_j, \xi_j^{i_j}, i_j) - M\partial f(\hat{X}_j, i_j)) \left(\frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right) \right\|^2 \\ \leqslant& \sum_{j=0}^{k} \mathbb{E} \|\partial g(\hat{X}_j, \xi_j^{i_j}, i_j) - M\partial f(\hat{X}_j, i_j)\|^2 \left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right\|^2 \\ \overset{\text{Lemma 3}}{\leqslant}& \frac{n-1}{n} \sum_{j=0}^{k} \mathbb{E} \|\partial g(\hat{X}_j, \xi_j^{i_j}, i_j) - M\partial f(\hat{X}_j, i_j)\|^2 \rho^{k-j} \\ \overset{\text{Assumption 1:5}}{\leqslant}& \frac{n-1}{n} M\sigma^2 \sum_{j=0}^{k} \rho^{k-j} \leqslant \frac{n-1}{n} \frac{M\sigma^2}{1-\rho}. \end{aligned}$$

A_4 can be bounded by another constant:

$$\begin{aligned} A_4 =& \sum_{k \geqslant j > j' \geqslant 0} \mathbb{E} \begin{matrix} \left\langle (\partial g(\hat{X}_j, \xi_j^{i_j}, i_j) - M\partial f(\hat{X}_j, i_j)) \left(\frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right), \right. \\ \left. (\partial g(\hat{X}_{j'}, \xi_{j'}^{i_{j'}}, i_{j'}) - M\partial f(\hat{X}_{j'}, i_{j'})) \left(\frac{\mathbf{1}_n}{n} - \prod_{q=j'+1}^{k} W_q e_i \right) \right\rangle \end{matrix} \\ \leqslant& \sum_{k \geqslant j > j' \geqslant 0} \begin{matrix} \mathbb{E} \|\partial g(\hat{X}_j, \xi_j^{i_j}, i_j) - M\partial f(\hat{X}_j, i_j)\| \left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right\| \times \\ \|\partial g(\hat{X}_{j'}, \xi_{j'}^{i_{j'}}, i_{j'}) - M\partial f(\hat{X}_{j'}, i_{j'})\| \left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j'+1}^{k} W_q e_i \right\| \end{matrix} \\ \leqslant& \mathbb{E} \sum_{k \geqslant j > j' \geqslant 0} \left(\begin{matrix} \frac{\left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right\|^2 \left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j'+1}^{k} W_q e_i \right\|^2}{2\alpha_{j,j'}} \\ + \frac{\|\partial g(\hat{X}_j, \xi_j^{i_j}, i_j) - M\partial f(\hat{X}_j, i_j)\|^2 \|\partial g(\hat{X}_{j'}, \xi_{j'}^{i_{j'}}, i_{j'}) - M\partial f(\hat{X}_{j'}, i_{j'})\|^2}{2/\alpha_{j,j'}} \end{matrix} \right), \forall \alpha_{j,j'} > 0 \\ \overset{(5)}{\leqslant}& \mathbb{E} \sum_{k \geqslant j > j' \geqslant 0} \left(\frac{\left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right\|^2 \left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j'+1}^{k} W_q e_i \right\|^2}{2\alpha_{j,j'}} + \frac{M^2\sigma^4}{2/\alpha_{j,j'}} \right), \forall \alpha_{j,j'} > 0 \\ \leqslant& \mathbb{E} \sum_{k \geqslant j > j' \geqslant 0} \left(\frac{n-1}{n} \frac{\left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j'+1}^{k} W_q e_i \right\|^2}{2\alpha_{j,j'}} + \frac{M^2\sigma^4}{2/\alpha_{j,j'}} \right), \forall \alpha_{j,j'} > 0 \\ \overset{\text{Lemma 3}}{\leqslant}& \mathbb{E} \sum_{k \geqslant j > j' \geqslant 0} \left(\left(\frac{n-1}{n} \right)^2 \frac{\rho^{k-j'}}{2\alpha_{j,j'}} + \frac{M^2\sigma^4}{2/\alpha_{j,j'}} \right), \forall \alpha_{j,j'} > 0. \end{aligned}$$

We can choose $\alpha_{j,j'} > 0$ to make the term in the last step become $\frac{n-1}{n} \sum_{k \geqslant j > j' \geqslant 0}^{k} \rho^{\frac{k-j'}{2}} M\sigma^2$ (by applying inequality of arithmetic and geometric means). Thus

$$\begin{aligned} A_4 \leqslant& \frac{n-1}{n} \sum_{k \geqslant j > j' \geqslant 0}^{k} \rho^{\frac{k-j'}{2}} M\sigma^2 \leqslant \frac{n-1}{n} M\sigma^2 \sum_{j'=0}^{k} \sum_{j=j'+1}^{k} \rho^{\frac{k-j'}{2}} \\ =& \frac{n-1}{n} M\sigma^2 \sum_{j'=0}^{k} (k-j') \rho^{\frac{k-j'}{2}} \leqslant \frac{n-1}{n} M\sigma^2 \frac{\sqrt{\rho}}{\left(1-\sqrt{\rho}\right)^2}. \end{aligned}$$

Putting A_3 and A_4 back into A_1 we obtain:

$$A_1 \leqslant \frac{n-1}{n} M\sigma^2 \left(\frac{1}{1-\rho} + \frac{2\sqrt{\rho}}{\left(1-\sqrt{\rho}\right)^2} \right) = M\sigma^2 \bar{\rho}. \tag{15}$$

We then start bounding A_2:

$$\begin{aligned}
A_2 =& \mathbb{E} \left\| \sum_{j=0}^{k} \partial f(\hat{X}_j, i_j) \left(\frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right) \right\|^2 \\
=& \mathbb{E} \sum_{j=0}^{k} \left\| \partial f(\hat{X}_j, i_j) \left(\frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right) \right\|^2 \\
&+ 2\mathbb{E} \sum_{j=0}^{k} \sum_{j'=j+1}^{k} \left\langle \partial f(\hat{X}_j, i_j) \left(\frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right), \partial f(\hat{X}_{j'}, i_{j'}) \left(\frac{\mathbf{1}_n}{n} - \prod_{q=j'+1}^{k} W_q e_i \right) \right\rangle \\
\overset{\text{Lemma 3,(1)}}{\leqslant}& \frac{n-1}{n} \mathbb{E} \sum_{j=0}^{k} \left(\sum_{i=1}^{n} p_i \|\nabla f_i(\hat{x}_j^i)\|^2 \right) \rho^{k-j} \\
&+ 2\mathbb{E} \sum_{j=0}^{k} \sum_{j'=j+1}^{k} \|\partial f(\hat{X}_j, i_j)\| \left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right\| \|\partial f(\hat{X}_{j'}, i_{j'})\| \left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j'+1}^{k} W_q e_i \right\|.
\end{aligned} \tag{16}$$

For the second term:

$$\begin{aligned}
&\mathbb{E} \sum_{j=0}^{k} \sum_{j'=j+1}^{k} \|\partial f(\hat{X}_j, i_j)\| \left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right\| \|\partial f(\hat{X}_{j'}, i_{j'})\| \left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j'+1}^{k} W_q e_i \right\| \\
\leqslant& \mathbb{E} \sum_{j=0}^{k} \sum_{j'=j+1}^{k} \left(\frac{\|\partial f(\hat{X}_j, i_j)\|^2 \|\partial f(\hat{X}_{j'}, i_{j'})\|^2}{2\alpha_{j,j'}} + \frac{\left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right\|^2 \left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j'+1}^{k} W_q e_i \right\|^2}{2/\alpha_{j,j'}} \right), \forall \alpha_{j,j'} > 0 \\
\overset{\text{Lemma 3}}{\leqslant}& \frac{1}{2} \mathbb{E} \sum_{j \neq j'}^{k} \left(\frac{\|\partial f(\hat{X}_j, i_j)\|^2 \|\partial f(\hat{X}_{j'}, i_{j'})\|^2}{2\alpha_{j,j'}} + \frac{\rho^{k-\min\{j,j'\}}}{2/\alpha_{j,j'}} \left(\frac{n-1}{n} \right)^2 \right), \quad \forall \alpha_{j,j'} > 0, \alpha_{j,j'} = \alpha_{j',j}.
\end{aligned}$$

By applying inequality of arithmetic and geometric means to the term in the last step we can choose $\alpha_{j,j'} > 0$ such that

$$\begin{aligned}
&\mathbb{E} \sum_{j=0}^{k} \sum_{j'=j+1}^{k} \|\partial f(\hat{X}_j, i_j)\| \left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j+1}^{k} W_q e_i \right\| \|\partial f(\hat{X}_{j'}, i_{j'})\| \left\| \frac{\mathbf{1}_n}{n} - \prod_{q=j'+1}^{k} W_q e_i \right\| \\
\leqslant& \frac{n-1}{2n} \mathbb{E} \sum_{j \neq j'}^{k} \left(\|\partial f(\hat{X}_j, i_j)\| \|\partial f(\hat{X}_{j'}, i_{j'})\| \rho^{\frac{k-\min\{j,j'\}}{2}} \right) \\
\leqslant& \frac{n-1}{2n} \mathbb{E} \sum_{j \neq j'}^{k} \left(\frac{\|\partial f(\hat{X}_j, i_j)\|^2 + \|\partial f(\hat{X}_{j'}, i_{j'})\|^2}{2} \rho^{\frac{k-\min\{j,j'\}}{2}} \right) \\
=& \frac{n-1}{2n} \mathbb{E} \sum_{j \neq j'}^{k} \left(\|\partial f(\hat{X}_j, i_j)\|^2 \rho^{\frac{k-\min\{j,j'\}}{2}} \right) = \frac{n-1}{n} \sum_{j=0}^{k} \sum_{j'=j+1}^{k} \left(\mathbb{E} \|\partial f(\hat{X}_j, i_j)\|^2 \rho^{\frac{k-j}{2}} \right) \\
=& \frac{n-1}{n} \sum_{j=0}^{k} \left(\sum_{i=1}^{n} p_i \mathbb{E} \|\nabla f_i(\hat{x}_j^i)\|^2 \right) (k-j) \rho^{\frac{k-j}{2}}.
\end{aligned} \tag{17}$$

It follows from (17) and (16) that

$$A_2 \leqslant \frac{n-1}{n} \mathbb{E} \sum_{j=0}^{k} \left(\sum_{i=1}^{n} p_i \|\nabla f_i(\hat{x}_j^i)\|^2 \right) \left(\rho^{k-j} + 2(k-j) \rho^{\frac{k-j}{2}} \right)$$

$$
\begin{aligned}
&\overset{\text{Lemma 5}}{\leqslant} \frac{n-1}{n}\sum_{j=0}^{k}\left(12L^2\mathbb{E}\hat{M}_j + 6\varsigma^2 + 2\mathbb{E}\left\|\sum_{j'=1}^{n} p_{j'}\nabla f_{j'}(\hat{x}_j^{j'})\right\|^2\right)\left(\rho^{k-j} + 2(k-j)\rho^{\frac{k-j}{2}}\right)\\
&\leqslant \frac{n-1}{n}\sum_{j=0}^{k}\left(12L^2\mathbb{E}\hat{M}_j + 2\mathbb{E}\left\|\sum_{j'=1}^{n} p_{j'}\nabla f_{j'}(\hat{x}_j^{j'})\right\|^2\right)\left(\rho^{k-j} + 2(k-j)\rho^{\frac{k-j}{2}}\right)\\
&\quad + 6\varsigma^2\underbrace{\frac{n-1}{n}\left(\frac{1}{1-\rho} + \frac{2\sqrt{\rho}}{\left(1-\sqrt{\rho}\right)^2}\right)}_{=\bar{\rho}}.
\end{aligned}
\tag{18}
$$

Finally from (15), (18) and (14) we obtain

$$
\begin{aligned}
&\mathbb{E}\left\|\frac{X_{k+1}\mathbf{1}_n}{n} - X_{k+1}e_i\right\|^2\\
\leqslant& 2\gamma^2 A_1 + 2M^2\gamma^2 A_2\\
\leqslant& 2\gamma^2 M\sigma^2\bar{\rho}\\
&+ 2\gamma^2 M^2\mathbb{E}\sum_{j=0}^{k}\left(12L^2\hat{M}_j + 2\mathbb{E}\left\|\sum_{i=1}^{n} p_i\nabla f_i(\hat{x}_j^i)\right\|^2\right)\left(\rho^{k-j} + 2(k-j)\rho^{\frac{k-j}{2}}\right) + 12\gamma^2 M^2\varsigma^2\bar{\rho}\\
=& 2\gamma^2(M\sigma^2 + 6M^2\varsigma^2)\bar{\rho}\\
&+ 2\frac{n-1}{n}M^2\gamma^2\mathbb{E}\sum_{j=0}^{k}\left(12L^2\hat{M}_j + 2\mathbb{E}\left\|\sum_{i=1}^{n} p_i\nabla f_i(\hat{x}_j^i)\right\|^2\right)\left(\rho^{k-j} + 2(k-j)\rho^{\frac{k-j}{2}}\right).
\end{aligned}
$$

This completes the proof.

□

Lemma 7. *While $C_1 > 0$, we have*

$$
\frac{\sum_{k=0}^{K-1}\mathbb{E}\hat{M}_k}{K} \leqslant \frac{2\gamma^2(M\sigma^2 + 6M^2\varsigma^2)\bar{\rho} + \frac{4\gamma^2M^2}{K}\left(T\frac{n-1}{n} + \bar{\rho}\right)\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n} p_i\nabla f_i(\hat{x}_k^i)\right\|^2}{C_1}, \quad \forall K \geq 1.
$$

Proof. From Lemma 6 and noting that $\hat{X}_k = X_{k-\tau_k}$, we have

$$
\begin{aligned}
&\mathbb{E}\left\|\frac{\hat{X}_k\mathbf{1}_n}{n} - \hat{X}_k e_i\right\|^2\\
\leqslant& 2\gamma^2(M\sigma^2 + 6M^2\varsigma^2)\bar{\rho}\\
&+ 2\frac{n-1}{n}M^2\gamma^2\sum_{j=0}^{k-\tau_k-1}\left(12L^2\mathbb{E}\hat{M}_j + 2\mathbb{E}\left\|\sum_{i=1}^{n} p_i\nabla f_i(\hat{x}_j^i)\right\|^2\right)\left(\rho^{k-\tau_k-1-j} + 2(k-\tau_k-1-j)\rho^{\frac{k-\tau_k-1-j}{2}}\right).
\end{aligned}
$$

By averaging from $i = 1$ to n with distribution $\mathcal{I}$ we obtain

$$
\begin{aligned}
&\mathbb{E}\hat{M}_k = \sum_{i=1}^{n} p_i\mathbb{E}\left\|\frac{\hat{X}_k\mathbf{1}_n}{n} - \hat{X}_k e_i\right\|^2\\
\leqslant& 2\gamma^2(M\sigma^2 + 6M^2\varsigma^2)\bar{\rho}\\
&+ 2M^2\gamma^2\frac{n-1}{n}\sum_{j=0}^{k-\tau_k-1}\left(12L^2\mathbb{E}\hat{M}_j + 2\mathbb{E}\left\|\sum_{i=1}^{n} p_i\nabla f_i(\hat{x}_j^i)\right\|^2\right)\left(\rho^{k-\tau_k-1-j} + 2(k-\tau_k-1-j)\rho^{\frac{k-\tau_k-1-j}{2}}\right).
\end{aligned}
$$

It follows that

$$
\frac{\sum_{k=0}^{K-1}\mathbb{E}\hat{M}_k}{K} \leqslant 2\gamma^2(M\sigma^2 + 6M^2\varsigma^2)\bar{\rho}
$$

$$
\begin{aligned}
&+\frac{2\gamma^2}{K}\frac{n-1}{n}M^2\sum_{k=0}^{K-1}\sum_{j=0}^{k-\tau_k-1}\left(12L^2\mathbb{E}\hat{M}_j+2\mathbb{E}\left\|\sum_{i=1}^n p_i\nabla f_i(\hat{x}_j^i)\right\|^2\right)\times\\
&\qquad\left(\rho^{k-\tau_k-1-j}+2(k-\tau_k-1-j)\rho^{\frac{k-\tau_k-1-j}{2}}\right)\\
=&2\gamma^2(M\sigma^2+6M^2\varsigma^2)\bar{\rho}\\
&+\frac{2\gamma^2}{K}\frac{n-1}{n}M^2\sum_{k=0}^{K-1}\sum_{j=0}^{k-\tau_k-1}\left(12L^2\mathbb{E}\hat{M}_j+2\mathbb{E}\left\|\sum_{i=1}^n p_i\nabla f_i(\hat{x}_j^i)\right\|^2\right)\times\\
&\qquad\left(\rho^{\max\{k-\tau_k-1-j,0\}}+2(\max\{k-\tau_k-1-j,0\})\rho^{\frac{\max\{k-\tau_k-1-j,0\}}{2}}\right)\\
\leqslant&2\gamma^2(M\sigma^2+6M^2\varsigma^2)\bar{\rho}\\
&+\frac{2\gamma^2}{K}\frac{n-1}{n}M^2\sum_{j=0}^{K-1}\sum_{k=j+1}^{\infty}\left(12L^2\mathbb{E}\hat{M}_j+2\mathbb{E}\left\|\sum_{i=1}^n p_i\nabla f_i(\hat{x}_j^i)\right\|^2\right)\times\\
&\qquad\left(\rho^{\max\{k-\tau_k-1-j,0\}}+2(\max\{k-\tau_k-1-j,0\})\rho^{\frac{\max\{k-\tau_k-1-j,0\}}{2}}\right)\\
\leqslant&2\gamma^2(M\sigma^2+6M^2\varsigma^2)\bar{\rho}\\
&+\frac{2\gamma^2}{K}\frac{n-1}{n}M^2\sum_{j=0}^{K-1}\left(12L^2\mathbb{E}\hat{M}_j+2\mathbb{E}\left\|\sum_{i=1}^n p_i\nabla f_i(\hat{x}_j^i)\right\|^2\right)\left(T+\sum_{h=0}^{\infty}\left(\rho^h+2h\rho^{\frac{h}{2}}\right)\right)\\
\leqslant&2\gamma^2(M\sigma^2+6M^2\varsigma^2)\bar{\rho}\\
&+\frac{2\gamma^2}{K}M^2\left(T\frac{n-1}{n}+\bar{\rho}\right)\sum_{j=0}^{K-1}\left(12L^2\mathbb{E}\hat{M}_j+2\mathbb{E}\left\|\sum_{i=1}^n p_i\nabla f_i(\hat{x}_{j,i})\right\|^2\right)\\
\leqslant&2\gamma^2(M\sigma^2+6M^2\varsigma^2)\bar{\rho}\\
&+\frac{4\gamma^2M^2}{K}\left(T\frac{n-1}{n}+\bar{\rho}\right)\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^n p_i\nabla f_i(\hat{x}_k^i)\right\|^2\\
&+\frac{24L^2\gamma^2M^2}{K}\left(T\frac{n-1}{n}+\bar{\rho}\right)\sum_{k=0}^{K-1}\mathbb{E}\hat{M}_k.
\end{aligned}
$$

By rearranging the terms we obtain

$$
\begin{aligned}
&\underbrace{\left(1-24L^2M^2\gamma^2\left(T\frac{n-1}{n}+\bar{\rho}\right)\right)}_{C_1}\frac{\sum_{k=0}^{K-1}\mathbb{E}\hat{M}_k}{K}\\
\leqslant&2\gamma^2(M\sigma^2+6M^2\varsigma^2)\bar{\rho}+\frac{4\gamma^2M^2}{K}\left(T\frac{n-1}{n}+\bar{\rho}\right)\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^n p_i\nabla f_i(\hat{x}_k^i)\right\|^2,
\end{aligned}
$$

we complete the proof. □

Lemma 8. *For all $k\geq 0$ we have*

$$
\mathbb{E}\left\|\frac{X_k\mathbf{1}_n-\hat{X}_k\mathbf{1}_n}{n}\right\|^2\leqslant\frac{\tau_k^2\gamma^2\sigma^2M}{n^2}+\tau_k\gamma^2\sum_{t=1}^{\tau_k}\left(\frac{M^2}{n^2}\sum_{i=1}^n p_i\mathbb{E}\|\nabla f_i(\hat{x}_{k-t}^i)\|^2\right).
$$

Proof.

$$
\begin{aligned}
\mathbb{E}\left\|\frac{X_k\mathbf{1}_n-\hat{X}_k\mathbf{1}_n}{n}\right\|^2&\overset{\text{Assumption 1-7}}{=}\mathbb{E}\left\|\frac{\sum_{t=1}^{\tau_k}\gamma\partial g(\hat{X}_{k-t};\xi_{k-t}^{i_{k-t}},i_{k-t})\mathbf{1}_n}{n}\right\|^2\\
&\leqslant\tau_k\sum_{t=1}^{\tau_k}\gamma^2\mathbb{E}\left\|\frac{\partial g(\hat{X}_{k-t};\xi_{k-t}^{i_{k-t}},i_{k-t})\mathbf{1}_n}{n}\right\|^2\\
&\overset{\text{Lemma 4}}{\leqslant}\tau_k\sum_{t=1}^{\tau_k}\gamma^2\left(\frac{\sigma^2M}{n^2}+\frac{M^2}{n^2}\sum_{i=1}^n p_i\mathbb{E}\left\|\nabla f_i(\hat{x}_{k-t}^i)\right\|^2\right),
\end{aligned}
$$

where the first step comes from any $n\times n$ doubly stochastic matrix multiplied by $\mathbf{1}_n$ equals $\mathbf{1}_n$ and Assumption 1-7. □

Proof to Corollary 2. To prove this result, we will apply Theorem 1. We first verify that all conditions can be satisfied in Theorem 1.

First $C_1 > 0$ can be satisfied by a stronger condition $C_1 \geq 1/2$ which can be satisfied by $\gamma \leqslant \frac{1}{4\sqrt{6}ML}\left(T\frac{n-1}{n}+\bar{\rho}\right)^{-1/2}$. Second $C_3 \leq 1$ can be satisfied by :

$$\gamma \leqslant \min\left\{\frac{n}{8MT^2L}, \frac{1}{8\sqrt{3}LM}\bar{\rho}^{-1/2}, \frac{1}{32nML}\bar{\rho}^{-1}, \frac{n^{1/3}}{8\sqrt{6}MLT^{2/3}}\bar{\rho}^{-1/3}\right\}$$

and $C_1 \geq 1/2$, which can be seen from

$$\begin{aligned} C_3 =& \frac{1}{2} + \frac{2\left(6\gamma^2L^2M^2 + \gamma nML + \frac{12M^3L^3T^2\gamma^3}{n}\right)\bar{\rho}}{C_1} + \frac{LT^2\gamma M}{n} \\ \overset{C_1\geqslant\frac{1}{2}}{\leqslant}& \frac{1}{2} + 24\gamma^2L^2M^2 + 4\gamma nML + \frac{48M^3L^3T^2\gamma^3}{n}\bar{\rho} + \frac{LT^2\gamma M}{n}. \end{aligned}$$

The requirements on γ are given by making each of the last four terms smaller than $1/8$:

$$\begin{aligned} \frac{LT^2\gamma M}{n} \leqslant& \frac{1}{8} \Longleftarrow \gamma \leqslant \frac{n}{8MT^2L}, \\ 24\gamma^2L^2M^2\bar{\rho} \leqslant& \frac{1}{8} \Longleftarrow \gamma \leqslant \frac{1}{8\sqrt{3}LM}\bar{\rho}^{-1/2}, \\ 4\gamma nML\bar{\rho} \leqslant& \frac{1}{8} \Longleftarrow \gamma \leqslant \frac{1}{32nML}\bar{\rho}^{-1}, \end{aligned}$$

and

$$\frac{48M^3L^3T^2\gamma^3}{n}\bar{\rho} \leqslant \frac{1}{8} \Longleftarrow \gamma \leqslant \frac{n^{1/3}}{8\sqrt{6}MLT^{2/3}}\bar{\rho}^{-1/3}.$$

Third $C_2 \geqslant 0$ can be satisfied by

$$\gamma \leq \min\left\{\frac{n}{10LM}, \frac{n}{2\sqrt{5}MLT}, \frac{n^{1/3}}{8ML}\left(T\frac{n-1}{n}+\bar{\rho}\right)^{-1/3}, \frac{1}{4\sqrt{5}LM}\left(T\frac{n-1}{n}+\bar{\rho}\right)^{-1/2}, \frac{n^{1/2}}{6MLT^{1/2}}\left(T\frac{n-1}{n}+\bar{\rho}\right)^{-1/4}\right\}$$

and $C_1 \geq 1/2$, which can be seen from

$$\begin{aligned} C_2 :=& \frac{\gamma M}{2n} - \frac{\gamma^2LM^2}{n^2} - \frac{2M^3L^2T^2\gamma^3}{n^3} - \left(\frac{6\gamma^2L^3M^2}{n^2} + \frac{\gamma M}{n}L^2 + \frac{12M^3L^4T^2\gamma^3}{n^3}\right)\frac{4M^2\gamma^2(T\frac{n-1}{n}+\bar{\rho})}{C_1} \geqslant 0 \\ \overset{C_1\geqslant\frac{1}{2}}{\Longleftarrow} 1 \geqslant& \frac{2\gamma LM}{n} + \frac{4M^2L^2T^2\gamma^2}{n^2} + \frac{96\gamma L^3M}{n} + 16L^2 + \frac{192M^2L^4T^2\gamma^2}{n^2}M^2\gamma^2(T\frac{n-1}{n}+\bar{\rho}). \end{aligned}$$

The last inequality is satisfied given the requirements on γ because each term on the RHS is bounded by $1/5$:

$$\begin{aligned} \frac{2\gamma LM}{n} \leqslant& \frac{1}{5} \Longleftarrow \gamma \leqslant \frac{n}{10LM}, \\ \frac{4M^2L^2T^2\gamma^2}{n^2} \leqslant& \frac{1}{5} \Longleftarrow \gamma \leqslant \frac{n}{2\sqrt{5}MLT}, \\ \frac{96\gamma L^3M}{n}M^2\gamma^2(T\frac{n-1}{n}+\bar{\rho}) \leqslant& \frac{1}{5} \Longleftarrow \gamma \leqslant \frac{n^{1/3}}{8ML}(T\frac{n-1}{n}+\bar{\rho})^{-1/3}, \\ 16L^2M^2\gamma^2(T\frac{n-1}{n}+\bar{\rho}) \leqslant& \frac{1}{5} \Longleftarrow \gamma \leqslant \frac{1}{4\sqrt{5}LM}(T\frac{n-1}{n}+\bar{\rho})^{-1/2}, \\ \frac{192M^2L^4T^2\gamma^2}{n^2}M^2\gamma^2(T\frac{n-1}{n}+\bar{\rho}) \leqslant& \frac{1}{5} \Longleftarrow \gamma \leqslant \frac{n^{1/2}}{6MLT^{1/2}}(T\frac{n-1}{n}+\bar{\rho})^{-1/4}. \end{aligned}$$

Combining all above the requirements on γ to satisfy $C_1 \geq 1/2, C_2 \geq 0$ and $C_3 \leq 1$ are

$$\gamma \leqslant \frac{1}{ML}\min\left\{\begin{array}{c} \frac{1}{4\sqrt{6}}\left(T\frac{n-1}{n}+\bar{\rho}\right)^{-1/2}, \frac{n}{8T^2}, \frac{1}{8\sqrt{3}}\bar{\rho}^{-1/2}, \\ \frac{1}{32n}\bar{\rho}^{-1}, \frac{n^{1/3}}{8\sqrt{6}T^{2/3}}\bar{\rho}^{-1/3}, \\ \frac{n}{10}, \frac{n}{2\sqrt{5}T}, \frac{n^{1/3}}{8}\left(T\frac{n-1}{n}+\bar{\rho}\right)^{-1/3}, \\ \frac{1}{4\sqrt{5}}\left(T\frac{n-1}{n}+\bar{\rho}\right)^{-1/2}, \frac{n^{1/2}}{6T^{1/2}}\left(T\frac{n-1}{n}+\bar{\rho}\right)^{-1/4} \end{array}\right\}.$$

Note that the RHS is larger than

$$U := \frac{1}{ML}\min\left\{\begin{array}{c} \frac{1}{8\sqrt{3}\sqrt{T\frac{n-1}{n}+\bar{\rho}}}, \frac{n}{8T^2}, \frac{1}{32n\bar{\rho}}, \frac{n}{10}, \frac{n^{1/12}(n-1)^{-1/4}}{\left(8\sqrt{6}T^{2/3}+8\right)\left(T+\bar{\rho}\frac{n}{n-1}\right)^{1/3}} \end{array}\right\}.$$

Let $\gamma = \frac{n}{10ML+\sqrt{\sigma^2+6M\varsigma^2}\sqrt{KM}}$ then if $\gamma \leqslant U$ we will have $C_1 \geq 1/2, C_2 \geq 0$ and $C_3 \leq 1$. Further investigation gives us

$$\begin{aligned}
\gamma = \frac{n}{10ML+\sqrt{\sigma^2+6M\varsigma^2}\sqrt{KM}} \leqslant & \frac{1}{ML}\min\left\{\begin{array}{c}\frac{1}{8\sqrt{3}\sqrt{T\frac{n-1}{n}+\bar{\rho}}}, \frac{n}{8T^2}, \frac{1}{32n\bar{\rho}}, \\ \frac{n}{10}, \frac{n^{1/12}(n-1)^{-1/4}}{\left(8\sqrt{6}T^{2/3}+8\right)\left(T+\bar{\rho}\frac{n}{n-1}\right)^{1/3}}\end{array}\right\} \\
\Longleftarrow 10ML+\sqrt{\sigma^2+6M\varsigma^2}\sqrt{KM} \geqslant & nML\max\left\{\begin{array}{c}8\sqrt{3}\sqrt{T\frac{n-1}{n}+\bar{\rho}}, \frac{8T^2}{n}, 32n\bar{\rho}, \\ \frac{\left(8\sqrt{6}T^{2/3}+8\right)\left(T+\bar{\rho}\frac{n}{n-1}\right)^{1/3}}{n^{1/12}(n-1)^{-1/4}}\end{array}\right\} \\
\Longleftarrow K \geqslant & \frac{ML^2n^2}{\sigma^2+6M\varsigma^2}\max\left\{\begin{array}{c}192\left(T\frac{n-1}{n}+\bar{\rho}\right), \frac{64T^4}{n^2}, 1024n^2\bar{\rho}^2, \\ \frac{\left(8\sqrt{6}T^{2/3}+8\right)^2\left(T+\bar{\rho}\frac{n}{n-1}\right)^{2/3}}{n^{1/6}(n-1)^{-1/2}}\end{array}\right\}.
\end{aligned}$$

It follows from Theorem 1 that if the last inequality is sastisfied and $\gamma = \frac{n}{10ML+\sqrt{\sigma^2+6M\varsigma^2}\sqrt{KM}}$, we have

$$\begin{aligned}
\frac{\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\frac{X_k\mathbf{1}_n}{n}\right)\right\|^2}{K} \leqslant & \frac{2(\mathbb{E}f(x_0)-f^*)n}{\gamma KM}+\frac{2\gamma L}{Mn}(M\sigma^2+6M^2\varsigma^2) \\
= & \frac{20(\mathbb{E}f(x_0)-f^*)L}{K}+\frac{2(\mathbb{E}f(x_0)-f^*)\sqrt{\sigma^2+6M\varsigma^2}}{\sqrt{KM}} \\
& +\frac{2L}{M\left(10ML+\sqrt{\sigma^2+6M\varsigma^2}\sqrt{KM}\right)}(M\sigma^2+6M^2\varsigma^2) \\
\leqslant & \frac{20(\mathbb{E}f(x_0)-f^*)L}{K}+\frac{2(\mathbb{E}f(x_0)-f^*+L)\sqrt{\sigma^2+6M\varsigma^2}}{\sqrt{KM}}.
\end{aligned}$$

This completes the proof. □