Proofs on the Convergence Property of Partial Reduce

February 19, 2021

Theorem 1 (**Convergence of Partial Reduce**). *We assume the bound of gradient variance σ^2 is in inverse proportion to the mini-batch size M. We define $\bar{\rho} = \frac{\rho}{1-\rho} + \frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2}$. For partial reduce with P, under Assumptions 1–2, if the learning rate satisfies*

$$\eta L + \frac{2N^3\eta^2\bar{\rho}}{P^2} \leq 1, \tag{1}$$

where $\eta = \frac{P}{N}\gamma$, and all local models are initialized at a same point $\mathbf{u}_1$, then the average-squared gradient norm after K iterations is bounded as follows

$$\mathbb{E}\Big[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\mathbf{u}_k)\|^2\Big] \leq \underbrace{\frac{2[F(\mathbf{u}_1) - F_{inf}]}{\eta K} + \frac{\eta L\sigma^2}{P}}_{\text{SGD error}} + \underbrace{\frac{2\eta^2L^2\sigma^2N^3\bar{\rho}}{P^2}}_{\text{network error}} \tag{2}$$

A Proofs on the Convergence Property of Partial Reduce

A.1 Connection with Cooperative SGD

Cooperative SGD framework [2] is a powerful framework that enables to obtain an integrated analysis of communication-efficient algorithms, such as periodic averaging SGD and decentralized SGD. The cooperative SGD algorithm is denoted by $\mathcal{A}(\tau, \mathbf{W}, v)$, where τ is the number of local updates, $\mathbf{W}$ is the model average matrix, and v is the number of auxiliary variables. At iteration k, the m workers have different local models $\mathbf{x}_k^1 \dots \mathbf{x}_k^m$. In addition, there are v auxiliary variables $\mathbf{z}_k^1 \dots \mathbf{z}_k^v$ that are stored at other v workers. In each iteration, the m workers evaluate the gradient $\mathbf{g}(\mathbf{x}_k^i)$ for one minibatch of data and update $\mathbf{x}_k^i$. However, the gradients for auxiliary variables are zero, i.e., $\mathbf{g}(\mathbf{x}_k^i) = \mathbf{0}, \forall j \in \{1 \dots v\}$. In iteration k, the local models and auxiliary variables are averaged according to mixing matrix $\mathbf{W_k}$. Define matrices $\mathbf{X}_k$ and $\mathbf{G}_k$ that concatenate all local models and gradients of $m + v$ workers, a general update rule in the framework is:

$$\mathbf{X_{k+1}} = (\mathbf{X_k} - \gamma\mathbf{G}_k)\mathbf{W}_k. \tag{3}$$

We notice that our proposed partial reduce algorithm's update rule can be connected to the rule in the equation (3) of cooperative SGD. Specifically, given a total of N workers with P-reduce operations, we treat $m = P$ and $m + v = N$. In iteration k, we treat P workers in a selected worker group $\mathcal{S}_k$ as workers perform local update and model average, whereas the local models of other $N - P$ workers out of $\mathcal{S}_k$ as auxiliary variables in this iteration.

For the convergence analysis, most assumptions made by cooperative SGD framework and our P-reduce are similar except the spectral gap condition. In particular, cooperative SGD framework assumes a fixed model averaging matrix $\mathbf{W}$ satisfying:

$$\max\{|\lambda_2(\mathbf{W})|, |\lambda_N(\mathbf{W})|\} < \zeta,\ \zeta^1 \in [0, 1) \tag{4}$$

By contrast, the model averaging matrix $\mathbf{W}_k$ in P-reduce algorithm can be different over iterations, depending on the dynamic membership in worker group $\mathcal{S}_k$, and the $\mathbb{E}[\mathbf{W}_k]$ satisfies the spectral gap condition of inequality (8). Clearly, the condition (4) is a special case of condition (8). Therefore, we prove our theorem by extending the cooperative SGD framework to enable variable model-average matrix $\mathbf{W}_k$.

A.2 Proof Preliminaries

Before our proof, we introduce the necessary notations, assumptions and supporting lemmas for Theorem 1.

A.2.1 Notations

The notations used in the proof are listed below.

[1] We use the notation ρ in our paper and proofs.

Number of workers	N
Size work group in partial reduce	P
Total iterations	K
Mixing matrix	$\mathbf{W}$
Spectral gap	ρ
Learning rate	γ,η
Lipschitz constant	L
Variance bounds for stochastic gradients	β, σ^2
Euclidean or vector norm	$\|\cdot\|$
Frobenius norm	$\|\cdot\|_{\mathrm{F}}$

Table 1: List of notations.

A.2.2 Assumptions

We recall the commonly used assumptions for the SGD convergence analysis [1], which are also used in cooperative SGD framework:
Assumption 1.

(1) **Lipschitzian gradient:** All functions $f_i(\cdot)$'s are smooth with L-Lipschitzian gradients.

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leqslant L\|\mathbf{x} - \mathbf{y}\| \tag{5}$$

(2) **Unbiased estimation:**

$$\mathbb{E}_{\xi|\mathbf{x}}[g(\mathbf{x})] = \nabla F(\mathbf{x}) \tag{6}$$

(3) **Bounded variance**[2]**:** Assume the variance of the stochastic gradient is bounded:

$$\mathbb{E}_{\xi|\mathbf{x}}[g(\mathbf{x}) - \nabla F(\mathbf{x})] \leq \beta \left\|\nabla F(\mathbf{x})\right\|^2 + \sigma^2 \tag{7}$$

Further, the model averaging with $\mathbf{W}_k$ in P-reduce algorithm satisfies the the following assumptions:
Assumption 2.

(1) **Stochastic averaging:** $\mathbf{W}_k$ is arbitrary doubly stochastic for all k and $\mathbf{W}_k = \mathbf{W}_k^\top$.
(2) **Dependence of random variables:** $\mathbf{W}_k$ is a random variable independent on ξ_k and k.
(3) **Spectral gap:** There exists a $\rho \in [0, 1)$ such that

$$\max\{|\lambda_2(\mathbb{E}[\mathbf{W}_k])|, |\lambda_n(\mathbb{E}[\mathbf{W}_k])|\} \leq \rho, \forall k. \tag{8}$$

A.2.3 Supporting Lemmas

Lemma 1. *The Forbenius norm defined for $A \in M_n$ by*

$$\|\mathbf{A}\|_F^2 = \|Tr(\mathbf{A}\mathbf{A}^\top)| = \sum_{i=1}^{n}\sum_{j=1}^{n}|a_{i,j}|^2 = \sum_{i=1}^{n}\left\|\mathbf{A}^{(i)}\right\|^2 \tag{9}$$

Lemma 2. *Under Assumption 1 (3), we have the following bound for the stochastic gradient:*

$$\mathbb{E}\left\|g(\mathbf{X}_k) - \nabla F(\mathbf{X}_k)\right\|^2 \leq \beta\left\|\nabla F(\mathbf{X}_k)\right\|^2 + P\sigma^2 \tag{10}$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left\|g(\mathbf{X}_k) - \nabla F(\mathbf{X}_k)\right\|^2 && (11)\\
\leq&\mathbb{E}\left\|g(\mathbf{X}_k) - \nabla F(\mathbf{X}_k)\right\|_{\mathrm{F}}^2 && (12)\\
=&\mathbb{E}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|g(\mathbf{x}_k^{(i)}) - \nabla F(\mathbf{x}_k^{(i)})\right\|^2 && (13)\\
\leq&\mathbb{E}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left[\beta\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 + \sigma^2\right] && (14)\\
=&\beta\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 + P\sigma^2 && (15)\\
=&\beta\left\|\nabla F(\mathbf{X}_k)\right\|_{\mathrm{F}}^2 + P\sigma^2 && (16)
\end{aligned}$$

where (12) comes from $\|\mathbf{A}\| \leq \|\mathbf{A}\|_{\mathrm{F}}$, (13) and (16) comes from Lemma 1, (14) follows Assumption 2 (3). □

[2]As stated in [2], constant β only influences the constraint on the learning rate () and will not appear in the expression of the gradient norm upper bound (). In order to get neater results, β is set as 0 in the main paper. The proofs are valid for arbitrary β.

Lemma 3. *Suppose there is a sequence of $N \times N$ matrices $\{\mathbf{W}_l\}_{l=s}^{k-1}$, $0 \leq s < k$ and each $\mathbf{W}_l$ satisfies Assumption 2. Then*

$$\mathbb{E}\left\|(\prod_{l=s}^{k-1}\mathbf{W}_l - \mathbf{J})\mathbf{e}_i\right\|^2 \leq \rho^{k-s} \tag{17}$$

where $\mathbf{J} = \mathbf{1}\mathbf{1}^\top/(\mathbf{1}^\top\mathbf{1})$, $\mathbf{e}_i$ is the standard basis vector and $\max\{|\lambda_2(\mathbb{E}[\mathbf{W}_l])|, |\lambda_N(\mathbb{E}[\mathbf{W}_l])|\} \leq \rho, \forall l$.

Proof. Let $\mathbf{y}_{k-1} = (\prod_{l=s}^{k-1}\mathbf{W}_l - \mathbf{J})\mathbf{e}_i$. Then noting that $\mathbf{y}_{k-1} = \mathbf{W}_{k-1}\mathbf{y}_{k-2}$ we have

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{y}_{k-1}\right\|^2 &= \mathbb{E}\left\|\mathbf{W}_{k-1}\mathbf{y}_{k-2}\right\|^2 &(18)\\
&= \mathbb{E}\langle \mathbf{W}_{k-1}\mathbf{y}_{k-2}, \mathbf{W}_{k-1}\mathbf{y}_{k-2}\rangle &(19)\\
&= \mathbb{E}\langle \mathbf{y}_{k-2}, \mathbf{W}_{k-1}^\top\mathbf{W}_{k-1}\mathbf{y}_{k-2}\rangle &(20)\\
&= \mathbb{E}\langle \mathbf{y}_{k-2}, \mathbb{E}(\mathbf{W}_{k-1}^\top\mathbf{W}_{k-1})\mathbf{y}_{k-2}\rangle. &(21)
\end{aligned}$$

We find that $\mathbf{W}_k^\top\mathbf{W}_k = \mathbf{W}_k$ according to the definition of $\mathbf{W}_k$, i.e., the (i,j)-th element in $\mathbf{W}_k^\top\mathbf{W}_k$ could also be calculated by:

$$\sum_{l=1}^{N}\mathbf{W}_k^\top(i,l)\mathbf{W}_k(l,j) = \sum_{l=1}^{N}\mathbf{W}_k(l,i)\mathbf{W}_k(l,j) = \begin{cases} 1/P, & \text{if workers } i,j \in \mathcal{S}_k,\\ 1, & \text{if worker } i \notin \mathcal{S}_k \text{ and } i = j,\\ 0, & \text{otherwise.}\end{cases} \tag{22}$$

Specifically, if workers $i, j \in \mathcal{S}_k$, $\mathbf{W}_k(l,i) = \mathbf{W}_k(l,j) = 1/P$ when $l \in \mathcal{S}_k$, otherwise $\mathbf{W}_k(l,i) = \mathbf{W}_k(l,j) = 0$. We have $\sum_{l=1}^{N}\mathbf{W}_k(l,i)\mathbf{W}_k(l,j) = \sum_{l=1}^{P}1/P^2 = 1/P$. If workers $i \notin \mathcal{S}_k$ and $i = j$, $\mathbf{W}_k(l,i) = \mathbf{W}_k(l,j) = 1$ when $l = i$, otherwise $\mathbf{W}_k(l,i) = \mathbf{W}_k(l,j) = 0$. We have $\sum_{l=1}^{N}\mathbf{W}_k(l,i)\mathbf{W}_k(l,j) = 1$. The result is consistent with the definition of $\mathbf{W}_k(i,j)$.

Since $\mathbf{W}_k^\top\mathbf{W}_k = \mathbf{W}_k$, the spectral gap condition Assumption 2 (2) can be rewritten as:

$$\max\{|\lambda_2(\mathbb{E}[\mathbf{W}_k^\top\mathbf{W}_k])|, |\lambda_N(\mathbb{E}[\mathbf{W}_k^\top\mathbf{W}_k])|\} \leq \rho, \forall k. \tag{23}$$

Note that $\mathbb{E}(\mathbf{W}_{k-1}^\top\mathbf{W}_{k-1})$ is symmetric and doubly stochastic and $\mathbf{1}_n$ is an an eigenvector of $\mathbb{E}(\mathbf{W}_{k-1}^\top\mathbf{W}_{k-1})$ with eigenvalue 1. Starting from $\mathbf{1}_n$ we construct a basis of $\mathbb{R}^n$ composed by the eigenvectors of $\mathbb{E}(\mathbf{W}_{k-1}^\top\mathbf{W}_{k-1})$, which is guaranteed to exist by the spectral theorem of Hermitian matrices. From (23) the magnitude of all other eigenvectors' associated eigenvalues should be smaller or equal to ρ. We decompose $\mathbf{y}_K$ using this constructed basis and it follows that

$$\mathbb{E}\|\mathbf{y}_{K-1}\|^2 \leqslant \rho\mathbb{E}\|\mathbf{y}_{K-2}\|^2.$$

Since $\mathbf{y}_s = (\mathbf{W}_s - \mathbf{J})\mathbf{e}_i = \mathbf{W}_s(\mathbf{I} - \mathbf{J})\mathbf{e}_i$, we define $\mathbf{y}_{s-1} = (\mathbf{I} - \mathbf{J})\mathbf{e}_i$ to make $\mathbf{y}_s = \mathbf{W}_s\mathbf{y}_{s-1}$. Noting that $\|\mathbf{y}_{s-1}\|^2 = \|(\mathbf{I} - \mathbf{J})\mathbf{e}_i\|^2 = \frac{(N-1)^2}{N^2} + \sum_{i=1}^{N-1}\frac{1}{N^2} = \frac{N^2-2N+1+N-1}{N^2} = \frac{N-1}{N}$, by induction, we complete the proof.

This lemma further explains the implication of the spectral gap Assumption 2 (3). Notice that the continued multiplication of W_l represents for multiple model averaging steps through our partial reduce. And the matrix $\mathbf{J}$ is the mode average matrix of fully synchronous SGD, where the local models are synchronized with all other workers (e.g., via AllReduce) after every iteration. So the provided Lemma 3 demonstrates the upper bound of the model transition difference between our partial reduce and AllReduce. □

A.3 Proof of Theorem 1

Before providing the proof of Theorem 1, we prefer to first present an important lemma that describes the basic convergence upper bound framework.

Lemma 4. *In partial reduce, under Assumption 1, the average-squared gradient norm after K iterations is bounded as follows*

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\mathbf{u}_k)\|^2\right] \leq & \frac{2\left[F(\mathbf{u}_1) - F_{inf}\right]}{\eta K} + \frac{\eta L\sigma^2}{P} + \frac{L^2}{KP}\sum_{k=1}^{K}\|\mathbf{X}_k(\mathbf{I} - \mathbf{J})\|_F^2 - \\
& \left[1 - \eta L\left(\frac{\beta}{P} + 1\right)\right]\frac{1}{K}\sum_{k=1}^{K}\frac{\mathbb{E}\left\|\nabla F(\mathbf{X}_k)\right\|_F^2}{P}.
\end{aligned} \tag{24}$$

Since the proof of Lemma 4 only relies on commonly used Assumption 1, our result is consistent with the cooperative SGD's intermediate lemma [2]. We remove the proof details of Lemma 4 to Sec. A.4, and focus on our extension on the cooperative SGD (under more generalized Assumption 2) below. Note that our Theorem 1 follows the same SGD error (i.e., the first two items) as Eq. (24). Our goal is to provide an upper bound for the left network error items, which make the proofs of our P-reduce different from the cooperative SGD framework.

A.3.1 Decomposition.

To provide an upper bound for the term $\frac{L^2}{KP}\sum_{k=1}^{K}\|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2$, we derive a specific expression for $\mathbf{X}_k(\mathbf{I}-\mathbf{J})$. According to the update rule (3), one can observe that

$$\begin{aligned}\mathbf{X}_k(\mathbf{I}-\mathbf{J}) =&(\mathbf{X}_{k-1}-\gamma\mathbf{G}_{k-1})\,\mathbf{W}_{k-1}(\mathbf{I}-\mathbf{J}) &(25)\\ =&\mathbf{X}_{k-1}(\mathbf{I}-\mathbf{J})\mathbf{W}_{k-1}-\gamma\mathbf{G}_{k-1}(\mathbf{W}_{k-1}-\mathbf{J}) &(26)\end{aligned}$$

where (26) follows the special property of doubly stochastic matrix: $\mathbf{W}_{k-1}\mathbf{J}=\mathbf{J}\mathbf{W}_{k-1}=\mathbf{J}$ and hence $(\mathbf{I}-\mathbf{J})\mathbf{W}_{k-1}=\mathbf{W}_{k-1}(\mathbf{I}-\mathbf{J})$. Then, expanding the expression of $\mathbf{X}_{k-1}$, we have

$$\begin{aligned}\mathbf{X}_k(\mathbf{I}-\mathbf{J}) =&[\mathbf{X}_{k-2}(\mathbf{I}-\mathbf{J})\mathbf{W}_{k-2}-\gamma\mathbf{G}_{k-2}(\mathbf{W}_{k-2}-\mathbf{J})]\,\mathbf{W}_{k-1}-\gamma\mathbf{G}_{k-1}(\mathbf{W}_{k-1}-\mathbf{J}) &(27)\\ =&\mathbf{X}_{k-2}(\mathbf{I}-\mathbf{J})\mathbf{W}_{k-2}\mathbf{W}_{k-1}-\gamma\mathbf{G}_{k-2}(\mathbf{W}_{k-2}\mathbf{W}_{k-1}-\mathbf{J})-\gamma\mathbf{G}_{k-1}(\mathbf{W}_{k-1}-\mathbf{J}) &(28)\end{aligned}$$

Repeating the same procedure for $\mathbf{X}_{k-2},\mathbf{X}_{k-3},\ldots,\mathbf{X}_2$, finally we get

$$\mathbf{X}_k(\mathbf{I}-\mathbf{J}) =\mathbf{X}_1(\mathbf{I}-\mathbf{J})\mathbf{\Phi}_{1,k-1}-\gamma\sum_{s=1}^{k-1}\mathbf{G}_s(\mathbf{\Phi}_{s,k-1}-\mathbf{J}) \tag{29}$$

where $\mathbf{\Phi}_{s,k-1}=\prod_{l=s}^{k-1}\mathbf{W}_l$. Since all optimization variables are initialized at the same point $\mathbf{X}_1(\mathbf{I}-\mathbf{J})=0$, the squared norm of the network error term can be directly written as

$$\mathbb{E}\,\|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2 =\gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1}\mathbf{G}_s(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_{\mathrm{F}}^2. \tag{30}$$

Note that the network error term can be decomposed into two parts:

$$\begin{aligned}\mathbb{E}\,\|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2 =&\gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1}\mathbf{G}_s(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_{\mathrm{F}}^2 &(31)\\ =&\gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1}(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J})+\sum_{s=1}^{k-1}\mathbf{Q}_s(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_{\mathrm{F}}^2 &(32)\\ \le&\underbrace{2\gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1}(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_{\mathrm{F}}^2}_{T_1}+\underbrace{2\gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1}\mathbf{Q}_s(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_{\mathrm{F}}^2}_{T_2} &(33)\end{aligned}$$

where $\mathbf{Q}_s=\nabla F(\mathbf{X}_s)$, (33) follows $\|a+b\|^2\le 2\|a\|^2+2\|b\|^2$. Next, we are going to separately provide bounds for T_1 and T_2. Recall that we are interested in the average of all iterates $\frac{L^2}{KP}\sum_{k=1}^{K}\|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2$. Accordingly, we will also derive the bounds for $\frac{L^2}{KP}\sum_{k=1}^{K}T_1$ and $\frac{L^2}{KP}\sum_{k=1}^{K}T_2$.

A.3.2 Bounding T_1.

For the first term T_1, we have

$$\begin{aligned}T_1 =&2\gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1}(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_{\mathrm{F}}^2 &(34)\\ =&2\gamma^2\mathbb{E}\sum_{i=1}^{N}\left\|\sum_{s=1}^{k-1}(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\mathbf{e_i}\right\|^2 &(35)\\ =&2\gamma^2\mathbb{E}\sum_{i=1}^{N}\left[\underbrace{\sum_{s=1}^{k-1}\|(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\mathbf{e_i}\|^2}_{A_1}+\underbrace{2\sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1}\langle(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\mathbf{e_i},(\mathbf{G}_l-\mathbf{Q}_l)(\mathbf{\Phi}_{l,k-1}-\mathbf{J})\mathbf{e_i}\rangle}_{A_2}\right] &(36)\end{aligned}$$

where (35) comes from Lemma 1 and (36) follows $(\sum_{i=1}^{N}a_i)^2=\sum_{i=1}^{N}a_i^2+2\sum_{i=1}^{N}\sum_{j=i+1}^{N}a_ia_j$.

A_1 can be bounded by:

$$A_1 = \sum_{s=1}^{k-1} \left\|(\mathbf{G}_s - \mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e_i}\right\|^2 \tag{37}$$

$$\leq \sum_{s=1}^{k-1} \left\|\mathbf{G}_s - \mathbf{Q}_s\right\|^2 \left\|(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e_i}\right\|^2 \tag{38}$$

$$\leq \sum_{s=1}^{k-1} \left[\beta \left\|\nabla F(\mathbf{X}_s)\right\|_{\mathrm{F}}^2 + P\sigma^2\right] \left\|(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e_i}\right\|^2 \tag{39}$$

$$\leq \sum_{s=1}^{k-1} \left[\beta \left\|\nabla F(\mathbf{X}_s)\right\|_{\mathrm{F}}^2 + P\sigma^2\right] \rho^{k-s} \tag{40}$$

$$\leq P\sigma^2 \frac{\rho}{1-\rho} + \beta \sum_{s=1}^{k-1} \rho^{k-s} \left\|\nabla F(\mathbf{X}_s)\right\|_{\mathrm{F}}^2 \tag{41}$$

where (39) comes from Lemma 2, (40) comes from Lemma 3 and (41) follows the summation formula of power series

$$\sum_{s=1}^{k-1} \rho^{k-s} \leq \sum_{s=-\infty}^{k-1} \rho^{k-s} \leq \frac{\rho}{1-\rho}. \tag{42}$$

The cross items A_2 can be bounded by:

$$A_2 = \sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} \langle (\mathbf{G}_s - \mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e_i}, (\mathbf{G}_l - \mathbf{Q}_l)(\mathbf{\Phi}_{l,k-1} - \mathbf{J})\mathbf{e_i} \rangle \tag{43}$$

$$\leq \sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} \left\|\mathbf{G}_s - \mathbf{Q}_s\right\| \left\|(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e_i}\right\| \left\|\mathbf{G}_l - \mathbf{Q}_l\right\| \left\|(\mathbf{\Phi}_{l,k-1} - \mathbf{J})\mathbf{e_i}\right\| \tag{44}$$

$$\leq \sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} \left[\frac{1}{2c_{s,l}} \left\|(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e_i}\right\|^2 \left\|(\mathbf{\Phi}_{l,k-1} - \mathbf{J})\mathbf{e_i}\right\|^2 + \frac{1}{2/c_{s,l}} \left\|\mathbf{G}_s - \mathbf{Q}_s\right\|^2 \left\|\mathbf{G}_l - \mathbf{Q}_l\right\|^2\right], \forall c_{s,l} > 0 \tag{45}$$

$$\leq \sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} \left[\frac{1}{2c_{s,l}} \rho^{k-s}\rho^{k-l} + \frac{1}{2/c_{s,l}} (\beta \left\|\nabla F(\mathbf{X}_s)\right\|_{\mathrm{F}}^2 + P\sigma^2)(\beta \left\|\nabla F(\mathbf{X}_l)\right\|_{\mathrm{F}}^2 + P\sigma^2)\right], \forall c_{s,l} > 0 \tag{46}$$

$$\leq \sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} (\frac{1}{2c_{s,l}} \rho^{k-s} + \frac{1}{2/c_{s,l}} P^2\sigma^4), \forall c_{s,l} > 0. \tag{47}$$

where (45) follows $ab \leq \frac{1}{2}(a^2/c + cb^2), \forall c > 0$, (46) comes from Lemma 2 and 3. We can choose $c_{s,l} > 0$ and make the term in the last step become $\sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1}(\rho^{(k-s)/2}P\sigma^2)$ (by applying inequality of arithmetic and geometric means). Note that we directly remove the items related with β in (47) for a neater formula. The simplification will not affect the final result since we set $\beta = 0$ at last. Thus

$$A_2 \leq \sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} (\rho^{(k-s)/2} P\sigma^2) \tag{48}$$

$$= \sum_{s=1}^{k-1} (k-s)\rho^{(k-s)/2} P\sigma^2 \leq \frac{\sqrt{\rho}}{(1-\sqrt{\rho})^2} P\sigma^2. \tag{49}$$

where (49) follows the summation formula of power series

$$\sum_{s=1}^{k-1} (k-s)\rho^{(k-s)/2} \leq \sum_{s=-\infty}^{k-1} (k-s)\rho^{(k-s)/2} \leq \frac{\sqrt{\rho}}{(1-\sqrt{\rho})^2}. \tag{50}$$

Substituting (41) (49) back into (36), we have

$$T_1 \leq 2\gamma^2 \mathbb{E} \sum_{i=1}^{N} \left[P\sigma^2 \frac{\rho}{1-\rho} + \beta \sum_{s=1}^{k-1} \rho^{k-s} \left\| \nabla F(\mathbf{X}_s) \right\|_{\mathrm{F}}^2 + \frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2} P\sigma^2 \right] \tag{51}$$

$$= 2NP\gamma^2\sigma^2 \left(\frac{\rho}{1-\rho} + \frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2} \right) + 2N\gamma^2\beta \sum_{s=1}^{k-1} \rho^{k-s} \left\| \nabla F(\mathbf{X}_s) \right\|_{\mathrm{F}}^2 \tag{52}$$

A.3.3 Bounding T_2.

$$T_2 = 2\gamma^2 \mathbb{E} \left\| \sum_{s=1}^{k-1} \mathbf{Q}_s (\mathbf{\Phi}_{s,k-1} - \mathbf{J}) \right\|_{\mathrm{F}}^2 \tag{53}$$

$$= 2\gamma^2 \sum_{i=1}^{N} \mathbb{E} \left\| \sum_{s=1}^{k-1} \mathbf{Q}_s (\mathbf{\Phi}_{s,k-1} - \mathbf{J}) \mathbf{e}_i \right\|^2 \tag{54}$$

$$= 2\gamma^2 \sum_{i=1}^{N} \left[\sum_{s=1}^{k-1} \mathbb{E} \left\| \mathbf{Q}_s (\mathbf{\Phi}_{s,k-1} - \mathbf{J}) \mathbf{e}_i \right\|^2 + 2 \sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} \mathbb{E} \langle \mathbf{Q}_s (\mathbf{\Phi}_{s,k-1} - \mathbf{J}) \mathbf{e}_i, \mathbf{Q}_l (\mathbf{\Phi}_{l,k-1} - \mathbf{J}) \mathbf{e}_i \rangle \right] \tag{55}$$

$$\leq 2\gamma^2 \sum_{i=1}^{N} \left[\sum_{s=1}^{k-1} \rho^{k-s} \mathbb{E} \left\| \nabla F(\mathbf{X}_s) \right\|_{\mathrm{F}}^2 + 2 \underbrace{\sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} \mathbb{E} \left\| \mathbf{Q}_s \right\| \left\| (\mathbf{\Phi}_{s,k-1} - \mathbf{J}) \mathbf{e}_i \right\| \left\| \mathbf{Q}_l \right\| \left\| (\mathbf{\Phi}_{l,k-1} - \mathbf{J}) \mathbf{e}_i \right\|}_{A_3} \right] \tag{56}$$

The cross items A_3 can be bounded by:

$$A_3 = \sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} \mathbb{E} \left\| \mathbf{Q}_s \right\| \left\| (\mathbf{\Phi}_{s,k-1} - \mathbf{J}) \mathbf{e}_i \right\| \left\| \mathbf{Q}_l \right\| \left\| (\mathbf{\Phi}_{l,k-1} - \mathbf{J}) \mathbf{e}_i \right\| \tag{57}$$

$$\leq \sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} \mathbb{E} \left[\frac{1}{2c_{s,l}} \left\| (\mathbf{\Phi}_{s,k-1} - \mathbf{J}) \mathbf{e}_i \right\|^2 \left\| (\mathbf{\Phi}_{l,k-1} - \mathbf{J}) \mathbf{e}_i \right\|^2 + \frac{1}{2/c_{s,l}} \left\| \mathbf{Q}_s \right\|^2 \left\| \mathbf{Q}_l \right\|^2 \right], \forall c_{s,l} > 0 \tag{58}$$

$$\leq \sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} \mathbb{E} \left[\frac{1}{2c_{s,l}} \rho^{k-s} + \frac{1}{2/c_{s,l}} \left\| \nabla F(\mathbf{X}_s) \right\|^2 \left\| \nabla F(\mathbf{X}_l) \right\|^2 \right], \forall c_{s,l} > 0 \tag{59}$$

$$\leq \sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} \mathbb{E} \left[\rho^{(k-s)/2} \left\| \nabla F(\mathbf{X}_s) \right\| \left\| \nabla F(\mathbf{X}_l) \right\| \right] \tag{60}$$

$$\leq \sum_{s=1}^{k-1} \sum_{l=s+1}^{k-1} \mathbb{E} \left[\rho^{(k-s)/2} \frac{\left\| \nabla F(\mathbf{X}_s) \right\|^2 + \left\| \nabla F(\mathbf{X}_l) \right\|^2}{2} \right] \tag{61}$$

$$\leq \sum_{s=1}^{k-1} \left[(k-s) \rho^{(k-s)/2} \mathbb{E} \left\| \nabla F(\mathbf{X}_s) \right\|^2 \right] \tag{62}$$

$$\leq \sum_{s=1}^{k-1} \left[(k-s) \rho^{(k-s)/2} \mathbb{E} \left\| \nabla F(\mathbf{X}_s) \right\|_{\mathrm{F}}^2 \right] \tag{63}$$

We have

$$T_2 \leq 2\gamma^2 \sum_{i=1}^{N} \left[\sum_{s=1}^{k-1} (\rho^{k-s} + 2(k-s)\rho^{(k-s)/2}) \mathbb{E} \left\| \nabla F(\mathbf{X}_s) \right\|_{\mathrm{F}}^2 \right] \tag{64}$$

$$\leq 2N\gamma^2 \left[\sum_{s=1}^{k-1} (\rho^{k-s} + 2(k-s)\rho^{(k-s)/2}) \mathbb{E} \left\| \nabla F(\mathbf{X}_s) \right\|_{\mathrm{F}}^2 \right] \tag{65}$$

We complete the second part.

A.3.4 Final result.

According to (33)(52)(65), setting $\beta = 0$, the network error can be bounded as

$$\frac{1}{KP}\sum_{k=1}^{K}\|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2 \leq \frac{1}{KP}\sum_{k=1}^{K}(T_1+T_2) \tag{66}$$

$$\leq 2\gamma^2\sigma^2 N(\frac{\rho}{1-\rho}+\frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2}) + 2\gamma^2\beta\frac{N}{KP}\sum_{k=1}^{K}\sum_{s=1}^{k-1}\rho^{k-s}\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2 + \frac{2N\gamma^2}{KP}\sum_{k=1}^{K}\sum_{s=1}^{k-1}(\rho^{k-s}+2(k-s)\rho^{(k-s)/2})\mathbb{E}\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2 \tag{67}$$

$$\leq 2\gamma^2\sigma^2 N(\frac{\rho}{1-\rho}+\frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2}) + 2\gamma^2\beta\frac{N}{KP}\sum_{s=1}^{K}\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2\sum_{k=s+1}^{+\infty}\rho^{k-s} + \frac{2N\gamma^2}{KP}\sum_{s=1}^{K}\mathbb{E}\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2\sum_{k=s+1}^{+\infty}(\rho^{k-s}+2(k-s)\rho^{(k-s)/2}) \tag{68}$$

$$\leq 2\gamma^2\sigma^2 N(\frac{\rho}{1-\rho}+\frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2}) + \frac{2N\gamma^2\beta\rho}{1-\rho}\frac{1}{K}\sum_{s=1}^{K}\frac{\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2}{P} + 2\gamma^2 N(\frac{\rho}{1-\rho}+\frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2})\frac{1}{K}\sum_{s=1}^{K}\mathbb{E}\frac{\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2}{P}. \tag{69}$$

Substituting the expression of network error back to inequality (101), we obtain

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla F(\mathbf{u}_k)\|^2 \leq \frac{2(F(\mathbf{x}_1)-F_{\inf})}{\eta K} + \frac{\eta L\sigma^2}{P} + 2\gamma^2L^2\sigma^2N(\frac{\rho}{1-\rho}+\frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2}) - \left[1-\eta L\left(\frac{\beta}{P}+1\right)-2N\gamma^2L^2(\frac{(\beta+1)\rho}{1-\rho}+\frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2})\right]\frac{1}{K}\sum_{k=1}^{K}\frac{\mathbb{E}\|\nabla F(\mathbf{X}_k)\|_{\mathrm{F}}^2}{P}. \tag{70}$$

When the learning rate satisfies

$$1-\eta L\left(\frac{\beta}{P}+1\right)-2N\gamma^2L^2(\frac{(\beta+1)\rho}{1-\rho}+\frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2}) \leq 0, \tag{71}$$

we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla F(\mathbf{u}_k)\|^2 \leq \frac{2(F(\mathbf{x}_1)-F_{\inf})}{\eta K} + \frac{\eta L\sigma^2}{P} + \frac{2\eta^2L^2\sigma^2N^3}{P^2}(\frac{\rho}{1-\rho}+\frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2}) \tag{72}$$

where $\eta = \frac{P}{N}\gamma$. Setting $\beta = 0$, the condition on learning rate (71) can be further simplified as follows:

$$\eta L\left(\frac{\beta}{P}+1\right)+2N\gamma^2(\frac{(\beta+1)\rho}{1-\rho}+\frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2}) \tag{73}$$

$$=\eta L+\frac{2N^3\eta^2}{P^2}(\frac{\rho}{1-\rho}+\frac{2\sqrt{\rho}}{(1-\sqrt{\rho})^2}) \leq 1, \tag{74}$$

Here, we complete the proof.

A.4 Proof of Lemma 4

For the ease of writing, we first define some notations. Let Ξ_k denote the set $\{\xi_k^{(1)},\ldots,\xi_k^{(m)}\}$ of mini-batches at m workers in iteration k. We use notation $\mathbf{E}_k$ to denote the conditional expectation $\mathbb{E}_{\Xi_K|\mathbf{X}_k}$. Besides, define averaged stochastic gradient and averaged full batch gradient in partial group $\mathcal{S}_k$ as follows:

$$\mathcal{G}_k = \frac{1}{P}\sum_{i=1,i\in\mathcal{S}_k}^{N} g(\mathbf{x}_k^{(i)}),\ \mathcal{H}_k = \frac{1}{P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\nabla F(\mathbf{x}_k^{(i)}). \tag{75}$$

A.4.1 Supporting Lemmas

Lemma 5. *Under Assumption 1, we have the following variance bound for the averaged stochastic gradient:*

$$\mathbb{E}_{\Xi_K|\mathbf{X}_k}\left[\|\mathcal{G}_k-\mathcal{H}_k\|^2\right]\leq\frac{\beta}{P^2}\|\nabla F(\mathbf{X}_k)\|_F^2+\frac{\sigma^2}{P}. \tag{76}$$

Proof. According to the definition of $\mathcal{G}_k, \mathcal{H}_k$ (75), we have

$$\mathbb{E}_{\Xi_K|\mathbf{X}_k}\left[\|\mathcal{G}_k-\mathcal{H}_k\|^2\right] \tag{77}$$

$$=\mathbb{E}_{\Xi_K|\mathbf{X}_k}\left\|\frac{1}{P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left[g(\mathbf{x}_k^{(i)})-\nabla F(\mathbf{x}_k^{(i)})\right]\right\|^2 \tag{78}$$

$$=\frac{1}{P^2}\mathbb{E}_{\Xi_K|\mathbf{X}_k}\left[\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|g(\mathbf{x}_k^{(i)})-\nabla F(\mathbf{x}_k^{(i)})\right\|^2+\sum_{j\neq l}\left\langle g(\mathbf{x}_k^{(j)})-\nabla F(\mathbf{x}_k^{(j)}),\, g(\mathbf{x}_k^{(l)})-\nabla F(\mathbf{x}_k^{(l)})\right\rangle\right] \tag{79}$$

$$=\frac{1}{P^2}\sum_{i=1,i\in\mathcal{S}_k}^{N}\mathbb{E}_{\xi_k^{(i)}|\mathbf{X}_k}\left\|g(\mathbf{x}_k^{(i)})-\nabla F(\mathbf{x}_k^{(i)})\right\|^2+\frac{1}{P^2}\sum_{j\neq l}\left\langle\mathbb{E}_{\xi_k^{(j)}|\mathbf{X}_k}\left[g(\mathbf{x}_k^{(j)})-\nabla F(\mathbf{x}_k^{(j)})\right],\,\mathbb{E}_{\xi_k^{(l)}|\mathbf{X}_k}\left[g(\mathbf{x}_k^{(l)})-\nabla F(\mathbf{x}_k^{(l)})\right]\right\rangle \tag{80}$$

where equation (80) is due to $\{\xi_k^{(i)}\}$ are independent random variables. Now, directly applying Assumption 3 and 4 to (80), one can observe that all cross terms are zero. Then, we have

$$\mathbb{E}_{\Xi_K|\mathbf{X}_k}\|\mathcal{G}_k-\mathcal{H}_k\|^2\leq\frac{1}{P^2}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left[\beta\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2+\sigma^2\right] \tag{81}$$

$$=\frac{\beta}{P}\frac{\|\nabla F(\mathbf{X}_k)\|_{\mathrm{F}}^2}{P}+\frac{\sigma^2}{P}. \tag{82}$$

□

Lemma 6. *Under Assumption 1, the expected inner product between stochastic gradient and full batch gradient can be expanded as*

$$\mathbf{E}_k\left[\langle\nabla F(\mathbf{u}_k),\,\mathcal{G}_k\rangle\right]=\frac{1}{2}\|\nabla F(\mathbf{u}_k)\|^2+\frac{1}{2P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2-\frac{1}{2P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{u}_k)-\nabla F(\mathbf{x}_k^{(i)})\right\|^2 \tag{83}$$

where $\mathbf{E}_k$ denotes the conditional expectation $\mathbb{E}_{\Xi_K|\mathbf{X}_k}$.

Proof.

$$\mathbf{E}_k\left[\langle\nabla F(\mathbf{u}_k),\,\mathcal{G}_k\rangle\right]=\mathbf{E}_k\left[\left\langle\nabla F(\mathbf{u}_k),\,\frac{1}{P}\sum_{i=1,i\in\mathcal{S}_k}^{N}g(\mathbf{x}_k^{(i)})\right\rangle\right] \tag{84}$$

$$=\frac{1}{P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\langle\nabla F(\mathbf{u}_k),\,\nabla F(\mathbf{x}_k^{(i)})\right\rangle \tag{85}$$

$$=\frac{1}{2P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left[\|\nabla F(\mathbf{u}_k)\|^2+\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2-\left\|\nabla F(\mathbf{u}_k)-\nabla F(\mathbf{x}_k^{(i)})\right\|^2\right] \tag{86}$$

$$=\frac{1}{2}\|\nabla F(\mathbf{u}_k)\|^2+\frac{1}{2P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2-\frac{1}{2P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{u}_k)-\nabla F(\mathbf{x}_k^{(i)})\right\|^2 \tag{87}$$

where equation (86) comes from $2\mathbf{a}^\top\mathbf{b}=\|\mathbf{a}\|^2+\|\mathbf{b}\|^2-\|\mathbf{a}-\mathbf{b}\|^2$. □

Lemma 7. *Under Assumption 1, the squared norm of stochastic gradient can be bounded as*

$$\mathbf{E}_k\left[\|\mathcal{G}_k\|^2\right]\leq\left(\frac{\beta}{P}+1\right)\frac{\|\nabla F(\mathbf{X}_k)\|_F^2}{P}+\frac{\sigma^2}{P}.$$

Proof. Since $\mathbf{E}_k[\mathcal{G}_k] = \mathcal{H}_k$, then we have

$$\mathbf{E}_k\left[\|\mathcal{G}_k\|^2\right] = \mathbf{E}_k\left[\|\mathcal{G}_k - \mathbf{E}_k[\mathcal{G}_k]\|^2\right] + \|\mathbf{E}_k[\mathcal{G}_k]\|^2 \tag{88}$$

$$= \mathbf{E}_k\left[\|\mathcal{G}_k - \mathcal{H}_k\|^2\right] + \|\mathcal{H}_k\|^2 \tag{89}$$

$$\leq \frac{\beta}{P}\frac{\|\nabla F(\mathbf{X}_k)\|_{\mathrm{F}}^2}{P} + \frac{\sigma^2}{P} + \|\mathcal{H}_k\|^2 \tag{90}$$

$$\leq \frac{\beta}{P}\frac{\|\nabla F(\mathbf{X}_k)\|_{\mathrm{F}}^2}{P} + \frac{\sigma^2}{P} + \frac{1}{P}\|\nabla F(\mathbf{X}_k)\|_{\mathrm{F}}^2 \tag{91}$$

$$= \left(\frac{\beta}{P} + 1\right)\frac{\|\nabla F(\mathbf{X}_k)\|_{\mathrm{F}}^2}{P} + \frac{\sigma^2}{P}, \tag{92}$$

where (90) follows (5) and (91) comes from the convexity of vector norm and Jensen's inequality:

$$\|\mathcal{H}_k\|^2 = \left\|\frac{1}{P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\nabla F(\mathbf{x}_k^{(i)})\right\|^2 \leq \frac{1}{P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 = \frac{1}{P}\|\nabla F(\mathbf{X}_k)\|_{\mathrm{F}}^2. \tag{93}$$

□

A.4.2 Proof of Lemma 4

We rewrite the update rule by multiplying $\mathbf{1}_N/N$ on both sides in (3), we get

$$\mathbf{X}_{k+1}\frac{\mathbf{1}_N}{N} = \mathbf{X}_k\frac{\mathbf{1}_N}{N} - \gamma\mathbf{G}_k\frac{\mathbf{1}_N}{N} \tag{94}$$

where $\mathbf{W}_k$ disappears due to the special property from Assumption 5: $\mathbf{W}_k\mathbf{1}_N = \mathbf{1}_N$. Then, define the average model and effective learning rate as

$$\mathbf{u}_k = \mathbf{X}_k\frac{\mathbf{1}_N}{N},\ \eta = \frac{P}{N}\gamma. \tag{95}$$

After rearranging, one can obtain

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \eta\left[\frac{1}{P}\sum_{i=1,i\in\mathcal{S}_k}^{N} g(\mathbf{x}_k^{(i)})\right] \tag{96}$$

Observe that the averaged model $\mathbf{u}_k$ is performing perturbed stochastic gradient descent. In the sequel, we will focus on the convergence of the averaged model $\mathbf{u}_k$, which is common practice in distributed optimization literature [3].

According to Lipschitz continuous gradient assumption, we have

$$\mathbf{E}_k\left[F(\mathbf{u}_{k+1})\right] - F(\mathbf{u}_k) \leq -\eta\mathbf{E}_k\left[\langle\nabla F(\mathbf{u}_k),\, \mathcal{G}_k\rangle\right] + \frac{\eta^2 L}{2}\mathbf{E}_k\left[\|\mathcal{G}_k\|^2\right]. \tag{97}$$

Combining with Lemma 6 and 7, we obtain

$$\mathbf{E}_k\left[F(\mathbf{u}_{k+1})\right] - F(\mathbf{u}_k) \leq -\frac{\eta}{2}\|\nabla F(\mathbf{u}_k)\|^2 - \frac{\eta}{2P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 + \frac{\eta}{2P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{u}_k) - \nabla F(\mathbf{x}_k^{(i)})\right\|^2 +$$

$$\frac{\eta^2 L}{2P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2\cdot\left(\frac{\beta}{P}+1\right) + \frac{\eta^2 L\sigma^2}{2P} \tag{98}$$

$$\leq -\frac{\eta}{2}\|\nabla F(\mathbf{u}_k)\|^2 - \frac{\eta}{2}\left[1 - \eta L\left(\frac{\beta}{P}+1\right)\right]\cdot\frac{1}{P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 +$$

$$\frac{\eta^2 L\sigma^2}{2P} + \frac{\eta L^2}{2P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\mathbf{u}_k - \mathbf{x}_k^{(i)}\right\|^2. \tag{99}$$

After minor rearranging and according to the definition of Frobenius norm, it is easy to show

$$\|\nabla F(\mathbf{u}_k)\|^2 \leq \frac{2\left[F(\mathbf{u}_k) - \mathbf{E}_k\left[F(\mathbf{u}_{k+1})\right]\right]}{\eta} + \frac{\eta L\sigma^2}{P} + \frac{L^2}{P}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\mathbf{u}_k - \mathbf{x}_k^{(i)}\right\|^2 -$$

$$\left[1 - \eta L\left(\frac{\beta}{P}+1\right)\right]\frac{1}{P}\|\nabla F(\mathbf{X}_k)\|_{\mathrm{F}}^2. \tag{100}$$

Taking the total expectation and averaging over all iterates, we have

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(\mathbf{u}_k)\right\|^2\right] \leq \frac{2\left[F(\mathbf{u}_1)-F_{\text{inf}}\right]}{\eta K} + \frac{\eta L\sigma^2}{P} + \frac{L^2}{KP}\sum_{k=1}^{K}\sum_{i=1,i\in\mathcal{S}_k}^{N}\mathbb{E}\left\|\mathbf{u}_k-\mathbf{x}_k^{(i)}\right\|^2 - \left[1-\eta L\left(\frac{\beta}{P}+1\right)\right]\frac{1}{K}\sum_{k=1}^{K}\frac{\mathbb{E}\left\|\nabla F(\mathbf{X}_k)\right\|_{\text{F}}^2}{P}. \tag{101}$$

If the effective learning rate satisfies $\eta L(\beta/P+1)\leq 1$, then

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(\mathbf{u}_k)\right\|^2\right] \leq \frac{2\left[F(\mathbf{u}_1)-F_{\text{inf}}\right]}{\eta K} + \frac{\eta L\sigma^2}{P} + \frac{L^2}{Km}\sum_{k=1}^{K}\sum_{i=1,i\in\mathcal{S}_k}^{N}\mathbb{E}\left\|\mathbf{u}_k-\mathbf{x}_k^{(i)}\right\|^2. \tag{102}$$

Recalling the definition $\mathbf{u}_k=\mathbf{X}_k\mathbf{1}_N/N$ and adding a non-negative term to the RHS, one can get

$$\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\mathbf{u}_k-\mathbf{x}_k^{(i)}\right\|^2 \leq \sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\mathbf{u}_k-\mathbf{x}_k^{(i)}\right\|^2 + \sum_{i=1,i\notin\mathcal{S}_k}^{N}\left\|\mathbf{u}_k-\mathbf{z}_k^{(i)}\right\|^2 \tag{103}$$

$$= \left\|\mathbf{u}\mathbf{1}_N^\top-\mathbf{X}_k\right\|_{\text{F}}^2 \tag{104}$$

$$= \left\|\mathbf{X}_k\frac{\mathbf{1}_N\mathbf{1}_N^\top}{N}-\mathbf{X}_k\right\|_{\text{F}}^2 = \left\|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\right\|_{\text{F}}^2 \tag{105}$$

where $\mathbf{I},\mathbf{J}$ are $N\times N$ matrices. Plugging the inequality (105) into (102), we complete the proof.

References

[1] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[2] Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. In *ICML Workshop*, 2019.

[3] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM J. Optim.*, 26(3):1835–1854, 2016.