A EXPERIMENTAL SETTINGS

We train BERT-base using the Adam optimizer and ResNet-50 using the Momentum optimizer. The batch size is 256 for both models. We train the ResNet-50 model using the learning rate of 1e-2, the momentum of 0.9, the batch size of 256, and the weight decay of 1e-4. The first epoch is used for warm-up. We adopted a similar learning rate tuning scheme as proposed in [1]. The learning rate decays by 0.2 for every 30 epochs. We train BERT-base using the learning rate of 1e-5, the batch size of 256, and the weight decay of 1e-2. The first 2000 iterations are used for warm-up.

B ADDITIONAL EXPERIMENTS

B.1 Synchronization Group Adaptivity

To better understand our adaptive group scheduling, Figure 9 illustrates generated group distribution of SDPipe (in terms of percentiles) for ResNet-50 under the worker configuration of 1×8. We record the group size for each synchronization operation during the training process. Recall that we require the sync-graph to be connected every P iterations. In the special case of $P = 1$, SDPipe is the same as All-Reduce, which requires global synchronization in every step. However, by slightly relaxing $P > 1$, SDPipe permits high flexibility in group size adapting to heterogeneous settings, rather than a fixed communication topology like All-Reduce. Such relaxation is essentially the tradeoff for high availability (without waiting stragglers) with little model inconsistency.

C PROOF

C.1 Proof Preliminaries

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

C.1.1 Notations. The notations used in the proof are listed below.

Number of stages	M
Number of pipelines	N
Total iterations	K
Synchronization matrix/tensor	$\mathbf{W}/\mathcal{W}$
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.

For any stage j, we suppose $1 \le n_1 \le \mathcal{S}_k^j(i) \le n_2 \le N$, where n_1 and n_2 are the maximum and the minimum worker group sizes during the entire training process respectively. In the following proof, we first introducing the convergence property of SDPipe under single stage situation (i.e., $M = 1$) and then extend it to multiple stages. With a single stage, the global view of model synchronization step can be viewed as:

$$\mathbf{X}_{k+1} = (\mathbf{X}_k - \eta \mathbf{G}_k)\mathbf{W}_k,$$

where matrices $\mathbf{X}_k$ and $\mathbf{G}_k$ contain the local model vector $\mathbf{x}_k^i$ and gradient vector $\mathbf{g}(\mathbf{x}_k^i)$ of each worker (i.e., pipeline) i at the k-th iteration. $\mathbf{W}_k$ is the synchronization matrix used for model averaging. For example, supposing that we have three workers at stage j, at iteration k, fast workers from pipeline 1,2 finish the gradient computation and then form the group $\mathcal{S}_k^j = \{1, 2\}$, which can be equivalently expressed by E.q. (1):

$$\mathbf{X}_{k+1}^j = \widehat{\mathbf{X}}_k^j \mathbf{W}_k^j = \begin{bmatrix}\mathbf{x}^{j,1} & \mathbf{x}^{j,2} & \mathbf{x}^{j,3}\end{bmatrix} \underbrace{\begin{bmatrix}1/2 & 1/2 & 0\\ 1/2 & 1/2 & 0\\ 0 & 0 & 1\end{bmatrix}}_{\text{update}(\mathbf{x}^{j,1}, \mathbf{x}^{j,2})} = \begin{bmatrix}\frac{\mathbf{x}^{j,1}+\mathbf{x}^{j,2}}{2}\\ \frac{\mathbf{x}^{j,1}+\mathbf{x}^{j,2}}{2}\\ \mathbf{x}^{j,3}\end{bmatrix},$$

where $\widehat{\mathbf{X}}_k^j = \mathbf{X}_k^j - \eta \mathbf{G}_k^j$. We see that two fast workers perform model synchronization without the slow one.

C.1.2 Assumptions. We suppose $\mathbf{W}_k$ is independent on of the data samples at the kth iteration, and only depends on the arrival of ready signals from workers to request the kth group. Moreover, the speed of workers to generate ready signals could vary significantly in cloud environment due to resource sharing and network latency, leading to high dynamics and randomness of forming groups at different iterations. This makes $\mathbf{W}_k$ *largely* uncorrelated with k. We make the following commonly used assumptions [13, 35, 36, 54]:

Assumption 2.

(1) **Lipschitzian gradient:** $\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leqslant L\|\mathbf{x} - \mathbf{y}\|$
(2) **Unbiased estimation:** $\mathbb{E}_{\xi|\mathbf{x}}[g(\mathbf{x})] = \nabla F(\mathbf{x})$
(3) **Bounded variance:** $\mathbb{E}_{\xi|\mathbf{x}}[g(\mathbf{x}) - \nabla F(\mathbf{x})] \le \sigma^2$
(4) **Stochastic averaging:** $\mathbf{W}_k$ is doubly stochastic for all k, i.e., $\mathbf{W}_k = \mathbf{W}_k^\top$, $\mathbf{W}_k \mathbf{1}_N = \mathbf{1}_N$.
(5) **Dependence of random variables:** $\mathbf{W}_k$ is a random variable independent on ξ_k and k.

Note that the unbiased estimation is satisfied through the distributed file system (e.g., HDFS) or storing a portion of dataset and shuffling the local data among the workers periodically [16, 63].

C.1.3 Supporting Lemmas.

Lemma 1. *The Frobenius norm defined for $\mathbf{A} \in \mathbf{M}_n$ by*

$$\|\mathbf{A}\|_F^2 = \|Tr(\mathbf{A}\mathbf{A}^\top)| = \sum_{i=1}^n \sum_{j=1}^n |a_{i,j}|^2 = \sum_{i=1}^n \left\|\mathbf{A}^{(i)}\right\|^2 \tag{6}$$

Lemma 2. *[43] Suppose there is a sequence of $N \times N$ matrices $\{\mathbf{W}_l\}_{l=s}^k$, $0 \le s \le k$ and each $\mathbf{W}_l$ satisfies Assumption 1 and 2. We denote $\prod_{l=s}^k \mathbf{W}_l$ as $\mathbf{\Phi}_{s,k}$, then we have*

$$|\mathbf{\Phi}_{s,k}(i,j) - \frac{1}{N}| \le 2\frac{1 + p^{-NP}}{1 - p^{NP}}(1 - p^{NP})^{(k-s)/NP} \tag{7}$$

where p is the smallest value of all synchronization matrices, i.e., $p = \arg\min \mathbf{W}_k(i,j) = 1/n_1, \forall k$ with $\mathbf{W}_k(i,j) > 0, \forall i, j$.

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

$$\mathbb{E}\,\|g(\mathbf{X}_k) - \nabla F(\mathbf{X}_k)\|^2 \le \beta\,\|\nabla F(\mathbf{X}_k)\|^2 + n_1\sigma^2 \tag{8}$$

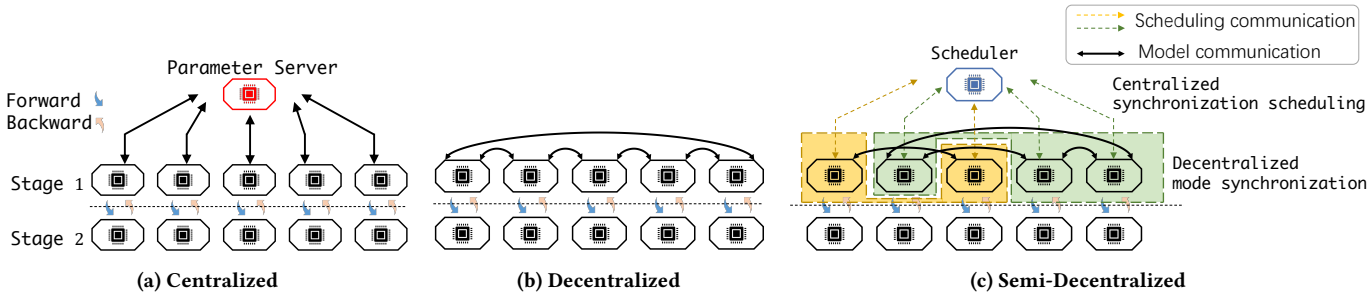

Figure 8: Comparison of different model synchronization schemes. The black arrows represent synchronization. The red node represents PS in centralized algorithms (Figure 8a). In Figure 8c, the blue node represents for the scheduler and the other dashed lines represent the scheduling communication. We only illustrate the synchronization for a single pipeline stage for simplicity.

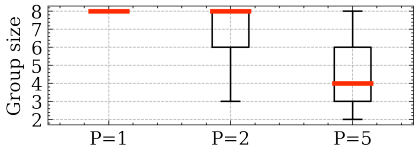

Figure 9: The distribution of worker numbers for each group-sync operation in SDPipe during training with 1×8 workers.

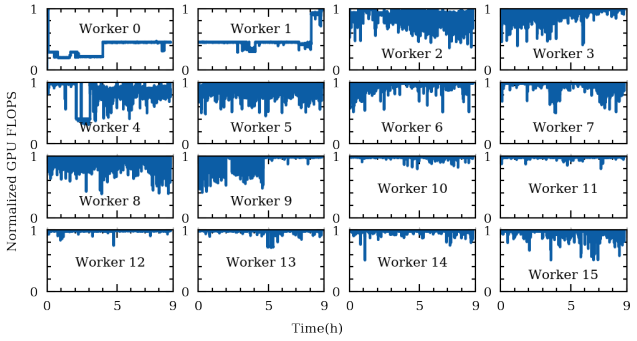

Figure 10: Illustration of real cloud GPU performance.

Proof.

$$\mathbb{E}\left\|g(\mathbf{X}_k) - \nabla F(\mathbf{X}_k)\right\|^2 \tag{9}$$

$$\leq \mathbb{E}\left\|g(\mathbf{X}_k) - \nabla F(\mathbf{X}_k)\right\|_{\mathrm{F}}^2 \tag{10}$$

$$= \mathbb{E}\sum_{i=1, i\in\mathcal{S}_k}^{N}\left\|g(\mathbf{x}_k^{(i)}) - \nabla F(\mathbf{x}_k^{(i)})\right\|^2 \tag{11}$$

$$\leq \mathbb{E}\sum_{i=1, i\in\mathcal{S}_k}^{N}\left[\beta\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 + \sigma^2\right] \tag{12}$$

$$= \beta\sum_{i=1, i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 + n_1\sigma^2 \tag{13}$$

$$= \beta\left\|\nabla F(\mathbf{X}_k)\right\|_{\mathrm{F}}^2 + n_1\sigma^2 \tag{14}$$

where (10) comes from $\|\mathbf{A}\| \leq \|\mathbf{A}\|_{\mathrm{F}}$, (11) and (14) comes from Lemma 1, comes from $|\mathcal{S}_k| \leq n_1$, (12) follows Assumption 2 (3). □

Lemma 4. *Suppose there is a sequence of $N \times N$ matrices $\{\mathbf{W}_l\}_{l=s}^k$, $0 \leq s \leq k$ and each $\mathbf{W}_l$ satisfies Assumption 2 and 1. We denote $\prod_{l=s}^k \mathbf{W}_l$ as $\mathbf{\Phi}_{s,k}$, then we have*

$$\mathbb{E}\left\|(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e}_i\right\|^2 \leq 4N\left(\frac{1+n_1^{NP}}{1-n_1^{-NP}}\right)^2(1-n_1^{-NP})^{2(k-s-1)/(NP)} \tag{15}$$

where $\mathbf{J} = \mathbf{1}\mathbf{1}^\top/(\mathbf{1}^\top\mathbf{1})$, $\mathbf{e}_i$ is the standard basis vector. For simplicity, in the following, we denote $4N(\frac{1+n_1^{NP}}{1-n_1^{-NP}})^2$ as t_1 and $(1-n_1^{-NP})^{2/(NP)}$ as t_2, then the RHS of Eq. (15) becomes $t_1 t_2^{k-s-1}$.

The proof of Lemma 4 could be easily obtained from Lemma 2.

C.2 Proof of Theorem 1

C.2.1 Proof of Single Stage. We first consider SDPipe with single stage, we have the following results:

Theorem 2 (**Convergence of SDPipe** (single stage)). *We assume the bound of gradient variance σ^2 is in inverse proportion to the mini-batch size. For semi-decentralized parallel SGD, under Assumptions 1–2, if the learning rate satisfies*

$$\eta L + \frac{2N^3\eta^2L^2t_1}{n_2^2}\left(\frac{1}{1-t_2} + \frac{2}{(1-\sqrt{t_2})^2}\right) \leq 1, \tag{16}$$

where $\eta = \frac{n_2\gamma}{N}$, $t_1 = 4N(\frac{1+n_1^{NP}}{1-n_1^{-NP}})^2$ and $t_2 = (1-n_1^{-NP})^{2/(NP)}$, 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

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(\mathbf{u}_k)\right\|^2\right] \leq \frac{2[F(\mathbf{u}_1) - F_{inf}]}{\eta K} + \frac{\eta L\sigma^2}{n_2} + \frac{2\eta^2L^2\sigma^2N^3n_1t_1}{n_2^3}\left(\frac{1}{1-t_2} + \frac{2}{(1-\sqrt{t_2})^2}\right). \tag{17}$$

Corollary 1. *Futhermore, if the learning rate is $\eta = \frac{1}{L}\sqrt{\frac{N}{K}}$, the average-squared gradient norm after K iterations is bounded by*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(\mathbf{u}_k)\right\|^2\right] \leq \frac{2L[F(\mathbf{u}_1) - F_{inf}] + \sigma^2}{\sqrt{NK}} + \frac{2N^2\sigma^2t_1}{K}\left(\frac{1}{1-t_2} + \frac{2}{(1-\sqrt{t_2})^2}\right) \tag{18}$$

if the total iterations K is sufficiently large: $K \geq \frac{12N^2t_1}{(1-\sqrt{t_2})^2}$. And if $K \geq \frac{36N^5t_1^2}{(1-\sqrt{t_2})^4}$, then the average-squared gradient norm will be bounded by $2[L(F(\mathbf{u}_1) - F_{inf}) + \sigma^2]/\sqrt{NK}$.

Before providing the proof of Theorem 2, we prefer to first present an important lemma that describes the basic convergence upper bound framework.

Lemma 5. *In SDPipe, under Assumption 2, the average-squared gradient norm after K iterations is bounded as follows*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\mathbf{u}_k)\|^2\right] \leq \frac{2\left[F(\mathbf{u}_1) - F_{inf}\right]}{\eta K} + \frac{\eta L\sigma^2}{n_2} + \frac{L^2}{Kn_2}\sum_{k=1}^{K}\|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\|_F^2 - \left[1-\eta L\left(\frac{\beta}{n_2}+1\right)\right]\frac{1}{K}\sum_{k=1}^{K}\frac{\mathbb{E}\|\nabla F(\mathbf{X}_k)\|_F^2}{n_2}. \tag{19}$$

Since the proof of Lemma 5 only relies on commonly used Assumption 2, our result is consistent with the analysis of fully synchronous distributed SGD. We remove the proof details of Lemma 5 to Sec. C.3, and focus on our extension on the SDPipe below. Note that our Theorem 2 follows the same SGD error (i.e., the first two items) as Eq. (19). Our goal is to provide an upper bound for the left network error items in the following.

C.2.2 Decomposition. To provide an upper bound for the term $\frac{L^2}{Kn_2}\sum_{k=1}^{K}\|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\|_F^2$, we derive a specific expression for $\mathbf{X}_k(\mathbf{I}-\mathbf{J})$. According to the SGD update rule, one can observe that

$$\mathbf{X}_k(\mathbf{I}-\mathbf{J}) = (\mathbf{X}_{k-1} - \gamma\mathbf{G}_{k-1})\,\mathbf{W}_{k-1}(\mathbf{I}-\mathbf{J}) \tag{20}$$

$$= \mathbf{X}_{k-1}(\mathbf{I}-\mathbf{J})\mathbf{W}_{k-1} - \gamma\mathbf{G}_{k-1}(\mathbf{W}_{k-1}-\mathbf{J}) \tag{21}$$

where (21) 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

$$\mathbf{X}_k(\mathbf{I}-\mathbf{J}) = [\mathbf{X}_{k-2}(\mathbf{I}-\mathbf{J})\mathbf{W}_{k-2} - \gamma\mathbf{G}_{k-2}(\mathbf{W}_{k-2}-\mathbf{J})]\,\mathbf{W}_{k-1} - \gamma\mathbf{G}_{k-1}(\mathbf{W}_{k-1}-\mathbf{J}) \tag{22}$$

$$= \mathbf{X}_{k-2}(\mathbf{I}-\mathbf{J})\mathbf{W}_{k-2}\mathbf{W}_{k-1} - \gamma\mathbf{G}_{k-2}(\mathbf{W}_{k-2}\mathbf{W}_{k-1}-\mathbf{J}) - \gamma\mathbf{G}_{k-1}(\mathbf{W}_{k-1}-\mathbf{J}) \tag{23}$$

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

$$\mathbf{X}_k(\mathbf{I}-\mathbf{J}) = \mathbf{X}_1(\mathbf{I}-\mathbf{J})\mathbf{\Phi}_{1,k-1} - \gamma\sum_{s=1}^{k-1}\mathbf{G}_s(\mathbf{\Phi}_{s,k-1}-\mathbf{J}) \tag{24}$$

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

$$\mathbb{E}\|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\|_F^2 = \gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1}\mathbf{G}_s(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_F^2. \tag{25}$$

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

$$\mathbb{E}\|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\|_F^2 = \gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1}\mathbf{G}_s(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_F^2 \tag{26}$$

$$= \gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1}(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J}) + \sum_{s=1}^{k-1}\mathbf{Q}_s(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_F^2 \tag{27}$$

$$\leq \underbrace{2\gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1}(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_F^2}_{T_1} + \underbrace{2\gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1}\mathbf{Q}_s(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_F^2}_{T_2} \tag{28}$$

where $\mathbf{Q}_s = \nabla F(\mathbf{X}_s)$, (28) follows $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. 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 $\frac{L^2}{Kn_2}\sum_{k=1}^{K}\|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\|_F^2$. Accordingly, we will also derive the bounds for $\frac{L^2}{Kn_2}\sum_{k=1}^{K}T_1$ and $\frac{L^2}{Kn_2}\sum_{k=1}^{K}T_2$.

C.2.3 Bounding T_1. For the first term T_1, we have

$$T_1 = 2\gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1}(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_F^2 \tag{29}$$

$$= 2\gamma^2\mathbb{E}\sum_{i=1}^{N}\left\|\sum_{s=1}^{k-1}(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\mathbf{e_i}\right\|^2 \tag{30}$$

$$= 2\gamma^2\mathbb{E}\sum_{i=1}^{N}\left[\underbrace{\sum_{s=1}^{k-1}\left\|(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\mathbf{e_i}\right\|^2}_{A_1} + \underbrace{2\sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1}\langle(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\mathbf{e_i}, (\mathbf{G}_l-\mathbf{Q}_l)(\mathbf{\Phi}_{l,k-1}-\mathbf{J})\mathbf{e_i}\rangle}_{A_2}\right] \tag{31}$$

where (30) comes from Lemma 1 and (31) 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:

$$A_1 = \sum_{s=1}^{k-1}\left\|(\mathbf{G}_s-\mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\mathbf{e_i}\right\|^2 \tag{32}$$

$$\leq \sum_{s=1}^{k-1}\|\mathbf{G}_s-\mathbf{Q}_s\|^2\left\|(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\mathbf{e_i}\right\|^2 \tag{33}$$

$$\leq \sum_{s=1}^{k-1}\left[\beta\|\nabla F(\mathbf{X}_s)\|_F^2 + n_1\sigma^2\right]\left\|(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\mathbf{e_i}\right\|^2 \tag{34}$$

$$\leq \sum_{s=1}^{k-1}\left[\beta\|\nabla F(\mathbf{X}_s)\|_F^2 + n_1\sigma^2\right]t_1t_2^{k-s-1} \tag{35}$$

$$\leq n_1\sigma^2t_1\frac{1}{1-t_2} + \beta\sum_{s=1}^{k-1}t_1t_2^{k-s-1}\|\nabla F(\mathbf{X}_s)\|_F^2 \tag{36}$$

where (34) comes from Lemma 3, (35) comes from Lemma 2 and (36) follows the summation formula of power series

$$\sum_{s=1}^{k-1}t_2^{k-s-1} \leq \sum_{s=-\infty}^{k-1}t_2^{k-s-1} \leq \frac{1}{1-t_2}. \tag{37}$$

The cross items A_2 can be bounded by:

$$A_2 = \sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} \langle (\mathbf{G}_s - \mathbf{Q}_s)(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e_i}, (\mathbf{G}_l - \mathbf{Q}_l)(\mathbf{\Phi}_{l,k-1} - \mathbf{J})\mathbf{e_i}\rangle \tag{38}$$

$$\leq \sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} \|\mathbf{G}_s - \mathbf{Q}_s\| \left\|(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e_i}\right\| \|\mathbf{G}_l - \mathbf{Q}_l\| \left\|(\mathbf{\Phi}_{l,k-1} - \mathbf{J})\mathbf{e_i}\right\| \tag{39}$$

$$\leq \sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} \left[\frac{1}{2c_{s,l}} \left\|(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e_i}\right\|^2 \left\|(\mathbf{\Phi}_{l,k-1} - \mathbf{J})\mathbf{e_i}\right\|^2 + \frac{1}{2/c_{s,l}} \|\mathbf{G}_s - \mathbf{Q}_s\|^2 \|\mathbf{G}_l - \mathbf{Q}_l\|^2 \right], \forall c_{s,l} > 0 \tag{40}$$

$$\leq \sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} \left[\frac{1}{2c_{s,l}} t_1^2 t_2^{k-s-1} t_2^{k-l-1} + \frac{1}{2/c_{s,l}} (\beta \|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2 + n_1\sigma^2)(\beta \|\nabla F(\mathbf{X}_l)\|_{\mathrm{F}}^2 + n_1\sigma^2) \right], \forall c_{s,l} > 0 \tag{41}$$

$$\leq \sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} \left(\frac{1}{2c_{s,l}} t_1^2 t_2^{k-s-1} + \frac{1}{2/c_{s,l}} n_1^2\sigma^4\right), \forall c_{s,l} > 0. \tag{42}$$

where (40) follows $ab \leq \frac{1}{2}(a^2/c + cb^2), \forall c > 0$, (41) comes from Lemma 3 and 2. We can choose $c_{s,l} > 0$ and make the term in the last step become $\sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1}(t_1 t_2^{(k-s-1)/2} n_1\sigma^2)$ (by applying inequality of arithmetic and geometric means). Note that we directly remove the items related with β in (42) for a neater formula. The simplification will not affect the final result since we set $\beta = 0$ at last. Thus

$$A_2 \leq \sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} (t_1 t_2^{(k-s-1)/2} n_1\sigma^2) \tag{43}$$

$$= \sum_{s=1}^{k-1} (k-s) t_1 t_2^{(k-s-1)/2} n_1\sigma^2 \leq \frac{1}{(1-\sqrt{t_2})^2} t_1 n_1\sigma^2. \tag{44}$$

where (44) follows the summation formula of power series

$$\sum_{s=1}^{k-1} (k-s) t_2^{(k-s-1)/2} \leq \sum_{s=-\infty}^{k-1} (k-s) t_2^{(k-s-1)/2} \leq \frac{1}{(1-\sqrt{t_2})^2}. \tag{45}$$

Substituting (36) (44) back into (31), we have

$$T_1 \leq 2\gamma^2\mathbb{E}\sum_{i=1}^{N}\left[n_1\sigma^2 t_1 \frac{1}{1-t_2} + \beta\sum_{s=1}^{k-1} t_1 t_2^{k-s-1} \|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2 + \frac{2}{(1-\sqrt{t_2})^2} n_1\sigma^2 t_1 \right] \tag{46}$$

$$= 2Nn_1\gamma^2\sigma^2 t_1 \left(\frac{1}{1-t_2} + \frac{2}{(1-\sqrt{t_2})^2}\right) + 2N\gamma^2\beta\sum_{s=1}^{k-1} t_1 t_2^{k-s-1} \|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2 \tag{47}$$

C.2.4 Bounding T_2.

$$T_2 = 2\gamma^2\mathbb{E}\left\|\sum_{s=1}^{k-1} \mathbf{Q}_s(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\right\|_{\mathrm{F}}^2 \tag{48}$$

$$= 2\gamma^2 \sum_{i=1}^{N} \mathbb{E}\left\|\sum_{s=1}^{k-1} \mathbf{Q}_s(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e}_i\right\|^2 \tag{49}$$

$$= 2\gamma^2 \sum_{i=1}^{N} \left[\sum_{s=1}^{k-1} \mathbb{E}\left\|\mathbf{Q}_s(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e}_i\right\|^2 + 2\sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} \mathbb{E}\langle \mathbf{Q}_s(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e}_i, \mathbf{Q}_l(\mathbf{\Phi}_{l,k-1} - \mathbf{J})\mathbf{e}_i\rangle \right] \tag{50}$$

$$\leq 2\gamma^2 \sum_{i=1}^{N} \left[\sum_{s=1}^{k-1} t_1 t_2^{k-s-1} \mathbb{E}\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2 + \underbrace{2\sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} \mathbb{E}\|\mathbf{Q}_s\| \left\|(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e}_i\right\| \|\mathbf{Q}_l\| \left\|(\mathbf{\Phi}_{l,k-1} - \mathbf{J})\mathbf{e}_i\right\|}_{A_3} \right] \tag{51}$$

The cross items A_3 can be bounded by:

$$A_3 = \sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} \mathbb{E}\|\mathbf{Q}_s\| \left\|(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e}_i\right\| \|\mathbf{Q}_l\| \left\|(\mathbf{\Phi}_{l,k-1} - \mathbf{J})\mathbf{e}_i\right\| \tag{52}$$

$$\leq \sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} \mathbb{E}\left[\frac{1}{2c_{s,l}} \left\|(\mathbf{\Phi}_{s,k-1} - \mathbf{J})\mathbf{e}_i\right\|^2 \left\|(\mathbf{\Phi}_{l,k-1} - \mathbf{J})\mathbf{e}_i\right\|^2 + \frac{1}{2/c_{s,l}} \|\mathbf{Q}_s\|^2 \|\mathbf{Q}_l\|^2 \right], \forall c_{s,l} > 0 \tag{53}$$

$$\leq \sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} \mathbb{E}\left[\frac{1}{2c_{s,l}} t_1^2 t_2^{k-s-1} + \frac{1}{2/c_{s,l}} \|\nabla F(\mathbf{X}_s)\|^2 \|\nabla F(\mathbf{X}_l)\|^2 \right], \forall c_{s,l} > 0 \tag{54}$$

$$\leq \sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} \mathbb{E}\left[t_1 t_2^{(k-s-1)/2} \|\nabla F(\mathbf{X}_s)\| \|\nabla F(\mathbf{X}_l)\| \right] \tag{55}$$

$$\leq \sum_{s=1}^{k-1}\sum_{l=s+1}^{k-1} \mathbb{E}\left[t_1 t_2^{(k-s-1)/2} \frac{\|\nabla F(\mathbf{X}_s)\|^2 + \|\nabla F(\mathbf{X}_l)\|^2}{2} \right] \tag{56}$$

$$\leq \sum_{s=1}^{k-1} \left[(k-s) t_1 t_2^{(k-s-1)/2} \mathbb{E}\|\nabla F(\mathbf{X}_s)\|^2 \right] \tag{57}$$

$$\leq \sum_{s=1}^{k-1} \left[(k-s) t_1 t_2^{(k-s-1)/2} \mathbb{E}\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2 \right] \tag{58}$$

We have

$$T_2 \le 2\gamma^2 \sum_{i=1}^{N}\left[\sum_{s=1}^{k-1}(t_2^{k-s-1}+2(k-s)t_2^{(k-s-1)/2})t_1\mathbb{E}\left\|\nabla F(\mathbf{X}_s)\right\|_{\mathrm{F}}^2\right] \tag{59}$$

$$\le 2N\gamma^2\left[\sum_{s=1}^{k-1}(t_2^{k-s-1}+2(k-s)t_2^{(k-s-1)/2})t_1\mathbb{E}\left\|\nabla F(\mathbf{X}_s)\right\|_{\mathrm{F}}^2\right] \tag{60}$$

We complete the second part.

C.2.5 Final result. According to (28)(47)(60), setting $\beta = 0$, the network error can be bounded as

$$\frac{1}{Kn_2}\sum_{k=1}^{K}\left\|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\right\|_{\mathrm{F}}^2 \le \frac{1}{Kn_2}\sum_{k=1}^{K}(T_1+T_2) \tag{61}$$

$$\begin{aligned}\le& 2\frac{n_1}{n_2}\gamma^2\sigma^2 N t_1\left(\frac{1}{1-t_2}+\frac{2}{(1-\sqrt{t_2})^2}\right)\\ &+2\frac{1}{n_2}N\gamma^2\beta\frac{1}{K}\sum_{k=1}^{K}\sum_{s=1}^{k-1}t_1 t_2^{k-s-1}\left\|\nabla F(\mathbf{X}_s)\right\|_{\mathrm{F}}^2\\ &+\frac{2N\gamma^2 t_1}{Kn_2}\sum_{k=1}^{K}\sum_{s=1}^{k-1}(t_2^{k-s-1}+2(k-s)t_2^{(k-s-1)/2})\mathbb{E}\left\|\nabla F(\mathbf{X}_s)\right\|_{\mathrm{F}}^2\end{aligned} \tag{62}$$

$$\begin{aligned}\le& 2\frac{n_1}{n_2}\gamma^2\sigma^2 N t_1\left(\frac{1}{1-t_2}+\frac{2}{(1-\sqrt{t_2})^2}\right)\\ &+2\frac{1}{n_2}N\gamma^2\beta\frac{t_1}{K}\sum_{s=1}^{K}\left\|\nabla F(\mathbf{X}_s)\right\|_{\mathrm{F}}^2\sum_{k=s+1}^{+\infty}t_2^{k-s-1}\\ &+\frac{2N\gamma^2 t_1}{Kn_2}\sum_{s=1}^{K}\mathbb{E}\left\|\nabla F(\mathbf{X}_s)\right\|_{\mathrm{F}}^2\sum_{k=s+1}^{+\infty}(t_2^{k-s-1}+2(k-s)t_2^{(k-s-1)/2})\end{aligned} \tag{63}$$

$$\begin{aligned}\le& 2\gamma^2\sigma^2 N t_1\frac{n_1}{n_2}\left(\frac{1}{1-t_1}+\frac{2}{(1-\sqrt{t_2})^2}\right)+\frac{2N\gamma^2\beta t_1}{1-t_2}\frac{1}{K}\sum_{s=1}^{K}\frac{\left\|\nabla F(\mathbf{X}_s)\right\|_{\mathrm{F}}^2}{n_2}\\ &+2\gamma^2 N t_1\left(\frac{1}{1-t_2}+\frac{2}{(1-\sqrt{t_2})^2}\right)\frac{1}{K}\sum_{s=1}^{K}\mathbb{E}\frac{\left\|\nabla F(\mathbf{X}_s)\right\|_{\mathrm{F}}^2}{n_2}.\end{aligned} \tag{64}$$

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

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla F(\mathbf{u}_k)\right\|^2 \le& \frac{2(F(\mathbf{x}_1)-F_{\inf})}{\eta K}+\frac{\eta L\sigma^2}{n_2}+\\ &2\gamma^2L^2\sigma^2 N t_1\frac{n_1}{n_2}\left(\frac{1}{1-t_2}+\frac{2}{(1-\sqrt{t_2})^2}\right)-\left[1-\eta L\left(\frac{\beta}{n_2}+1\right)\right.\\ &\left.-2N\gamma^2L^2t_1\left(\frac{\beta+1}{1-t_2}+\frac{2}{(1-\sqrt{t_2})^2}\right)\right]\frac{1}{K}\sum_{k=1}^{K}\frac{\mathbb{E}\left\|\nabla F(\mathbf{X}_k)\right\|_{\mathrm{F}}^2}{n_2}.\end{aligned} \tag{65}$$

When the learning rate satisfies

$$1-\eta L\left(\frac{\beta}{n_2}+1\right)-2N\gamma^2L^2t_1\left(\frac{\beta+1}{1-t_2}+\frac{2}{(1-\sqrt{t_2})^2}\right)\le 0, \tag{66}$$

we have

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla F(\mathbf{u}_k)\right\|^2 \le& \frac{2(F(\mathbf{x}_1)-F_{\inf})}{\eta K}+\frac{\eta L\sigma^2}{n_2}\\ &+2\eta^2L^2\sigma^2\frac{N^3n_1}{n_2^3}t_1\left(\frac{1}{1-t_2}+\frac{2}{(1-\sqrt{t_2})^2}\right),\end{aligned} \tag{67}$$

where $\eta = \frac{n_2\gamma}{N}$. Setting $\beta = 0$, the condition on learning rate (66) can be further simplified as follows:

$$\eta L\left(\frac{\beta}{2}+1\right)+2N\gamma^2L^2t_1\left(\frac{\beta+1}{1-t_2}+\frac{2}{(1-\sqrt{t_2})^2}\right) \tag{68}$$

$$=\eta L+2\frac{N^3\eta^2}{n_2^2}L^2t_1\left(\frac{1}{1-t_2}+\frac{2}{(1-\sqrt{t_2})^2}\right)\le 1, \tag{69}$$

Here, we complete the proof.

C.2.6 Extending to Multi-Stage. When extending to multiple stages situation, the synchronization step could be replaced by:

$$\mathbf{X}_{k+1} = \mathbf{X}_{k+1/2}\times_M \mathcal{W}_k, \tag{70}$$

where $\mathcal{W}$ is the synchronization tensor with the size of $M\times N\times N$. The product $\times_M$ represents to perform a M-way matrix multiplication in the stage level (i.e., each stage is $\mathbf{X}_{k+1/2}^j\times\mathcal{W}_k^j$) and concatenate them into the tensor $\mathbf{X}_{k+1}$.

Based on the above formulation, the major difference in the proofs come from Lemma 4. The RHS of E.q. (25) in the decomposition step could be

$$\mathbb{E}\left\|\sum_{s=1}^{k-1}\mathbf{G}_s\times_M(\mathcal{W}_s\times_M\mathcal{W}_{s+1}\times_M\dots\mathcal{W}_{k-1}-\mathbf{J}_M)\right\|_{\mathrm{F}}^2 \tag{71}$$

$$=M\mathbb{E}\left\|\sum_{s=1}^{k-1}\mathbf{G}_s(\mathbf{\Phi}_{s,k-1}-\mathbf{J})\right\|_{\mathrm{F}}^2. \tag{72}$$

Therefore, the previous proofs could be reused except for adding a scale factor of M for the squared norm network error term. We can further obtain the final convergence bound as

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla F(\mathbf{u}_k)\right\|^2 \le& \frac{2(F(\mathbf{x}_1)-F_{\inf})}{\eta K}+\frac{\eta L\sigma^2}{n_2}\\ &+2M\eta^2L^2\sigma^2\frac{N^3n_1}{n_2^3}t_1\left(\frac{1}{1-t_2}+\frac{2}{(1-\sqrt{t_2})^2}\right).\end{aligned} \tag{73}$$

Discussion: Our analysis first provides the proof of the convergence property of SDPipe with a single stage. And then we extend them to multiple stages. In this step, we assume the single pipeline training algorithm itself will not break the convergence property like vanilla standalone SGD training. For example, PipeDream uses a weight stashing technique to avoid a fundamental mismatch between the version of weights used in the forward step and the adjacent backward step. Although it doesn't provide strict theoretical proof, the experimental results on a variety of workloads have shown that it can prevent model convergence and even match the standard model quality. Therefore, in our approach, we didn't involve these effects from pipeline parallel training in our analysis.

We manage to study what impact is the sync-graph-guided semi-decentralized training has on pipeline parallelism. We are also glad to extend our approach and even the theoretical analysis to other pipeline parallel training methods in the future.

C.3 Proof of Lemma 5

For the ease of writing, we first define some notations. Let Ξ_k denote the set $\{\xi_k^{(1)}, \ldots, \xi_k^{(N)}\}$ of mini-batches at N 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:

$$\mathcal{G}_k = \frac{1}{|\mathcal{S}_k|}\sum_{i=1,i\in\mathcal{S}_k}^{N} g(\mathbf{x}_k^{(i)}),\ \mathcal{H}_k = \frac{1}{|\mathcal{S}_k|}\sum_{i=1,i\in\mathcal{S}_k}^{N} \nabla F(\mathbf{x}_k^{(i)}). \quad (74)$$

C.3.1 Supporting Lemmas.

Lemma 6. *Under Assumption 2, we have the following variance bound for the averaged stochastic gradient:*

$$\mathbb{E}_{\Xi_K|\mathbf{X}_k}\left[\|\mathcal{G}_k - \mathcal{H}_k\|^2\right] \le \frac{\beta}{n_2^2}\|\nabla F(\mathbf{X}_k)\|_F^2 + \frac{\sigma^2}{n_2}. \quad (75)$$

Proof. According to the definition of $\mathcal{G}_k, \mathcal{H}_k$ (74), we have

$$\mathbb{E}_{\Xi_K|\mathbf{X}_k}\left[\|\mathcal{G}_k - \mathcal{H}_k\|^2\right] \quad (76)$$

$$= \mathbb{E}_{\Xi_K|\mathbf{X}_k}\left\|\frac{1}{|\mathcal{S}_k|}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left[g(\mathbf{x}_k^{(i)}) - \nabla F(\mathbf{x}_k^{(i)})\right]\right\|^2 \quad (77)$$

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

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

where equation (79) is due to $\{\xi_k^{(i)}\}$ are independent random variables. Now, directly applying Assumption 3 and 4 to (79), one can observe that all cross terms are zero. Then, since $\forall k, n_2 \le |\mathcal{S}_k|$, we have

$$\mathbb{E}_{\Xi_K|\mathbf{X}_k}\|\mathcal{G}_k - \mathcal{H}_k\|^2 \le \frac{1}{|\mathcal{S}_k|^2}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left[\beta\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 + \sigma^2\right] \quad (80)$$

$$\le \frac{\beta}{n_2^2}\|\nabla F(\mathbf{X}_k)\|_F^2 + \frac{\sigma^2}{n_2}. \quad (81)$$

□

Lemma 7. *Under Assumption 2, the expected inner product between stochastic gradient and full batch gradient can be expanded as*

$$\mathbf{E}_k\left[\langle\nabla F(\mathbf{u}_k), \mathcal{G}_k\rangle\right] = \frac{1}{2}\|\nabla F(\mathbf{u}_k)\|^2 + \frac{1}{2|\mathcal{S}_k|}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 - \frac{1}{2|\mathcal{S}_k|}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{u}_k) - \nabla F(\mathbf{x}_k^{(i)})\right\|^2 \quad (82)$$

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

Proof.

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

$$= \frac{1}{|\mathcal{S}_k|}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\langle\nabla F(\mathbf{u}_k), \nabla F(\mathbf{x}_k^{(i)})\right\rangle \quad (84)$$

$$= \frac{1}{2|\mathcal{S}_k|}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left[\|\nabla F(\mathbf{u}_k)\|^2 + \left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 - \left\|\nabla F(\mathbf{u}_k) - \nabla F(\mathbf{x}_k^{(i)})\right\|^2\right] \quad (85)$$

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

where equation (85) comes from $2\mathbf{a}^\top\mathbf{b} = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a}-\mathbf{b}\|^2$. □

Lemma 8. *Under Assumption 2, the squared norm of stochastic gradient can be bounded as*

$$\mathbf{E}_k\left[\|\mathcal{G}_k\|^2\right] \le \left(\frac{\beta}{n_2} + 1\right)\frac{\|\nabla F(\mathbf{X}_k)\|_F^2}{n_2} + \frac{\sigma^2}{n_2}.$$

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

$$\mathbf{E}_k\left[\|\mathcal{G}_k\|^2\right] = \mathbf{E}_k\left[\|\mathcal{G}_k - \mathbf{E}_k[\mathcal{G}_k]\|^2\right] + \|\mathbf{E}_k[\mathcal{G}_k]\|^2 \quad (87)$$

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

$$\le \frac{\beta}{n_2}\frac{\|\nabla F(\mathbf{X}_k)\|_F^2}{n_2} + \frac{\sigma^2}{N} + \|\mathcal{H}_k\|^2 \quad (89)$$

$$\le \frac{\beta}{n_2}\frac{\|\nabla F(\mathbf{X}_k)\|_F^2}{n_2} + \frac{\sigma^2}{n_2} + \frac{1}{n_2}\|\nabla F(\mathbf{X}_k)\|_F^2 \quad (90)$$

$$= \left(\frac{\beta}{n_2} + 1\right)\frac{\|\nabla F(\mathbf{X}_k)\|_F^2}{n_2} + \frac{\sigma^2}{n_2}, \quad (91)$$

where (89) follows (6) and (90) comes from the convexity of vector norm and Jensen's inequality:

$$\|\mathcal{H}_k\|^2 = \left\|\frac{1}{|\mathcal{S}_k|}\sum_{i=1,i\in\mathcal{S}_k}^{N}\nabla F(\mathbf{x}_k^{(i)})\right\|^2 \le \frac{1}{|\mathcal{S}_k|}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 \le \frac{1}{n_2}\|\nabla F(\mathbf{X}_k)\|_F^2. \quad (92)$$

□

C.3.2 Proof of Lemma 5. We rewrite the update rule by multiplying $\mathbf{1}_N/N$ on both sides in E.q. (1), we get

$$\mathbf{X}_{k+1}\frac{\mathbf{1}_N}{N} = \mathbf{X}_k\frac{\mathbf{1}_N}{N} - \gamma \mathbf{G}_k\frac{\mathbf{1}_N}{N} \tag{93}$$

where $\mathbf{W}_k$ disappears due to the special property from Assumption 2.(4): $\mathbf{W}_k\mathbf{1}_N = \mathbf{1}_N$. Then, define the average model as

$$\mathbf{u}_k = \mathbf{X}_k\frac{\mathbf{1}_N}{N}, \eta = \frac{n_2\gamma}{N}. \tag{94}$$

After rearranging, one can obtain

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \eta\left[\frac{1}{n_2}\sum_{i=1,i\in\mathcal{S}_k}^{N} g(\mathbf{x}_k^{(i)})\right] \tag{95}$$

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 [58].

According to Lipschitz continuous gradient assumption, we have

$$\mathbf{E}_k\left[F(\mathbf{u}_{k+1})\right] - F(\mathbf{u}_k) \le -\eta\mathbf{E}_k\left[\langle\nabla F(\mathbf{u}_k), \mathcal{G}_k\rangle\right] + \frac{\eta^2 L}{2}\mathbf{E}_k\left[\|\mathcal{G}_k\|^2\right]. \tag{96}$$

Combining with Lemma 7 and 8, we obtain

$$\begin{aligned}
\mathbf{E}_k\left[F(\mathbf{u}_{k+1})\right] - F(\mathbf{u}_k) &\le -\frac{\eta}{2}\|\nabla F(\mathbf{u}_k)\|^2 - \frac{\eta}{2n_2}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 \\
&+ \frac{\eta}{2n_2}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{u}_k) - \nabla F(\mathbf{x}_k^{(i)})\right\|^2 \\
&+ \frac{\eta^2 L}{2n_2}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2\cdot\left(\frac{\beta}{n_2}+1\right) + \frac{\eta^2 L\sigma^2}{2n_2} \qquad (97)\\
&\le -\frac{\eta}{2}\|\nabla F(\mathbf{u}_k)\|^2 - \frac{\eta}{2}\left[1-\eta L\left(\frac{\beta}{n_2}+1\right)\right]\cdot\frac{1}{n_2}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\nabla F(\mathbf{x}_k^{(i)})\right\|^2 \\
&+ \frac{\eta^2 L\sigma^2}{2n_2} + \frac{\eta L^2}{2n_2}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\mathbf{u}_k - \mathbf{x}_k^{(i)}\right\|^2. \qquad (98)
\end{aligned}$$

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

$$\|\nabla F(\mathbf{u}_k)\|^2 \le \frac{2\left[F(\mathbf{u}_k) - \mathbf{E}_k\left[F(\mathbf{u}_{k+1})\right]\right]}{\eta} + \frac{\eta L\sigma^2}{n_2} \tag{99}$$

$$+ \frac{L^2}{n_2}\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\mathbf{u}_k - \mathbf{x}_k^{(i)}\right\|^2 - \left[1-\eta L\left(\frac{\beta}{n_2}+1\right)\right]\frac{1}{n_2}\|\nabla F(\mathbf{X}_k)\|_{\mathrm{F}}^2. \tag{100}$$

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

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\mathbf{u}_k)\|^2\right] \le\ & \frac{2\left[F(\mathbf{u}_1) - F_{\inf}\right]}{\eta K} + \frac{\eta L\sigma^2}{n_2} + \\
& \frac{L^2}{Kn_2}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{u}_k - \mathbf{x}_k^{(i)}\right\|^2 - \left[1-\eta L\left(\frac{\beta}{n_2}+1\right)\right]\frac{1}{K}\sum_{k=1}^{K}\frac{\mathbb{E}\|\nabla F(\mathbf{X}_k)\|_{\mathrm{F}}^2}{n_2}.
\end{aligned} \tag{101}$$

If the effective learning rate satisfies $\eta L(\beta/n_2+1) \le 1$, then

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\mathbf{u}_k)\|^2\right] \le & \frac{2\left[F(\mathbf{u}_1) - F_{\inf}\right]}{\eta K} + \frac{\eta L\sigma^2}{n_2} + \\
& \frac{L^2}{Kn_2}\sum_{k=1}^{K}\sum_{i=1,i\in\mathcal{S}_k}^{N}\mathbb{E}\left\|\mathbf{u}_k - \mathbf{x}_k^{(i)}\right\|^2.
\end{aligned} \tag{102}$$

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

$$\sum_{i=1,i\in\mathcal{S}_k}^{N}\left\|\mathbf{u}_k - \mathbf{x}_k^{(i)}\right\|^2 \le \left\|\mathbf{u}\mathbf{1}_N^\top - \mathbf{X}_k\right\|_{\mathrm{F}}^2 \tag{103}$$

$$= \left\|\mathbf{X}_k\frac{\mathbf{1}_N\mathbf{1}_N^\top}{N} - \mathbf{X}_k\right\|_{\mathrm{F}}^2 = \|\mathbf{X}_k(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2 \tag{104}$$

where $\mathbf{I}, \mathbf{J}$ are $N\times N$ matrices. Plugging the inequality (104) into (102), we complete the proof.