A. Additional Experiment Results

As it was mentioned in Section 5 to show the pure effect of redundancy in distributed training we start with a warm-up experiment. In this experiment RI-SGD has only 1 round of communication at the end of training, hence each device does the training locally with a bit of redundancy infused. We choose different redundancy values from $\mu \in \{0.0, 0.05, 0.1, 0.25\}$ and do the training and compare the results with syncSGD. Please note that when $\mu = 0$, our algorithm is equivalant to PR-SGD (Yu et al., 2018). Figure 5 shows the result of this experiment. When we do not have any redundancy, there is a gap between training error of this setting and SyncSGD, which is evident. However, as we add more redundancy this gap is diminished and we can reach smaller error rate way sooner than SyncSGD. That is the most interesting part that we are gaining the speed-up with respect to SyncSGD (almost twice as fast as SyncSGD) with only adding redundancy to the training process. Increasing the redundancy rate would slightly increase the time of training, however, reduces the final error. Hence, as we move from fully synchronous SGD to distributed local SGD we are trading accuracy with speed. On the other hand, redundancy can increase the accuracy, with roughly the same speed-up, thus we can benefit from the advantages of the two settings in our RI-SGD.

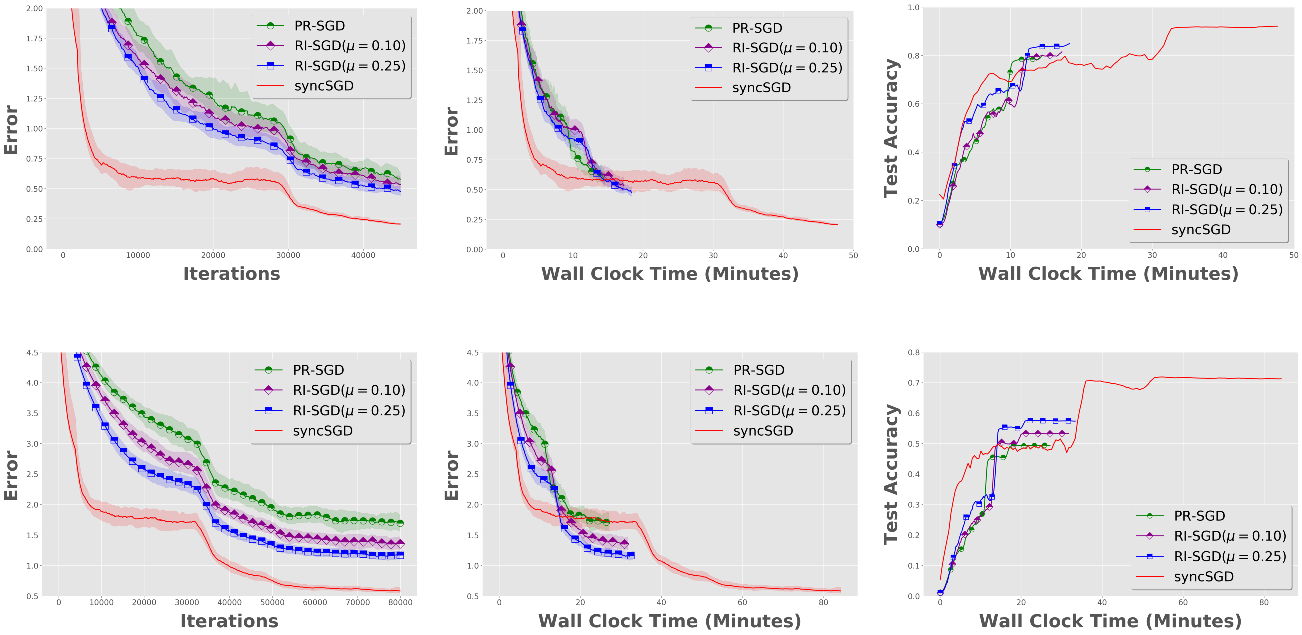
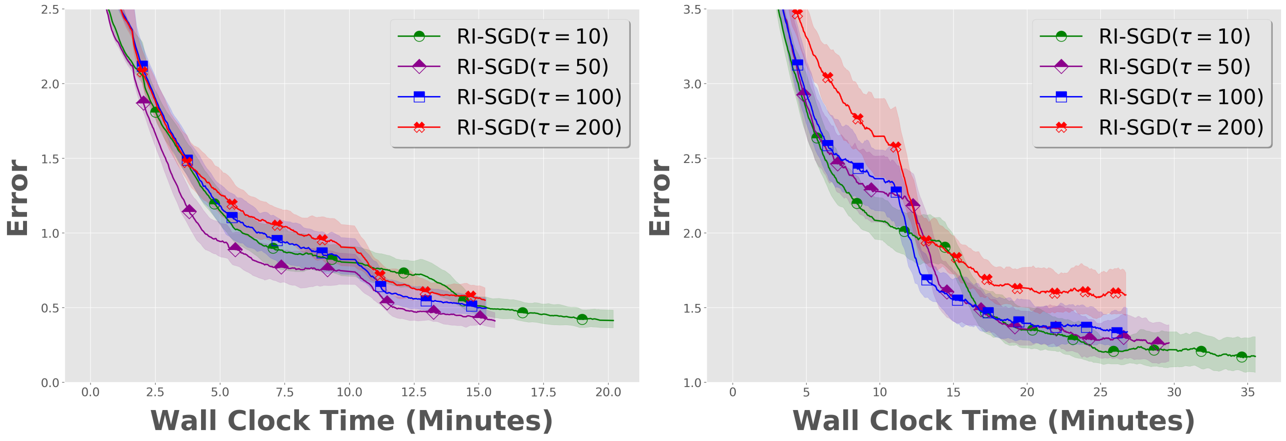


Figure 5. Training error and test accuracy for CIFAR10 and CIFAR100 on ResNet using syncSGD and RI-SGD with $\mu \in \{0.0, 0.1, 0.25\}$ and τ equals to the number of iterations for each experiment, which means each device train on its data locally. Top row is for CIFAR10 dataset, and bottom row is for CIFAR100 dataset.

In the main experiment in Section 5, we need to find the best number of local updates τ. Hence, we keep the redundancy $\mu = 0$, and change the $\tau \in \{10, 50, 100, 200\}$ and compare the trade-off between training time and error rate. Figure 6 shows the results for this experiments, which we can see that for both datasets $\tau = 50$ seems to be an optimum number of local updates for this setup.

In the next experiment, we show how RI-SGD would respond when a node fails. When a node is failing, it means that it cannot respond to other nodes in averaging steps. Since, we are adding redundancy, intuition suggests that RI-SGD is more robust to such a failure in nodes. This is corroborated in Figure 7, where the gap between training error of RI-SGD with $\mu = 0.5$ is infinitesimal compared to this gap when we have no redundancy at all ($\mu = 0$).

In the final experiment, we will investigate the efficacy of RI-SGD on large scale datasets like ImageNet, which includes $1,281,167$ training examples over 1000 classes and $50,000$ test examples. For this dataset, we choose ResNet50 and train the model on 4 and 8 GPUs, with $\tau = 50$ and $\mu \in \{0.125, 0.25\}$ for RI-SGD ($\mu = 0.125$ is only for 8 GPUs setting). The training error depicted in Figure 8, shows that RI-SGD can reach to almost same error rate twice as fast as Sync-SGD on 4 GPUs and three times faster in 8 GPUs setting. Hence, the speedup gain using RI-SGD can significantly impact the training



(a) CIFAR10

(b) CIFAR100

Figure 6. Finding the best number of local updates, τ. We try different numbers from $\tau \in \{10, 50, 100, 200\}$, and find the best one based on the trade-off between error rate and time of execution. For CIFAR10 we choose $\tau = 50$ for both datasets.

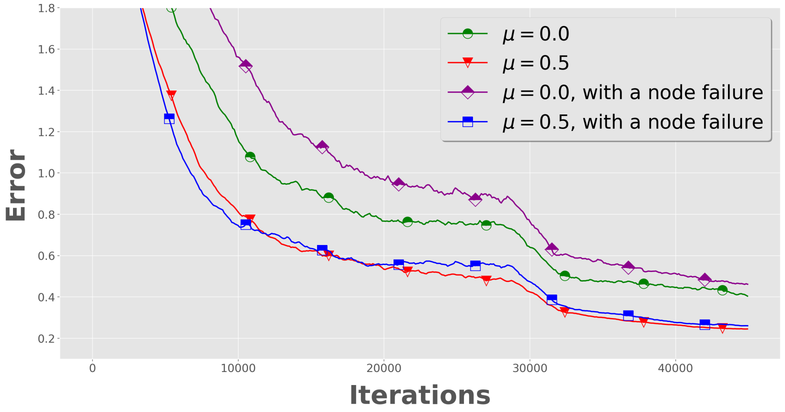


Figure 7. Parallel taining CIFAR10 on ResNet44 with 4 GPUs with and without failure. The failure is when a node is not responding or take infinite time to respond. Hence, it would not not communicate with others in averaging steps. When redundancy presents training is resilient against node failures.

time, while the final accuracy is almost the same.

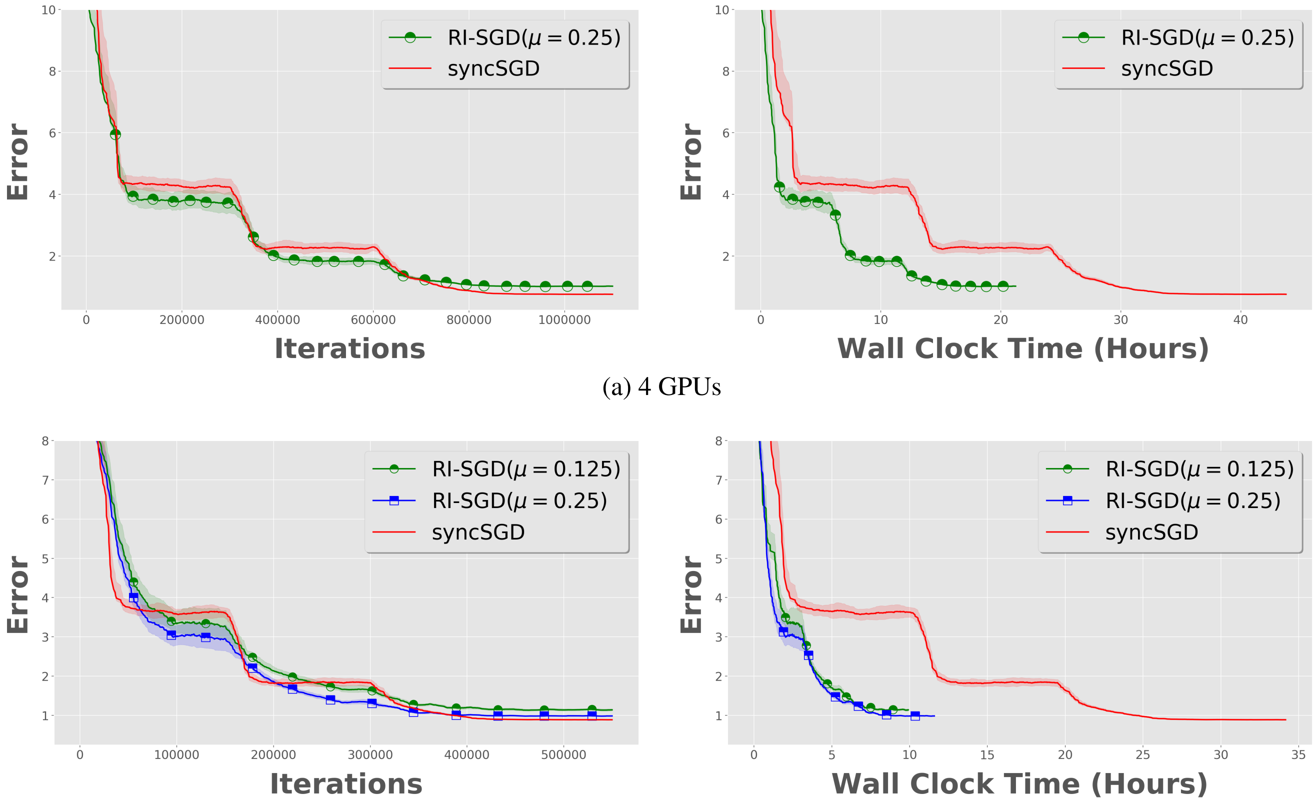
B. Notation

For convenience of the reader we presents all the parameters used in entire this paper at Table 1.

For the clarity in presentation, we define some notations. Let

$$\begin{aligned}
\mathbf{X}^{(t)} &= \begin{bmatrix} \mathbf{x}_1^{(t)} & \dots & \mathbf{x}_p^{(t)} \end{bmatrix} \\
\tilde{\xi}_j^{(t)} &= \{\tilde{\xi}_{j,1}^{(t)}, \dots, \tilde{\xi}_{j,q}^{(t)}\} \\
\tilde{\xi}^{(t)} &= \{\tilde{\xi}_1^{(t)}, \dots, \tilde{\xi}_p^{(t)}\}
\end{aligned}$$

We use notation $\mathbb{E}$ to denote the conditional expectation $\mathbb{E}_{\tilde{\xi}^{(t)}|\mathbf{X}^{(t)}}$. Note that while the data partition $\mathcal{D}_i$ is fix through entire iterations, mini batch samples from data data partition $\mathcal{D}_i$ varies from iteration to iteration and that is why we indicate mini-batch sample with superscript (t) at each iteration, $\xi_{j,i}^{(t)}$. We use the same notation in both tables for the rest of the Appendix.



(a) 4 GPUs

(b) 8 GPUs

Figure 8. Training ImageNet using ResNet50 with $\tau = 50$ and $\mu = 0.25$ for RI-SGD on 4 GPUs (top row), and $\tau = 50$ and $\mu \in \{0.125, 0.25\}$ for RI-SGD on 8 GPUs (bottom row). Comparing the training error of this scheme with Sync-SGD on 4 and 8 GPUs, the gained speedup in RI-SGD is clear (in 4 GPUs more than twice faster, and in 8 GPUs more than three times faster), while the gap between final errors is insignificant.

Table 1. Summary of notations used in the convergence analysis.

p	The number of workers
n	The number of data samples
$F(\mathbf{x})$	The global cost function
$f(\mathbf{x}, \mathcal{D}_i)$	The cost function evaluation over ith data partition
$\mathbf{x}_j^{(i)}$	Model at worker j and iteration i
$\mathbf{x}^{(i)}$	Global model at i-th communication round
$\mathcal{D}_i$	Set of the data points stored at worker i for naive gradient descent
$\tilde{\mathcal{D}}_i$	Set of data samples stored at worker i when there is a redundancy
q	$\lvert\tilde{\mathcal{D}_i}\rvert = \lvert\{\mathcal{D}_{j,1}, \ldots, \mathcal{D}_{j,q}\}\rvert$
$\xi_{j,i}^{(k)}$	Mini-batch sampled from data partition $\mathcal{D}_i$ at iteration k and worker j.
γ	The budget at each worker
η	The learning rate
τ	Number of local updates

Table 2. Notation for gradients

$\mathbf{g}_{j,i}^{(t)} \triangleq \nabla f(\mathbf{x}_j^{(t)}, \mathcal{D}_i),$	$\underline{\mathbf{g}}_{j,i}^{(t)} \triangleq \nabla f(\mathbf{x}_j^{(t)}, \mathcal{D} \setminus \mathcal{D}_i)$
$\mathbf{g}_j^{(t)} \triangleq \sum_{i=1}^q \mathbf{g}_{j,i}^{(t)}$	$\underline{\mathbf{g}}_j^{(t)} \triangleq \sum_{i=1}^q \underline{\mathbf{g}}_{j,i}^{(t)}$
$\tilde{\mathbf{g}}_{j,i}^{(t)} \triangleq \nabla f(\mathbf{x}_j^{(t)}, \xi_i^{(t)}), \xi_i^{(t)} \subset \mathcal{D}$	$\underline{\tilde{\mathbf{g}}}_{j,i}^{(t)} \triangleq \nabla f(\mathbf{x}_j^{(t)}, \xi_i^{(t)}), \xi_i^{(t)} \subset \mathcal{D} \setminus \mathcal{D}_i,$
$\tilde{\mathbf{g}}_j^{(t)} \triangleq \sum_{i=1}^q \tilde{\mathbf{g}}_{j,i}^{(t)}$	$\underline{\tilde{\mathbf{g}}}_j^{(t)} \triangleq \sum_{i=1}^q \underline{\tilde{\mathbf{g}}}_{j,i}^{(t)}$
$\mathbf{g}^{(t)} \triangleq \frac{1}{p} \sum_{j=1}^p \mathbf{g}_j,$	$\underline{\mathbf{g}}^{(t)} \triangleq \frac{1}{p} \sum_{j=1}^p \underline{\mathbf{g}}_j$
$\tilde{\mathbf{g}}^{(t)} \triangleq \frac{1}{p} \sum_{j=1}^p \tilde{\mathbf{g}}_j$	$\underline{\tilde{\mathbf{g}}}^{(t)} \triangleq \frac{1}{p} \sum_{j=1}^p \underline{\tilde{\mathbf{g}}}_j$
$\tilde{\mathcal{G}}_j^{(t)} \triangleq \sum_{i=1}^q \lVert\tilde{\mathbf{g}}_{j,i}^{(t)}\rVert^2$	$\underline{\tilde{\mathcal{G}}}_j^{(t)} \triangleq \sum_{i=1}^q \lVert\underline{\tilde{\mathbf{g}}}_{j,i}^{(t)}\rVert^2$
$\mathcal{G}_j^{(t)} \triangleq \sum_{i=1}^q \lVert\mathbf{g}_{j,i}^{(t)}\rVert^2$	$\underline{\mathcal{G}}_j^{(t)} \triangleq \sum_{i=1}^q \lVert\underline{\mathbf{g}}_{j,i}^{(t)}\rVert^2$
$\tilde{\mathbf{G}}_j^{(t)} \triangleq \min_i \lVert\underline{\mathbf{g}}_{j,i}^{(t)}\rVert^2$	$\tilde{\mathbf{G}}^{(t)} \triangleq \sum_{j=1}^p \tilde{\mathbf{G}}_j^{(t)}$

C. Detailed Comparison with Related Schemes

In this section, we compare our results with existing algorithms in terms of communication cost, convergence error as well as data budget. First of all note that the number of communication rounds is $\frac{T}{\tau}$. Fixing learning rate to be $\eta = \frac{\sqrt{p}}{L\sqrt{T}}$, in Parallel restarted-SGD Algorithm (PR-SGD) (Yu et al., 2018) convergence rate $O(\frac{1}{\sqrt{pT}})$ can be achieved with $\tau_{\text{PR-SGD}} = \frac{T^{0.25}}{p^{0.75}}$, while for our algorithm, provided with sufficient data redundancy, the same rate can be achieved with $\tau_{\text{RI-SGD}} = \frac{\sqrt{T}}{p^{1.5}}$. Therefore, for usual case of $T > p$, since $\frac{T}{\tau_{\text{RI-SGD}}} \leq \frac{T}{\tau_{\text{PR-SGD}}}$, our algorithm achieves the same convergence rate with less number of communication rounds, thus faster. Additionally, our convergence analysis is based on the bounding assumption over inner product of partial gradients which is weaker assumption than uniformly bounded gradient assumption.

Comparing with (Wang & Joshi, 2018), even though we make the additional Assumption of bounded inner product between different partial gradient, the main advantage of our work over (Wang & Joshi, 2018) is that each work, samples mini-batch from its own data which is the usual framework in the distributed systems. Furthermore, our algorithm is also extended to analyze the heterogeneous mini-batch samples framework. The results corresponding to the convergence analysis of various algorithms can be seen in Table. 3. In this table $\tilde{\mathbf{G}}_k^{(t)} \triangleq \min_{\mathcal{D}_i \in \mathcal{D} \setminus \tilde{\mathcal{D}}_k} \|\mathbf{g}_{k,i}^{(t)}\|^2$ and $\tilde{\mathbf{G}}^{(t)} \triangleq \frac{1}{p} \sum_{k=1}^p \tilde{\mathbf{G}}_k^{(t)}$. Algorithm Parallel Restarted SGD in Table .3 is based on the Assumption $\mathbb{E}_{\xi_{j,i}} \left[\|\hat{\mathbf{g}}_{j,i} - \mathbf{g}_{j,i}\|^2 |\mathbf{x}_j\right] \leq \sigma^2$, $\mathbb{E}_{\xi_{j,i}} \left[\|\hat{\mathbf{g}}_{j,i}\|^2 |\mathbf{x}_j\right] \leq \Delta$ and $\frac{1}{p} \leq \gamma \leq 1$.

Table 3. Comparison of different SGD based algorithms.

Strategy	Convergence error	Data Access
SGD	$\frac{2\left[F(\mathbf{x}^{(1)})-F^*\right]}{\eta T}+\frac{\eta LC_2^2}{p}$	$\mathcal{D}$
(Yu et al., 2018; Zhou & Cong, 2017)	$\frac{2\left[F(\mathbf{x}^{(1)})-F^*\right]}{\eta T}+\frac{\eta L\sigma^2}{p}+\left[4\eta^2\tau^2L^2\Delta\right]$	$\mathcal{D}_j, 1\leq j\leq p$
(Wang & Joshi, 2018)	$\frac{2\left[F(\mathbf{x}^{(1)})-F^*\right]}{\eta T}+\frac{\eta LC_2^2}{p}+\left[\eta^2L^2C_2^2(\tau-1)\right]$	$\mathcal{D}$
RI-SGD(τ)	$\frac{2\left[F(\mathbf{x}^{(1)})-F^*\right]}{\eta T}+\frac{\eta LC_2^2\gamma}{p}+\frac{p^2\eta}{2}(1-\gamma)\beta+(\frac{p+1}{p})\gamma\eta^2L^2(\tau-1)C_2^2$	$\tilde{\mathcal{D}}_j, 1\leq j\leq p$

D. Heterogeneous Mini-batch Samples

In many applications/networks, workers may have asymmetric and non-uniform data size (e.g., Federated Averaging (FedAvg) (McMahan et al., 2016)). Our algorithm and results apply for this case as well. We assume that the jth worker has q_j data chunks denoted by $\tilde{\mathcal{D}}_j=\{\mathcal{D}_{1,j},\dots,\mathcal{D}_{1,q_j}\}$. Therefore at each iteration jth worker node evaluates gradient over q_j mini-batches from its available data chunks. We denote the updating rule based on heterogeneous mini-batch data with RI-SGD$(\tau,q_1,\dots,q_p)$ which for local updates is same as (3) with the difference that $\tilde{\mathbf{g}}_j^{(t)}\triangleq\sum_{i=1}^{q_j}\tilde{\mathbf{g}}_{j,i}^{(t)}$ and similar averaging step to Algorithm 1.

The following theorem shows the convergence rate for heterogeneous sample size setting.

Theorem D.1. *For HRI-SGD$(\tau,q_1,\dots,q_p)$, under Assumptions 1-4, where $q_j\leq p$ if the learning rate satisfies $\max_{1\leq j\leq p}\left(\eta^2L^2(2\tau C_1+\tau(\tau-1))+\frac{\eta L(C_1+q_jp)}{p}\right)\leq 1$ and all local model parameters are initialized at the same point $\mathbf{x}^{(1)}$, the average-squared gradient after τ iterations is bounded as follows*

$$
\begin{aligned}
\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\Big]\leq{}&\frac{2\big[F(\mathbf{x}^{(1)})-F^*\big]}{\eta T}+\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}\frac{q_j-p}{p}\tilde{\mathbf{G}}_j^{(t)}\\
&+\frac{\beta}{2}\Big[\frac{1}{p}\sum_{j=1}^{p}(3q_j^2-2pq_j+p^2-p)+2\sum_{j=1}^{p}(p-q_j)\Big]\\
&+(\frac{p+1}{p})\sum_{j=1}^{p}\frac{\eta^2L^2q_j}{p^2}(\tau-1)C_2^2+\sum_{j=1}^{p}\frac{L\eta q_jC_2^2}{p^3}
\end{aligned}
$$

Remark 3. *The heterogeneous data allocation policy also includes data allocation over a given network in which workers are connected to other worker nodes based on the connectivity of underlying graph, and workers can only share their data with their neighbours in the graph.*

E. Redundancy and Intra-gradient Diversity

As we discussed in the convergence of RI-SGD, by inducing redundancy among workers, we can effectively reduce the variance in local updates and reduce the number of communications, while enjoying faster convergence. In this section, we investigate the effectiveness of redundancy infusion from gradient diversity perspective. Specifically, we show that infusing redundancy can increase the *intra-node* gradient diversity. A direct implication is that the effective size of mini-batch can be increased without any saturation or decay in performance.

Let us first review the definitions of gradient diversity and batch-bound that are introduced in (Yin et al., 2018) to understand the speedup saturation in distributed optimization. We then elaborate how redundancy affects these quantities.

Definition 1 (Gradient Diversity)**.** *Let $\mathbf{x}$ denote a fixed solution. The gradient diversity of objective at point $\mathbf{x}$ over the data set $\mathcal{D}$ is defined as:*

$$
\Delta_{\mathcal{D}}(\mathbf{x})\triangleq\frac{\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x})\|_2^2}{\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x})\|_2^2+\sum_{i\neq j}\langle\nabla f_i(\mathbf{x}),\nabla f_j(\mathbf{x})\rangle}
$$

The gradient diversity can be used to measure the dissimilarity between concurrent gradient updates. We note that when the partial gradients are orthogonal, the gradient diversity becomes larger. Having the definition of gradient diversity in place, we define the batch size bound at any point $\mathbf{x}$ with respect to training data $\mathcal{D}$ as follows.

Definition 2 (Batch-size Bound)**.** *Batch size bound is defined as* $\mathsf{B}_{\mathcal{D}}(\mathbf{x}) \triangleq n\Delta_{\mathcal{D}}(\mathbf{x})$.

A direct implication of batch size bound is that the effective mini-batch size at intermediate solutions is proportional to to gradient diversity at those solutions. In (Yin et al., 2018) it is shown that for both convex and non-convex optimization problems if the batch size satisfies $B \leq 1 + \delta_0 \mathsf{B}_{\mathcal{D}}(\mathbf{x})$ for some constant δ_0, saturation of performance can be avoided. Then, the key challenge to avoid saturation in mini-batch SGD for a *fixed* batch size is to introduce diversity-inducing mechanisms. Here we demonstrate that infusing redundancy results in infusing diversity. We note that the main distinguishing feature between our algorithm and (Yin et al., 2018) is that since our algorithm performs local SGD, the choice of mini-batch must be related to the *intra-node* gradient diversity, i.e., the gradient diversity of each worker's local data, rather than (inter) gradient diversity. We make the statement precise in the following result.

Lemma E.1. *Suppose, the total number of data points at each data chunk is* n, $|\mathcal{D}_j| = n$, *and suppose that for all* $1 \leq j \leq p$, *the norm of gradients at each chunk satisfies* $\|\nabla f(\mathbf{x}, \mathcal{D}_j)\|_2^2 = G$. *Also, assume the following conditions hold on the pairwise correlation of gradients in each individual chunk and each pair of chunks:*

$$\sum_{i \neq j, \{D_i, D_j\} \in \mathcal{D}_j} \Big\langle \nabla f(\mathbf{x}, D_i), \nabla f(\mathbf{x}, D_j) \Big\rangle = H,$$

$$\sum_{i \neq j, D_i \in \mathcal{D}_k, D_j \in \mathcal{D}_l, \mathcal{D}_k \neq \mathcal{D}_l} \Big\langle \nabla f(\mathbf{x}, D_i), \nabla f(\mathbf{x}, D_j) \Big\rangle = J.$$

Then the following holds on the batch size bound

$$\mathsf{B}_{\tilde{\mathcal{D}}_j}(\mathbf{x}) \geq \mathsf{B}_{\mathcal{D}_j}(\mathbf{x})\ ,\ 1 \leq j \leq p \tag{9}$$

as long as the observed partial gradient vectors over various data chunks are almost orthogonal, or more precisely if $G + H \geq \frac{n^2}{2} J$.

Proof. By expanding the definition of gradient diversity for all chunks of each worker, it follows that

$$\begin{aligned}
\mathsf{B}_{\tilde{\mathcal{D}}_j}(\mathbf{x}) &\overset{\text{①}}{=} nq\frac{qG}{qG + qH + \binom{q}{2}n^2 J} \\
&= nq\frac{G}{G + H + \frac{q-1}{2}n^2 J} \\
&\overset{\text{②}}{\geq} n\frac{G}{G + H} \qquad (10) \\
&= \mathsf{B}_{\mathcal{D}_j}(\mathbf{x}) \qquad (11)
\end{aligned}$$

where in ① we used the fact the number of training that at each worker is nq and the denominator in the definition of gradient diversity $\Delta_{\tilde{\mathcal{D}}_j}(\mathbf{x})$ can be decomposed into three terms (norm of gradients of all individual chunks, pairwise correlation of gradients of all individual chunks, and pairwise correlation of gradients of all pairs of q chunks), and ② follows from $G + H \geq \frac{n^2}{2} J$. □

An immediate implication of about lemma is that while increasing the mini-batch size does not effect the batch-size bound, adding redundancy can increase the intra-gradient diversity– allowing to use larger mini-batch sizes. This claim is verified empirically in Figures 3 and 4 in which for both experiment we have used equal size mini-batch, in which it is shown that adding redundancy increased the batch-sized bound and thus leads to improved performance.

Remark 4. *As it pointed out in (Lin et al., 2018), local updates can be seen as updating model with structured noise at each iteration. Therefore, as discussed in (Yin et al., 2018) adding independent noise could also improve the gradient diversity. As a consequence, the RI-SGD algorithm can benefit from higher gradient diversity due to factors: 1) Increasing intra-gradient diversity because of the infused redundancy. 2) Noisy model update due to local updates.*

F. Distribution-aware Convergence

As alluded to in the convergence of RI-SGD, the convergence rates based on Assumption 4 where we consider an upper bound β for the correlation between each pair of local full gradients of each data chunk $\mathbf{g}_{j,i}$ do not clearly reflect the effects of each pair individually. Here we provide an improved analysis of convergence rate that explicitly reflects the impact of distribution of data chunks on the final bound based on the correlation of each pair independently. To this end, we restate the assumption we made about the correlation of gradients of pair of data chunks for completeness:

Assumption 6. *For $1 \leq j \leq p$ we can have an upper bound on the inner products of gradients of each worker's data chunk and data chunks not presented in that worker as follows:*

$$|\langle \mathbf{g}_j, \underline{\mathbf{g}}_{j,i} \rangle| \leq \beta_{j,i}, \tag{12}$$

where $\mathbf{g}_j = \nabla f(\cdot, \mathcal{D}_{j,q})$ is gradient of the main data chunk at the jth worker and $\underline{\mathbf{g}}_{j,i} = \nabla f(\cdot, \mathcal{D}_i), \mathcal{D}_i \notin \tilde{\mathcal{D}}_j$ is the gradient of data chunk not present in the jth worker.

We note that by setting $\max_{j,i} \beta_{j,i} = \beta$, we can get the results which only depends on the number of mini-batches rather than the chunks.

Now we derive the convergence of RI-SGD based on above assumption.

Corollary F.0.1. *For RI-SGD(τ), if for all $1 \leq t \leq T$, there exists constants $\rho^{(t)} > 0$ such that $\sum_{j=1}^{p} \| \sum_i \mathbf{g}_{j,i}^{(t)} + \sum_i \underline{\mathbf{g}}_{j,i}^{(t)} \|^2 \geq \rho^{(t)} \sum_{j=1}^{p} \Big(\sum_i \|\mathbf{g}_{j,i}^{(t)}\|^2 \Big)$, under Assumptions 1-3, and if the learning rate satisfies $0 < \eta^2 L^2 \Big(2\tau C_1(\frac{p+1}{p}) + \tau(\tau-1)\Big) + \frac{\eta L (C_1 + \gamma p^2)}{p} < \rho^{(t)}$ for $1 \leq t \leq T$ and all local model parameters are initialized at the same point $\bar{\mathbf{x}}^{(1)}$, for the given budget $\frac{1}{p} \leq \gamma \leq 1$ the average-squared gradient after τ iterations is bounded as follows:*

$$\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\Big] \leq \frac{2\big[F(\bar{\mathbf{x}}^{(1)}) - F^*\big]}{\eta T} + \eta L C_2^2 \frac{\gamma}{p} + (\frac{p+1}{p})\gamma \eta^2 L^2 C_2^2 (\tau - 1) + \frac{2}{p}\sum_{j=1}^{p}\sum_{i=1}^{p-q}\sum_{l=1}^{p} \beta_{j,i} \tag{13}$$

Compared to previous convergence rates, the bound in Eq. (13) indicates an explicit dependency of the rate to the gradient dissimilarity between different partitions of the data. Interestingly, the final term in Eq. (13) can be interpreted in two equivalent ways that makes an intimate connection with diversity-infusing viewpoint of RI-SGD discussed before. First, it indicates that the error decreases as the inter-node gradient diversity reduces. Alternatively, it demonstrates that the error decreases as the *intra-node* gradient diversity increases. For a given data set/partition, redundancy increases the intra-node gradient-diversity, and conversely reduces inter-node gradient diversity. This is how redundancy reduces the overall error.

Remark 5. *Increasing the mini-batch size rather than having redundant data leads to lower diversity and would lead to a larger error in comparison. Our analysis can be used to examine this case as well, e.g, storing q copies of Di at node i is equivalent to larger minbatch sizes. Our analysis in (13) shows that this has lower diversity as compared to storing q different data chunks. The error saturation with larger mini-batches has been observed in (Lin et al., 2018), and is corroborated both by our analysis, as well as the experiment in Fig. 3.*

G. Proof of Convergence Rate

Here we show that, when the redundancy is full, our algorithm achieves linear speedup in terms of number of workers. To see this, we note that by setting $\eta = \frac{\sqrt{p}}{L\sqrt{T}}$, for $\tau \leq (\frac{\sqrt{T}}{p^{1.5}} + 1)$ in Theorem 4.2, and by choosing $\gamma = 1 - \frac{1}{p^{2.5}\sqrt{T}}$ the average convergence rate becomes as:

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big]\Big] &\leq \frac{2\big[F(\bar{\mathbf{x}}^{(1)})-F^*\big]}{\eta T}+\eta L C_2^2\frac{\gamma}{p}+(\frac{p+1}{p})\gamma\eta^2L^2(\tau-1)+2p^2\Big[(1-\gamma)\Big]\beta \\ &= \frac{2L[F(\bar{\mathbf{x}})-F^*]}{\sqrt{pT}}+\frac{C_2^2}{\sqrt{pT}}(1-\frac{1}{p^{2.5}\sqrt{T}})+2p^2\frac{1}{p^{2.5}\sqrt{T}}\beta \\ &\quad+(\frac{p+1}{p}(1-\frac{1}{p^{2.5}\sqrt{T}}))L^2(\frac{p}{L^2T})(\frac{\sqrt{T}}{p^{1.5}}) \\ &= O(\frac{1}{\sqrt{pT}})\end{aligned}$$

Note that this assumption is based on the condition of existence of $\rho^{(t)} > 0$ and the corresponding condition that $\eta = \frac{\sqrt{p}}{L\sqrt{T}}$ satisfies.

H. Proof of Theorem 4.1

Our proof is based on Lipschitz continuous gradient assumption, which gives us

$$\mathbb{E}[F(\bar{\mathbf{x}}^{(t+1)})]-F(\bar{\mathbf{x}}^{(t)}) \leq -\eta\mathbb{E}\big[\big\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\tilde{\mathbf{g}}^{(t)}\big\rangle\big]+\frac{\eta^2L}{2}\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}\|^2\Big] \tag{14}$$

By taking the average of above inequality for all iterations, we get

$$\frac{1}{T}\sum_{t=1}^{T}\Big[\mathbb{E}[F(\bar{\mathbf{x}}^{(t+1)})]-F(\bar{\mathbf{x}}^{(t)})\Big] \leq \frac{1}{T}\sum_{t=1}^{T}\big(-\eta\mathbb{E}\big[\big\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\tilde{\mathbf{g}}^{(t)}\big\rangle\big]\big)+\frac{1}{T}\sum_{t=1}^{T}\frac{\eta^2L}{2}\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}\|^2\Big] \tag{15}$$

We prove Theorem 4.1 by bounding components in (15) using the following Lemmas:

Lemma H.1. *Under Assumptions 1,2 and 3, we have the following variance bound from the averaged stochastic gradient:*

$$\mathbb{E}_t\Big[\|\tilde{\mathbf{g}}^{(t)}-\mathbf{g}^{(t)}\|^2\Big] \leq \frac{C_1}{p^2}\sum_{j=1}^{p}\mathcal{G}_j^{(t)}+\frac{q}{p^2}C_2^2 \tag{16}$$

where $\mathcal{G}_j^{(t)} \triangleq \sum_{i=1}^{q}\|\mathbf{g}_{j,i}^{(t)})\|^2$.

Using Lemma H.1 we can bound the $\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}\|^2\Big]$ as follows:

Lemma H.2. *Under Assumptions 1 to 4, the squared norm of stochastic gradient can be bounded as*

$$\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}\|^2\Big] \leq (\frac{C_1+qp}{p^2})\sum_{j=1}^{p}\mathcal{G}_j^{(t)}+\frac{qC_2^2}{p^2} \tag{17}$$

Corollary H.2.1. *From Lemma H.2, by summing over all steps, we conclude that:*

$$\frac{\eta^2L}{2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}\|^2\Big] \leq \frac{\eta^2L}{2}(\frac{C_1+qp}{p^2})\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}\mathcal{G}_j^{(t)}+\frac{\eta^2L}{2}\frac{qC_2^2}{p^2} \tag{18}$$

Lemma H.3. *Under Assumptions 1 to 4 we have:*

$$\frac{L^2}{pT}\sum_{t=1}^{T}\mathbb{E}\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}_j^2\|^2 \leq \frac{\eta^2L^2}{pT}\Big[\Big(2\tau C_1(\frac{p+1}{p})+\tau(\tau-1)\Big)\sum_{t=1}^{T}\sum_{j=1}^{p}\mathcal{G}_j^{(t)}+(\frac{p+1}{p})qT(\tau-1)C_2^2\Big] \tag{19}$$

Lemma H.4. *Under Assumptions 2, and according to the data allocation policy of Algorithm 1 the expected inner product between stochastic gradient and full batch gradient can be expanded as:*

$$\begin{aligned}-\eta\mathbb{E}\Big[\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\tilde{\mathbf{g}}^{(t)}\rangle\Big] \leq& -\frac{\eta}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2}\frac{1}{p}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} - \frac{\eta(1-\gamma)}{2}\tilde{\mathbf{G}}^{(t)}\\ &+ \frac{L^2}{2p}\mathbb{E}\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}_j^{(t)}\|^2 + \frac{p^2\eta}{2}\Big(\Big[3\gamma^2-4\gamma+(3-\frac{1}{p})\Big]\beta\Big)\end{aligned} \tag{20}$$

where $\tilde{\mathbf{G}}_j^{(t)} \triangleq \min_i \|\underline{\mathbf{g}}_{j,i}^{(t)}\|^2$ *and* $\tilde{\mathbf{G}}^{(t)} \triangleq \sum_{j=1}^{p}\tilde{\mathbf{G}}_j^{(t)}$.

Corollary H.4.1. *From Lemma H.4 and by taking summation over the iterations we have:*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}-\eta\mathbb{E}\Big[\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\tilde{\mathbf{g}}^{(t)}\rangle\Big] \leq& -\frac{\eta}{2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2}\frac{1}{p}\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} - \frac{\eta(1-\gamma)}{2}\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{G}}^{(t)}\\ &+ \frac{\eta^3L^2}{2pT}\Big[\Big(2\tau C_1(\frac{p+1}{p})+\tau(\tau-1)\Big)\sum_{t=1}^{T}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} + qT(\frac{p+1}{p})(\tau-1)C_2^2\Big]\\ &+ \frac{p^2\eta}{2}\Big(\Big[3\gamma^2-4\gamma+(3-\frac{1}{p})\Big]\beta\Big)\end{aligned} \tag{21}$$

Final step:

By plugging all the Lemmas in (15), we get:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\Big[\mathbb{E}[F(\bar{\mathbf{x}}^{(t+1)})] - F(\bar{\mathbf{x}}^{(t)})\Big] \leq& \frac{1}{T}\sum_{t=1}^{T}\big(-\eta\mathbb{E}\big[\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\tilde{\mathbf{g}}^{(t)}\rangle\big]\big) + \frac{1}{T}\sum_{t=1}^{T}\frac{\eta^2L}{2}\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}\|^2\Big]\\ \leq& -\frac{\eta}{2}\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big]\Big] - \frac{\eta}{2}\frac{1}{p}\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} - \frac{\eta(1-\gamma)}{2}\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{G}}^{(t)}\\ &+ \frac{\eta^3L^2}{2pT}\Big[\Big(2\tau C_1(\frac{p+1}{p})+\tau(\tau-1)\Big)\sum_{t=1}^{T}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} + q(\frac{p+1}{p})\gamma(\tau-1)C_2^2\Big]\\ &+ \frac{p^2\eta}{2}\Big(\Big[3\gamma^2-4\gamma+(3-\frac{1}{p})\Big]\beta + (\frac{\eta^2L(C_1+qp)}{2p^2})\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} + \frac{L\eta^2qC_2^2}{2p^2}\\ =& -\frac{\eta}{2}\big[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big]\big] - \frac{\eta(1-\gamma)}{2}\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{G}}^{(t)}\\ &+ \frac{p^2\eta}{2}\Big(\Big[3\gamma^2-4\gamma+(3-\frac{1}{p})\Big]\beta\\ &+ \frac{\eta}{2pT}\Big[\eta^2L^2\Big(2\tau C_1(\frac{p+1}{p})+\tau(\tau-1)\Big) - 1 + \frac{\eta L(C_1+qp)}{p}\Big]\sum_{t=1}^{T}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} \qquad (22)\\ &+ \frac{\eta^3L^2\gamma}{2}(\frac{p+1}{p})(\tau-1)C_2^2\\ \overset{\textcircled{1}}{\leq}& -\frac{\eta}{2}\big[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big]\big] - \frac{\eta(1-\gamma)}{2}\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{G}}^{(t)}\\ &+ \frac{\eta^3L^2\gamma}{2}(\frac{p+1}{p})(\tau-1)C_2^2 + \frac{p^2\eta}{2}\Big(\Big[3\gamma^2-4\gamma+(3-\frac{1}{p})\Big]\beta + \frac{L\eta^2\gamma C_2^2}{2p} \qquad (23)\end{aligned}$$

where ① comes from the learning rate choice of

$$\eta^2 L^2\Big(2\tau C_1(\frac{p+1}{p})+\tau(\tau-1)\Big)+\frac{\eta L(C_1+\gamma p^2)}{p}\leq 1$$

By rearranging (23) we get:

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big]\Big] &\leq \frac{2\big[F(\bar{\mathbf{x}}^{(1)})-F^*\big]}{\eta T}+\eta L C_2^2\frac{\gamma}{p}-(1-\gamma)\frac{1}{T}\sum_{k=1}^{T}\tilde{\mathbf{G}}^{(t)}+\gamma(\frac{p+1}{p})\eta^2L^2(\tau-1)\\ &+\frac{p^2\eta}{2}\Big(\Big[3\gamma^2-4\gamma+(3-\frac{1}{p})\Big]\beta\end{aligned}\tag{24}$$

I. Proof of Lemmas

I.1. Proof of Lemma.H.1

Proof. Following the definitions in Table 2, we can write

$$\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}-\mathbf{g}^{(t)}\|^2\Big]\overset{\textcircled{1}}{=}\mathbb{E}\Big[\|\frac{1}{p}\sum_{j=1}^{p}\sum_{i}\Big(\tilde{\mathbf{g}}_{j,i}^{(t)}-\mathbf{g}_{j,i}^{(t)}\Big)\|^2\Big]\tag{25}$$

$$=\frac{1}{p^2}\mathbb{E}\Big[\sum_{j=1}^{p}\sum_{i}\|\Big(\tilde{\mathbf{g}}_{j,i}^{(t)}-\mathbf{g}_{j,i}^{(t)})\Big)\|^2\tag{26}$$

$$+\sum_{i\neq j\,\vee\,l\neq m}\Big\langle\tilde{\mathbf{g}}_{i,l}^{(t)}-\mathbf{g}_{i,l}^{(t)}),\tilde{\mathbf{g}}_{j,m}^{(t)}-\mathbf{g}_{j,m}^{(t)}\Big\rangle\Big]\tag{27}$$

$$=\frac{1}{p^2}\Big[\sum_{j=1}^{p}\sum_{i\in\tilde{\mathcal{D}}_j}\mathbb{E}_{\xi_{j,i}^{(t)}|\mathbf{X}^{(t)}}\|\Big(\tilde{\mathbf{g}}_{j,i}^{(t)}-\mathbf{g}_{j,i}^{(t)}\Big)\|^2\tag{28}$$

$$+\sum_{i\neq j\,\vee\,l\neq m}\mathbb{E}\big[\langle\tilde{\mathbf{g}}_{j,l}^{(t)}-\mathbf{g}_{j,l}^{(t)},\tilde{\mathbf{g}}_{i,m}^{(t)}-\mathbf{g}_{i,m}^{(t)}\rangle\big]\Big]\tag{29}$$

$$\overset{\textcircled{2}}{=}\frac{1}{p^2}\Big[\sum_{j=1}^{p}\sum_{i}\mathbb{E}_{\xi_{j,i}^{(t)}|\mathbf{X}^{(t)}}\|\Big(\tilde{\mathbf{g}}_{j,i}^{(t)}-\mathbf{g}_{j,i}^{(t)}\Big)\|^2\tag{30}$$

$$+\sum_{i\neq j\,\vee\,l\neq m}\Big\langle\mathbb{E}_{\xi_{j,i}^{(t)}|\mathbf{X}_l^{(t)}}\big[\tilde{\mathbf{g}}_{j,i}^{(t)}-\mathbf{g}_{j,i}^{(t)}\big],\mathbb{E}_{\xi_{j,i}^{(t)}|\mathbf{X}_m^{(t)}}\big[\tilde{\mathbf{g}}_{m,l}^{(t)}-\mathbf{g}_{m,l}^{(t)}\big]\rangle\tag{31}$$

$$\overset{\textcircled{3}}{\leq}\frac{1}{p^2}\Big[\sum_{j=1}^{p}\sum_{i\in\tilde{\mathcal{D}}_j}\Big[C_1\|\mathbf{g}_{j,i}^{(t)}\|^2+\frac{C_2^2}{p}\Big]\tag{32}$$

$$=\frac{1}{p^2}\sum_{j=1}^{p}\sum_{i\in\tilde{\mathcal{D}}_j}C_1\|\mathbf{g}_{j,i}^{(t)}\|^2+\frac{1}{p^2}|\tilde{\mathcal{D}}_j^{(t)}|C_2^2\tag{33}$$

$$\overset{\textcircled{4}}{=}\frac{1}{p^2}\sum_{j=1}^{p}\sum_{i\in\tilde{\mathcal{D}}_j}C_1\|\mathbf{g}_{j,i}^{(t)}\|^2+\frac{q}{p^2}C_2^2\tag{34}$$

where in ① we use the definition of $\tilde{\mathbf{g}}^t$ and $\mathbf{g}^t$, in ② we use the fact that mini batches are chosen in i.i.d. manner, and ③ immediately follows from Assumptions 1 and 3. Finally, ④ is due to the symmetric data allocation policy in Algorithm 1. □

I.2. Proof of Lemma.H.2

Proof. Assumption 2 implies that $\mathbb{E}[\tilde{\mathbf{g}}^{(t)}] = \mathbf{g}^{(t)}$, then we have

$$
\begin{aligned}
\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}\|^2\Big] &= \mathbb{E}\big[\|\tilde{\mathbf{g}}^t - \mathbb{E}\big[\tilde{\mathbf{g}}^t\big]\|^2\Big] + \|\mathbb{E}\big[\tilde{\mathbf{g}}^t\big]\|^2 && (35)\\
&= \mathbb{E}\big[\|\tilde{\mathbf{g}}^t - \mathbf{g}^t\big]\|^2\Big] + \|\mathbf{g}^t\|^2 && (36)\\
&\le \frac{1}{p^2}\sum_{j=1}^{p}\sum_{i} C_1\|\mathbf{g}_{j,i}^{(t)}\|^2 + \frac{q}{p^2}C_2^2 + \|\mathbf{g}^t\|^2 && (37)\\
&= \frac{1}{p^2}\sum_{j=1}^{p}\sum_{i} C_1\|\mathbf{g}_{j,i}^{(t)}\|^2 + \frac{q}{p^2}C_2^2 + \|\frac{1}{p}\sum_{j=1}^{p}\sum_{i}\mathbf{g}_{j,i}^{(t)}\|^2 && (38)\\
&\overset{①}{\le} \frac{C_1}{p^2}\sum_{j=1}^{p}\sum_{i}\|\mathbf{g}_{j,i}^{(t)}\|^2 + \frac{q}{p^2}C_2^2 + \frac{1}{p}\sum_{j=1}^{p}\|\sum_{i}\mathbf{g}_{j,i}^{(t)}\|^2 && (39)\\
&\overset{②}{\le} (\frac{C_1}{p^2})\sum_{j=1}^{p}\sum_{i}\|\mathbf{g}_{j,i}^{(t)})\|^2 + \frac{q}{p^2}C_2^2 + \frac{1}{p}\sum_{j=1}^{p}|\tilde{\mathcal{D}}_j|\sum_{i}\|\mathbf{g}_{j,i}^{(t)}\|^2 && (40)\\
&= (\frac{C_1 + qp}{p^2})\sum_{j=1}^{p}\sum_{i}\|\mathbf{g}_{j,i}^{(t)}\|^2 + \frac{qC_2^2}{p^2} && (41)
\end{aligned}
$$

Where ① and ② follows from the fact that $\|\sum_{i=1}^{m}\mathbf{a}_i\|^2 \le m\sum_{i=1}^{m}\|\mathbf{a}_i\|^2$ where $\mathbf{a}_i \in \mathbb{R}^n$. □

I.3. Proof of LemmaH.4

Proof. In order to prove Lemma H.4, we need the following results.

Lemma I.1. *Under Assumptions 2, and according to the data allocation policy of Algorithm.1 the expected inner product between stochastic gradient and full batch gradient can be expanded as:*

$$
\begin{aligned}
\mathbb{E}\Big[\langle\nabla F(\bar{\mathbf{x}}^{(t)}), \tilde{\mathbf{g}}^t\rangle\Big] = \frac{1}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] + \frac{1}{2}\frac{1}{p}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} + \frac{1}{2}\frac{1}{p}\sum_{j=1}^{p}\underline{\mathcal{G}}_j^{(t)}\\
+ \frac{1}{2p}\sum_{j}\Big[\sum_{i\neq m}\Big\langle\mathbf{g}_{j,i}^{(t)}, \mathbf{g}_{j,m}^{(t)}\Big\rangle + \sum_{i\neq m}\Big\langle\underline{\mathbf{g}}_{j,i}^{(t)}, \underline{\mathbf{g}}_{j,m}^{(t)}\Big\rangle + \sum_{i,m}\Big\langle\mathbf{g}_{j,i}^{(t)}, \underline{\mathbf{g}}_{j,m}^{(t)}\Big\rangle\Big]\\
- \frac{1}{2p}\sum_{j=1}^{p}\|\nabla F(\bar{\mathbf{x}}^{(t)}) - \nabla F(\mathbf{x}_j^{(t)})\|^2 - \frac{1}{p}\sum_{j=1}^{p}\sum_{i}\Big\langle\nabla F(\bar{\mathbf{x}}^{(t)}), \underline{\mathbf{g}}_{j,i}^{(t)}\Big\rangle \quad (42)
\end{aligned}
$$

Proof.

$$\begin{aligned}
\mathbb{E}\Big[\Big\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \tilde{\mathbf{g}}^{(t)}\Big\rangle\Big] &= \mathbb{E}\Big[\Big\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \frac{1}{p}\sum_{j=1}^{p}\sum_{i}\tilde{\mathbf{g}}_{j,i}^{(t)}\Big\rangle\Big] && (43)\\
&\overset{\textcircled{1}}{=} \mathbb{E}\Big[\Big\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \frac{1}{p}\sum_{j=1}^{p}\sum_{i=1}^{p}\tilde{\mathbf{g}}_{j,i}^{(t)} - \frac{1}{p}\sum_{j=1}^{p}\sum_{i=1}^{q}\underline{\tilde{\mathbf{g}}}_{j,i}^{(t)})\Big\rangle\Big] && (44)\\
&\overset{\textcircled{2}}{=} \Big\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \frac{1}{p}\sum_{j=1}^{p}\nabla F(\mathbf{x}_j^{(t)})\Big\rangle - \Big\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \frac{1}{p}\sum_{j=1}^{p}\sum_{i=1}^{q}\underline{\mathbf{g}}_{j,i}^{(t)}\Big\rangle && (45)\\
&= \frac{1}{p}\sum_{j=1}^{p}\Big\langle \nabla F(u^{(t)}), \nabla F(\mathbf{x}_j^{(t)})\Big\rangle - \frac{1}{p}\sum_{j=1}^{p}\sum_{i}\Big\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \underline{\mathbf{g}}_{j,i}^{(t)}\Big\rangle && (46)\\
&\overset{\textcircled{3}}{=} \frac{1}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] + \frac{1}{2}\frac{1}{p}\sum_{j=1}^{p}\|\nabla F(\mathbf{x}_j^{(t)})\|^2 - \frac{1}{2}\frac{1}{p}\sum_{j=1}^{p}\|\nabla F(\bar{\mathbf{x}}^{(t)}) - \nabla F(\mathbf{x}_j^{(t)})\|^2 && (47)\\
&\qquad - \frac{1}{p}\sum_{j=1}^{p}\sum_{i}\Big\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \underline{\mathbf{g}}_{j,i}^{(t)}\Big\rangle && (48)\\
&\overset{\textcircled{4}}{=} \frac{1}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] + \frac{1}{2p}\sum_{j=1}^{p}\|\sum_{i}\mathbf{g}_{j,i}^{(t)} + \sum_{i}\underline{\mathbf{g}}_{j,i}^{(t)}\|^2 && (49)\\
&\quad - \frac{1}{2p}\sum_{j=1}^{p}\|\nabla F(\bar{\mathbf{x}}^{(t)}) - \nabla F(\mathbf{x}_j^{(t)})\|^2 - \frac{1}{p}\sum_{j=1}^{p}\sum_{i}\Big\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \mathbf{g}_{j,i}^{(t)}\Big\rangle && (50)\\
&\overset{\textcircled{5}}{=} \frac{1}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] + \frac{1}{2}\frac{1}{p}\sum_{j=1}^{p}\sum_{i}\|\mathbf{g}_{j,i}^{(t)}\|^2 + \frac{1}{2}\frac{1}{p}\sum_{j=1}^{p}\sum_{l}\|\underline{\mathbf{g}}_{j,l}^{(t)}\|^2 && (51)\\
&\quad + \frac{1}{2p}\sum_{j=1}^{p}\Big[\sum_{i\neq m}\Big\langle \mathbf{g}_{j,i}^{(t)}, \mathbf{g}_{j,m}^{(t)}\Big\rangle + \sum_{i\neq m}\Big\langle \underline{\mathbf{g}}_{j,i}^{(t)}, \underline{\mathbf{g}}_{j,m}^{(t)}\Big\rangle + \sum_{i,m}\Big\langle \mathbf{g}_{j,i}^{(t)}, \underline{\mathbf{g}}_{j,m}^{(t)}\Big\rangle\Big]\\
&\qquad - \frac{1}{2p}\sum_{j=1}^{p}\|\nabla F(\bar{\mathbf{x}}^{(t)}) - \nabla F(\mathbf{x}_j^{(t)})\|^2 - \frac{1}{p}\sum_{j=1}^{p}\sum_{i}\Big\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \underline{\mathbf{g}}_{j,i}^{(t)}\Big\rangle, && (52)
\end{aligned}$$

where ① follows from the Algorithm 1, and ② is implied by Assumption 1. In ③ we use the equality $2\langle \mathbf{a}, \mathbf{b}\rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2$. In ④ we use the fact that

$$\nabla F(\mathbf{x}_j^{(t)}) = \sum_{i}\mathbf{g}_{j,i}^{(t)} + \sum_{i}\underline{\mathbf{g}}_{j,i}^{(t)} \tag{53}$$

Finally, ⑤ holds because of applying and expanding (53). □

From Lemma I.1 we have:

$$\begin{aligned}
-\eta\mathbb{E}\Big[\langle\nabla F(\bar{\mathbf{x}}^{(t)}), \tilde{\mathbf{g}}^{(t)}\rangle\Big] &= -\frac{\eta}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2}\frac{1}{p}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} - \frac{\eta}{2}\frac{1}{p}\sum_{j=1}^{p}\underline{\mathcal{G}}_j^{(t)} \\
&\quad - \frac{\eta}{2p}\sum_j\Big[\sum_{i\neq m}\Big\langle\mathbf{g}_{j,i}^{(t)}, \mathbf{g}_{j,m}^{(t)}\Big\rangle + \sum_{i\neq m}\Big\langle\underline{\mathbf{g}}_{j,i}^{(t)}, \underline{\mathbf{g}}_{j,m}^{(t)}\Big\rangle + \sum_{i,m}\Big\langle\mathbf{g}_{j,i}^{(t)}, \underline{\mathbf{g}}_{j,m}^{(t)}\Big\rangle\Big] \\
&\qquad + \frac{\eta}{2}\frac{1}{p}\sum_{j=1}^{p}\|\nabla F(\bar{\mathbf{x}}^{(t)}) - \nabla F(\mathbf{x}_j^{(t)})\|^2 + \frac{\eta}{p}\sum_{j=1}^{p}\sum_i\Big\langle\nabla F(\bar{\mathbf{x}}^{(t)}), \underline{\mathbf{g}}_{j,i}^{(t)}\Big\rangle \\
&\overset{①}{\leq} -\frac{\eta}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2}\frac{1}{p}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} - \frac{\eta}{2}\frac{1}{p}\sum_{j=1}^{p}\underline{\mathcal{G}}_j^{(t)} \\
&\quad + \frac{\eta}{2p}\sum_j\Big[\sum_{i\neq m}|\Big\langle\mathbf{g}_{j,i}^{(t)}, \mathbf{g}_{j,m}^{(t)}\Big\rangle| + |\sum_{i\neq m}\Big\langle\underline{\mathbf{g}}_{j,i}^{(t)}, \underline{\mathbf{g}}_{j,m}^{(t)}\Big\rangle| + \sum_{i,m}|\Big\langle\mathbf{g}_{j,i}^{(t)}, \underline{\mathbf{g}}_{j,m}^{(t)}\Big\rangle|\Big] \\
&\qquad + \frac{L^2\eta}{2p}\mathbb{E}\Big(\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2\Big) + \eta p^2(1-\gamma)\beta \\
&\overset{②}{\leq} -\frac{\eta}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2}\frac{1}{p}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} - \frac{\eta(p-q)}{2p}\sum_{j=1}^{p}\tilde{\mathbf{G}}_j^{(t)} \\
&\quad + \frac{\eta}{2p}\sum_j\Big[\sum_{i\neq m}|\Big\langle\mathbf{g}_{j,i}^{(t)}, \mathbf{g}_{j,m}^{(t)}\Big\rangle| + \sum_{i\neq m}|\Big\langle\underline{\mathbf{g}}_{j,i}^{(t)}, \underline{\mathbf{g}}_{j,m}^{(t)}\Big\rangle| + \sum_{i,m}|\Big\langle\mathbf{g}_{j,i}^{(t)}, \underline{\mathbf{g}}_{j,m}^{(t)}\Big\rangle|\Big] \\
&\qquad + \frac{L^2\eta}{2p}\mathbb{E}\Big(\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2\Big) + \eta p^2(1-\gamma)\beta \\
&\overset{③}{\leq} -\frac{\eta}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2}\frac{1}{p}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} - \frac{\eta(p-q)}{2p}\tilde{\mathbf{G}}^{(t)} \\
&\quad + \frac{p\eta}{2p}\Big(\Big[q(q-1) + (p-q)(p-q-1) + q(q)\Big]\beta\Big) + \frac{L^2\eta}{2p}\mathbb{E}\Big(\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2\Big) \\
&\qquad\qquad + \eta p^2(1-\gamma)\beta \\
&= -\frac{\eta}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2}\frac{1}{p}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} - \frac{\eta(p-q)}{2p}\tilde{\mathbf{G}}^{(t)} \\
&\quad + \frac{\eta}{2}\Big(p^2\Big[3\gamma^2 - 2\gamma + (1 - \frac{1}{p})\Big]\beta\Big) + \frac{L^2\eta}{2p}\mathbb{E}\Big(\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2\Big) \\
&\qquad\qquad + \eta p^2(1-\gamma)\beta \\
&= -\frac{\eta}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2}\frac{1}{p}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} - \frac{\eta(p-q)}{2p}\tilde{\mathbf{G}}^{(t)} \\
&\quad + \frac{p^2\eta}{2}\Big(\Big[3\gamma^2 - 4\gamma + (3 - \frac{1}{p})\Big]\beta\Big) + \frac{L^2\eta}{2p}\mathbb{E}\Big(\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2\Big)
\end{aligned} \tag{54}$$

where ① comes from Assumption 4 and noting that

$$\frac{1}{p}\sum_{j=1}^{p}\|\nabla F(\bar{\mathbf{x}}^{(t)}) - \nabla F(\mathbf{x}_j^{(t)})\|^2 \leq \frac{L^2}{p}\Big(\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2\Big),$$

and ② follows from

$$\sum_{\mathcal{D}_i\in\mathcal{D}\setminus\tilde{\mathcal{D}}_j}\|\underline{\mathbf{g}}_{j,i}^{(t)}\|^2 \geq |\mathcal{D}\setminus\tilde{\mathcal{D}}_j| \times \min_{\mathcal{D}_i\in\mathcal{D}\setminus\tilde{\mathcal{D}}_j}\|\underline{\mathbf{g}}_{j,i}^{(t)}\|^2 = (p-q)\min_{\mathcal{D}_i\in\mathcal{D}\setminus\tilde{\mathcal{D}}_j}\|\underline{\mathbf{g}}_{j,i}^{(t)}\|^2,$$

and the definition $\tilde{\mathbf{G}}_j^{(t)} \triangleq \min_i \|\underline{\mathbf{g}}_{j,i}^{(t)}\|^2$ and $\tilde{\mathbf{G}}^{(t)} \triangleq \sum_{j=1}^p \tilde{\mathbf{G}}_j^{(t)}$ and finally ③ is due to bounded gradient Assumption. □

I.4. Proof of Lemma H.3

Let $t_c = s\tau$ denote the times when communication occurs. Therefore, according to Algorithm 1 we have:

$$\bar{\mathbf{x}}^{(t_c+1)} = \frac{1}{p}\sum_{j=1}^{p} \mathbf{x}_j^{(t_c+1)} \tag{55}$$

for $1 \leq j \leq p$. Then, according to the update rule of Algorithm 1, we can rewrite update rule as follows:

$$\begin{aligned}\mathbf{x}_j^{(t)} &= \mathbf{x}_j^{(t-1)} - \eta \sum_i \tilde{\mathbf{g}}_{j,i}^{(t-1)}\\ &\overset{①}{=} \mathbf{x}_j^{(t-2)} - \eta \Big[\sum_i \tilde{\mathbf{g}}_{j,i}^{(t-2)} + \sum_i \tilde{\mathbf{g}}_{j,i}^{(t-1)}\Big]\\ &\vdots\\ &= \bar{\mathbf{x}}^{(t_c+1)} - \eta \Big[\sum_{k=t_c}^{t} \sum_i \tilde{\mathbf{g}}_{j,i}^{(k)}\Big]\end{aligned} \tag{56}$$

where ① comes from the update rule of our Algorithm. Now, from (56) we compute the average model as follows:

$$\bar{\mathbf{x}}^{(t)} = \bar{\mathbf{x}}^{(t_c+1)} - \eta \Big[\frac{1}{p}\sum_{j=1}^{p}\sum_{k=t_c}^{t}\sum_i \tilde{\mathbf{g}}_{j,i}^{(k)}\Big] \tag{57}$$

First, without loss of generality, suppose $t = s\tau + r$ where s and l denotes the indices of communication round and local updates respectively. Next consider that for $t_c + 1 < t \leq t_c + \tau$, $\mathbb{E}_t\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2$ does not depend on time $t \leq t_c$ for $1 \leq j \leq p$. Therefore, for all iterations $1 \leq t \leq T$ we can write:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}\mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2 = \frac{1}{T}\sum_{s=0}^{\frac{T}{\tau}-1}\sum_{r=1}^{\tau}\sum_{j=1}^{p}\mathbb{E}\|\bar{\mathbf{x}}^{(s\tau+r)} - \mathbf{x}_j^{(s\tau+r)}\|^2 \tag{58}$$

Now, we bound $\mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_l^{(t)}\|^2$ for $t_c + 1 \leq t = s\tau + r \leq t_c + \tau$ as follows:

$$
\begin{aligned}
\mathbb{E}\|\bar{\mathbf{x}}^{(s\tau+r)} - \mathbf{x}_l^{(s\tau+r)}\|^2 &= \mathbb{E}\|\bar{\mathbf{x}}^{(t_c+1)} - \eta\Big[\sum_{k=t_c}^{t}\sum_{i}\tilde{\mathbf{g}}_{l,i}^{(k)}\Big] - \bar{\mathbf{x}}^{(t_c+1)} + \eta\Big[\frac{1}{p}\sum_{j=1}^{p}\sum_{k=t_c}^{t}\sum_{i}\tilde{\mathbf{g}}_{j,i}^{(k)}\Big]\|^2 \\
&\overset{①}{=} \eta^2\mathbb{E}\|\sum_{k=1}^{r}\sum_{i}\tilde{\mathbf{g}}_{l,i}^{(s\tau+k)} - \frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\tilde{\mathbf{g}}_{j,i}^{(s\tau+k)}\|^2 \\
&\overset{②}{\leq} 2\eta^2\Big[\mathbb{E}\|\sum_{k=1}^{r}\sum_{i}\tilde{\mathbf{g}}_{l,i}^{(s\tau+k)}\|^2 + \mathbb{E}\|\frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\tilde{\mathbf{g}}_{j,i}^{(s\tau+k)}\|^2\Big] \\
&\overset{③}{=} 2\eta^2\Big[\mathbb{E}_t\|\sum_{k=1}^{r}\sum_{i}\tilde{\mathbf{g}}_{l,i}^{(s\tau+k)} - \mathbb{E}\big[\sum_{k=1}^{r}\sum_{i}\tilde{\mathbf{g}}_{l,i}^{(s\tau+k)}\big]\|^2 + \|\mathbb{E}\big[\sum_{k=1}^{r}\sum_{i}\tilde{\mathbf{g}}_{l,i}^{(s\tau+k)}\big]\|^2 \\
&+ \mathbb{E}\|\frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\tilde{\mathbf{g}}_{j,i}^{(s\tau+k)} - \mathbb{E}\big[\frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\tilde{\mathbf{g}}_{j,i}^{(s\tau+k)}\big]\|^2\Big] + \|\mathbb{E}\big[\frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\tilde{\mathbf{g}}_{j,i}^{(s\tau+k)}\big]\|^2 \\
&\overset{④}{=} 2\eta^2\mathbb{E}_t\Big(\Big[\|\sum_{k=1}^{r}\sum_{i}\Big[\tilde{\mathbf{g}}_{l,i}^{(s\tau+k)} - \mathbf{g}_{l,i}^{(s\tau+k)}\Big]\|^2 + \|\sum_{k=1}^{r}\sum_{i}\mathbf{g}_{l,i}^{(s\tau+k)}\|^2\Big] \\
&\quad + \|\frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\Big[\tilde{\mathbf{g}}_{j,i}^{(s\tau+k)} - \mathbf{g}_{j,i}^{(s\tau+k)}\Big]\|^2 + \|\frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\mathbf{g}_{j,i}^{(s\tau+k)}\|^2\Big) \\
&= 2\eta^2\mathbb{E}\Big(\Big[\sum_{k=1}^{r}\sum_{i}\|\tilde{\mathbf{g}}_{l,i}^{(s\tau+k)} - \mathbf{g}_{l,i}^{(s\tau+k)}\|^2 \\
&+ \sum_{w\neq z\vee l\neq v\vee i\neq b}\Big\langle\tilde{\mathbf{g}}_{l,i}^{(w)} - \mathbf{g}_{l,i}^{(w)}, \tilde{\mathbf{g}}_{v,b}^{(z)} - \mathbf{g}_{v,b}^{(z)}\Big\rangle + \|\sum_{k=1}^{r}\sum_{i}\mathbf{g}_{l,i}^{(s\tau+k)}\|^2\Big] \\
&\quad + \frac{1}{p^2}\sum_{l=1}^{p}\sum_{k=1}^{r}\sum_{i}\|\tilde{\mathbf{g}}_{l,i}^{(s\tau+k)} - \mathbf{g}_{l,i}^{(s\tau+k)}\|^2 \\
&+ \frac{1}{p^2}\sum_{w\neq z\vee l\neq v\vee i\neq b}\Big\langle\tilde{\mathbf{g}}_{l,i}^{(w)} - \mathbf{g}_{l,i}^{(w)}, \tilde{\mathbf{g}}_{v,b}^{(z)} - \mathbf{g}_{v,b}^{(z)}\Big\rangle + \|\frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\mathbf{g}_{j,i}^{(s\tau+k)}\|^2\Big) \\
&\overset{⑤}{=} 2\eta^2\mathbb{E}\Big(\Big[\sum_{k=1}^{r}\sum_{i}\|\Big[\tilde{\mathbf{g}}_{l,i}^{(s\tau+k)} - \mathbf{g}_{l,i}^{(s\tau+k)}\|^2 + \|\sum_{k=1}^{r}\sum_{i}\mathbf{g}_{l,i}^{(s\tau+k)}\|^2\Big] \\
&\quad + \frac{1}{p^2}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\|\Big[\tilde{\mathbf{g}}_{j,i}^{(s\tau+k)} - \mathbf{g}_{j,i}^{(s\tau+k)}\|^2 + \|\frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\mathbf{g}_{j,i}^{(s\tau+k)}\|^2\Big) \\
&\overset{⑥}{\leq} 2\eta^2\mathbb{E}_t\Big(\Big[\sum_{k=1}^{r}\sum_{i}\|\tilde{\mathbf{g}}_{j,i}^{(s\tau+k)} - \mathbf{g}_{l,i}^{(s\tau+k)}\|^2 + q(r-1)\sum_{k=1}^{r}\sum_{i}\|\mathbf{g}_{l,i}^{(s\tau+k)}\|^2\Big] \\
&\quad + \frac{1}{p^2}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\|\tilde{\mathbf{g}}_{j,i}^{(s\tau+k)} - \mathbf{g}_{j,i}^{(s\tau+k)}\|^2 + \frac{q(r-1)}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\|\mathbf{g}_{j,i}^{(s\tau+k)}\|^2\Big) \\
&= 2\eta^2\Big(\Big[\sum_{k=1}^{r}\sum_{i}\mathbb{E}\|\tilde{\mathbf{g}}_{l,i}^{(s\tau+k)} - \mathbf{g}_{l,i}^{(s\tau+k)}\|^2 + q(r-1)\sum_{k=1}^{r}\sum_{i}\mathbb{E}\|\mathbf{g}_{l,i}^{(s\tau+k)}\|^2\Big] \\
&\quad + \frac{1}{p^2}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\mathbb{E}\|\tilde{\mathbf{g}}_{j,i}^{(s\tau+k)} - \mathbf{g}_{j,i}^{(s\tau+k)}\|^2 + \frac{q(r-1)}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\mathbb{E}\|\mathbf{g}_{j,i}^{(s\tau+k)}\|^2\Big) \qquad (59)
\end{aligned}
$$

where ① holds because $t = s\tau + r \leq t_c + \tau$, ② is due to $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2(\|\mathbf{a}\|^2 + \|\mathbf{a}\|^2)$, ③ comes from $\mathbb{E}[\mathbf{X}^2] = \mathbb{E}[[\mathbf{X} - \mathbb{E}[\mathbf{X}]]^2] + \mathbb{E}^2[\mathbf{X}]$, ④ comes from unbiased estimation Assumption. ⑤ is due to independent mini-batch sampling as well as unbiased estimation Assumption. ⑥ follow from inequality $\|\sum_{i=1}^{m}\mathbf{a}_i\|^2 \leq m\sum_{i=1}^{m}\|\mathbf{a}_i\|^2$. Next step is to bound the terms in (59) using Assumption 3 as follow:

$$\begin{aligned}\mathbb{E}\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}_l^{(t)}\|^2 &\leq 2\eta^2\Big(\Big[\sum_{k=1}^{r}\sum_{i}\Big[C_1\|\mathbf{g}_i(\mathbf{x}_l^{(s\tau+k)})\|^2+\frac{C_2^2}{p}\Big]+q(r-1)\sum_{k=1}^{r}\sum_{i}\|\Big[\mathbf{g}_i(\mathbf{x}_l^{(s\tau+k)})\Big]\|^2\Big]\\ &\quad+\frac{1}{p^2}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\Big[C_1\|\mathbf{g}_i(\mathbf{x}_j^{(s\tau+k)})\|^2+\frac{C_2^2}{p}\Big]+\frac{q(r-1)}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\mathbb{E}\|\Big[\mathbf{g}_i(\mathbf{x}_j^{(s\tau+k)})\Big]\|^2\Big)\\ &=2\eta^2\Big(\Big[\sum_{k=1}^{r}\sum_{i}C_1\|\mathbf{g}_{j,l}^{(s\tau+k)}\|^2+q(r-1)\frac{C_2^2}{p}+q(r-1)\sum_{k=1}^{r}\sum_{i}\|\mathbf{g}_{j,l}^{(s\tau+k)}\|^2\Big]\\ &\quad+\frac{1}{p^2}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}C_1\|\mathbf{g}_{j,i}^{(s\tau+k)}\|^2+\frac{q(r-1)C_2^2}{p^2}+\frac{q(r-1)}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\mathbb{E}\|\mathbf{g}_{j,i}^{(s\tau+k)}\|^2\Big)\end{aligned}\tag{60}$$

where ① is due to independent mini-batch sampling and Assumption 1. Now plugging (60) in (59) and after summation over worker indices we get:

$$\begin{aligned}\mathbb{E}\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}_j^{(t)}\|^2 &\leq 2\eta^2\Big(\Big[\sum_{l=1}^{p}\sum_{k=1}^{r}\sum_{i}C_1\|\mathbf{g}_{l,i}^{(s\tau+k)}\|^2+q(r-1)C_2^2+q(r-1)\sum_{l=1}^{p}\sum_{k=1}^{r}\sum_{i}\|\mathbf{g}_{l,i}^{(s\tau+k)}\|^2\Big]\\ &\quad+\frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}C_1\|\mathbf{g}_{j,i}^{(s\tau+k)}\|^2+\frac{q(r-1)C_2^2}{p}+q(r-1)\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\|\mathbf{g}_{j,i}^{(s\tau+k)}\|^2\Big)\\ &=2\eta^2\Big(\Big[(\frac{p+1}{p})\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}C_1\|\mathbf{g}_{j,i}^{(s\tau+k)}\|^2+\frac{q(p+1)(r-1)C_2^2}{p}\\ &\quad+2q(r-1)\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\|\mathbf{g}_{j,i}^{(s\tau+k)}\|^2\Big)\\ &=2\eta^2\Big(\Big[C_1(\frac{p+1}{p})+2q(r-1)\Big]\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\|\mathbf{g}_{j,i}^{(s\tau+k)}\|^2+\frac{q(p+1)(r-1)C_2^2}{p}\Big)\end{aligned}\tag{61}$$

Next, we first sum over local updates in (61) as follows:

$$\begin{aligned}\sum_{r=1}^{\tau}\mathbb{E}_t\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(s\tau+r)}-\mathbf{x}_j^{(s\tau+r)}\|^2 &\leq \sum_{r=1}^{\tau}2\eta^2\Big(\Big[C_1(\frac{p+1}{p})+2q(r-1)\Big]\sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{i}\|\mathbf{g}_{j,i}^{(s\tau+k)}\|^2+\frac{q(p+1)(r-1)C_2^2}{p}\Big)\\ &\leq \eta^2\Big(\Big[C_1(\frac{p+1}{p})+2q(\tau-1)\Big]\sum_{j=1}^{p}\sum_{k=1}^{\tau}\sum_{i}\|\mathbf{g}_{j,i}^{(s\tau+k)}\|^2+\frac{q(p+1)\tau(\tau-1)C_2^2}{2p}\Big)\end{aligned}\tag{62}$$

Final step is to sum over the communication step period over (62) as follows:

$$\begin{aligned}\sum_{s=0}^{\frac{T}{\tau}-1}\sum_{r=1}^{\tau}\mathbb{E}\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(s\tau+r)}-\mathbf{x}_j^{(s\tau+r)}\|^2 &\leq \sum_{s=0}^{\frac{T}{\tau}-1}2\eta^2\Big(\Big[C_1(\frac{p+1}{p})+2q(\tau-1)\Big]\sum_{j=1}^{p}\sum_{k=t_c}^{s\tau+r}\sum_{i}\|\mathbf{g}_{j,i}^{(k)}\|^2\\ &\quad+\frac{q(p+1)\tau(\tau-1)C_2^2}{2p}\Big)\\ &=2\eta^2\Big(\Big[C_1(\frac{p+1}{p})+2q(\tau-1)\Big]\sum_{j=1}^{p}\sum_{k=1}^{T}\sum_{i}\|\mathbf{g}_{j,i}^{(k)}\|^2\\ &\quad+\frac{q(p+1)T(\tau-1)C_2^2}{2p}\Big)\end{aligned}\tag{63}$$

J. Proof of Theorem 4.2

Note that the learning rate for this section $\eta^{(t)}$ depends on the constant $\rho^{(t)} > 0$ and for simplicity in presentation we drop the dependency on (t) and indicate it with η.

If for all $1 \leq t \leq T$, there exists a constant $\rho^{(t)} > 0$ such that

$$\sum_{j=1}^{p} \| \sum_{i} \mathbf{g}_{j,i}^{(t)} + \sum_{i} \underline{\mathbf{g}}_{j,i}^{(t)} \|^2 \geq \rho^{(t)} \sum_{j=1}^{p} \Big(\sum_{i} \|\mathbf{g}_{j,i}^{(t)}\|^2 \Big) \tag{64}$$

Using (64), we only need to rewrite Lemma I.1 as follows:

Lemma J.1. *Under Assumptions 2, and according to the data allocation policy of Algorithm 1 the expected inner product between stochastic gradient and full batch gradient can be expanded as:*

$$\begin{aligned}\mathbb{E}\Big[\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \tilde{\mathbf{g}}^{(t)} \rangle\Big] \geq \frac{1}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] + \frac{\rho^{(t)}}{2}\Big(\frac{1}{p^2}\sum_{j=1}^{p} \mathcal{G}_j^{(t)}\Big) \\ + \frac{1}{2}\|\nabla F(\bar{\mathbf{x}}^{(t)}) - \frac{1}{p}\sum_{j=1}^{p} \nabla F(\mathbf{x}_j^{(t)})\|^2 - \frac{1}{p}\sum_{j=1}^{p}\sum_{i} \Big\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \underline{\mathbf{g}}_{j,i}^{(t)} \Big\rangle\end{aligned} \tag{65}$$

which leads to

$$-\eta\mathbb{E}\Big[\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \tilde{\mathbf{g}}^{t} \rangle\Big] \leq -\frac{\eta}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2}\frac{\rho^{(t)}}{p}\sum_{j=1}^{p} \mathcal{G}_j^{(t)} + \frac{L^2}{2p}\mathbb{E}\sum_{j=1}^{p} \|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}|^2 + \eta p^2 \beta\big[(1-\gamma)\big].$$

Proof. The proof simply follows by combining Lemma I.1 and additional condition (64).

Now, using Lemmas J.1 and H.2, we continue from (14) as follows:

$$\mathbb{E}[F(\bar{\mathbf{x}}^{(t+1)})] - F(\bar{\mathbf{x}}^{(t)}) \leq -\eta\mathbb{E}\big[\langle \nabla F(\bar{\mathbf{x}}^{(t)}), \tilde{\mathbf{g}}^{(t)} \rangle\big] + \frac{\eta^2 L}{2}\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}\|^2\Big] \tag{66}$$

in which by summation over total iterations results in:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\Big[\mathbb{E}[F(\bar{\mathbf{x}}^{(t+1)})] - F(\bar{\mathbf{x}}^{(t)})\Big] &\leq -\frac{\eta}{2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta(1-\gamma)}{2}\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{G}}^{(t)} + \frac{p^2\eta}{2}(1-\frac{1}{p^2})\beta \\ &\quad + \frac{\eta}{2pT}\Big[\eta^2 L^2\Big(2\tau C_1(\frac{p+1}{p}) + \tau(\tau-1)\Big) - \rho^{(t)} + \frac{\eta L(C_1 + qp)}{p}\Big]\sum_{t=1}^{T}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} \qquad (67) \\ &\quad + \frac{\eta^3 L^2 \gamma}{2}(\tau-1)(\frac{p+1}{p})C_2^2 + \eta p^2(1-\gamma)\beta \\ &\overset{①}{\leq} -\frac{\eta}{2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta(1-\gamma)}{2}\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{G}}^{(t)} \\ &\quad + (\frac{p+1}{p})\frac{\eta^3 L^2\gamma}{2}(\tau-1)C_2^2 + \eta p^2\beta\Big(1-\gamma)\Big) + \frac{L\eta^2\gamma C_2^2}{2p} \qquad (68)\end{aligned}$$

where ① comes from the learning rate choice of

$$0 < \eta^2 L^2\Big(2\tau C_1(\frac{p+1}{p}) + \tau(\tau-1)\Big) + \frac{\eta L(C_1 + \gamma p^2)}{p} - \rho^{(t)} < 1 \tag{69}$$

Finally rearranging (23) we get:

$$\mathbb{E}\Big[\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big]\Big] \leq \frac{2\big[F(\bar{\mathbf{x}}^{(1)})-F^*\big]}{\eta T} + \eta L C_2^2\frac{\gamma}{p} + (\frac{p+1}{p})\gamma\eta^2L^2C_2^2(\tau-1) + 2p^2\Big[(1-\gamma)\Big]\beta \tag{70}$$

□

K. Proof of Theorem D.1

The proof is very similar to the proof of Theorem 4.2 with only one distinction that $\tilde{\mathcal{D}}_j = \{\mathcal{D}_{j,1},\dots,\mathcal{D}_{j,q_j}\}$ is equivalent to say $|\tilde{\mathcal{D}}_j| = q_j$. With this distinction in mind, we mention the counterparts of the lemmas in Subsection H.

Remark 6. *In this section, all the summation over indices i is from $i=1$ to $i=q_j$. For instance, for this section we have $\mathcal{G}_j^{(t)} \triangleq \sum_{i=1}^{q_j}\|\mathbf{g}_{j,i}^{(t)})\|^2$*

Proof.

Lemma K.1. *Under Assumptions 1,2 and 3, we have the following variance bound fro the averaged stochastic gradient:*

$$\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}-\mathbf{g}^{(t)}\|^2\Big] \leq \frac{C_1}{p^2}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} + \frac{\sum_{j=1}^{p}q_j}{p^3}C_2^2 \tag{71}$$

Lemma K.2. *Under Assumptions 1 to 4, the squared norm of stochastic gradient can be bounded as*

$$\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}\|^2\Big] \leq \sum_{j=1}^{p}(\frac{C_1+q_jp}{p^2})\mathcal{G}_j^{(t)} + \frac{\sum_{j=1}^{p}q_jC_2^2}{p^3} \tag{72}$$

We extend Lemma H.3 to heterogeneous

Lemma K.3. *Under Assumptions 1 to 4, we have:*

$$\frac{L^2}{pT}\sum_{t=1}^{T}\mathbb{E}\sum_{j=1}^{p}\|\mathbf{u}^{(\mathbf{t})}-\mathbf{x}_j^{(t)}\|^2 \leq \frac{\eta^2L^2}{pT}\Big[\Big(2\tau(\frac{p+1}{p})C_1+\tau(\tau-1)\Big)\sum_{j=1}^{p}\sum_{t=1}^{T}\mathcal{G}_j^{(t)} + \sum_{j=1}^{p}\frac{q_j}{p}T(\tau-1)C_2\Big] \tag{73}$$

which leads us to

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}-\eta\mathbb{E}\Big[\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\tilde{\mathbf{g}}^{t}\rangle\Big] \leq &-\frac{\eta}{2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2}\frac{1}{p}\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} - \frac{1}{2}\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}\frac{p-q_j}{p}\tilde{\mathbf{G}}_j^{(t)}\\ &+\frac{\eta^3L^2}{2pT}\Big[\Big(2\tau(\frac{p+1}{p})C_1+\tau(\tau-1)\Big)\sum_{j=1}^{p}\mathcal{G}_j^{(t)} + (\frac{p+1}{p})\sum_{j=1}^{p}\frac{q_j}{p}T(\tau-1)C_2\Big]\\ &+\frac{\eta\beta}{2}\Big[\frac{1}{p}\sum_{j=1}^{p}(3q_j^2-2pq_j+p^2-p)+2\sum_{j=1}^{p}(p-q_j)\Big]\end{aligned} \tag{74}$$

where $\tilde{\mathbf{G}}_j^{(t)} \triangleq \min_{\mathcal{D}_i\in\mathcal{D}\setminus\mathcal{D}_j}\|\mathbf{g}_{j,i}^{(t)}\|^2$.

Next, combining the results from Lemmas K.1, K.2, and K.3 results in:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\Big[\mathbb{E}[F(\bar{\mathbf{x}}^{(t+1)})] - F(\bar{\mathbf{x}}^{(t)})\Big] &\leq \frac{1}{T}\sum_{t=1}^{T}\big(-\eta\mathbb{E}\big[\big\langle\nabla F(\bar{\mathbf{x}}^{(t)}), \tilde{\mathbf{g}}^{(t)}\big\rangle\big]\big) + \frac{1}{T}\sum_{t=1}^{T}\frac{\eta^2 L}{2}\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}\|^2\Big] \\
&\leq -\frac{\eta}{2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2}\frac{1}{p}\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} - \frac{\eta}{2T}\sum_{t=1}^{T}\sum_{j=1}^{p}\frac{p-q_j}{p}\tilde{\mathbf{G}}_j^{(t)} \\
&\quad + \frac{\eta^3 L^2}{2T}\Big[\Big(2\tau C_1(\frac{p+1}{p}) + \tau(\tau-1)\Big)\sum_{t=1}^{T}\sum_{j=1}^{p}\mathcal{G}_j^{(t)} + (\frac{p+1}{p})\sum_{j=1}^{p}\frac{q_j}{p}\gamma(\tau-1)C_2^2\Big] \\
&\quad + \frac{\eta\beta}{2}\Big[\frac{1}{p}\sum_{j=1}^{p}(3q_j^2 - 2pq_j + p^2 - p) + 2\sum_{j=1}^{p}(p-q_j)\Big] \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}(\frac{\eta^2 L(C_1+q_j p)}{2p^2})\mathcal{G}_j^{(t)} \\
&\quad + \sum_{j=1}^{p}\frac{L\eta^2 q_j C_2^2}{2p^3} \\
&= -\frac{\eta}{2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2T}\sum_{t=1}^{T}\sum_{j=1}^{p}\frac{p-q_j}{p}\tilde{\mathbf{G}}_j^{(t)} \\
&+ \frac{\eta\beta}{2}\Big[\frac{1}{p}\sum_{j=1}^{p}(3q_j^2 - 2pq_j + p^2 - p) + 2\sum_{j=1}^{p}(p-q_j)\Big] \\
&\quad + \frac{\eta}{2pT}\sum_{t=1}^{T}\sum_{j=1}^{p}\Big[\eta^2 L^2\Big(2\tau(\frac{p+1}{p})C_1 + \tau(\tau-1)\Big) - 1 + \frac{\eta L(C_1+q_j p)}{p}\Big]\mathcal{G}_j^{(t)} \\
&\quad + (\frac{p+1}{p})\sum_{j=1}^{p}\frac{\eta^3 L^2 q_j}{2p^2}(\tau-1)C_2^2 + \sum_{j=1}^{p}\frac{L\eta^2 q_j C_2^2}{2p^3} \\
&\overset{①}{\leq} -\frac{\eta}{2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big] - \frac{\eta}{2T}\sum_{t=1}^{T}\sum_{j=1}^{p}\frac{p-q_j}{p}\tilde{\mathbf{G}}_j^{(t)} \\
&\quad + \frac{\eta\beta}{2}\Big[\frac{1}{p}\sum_{j=1}^{p}(3q_j^2 - 2pq_j + p^2 - p) + 2\sum_{j=1}^{p}(p-q_j)\Big] \\
&\quad + (\frac{p+1}{p})\sum_{j=1}^{p}\frac{\eta^3 L^2 q_j}{2p^2}(\tau-1)C_2^2 \\
&\quad + \sum_{j=1}^{p}\frac{L\eta^2 q_j C_2^2}{2p^3},
\end{aligned}
\tag{75}
$$

where ② is because of the learning rate choice of

$$
\max_{1\leq j\leq p}\Big(\eta^2 L^2(2\tau C_1(\frac{p+1}{p}) + \tau(\tau-1)) + \frac{\eta L(C_1+q_j p)}{p}\Big) \leq 1
$$

.

Now, if we rearrange (75) and take the average over T iterations we get:

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\big[\mathbb{E}[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big]\Big] \leq \frac{2\big[F(\mathbf{x}^{(1)})-F^*\big]}{\eta T} &- \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}\frac{p-q_j}{p}\tilde{\mathbf{G}}_j^{(t)} \\ &+\frac{\eta\beta}{2}\Big[\frac{1}{p}\sum_{j=1}^{p}(3q_j^2-2pq_j+p^2-p)+2\sum_{j=1}^{p}(p-q_j)\Big] \\ &+\frac{p+1}{p}\sum_{j=1}^{p}\frac{\eta^2L^2q_j}{p^2}(\tau-1)C_2^2+\sum_{j=1}^{p}\frac{L\eta q_jC_2^2}{p^3} \qquad (76)\end{aligned}$$

□

L. Proof of Lemmas

L.1. Proof of Lemma K.1

Proof. Due to similarity with Lemma H.1, we continue from (32):

$$\begin{aligned}\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}-\mathbf{g}^{(t)}\|^2\Big] &\leq \frac{1}{p^2}\Big[\sum_{j=1}^{p}\sum_{i}\Big[C_1\|\mathbf{g}_{j,i}^{(t)}\|^2+\frac{C_2^2}{p}\Big] && (77)\\ &=\frac{1}{p^2}\sum_{j=1}^{p}\Big[\sum_{i}C_1\|\mathbf{g}_{j,i}^{(t)}\|^2+\frac{|\tilde{\mathcal{D}}_j^{(t)}|C_2^2}{p}\Big] && (78)\\ &=\frac{1}{p^2}\sum_{j=1}^{p}\sum_{i}C_1\|\mathbf{g}_{j,i}^{(t)}\|^2+\frac{\sum_{j=1}^{p}q_jC_2^2}{p^3} && (79)\end{aligned}$$

□

L.2. Proof of Lemma K.2

Proof. Assumption 2 implies that $\mathbb{E}[\tilde{\mathbf{g}}^{(t)}]=\mathbf{g}^{(t)}$, then we have

$$\begin{aligned}\mathbb{E}\Big[\|\tilde{\mathbf{g}}^{(t)}\|^2\Big] &= \mathbb{E}\big[\|\tilde{\mathbf{g}}^{(t)}-\mathbb{E}\big[\tilde{\mathbf{g}}^{(t)}\big]\|^2\Big]+\|\mathbb{E}\big[\tilde{\mathbf{g}}^{(t)}\big]\|^2 && (80)\\ &=\mathbb{E}\big[\|\tilde{\mathbf{g}}^{(t)}-\mathbf{g}^{(t)}\big]\|^2\Big]+\|\mathbf{g}^{(t)}\|^2 && (81)\\ &\leq\frac{1}{p^2}\sum_{j=1}^{p}\sum_{i}C_1\|\mathbf{g}_{j,i}^{(t)}\|^2+\frac{\sum_{j=1}^{p}q_jC_2^2}{p^3}+\|\mathbf{g}^{(t)}\|^2 && (82)\\ &=\frac{1}{p^2}\sum_{j=1}^{p}\sum_{i}C_1\|\mathbf{g}_{j,i}^{(t)}\|^2+\frac{\sum_{j=1}^{p}q_jC_2^2}{p^3}+\|\frac{1}{p}\sum_{j=1}^{p}\sum_{i}\mathbf{g}_{j,i}^{(t)}\|^2 && (83)\\ &\overset{①}{\leq}\frac{C_1}{p^2}\sum_{j=1}^{p}\sum_{i}\|\mathbf{g}_{j,i}^{(t)}\|^2+\frac{\sum_{j=1}^{p}q_jC_2^2}{p^3}+\frac{1}{p}\sum_{j=1}^{p}\|\sum_{i}\mathbf{g}_{j,i}^{(t)}\|^2 && (84)\\ &\overset{②}{\leq}(\frac{C_1}{p^2})\sum_{j=1}^{p}\sum_{i}\|\mathbf{g}_{j,i}^{(t)}\|^2+\frac{\sum_{j=1}^{p}q_jC_2^2}{p^3}+\frac{1}{p}\sum_{j=1}^{p}|\tilde{\mathcal{D}}_j^{(t)}|\sum_{i}\|\mathbf{g}_{j,i}^{(t)}\|^2 && (85)\\ &=\sum_{j=1}^{p}(\frac{C_1}{p^2})\sum_{i}\|\mathbf{g}_{j,i}\|^2+\frac{\sum_{j=1}^{p}q_jC_2^2}{p^3}+\frac{1}{p}\sum_{j=1}^{p}q_j\sum_{i}\|\mathbf{g}_{j,i}(t)\|^2 && (86)\\ &=\sum_{j=1}^{p}(\frac{C_1+q_jp}{p^2})\sum_{i}\|\mathbf{g}_{j,i}^{(t)}\|^2+\frac{\sum_{j=1}^{p}q_jC_2^2}{p^3} && (87)\end{aligned}$$

Where ① comes from Jensen inequality and ② follows by the fact that $\|\sum_{i=1}^{m}\mathbf{a}_i\|^2 \leq m\sum_{i=1}^{m}\|\mathbf{a}_i\|^2$. □

Proof of Lemma K.3 due to similarity with the proof of Lemma H.4 is omitted.

M. Proof of Corollary F.0.1

Proof. First of all, consider that only term $\frac{1}{p}\sum_{j=1}^{p}\sum_i\left\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\underline{\mathbf{g}}_{j,i}^{(t)}\right\rangle$ depends on $\beta_j^{(l)}$, therefore we only need to consider the lower bound over $\frac{1}{p}\sum_{j=1}^{p}\sum_i\left\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\underline{\mathbf{g}}_{j,i}^{(t)}\right\rangle$ in $\mathbb{E}\Big[\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\tilde{\mathbf{g}}^{(t)}\rangle\Big]$. Under Assumption 1 to 3 and Assumption 5, if for all $1 \leq t \leq T$, there exists a constant $\rho^{(t)} > 0$ such that

$$\sum_{j=1}^{p}\|\sum_i\mathbf{g}_{j,i}^{(t)}+\sum_i\underline{\mathbf{g}}_{j,i}^{(t)}\|^2 \geq \rho^{(t)}\sum_{j=1}^{p}\Big(\sum_i\|\mathbf{g}_{j,i}^{(t)}\|^2\Big). \tag{88}$$

We change the Lemma J.1 as follows:

Lemma M.1. *Under Assumptions 2, and according to Algorithm 1 the expected inner product between stochastic gradient and full batch gradient can be expanded as:*

$$\begin{aligned}\mathbb{E}\Big[\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\tilde{\mathbf{g}}^{(t)}\rangle\Big] &\geq \frac{1}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big]+\frac{\rho^{(t)}}{2}\Big(\frac{1}{p^2}\sum_{j=1}^{p}\mathcal{G}_j^{(t)}\Big)\\ &\quad+\frac{1}{2}\|\nabla F(\bar{\mathbf{x}}^{(t)})-\frac{1}{p}\sum_{j=1}^{p}\nabla F(\mathbf{x}_j^{(t)})\|^2-\frac{1}{p}\sum_{j=1}^{p}\sum_i\left\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\underline{\mathbf{g}}_{j,i}^{(t)}\right\rangle\end{aligned} \tag{89}$$

which leads to

$$\begin{aligned}-\eta\mathbb{E}\Big[\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\tilde{\mathbf{g}}^{t}\rangle\Big] &\leq -\frac{\eta}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big]-\frac{\eta}{2}\frac{\rho^{(t)}}{p}\sum_{j=1}^{p}\mathcal{G}_j^{(t)}+\frac{L^2}{2p}\mathbb{E}\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}_j^{(t)}\|^2\\ &\quad+\frac{\eta}{p}\sum_{j=1}^{p}\sum_i\left\langle\nabla F(\bar{\mathbf{x}}^{(t)}),\underline{\mathbf{g}}_{j,i}^{(t)}\right\rangle\\ &= -\frac{\eta}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big]-\frac{\eta}{2}\frac{\rho^{(t)}}{p}\sum_{j=1}^{p}\mathcal{G}_j^{(t)}+\frac{L^2}{2p}\mathbb{E}\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}_j^{(t)}\|^2\\ &\quad+\frac{\eta}{p}\sum_{j=1}^{p}\sum_i\sum_{l=1}^{p}\left\langle\mathbf{g}_{l,l}(\bar{\mathbf{x}}^{(t)}),\underline{\mathbf{g}}_{j,i}^{(t)}\right\rangle\\ &\overset{①}{\leq} -\frac{\eta}{2}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big]-\frac{\eta}{2}\frac{\rho^{(t)}}{p}\sum_{j=1}^{p}\mathcal{G}_j^{(t)}+\frac{L^2}{2p}\mathbb{E}\sum_{j=1}^{p}\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}_j^{(t)}\|^2\\ &\quad+\frac{\eta}{p}\sum_{j=1}^{p}\sum_{i=1}^{q-p}\sum_{l=1}^{p}\beta_{j,i},\end{aligned} \tag{90}$$

where ① comes from Assumption 5. Similar to (68) and under condition (69), we obtain:

$$\mathbb{E}\Big[\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\big[\|\nabla F(\bar{\mathbf{x}}^{(t)})\|^2\big]\Big] \leq \frac{2\big[F(\bar{\mathbf{x}}^{(1)})-F^*\big]}{\eta T}+\eta LC_2^2\frac{\gamma}{p}+(\frac{p+1}{p})\gamma\eta^2L^2C_2^2(\tau-1)+\frac{2}{p}\sum_{j=1}^{p}\sum_{i=1}^{q-p}\sum_{l=1}^{p}\beta_{j,i} \tag{91}$$

□