

Part I

Appendix

A More Details about Experiments

A.1 Implementation Details for Experiments

First, the perturbed dataset is constructed as follows: according to Definition 1, we randomly select a client, then randomly choose a data point from this client’s training set and swap it with a data point from the test set. This process creates the perturbed dataset. To enhance the robustness of our experimental results, all reported metrics are averaged over three independent runs with different random seeds. Second, for the other experimental hyperparameters, aside from the experimental parameters to be explored, we use the following default settings: client number = 50, learning rate = 0.04, learning rate decay factor = 0.995, base communication topology = Ring, multiple gossip steps = 1, Dirichlet distribution coefficient (to control data heterogeneity) = 0.3, and communication rounds = 1200.

A.2 Experimental Validation of Excess Error.

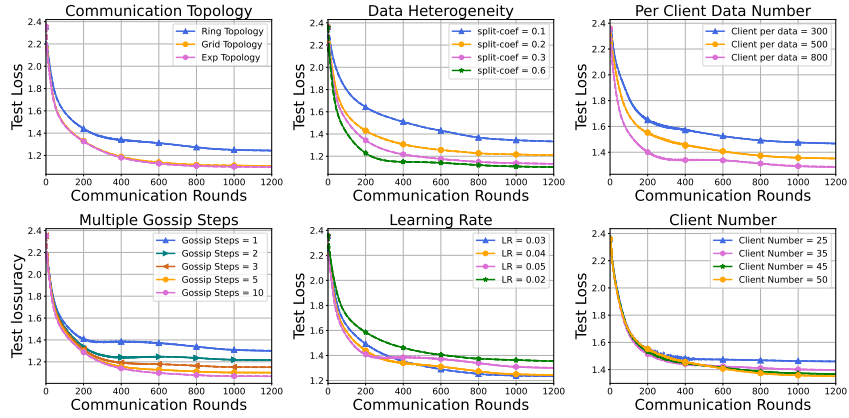


Figure 3: The test loss of the DSGD-MGS algorithm on the cifar10 test dataset.

According to the definition of excess error, $\mathbb{E}_{A,S}[R(A(S)) - R(\theta^*)]$, since θ^* represents the global optimal solution and is independent of the dataset S (i.e., $\mathbb{E}_{A,S}[R(\theta^*)]$ is a constant), the magnitude relationship of $\mathbb{E}_{A,S}[R(A(S)) - R(\theta^*)]$ is equivalent to that of $\mathbb{E}_{A,S}[R(A(S))]$. Therefore, the relative test errors $\mathbb{E}_{A,S}[R(A(S))]$ for different parameter settings can directly reflect the corresponding relationships in excess error.

Notably, our theoretical results provide, for the first time, excess error bounds for DSGD-MGS under non-convex assumptions and heterogeneous data, making them more general than the strongly convex results in [23]. As shown in the "Data Heterogeneity" section of Figure 3, the experimental findings align perfectly with our theoretical predictions, demonstrating that increasing data heterogeneity leads to higher excess error (see Theorem 4 and Remark 5).

Additionally, we conducted experiments with a fixed number of clients but varying per-client data sizes. As illustrated in the "Per Client Data Number" subplot of Figure 3, increasing the amount of data per client (i.e., n in Theorem 4) reduces excess error, exhibiting a nearly linear speedup, which is consistent with our theoretical analysis. It is also worth noting the learning rate experiments. Comparing the "Learning Rate" results in Figure 3 and Figure 2, we observe that generalization error decreases with smaller learning rates, while excess error shows a more nuanced trade-off, aligning well with our discussion in Remark 7 of Theorem 4.

A.3 More Experiments Results of DSGD-MGS on CIFAR10

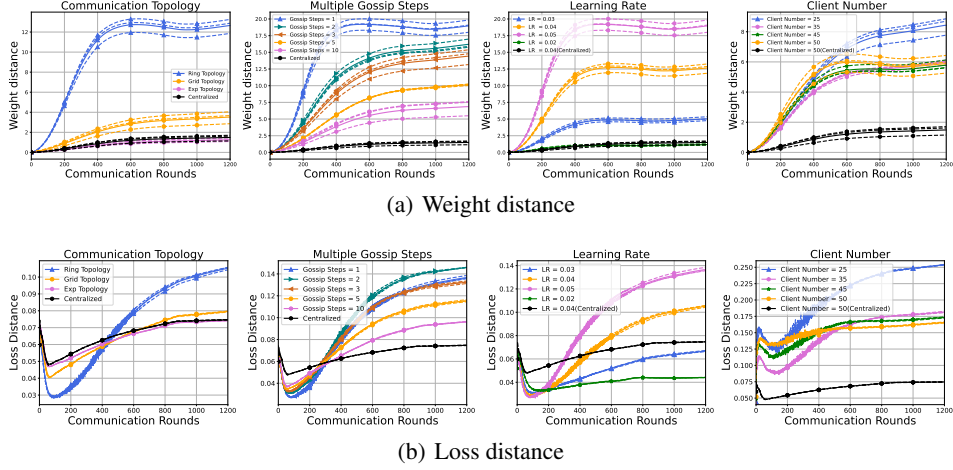


Figure 4: A comparison of the l_2 weight distance and Loss distance (i.e. test loss - train loss) for the DSGD-MGS algorithm on the cifar10 dataset with centralized methods.

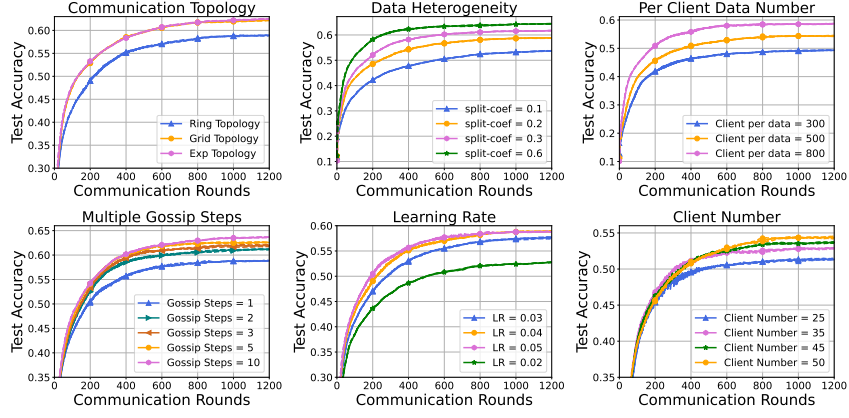


Figure 5: The accuracy of the DSGD-MGS algorithm on the cifar10 test dataset.

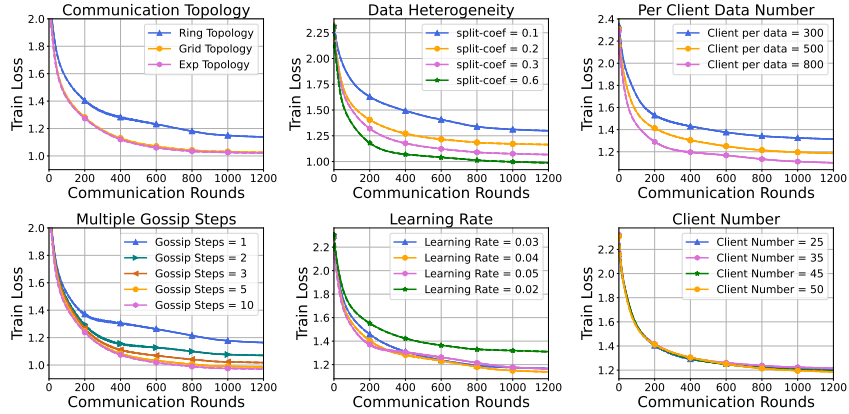


Figure 6: The loss of the DSGD-MGS algorithm on the cifar10 train dataset.

B Proof of generalization error for DSGD-MGS

We first present an important property of smooth functions, followed by the proof of Lemma 1. Subsequently, we provide the proofs of other lemmas and theorems appearing in the main text.

Lemma 3. [Case of $\alpha = 1$ in [33]] Assume that for all $z \in \mathcal{Z}$, the mapping $\theta \mapsto \ell(\theta; z)$ is non-negative, and its gradient $\theta \mapsto \nabla \ell(\theta; z)$ is $(1, \beta)$ -Hölder continuous (Assumption 1). Then there exists a constant $c_{1,1} = \sqrt{2\beta}$, such that for all $\theta \in \mathbb{R}^q$ and $z \in \mathcal{Z}$,

$$\|\nabla \ell(\theta; z)\|_2 \leq c_{1,1} \cdot \sqrt{\ell(\theta; z)}.$$

The above lemma 3 is also known as the self-bounding property of the function ℓ . Next, to prove Lemma 1, we introduce a useful inequality for β -smooth functions ℓ .

$$\ell(\theta; Z) \leq \ell(\tilde{\theta}; Z) + \langle \theta - \tilde{\theta}, \nabla \ell(\tilde{\theta}; Z) \rangle + \frac{\beta \|\theta - \tilde{\theta}\|_2^2}{2}. \quad (\text{B.1})$$

Proof of Lemma 1. Due to the symmetry, we know

$$\begin{aligned} \mathbb{E}_{S,A} [R(A(S)) - R_S(A(S))] &= \mathbb{E}_{S,\tilde{S},A} \left[\frac{1}{nm} \sum_{i,j} \left(R(A(S^{(ij)})) - R_S(A(S)) \right) \right] \\ &= \mathbb{E}_{S,\tilde{S},A} \left[\frac{1}{nm} \sum_{i,j} \left(\ell(A(S^{(ij)}); Z_{ij}) - \ell(A(S); Z_{ij}) \right) \right] \end{aligned} \quad (\text{B.2})$$

Since the loss function ℓ satisfies β -smoothness (B.1), we have:

$$\begin{aligned} \frac{1}{nm} \sum_{i,j} \ell(A(S^{(ij)}); Z_{ij}) &\leq \frac{1}{nm} \sum_{i,j} \ell(A(S); Z_{ij}) + \frac{1}{nm} \sum_{i,j} \left\langle A(S^{(ij)}) - A(S), \nabla \ell(A(S); Z_{ij}) \right\rangle \\ &\quad + \frac{\beta}{2nm} \sum_{i,j} \|A(S^{(ij)}) - A(S)\|^2 \end{aligned} \quad (\text{B.3})$$

That is

$$\begin{aligned} \mathbb{E}_{S,A} [R(A(S)) - R_S(A(S))] &\leq \frac{1}{nm} \sum_{i,j} \left\langle A(S^{(ij)}) - A(S), \nabla \ell(A(S); Z_{ij}) \right\rangle \\ &\quad + \frac{\beta}{2nm} \sum_{i,j} \|A(S^{(ij)}) - A(S)\|^2 \end{aligned}$$

According to the Schwartz's inequality we know

$$\begin{aligned} \left\langle A(S^{(ij)}) - A(S), \nabla \ell(A(S); Z_{ij}) \right\rangle &\leq \|A(S^{(ij)}) - A(S)\|_2 \|\nabla \ell(A(S); Z_{ij})\|_2 \\ &\leq \frac{\gamma}{2} \|A(S^{(ij)}) - A(S)\|_2^2 + \frac{1}{2\gamma} \|\nabla \ell(A(S); Z_{ij})\|_2^2 \end{aligned}$$

Combining the above two inequalities together, we derive

$$\begin{aligned} \mathbb{E}_{S,A} [R(A(S)) - R_S(A(S))] &\leq \frac{1}{2mn\gamma} \sum_{i,j} \mathbb{E}_{A,S} [\|\nabla \ell(A(S); Z_{ij})\|^2] \\ &\quad + \frac{\beta + \gamma}{2mn} \sum_{i,j} \mathbb{E}_{A,\tilde{A},S} [\|A(S) - A(S^{(ij)})\|^2] \end{aligned}$$

Thus, we have completed the proof.

B.1 Proof of Important Lemma

Our analysis for the non-convex case relies on l_2 on-average model stability and leverages the fact that D-SGD can make several steps before using the one example that has been swapped. This idea is summarized in the following lemma.

Lemma 4. Assume that the loss function $\ell(\cdot, z)$ is nonnegative for all z . For all $i = 1, \dots, n$ and $j = 1, \dots, m$, let $\{\theta_k^{(t)}\}_{t=0}^T$ and $\{\tilde{\theta}_k^{(t)}(i, j)\}_{t=0}^T$, the iterates of agent $k = 1, \dots, m$ for DSGD-MGS run on S and $S^{(ij)}$ respectively. Then, for every $t_0 \in \{0, 1, \dots, T\}$ we have:

$$\begin{aligned} |\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| &\leq \frac{t_0}{n} \sup_{\theta, z} \ell(\theta; z) + \frac{1}{2mn\gamma} \sum_{i=1}^n \sum_{j=1}^m \mathbb{E}[\|\nabla \ell(A_k(S); Z_{ij})\|^2] \\ &\quad + \frac{\gamma + \beta}{2mn} \sum_{i=1}^n \sum_{j=1}^m \mathbb{E}[\delta_k^{(T)}(i, j) | \delta^{(t_0)}(i, j) = \mathbf{0}] \end{aligned}$$

where $\delta^{(t)}(i, j)$ is the vector containing $\forall k = 1, \dots, m$, $\delta_k^{(t)}(i, j) = \|\theta_k^{(t)} - \tilde{\theta}_k^{(t)}(i, j)\|_2^2$.

Proof. Consider the notation of Def. 1 and notice that

$$R(A_k(S)) = \frac{1}{m} \sum_{j=1}^m \mathbb{E}_{Z \sim \mathcal{D}_j}[\ell(A_k(S); Z)] = \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{\tilde{S}}[\ell(A_k(S); \tilde{Z}_{ij})].$$

Then, for all $k = 1, \dots, m$, by linearity of expectation we have

$$\begin{aligned} \mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))] &= \mathbb{E}_{A,S,\tilde{S}} \left[\frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n \left(\ell(A_k(S); \tilde{Z}_{ij}) - \ell(A_k(S); Z_{ij}) \right) \right] \\ &= \mathbb{E}_{A,S,\tilde{S}} \left[\frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n \left(\ell(A_k(S^{(ij)}); Z_{ij}) - \ell(A_k(S); Z_{ij}) \right) \right]. \end{aligned}$$

Hence,

$$\begin{aligned} |\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| &\leq \mathbb{E}_{A,S,\tilde{S}} \left[\frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n \left| \ell(A_k(S^{(ij)}); Z_{ij}) - \ell(A_k(S); Z_{ij}) \right| \right] \\ &= \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{A,S,\tilde{S}} \left[\left| \ell(A_k(S^{(ij)}); Z_{ij}) - \ell(A_k(S); Z_{ij}) \right| \right] \end{aligned}$$

Let the event $\mathcal{E}(i, j) = \{\delta^{(t_0)}(i, j) = \mathbf{0}\}$, we have $\forall i, j$:

$$\begin{aligned} \mathbb{E}_{A,S,\tilde{S}} \left[\left| \ell(A_k(S^{(ij)}); Z_{ij}) - \ell(A_k(S); Z_{ij}) \right| \right] &= \mathbb{P}(\mathcal{E}(i, j)) \mathbb{E}[\ell(A_k(S^{(ij)}); Z_{ij}) - \ell(A_k(S); Z_{ij}) | \mathcal{E}(i, j)] \\ &\quad + \mathbb{P}(\mathcal{E}(i, j)^c) \mathbb{E}[\ell(A_k(S^{(ij)}); Z_{ij}) - \ell(A_k(S); Z_{ij}) | \mathcal{E}(i, j)^c] \\ &\leq \mathbb{E}[\ell(A_k(S^{(ij)}); Z_{ij}) - \ell(A_k(S); Z_{ij}) | \mathcal{E}(i, j)] + \mathbb{P}(\mathcal{E}(i, j)^c) \cdot \sup_{\theta, z} \ell(\theta; z) \end{aligned} \tag{B.4}$$

Considering the smoothness of ℓ , we have:

$$\begin{aligned} \ell(A_k(S^{(ij)}); Z_{ij}) - \ell(A_k(S); Z_{ij}) &= \left\langle A_k(S^{(ij)}) - A_k(S), \nabla \ell(A_k(S); Z_{ij}) \right\rangle + \frac{\beta}{2} \|A_k(S^{(ij)}) - A_k(S)\|^2 \\ &\leq \frac{\gamma}{2} \|A_k(S^{(ij)}) - A_k(S)\|^2 + \frac{1}{2\gamma} \|\nabla \ell(A_k(S); Z_{ij})\|^2 + \frac{\beta}{2} \|A_k(S^{(ij)}) - A_k(S)\|^2 \\ &= \frac{\gamma + \beta}{2} \|A_k(S^{(ij)}) - A_k(S)\|^2 + \frac{1}{2\gamma} \|\nabla \ell(A_k(S); Z_{ij})\|^2 \end{aligned} \tag{B.5}$$

Where the second inequality uses the bound $\langle a, b \rangle \leq \frac{\gamma}{2} \|a\|^2 + \frac{1}{2\gamma} \|b\|^2$, which holds for any $\gamma > 0$. Combining the above inequality (B.5) with inequality (B.4), we obtain:

$$\begin{aligned}
& \mathbb{E}_{A,S,\tilde{S}} \left[\left| \ell(A_k(S^{(ij)}); Z_{ij}) - \ell(A_k(S); Z_{ij}) \right| \right] \\
& \leq \mathbb{E}[\ell(A_k(S^{(ij)}); Z_{ij}) - \ell(A_k(S); Z_{ij}) | \mathcal{E}(i, j)] + \mathbb{P}(\mathcal{E}(i, j)^c) \cdot \sup_{\theta, z} \ell(\theta; z) \\
& \leq \frac{\gamma + \beta}{2} \mathbb{E}[\|A_k(S) - A_k(S^{(ij)})\|^2 | \mathcal{E}(i, j)] \\
& \quad + \frac{1}{2\gamma} \mathbb{E}[\|\nabla \ell(A_k(S); Z_{ij})\|^2 | \mathcal{E}(i, j)] + \mathbb{P}(\mathcal{E}(i, j)^c) \cdot \sup_{\theta, z} \ell(\theta; z) \\
& = \frac{\gamma + \beta}{2} \mathbb{E}[\delta_k^{(T)}(i, j) | \mathcal{E}(i, j)] + \frac{1}{2\gamma} \mathbb{E}[\|\nabla \ell(A_k(S); Z_{ij})\|^2] + \mathbb{P}(\mathcal{E}(i, j)^c) \cdot \sup_{\theta, z} \ell(\theta; z)
\end{aligned} \tag{B.6}$$

The last equality follows from the independence between $\|\nabla \ell(A_k(S); Z_{ij})\|^2$ and $\mathcal{E}(i, j)$. It remains to bound $\mathbb{P}(\mathcal{E}(i, j)^c)$. Let T_0 be the random variable of the first time step DSGD-MGS uses the swapped example. Since we necessarily have $\{T_0 > t_0\} \subset \mathcal{E}(i, j)$, we have $\mathcal{E}(i, j)^c \subset \{T_0 \leq t_0\}$ and therefore $\mathbb{P}(\mathcal{E}(i, j)^c) \leq \mathbb{P}(T_0 \leq t_0) = \sum_{t=1}^{t_0} \mathbb{P}(T_0 = t) \leq \sum_{t=1}^{t_0} \frac{1}{n} = \frac{t_0}{n}$. Averaging over i and j completes the proof. \square

We can now move on to the proof of the main theorem. We first apply Lemma 4 and the fact that, by assumption, $\ell \in [0, 1]$, so that for any $t_0 \in \{0, 1, \dots, T\}$ and any $k = 1, \dots, m$, we have:

$$\begin{aligned}
|\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| & \leq \frac{t_0}{n} + \frac{1}{2mn\gamma} \sum_{i,j} \mathbb{E}[\|\nabla \ell(A_k(S); Z_{ij})\|^2] \\
& \quad + \frac{\gamma + \beta}{2mn} \sum_{i,j} \mathbb{E}[\delta_k^{(T)}(i, j) | \delta^{(t_0)}(i, j) = \mathbf{0}]
\end{aligned} \tag{B.7}$$

It remains to control the right-hand term of Equation (B.7). We start with the proof for DSGD-MGS.

B.2 Proof of l_2 on average model stability of DSGD-MGS.

For a fixed couple (i, j) , we are first going to control the vector $\Delta^{(t)}(i, j) \triangleq \mathbb{E}[\delta^{(t)}(i, j) | \delta^{(t_0)}(i, j) = \mathbf{0}]$, where $\delta^{(t)}(i, j)$ is the vector containing $\forall k = 1, \dots, m$, $\delta_k^{(t)}(i, j) = \|\theta_k^{(t)} - \tilde{\theta}_k^{(t)}(i, j)\|_2^2$. When it is clear from context, we simply write $\tilde{\theta}_k^{(t)}(i, j) = \tilde{\theta}_k^{(t)}$.

We first estimate $\|\theta_k^{(t+1)} - \tilde{\theta}_k^{(t+1)}\|_2^2$.

$$\begin{aligned}
\|\theta_k^{(t+1)} - \tilde{\theta}_k^{(t+1)}\|_2^2 & = \left\| \sum_{l=1}^m W_{kl} \left[\theta_l^{(t)} - \tilde{\theta}_l^{(t)} + \eta_t \left(\nabla \ell(\tilde{\theta}_l^{(t)}; Z_{I_l^t l}) - \nabla \ell(\theta_l^{(t)}; Z_{I_l^t l}) \right) \right] \right\|^2 \\
& \leq \sum_{l=1}^m W_{kl} \left\| \theta_l^{(t)} - \tilde{\theta}_l^{(t)} + \eta_t \left(\nabla \ell(\tilde{\theta}_l^{(t)}; Z_{I_l^t l}) - \nabla \ell(\theta_l^{(t)}; Z_{I_l^t l}) \right) \right\|^2 \\
& \leq \sum_{l \neq j}^m W_{kl} \left\| \theta_l^{(t)} - \tilde{\theta}_l^{(t)} + \eta_t \left(\nabla \ell(\tilde{\theta}_l^{(t)}; Z_{I_l^t l}) - \nabla \ell(\theta_l^{(t)}; Z_{I_l^t l}) \right) \right\|^2 \\
& \quad + W_{kj} \left\| \theta_j^{(t)} - \tilde{\theta}_j^{(t)} + \eta_t \left(\nabla \ell(\tilde{\theta}_j^{(t)}; Z_{I_j^t j}) - \nabla \ell(\theta_j^{(t)}; Z_{I_j^t j}) \right) \right\|^2 \\
& \leq (1 + \eta_t \beta)^2 \sum_{l \neq j}^m W_{kl} \|\theta_k^{(t)} - \tilde{\theta}_k^{(t)}\|_2^2 \\
& \quad + W_{kj} \left\| \theta_j^{(t)} - \tilde{\theta}_j^{(t)} + \eta_t \left(\nabla \ell(\tilde{\theta}_j^{(t)}; Z_{I_j^t j}) - \nabla \ell(\theta_j^{(t)}; Z_{I_j^t j}) \right) \right\|^2
\end{aligned} \tag{B.8}$$

The first inequality in the above expression follows from Jensen's inequality, and the last inequality follows from the $(1 + \eta_t \beta)$ -expansiveness of ℓ [41] when $l \neq j$. Next, we perform a analysis of the second term on the right-hand side of the above inequality.

With probability $1 - \frac{1}{n}$, $I_j^t \neq i$ so $Z_{I_j^t j} = Z'_{I_j^t j}$. We have

$$\left\| \theta_j^{(t)} - \tilde{\theta}_j^{(t)} + \eta_t \left(\nabla \ell(\tilde{\theta}_j^{(t)}; Z_{I_j^t j}) - \nabla \ell(\theta_j^{(t)}; Z_{I_j^t j}) \right) \right\|^2 \leq (1 + \eta_t \beta)^2 \|\theta_j^{(t)} - \tilde{\theta}_j^{(t)}\|^2 \quad (\text{B.9})$$

With probability $\frac{1}{n}$, $I_j^t = i$ and in that case $Z_{I_j^t j} = Z_{ij} \neq \tilde{Z}_{ij} = Z'_{I_j^t j}$.

$$\begin{aligned} \left\| \theta_j^{(t)} - \tilde{\theta}_j^{(t)} + \eta_t \left(\nabla \ell(\tilde{\theta}_j^{(t)}; Z'_{ij}) - \nabla \ell(\theta_j^{(t)}; Z_{ij}) \right) \right\|^2 &\leq (1 + p) \|\theta_j^{(t)} - \tilde{\theta}_j^{(t)}\|^2 \\ &\quad + 2\eta_t^2(1 + p^{-1}) \|\nabla \ell(\tilde{\theta}_j^{(t)}; Z'_{ij})\|^2 + 2\eta_t^2(1 + p^{-1}) \|\nabla \ell(\theta_j^{(t)}; Z_{ij})\|^2 \end{aligned} \quad (\text{B.10})$$

Considering that I_k^t follows a uniform distribution ($I_k^t \sim \mathcal{U}\{1, \dots, n\}$), we get

$$\begin{aligned} \left\| \theta_j^{(t)} - \tilde{\theta}_j^{(t)} + \eta_t \left(\nabla \ell(\tilde{\theta}_j^{(t)}; Z'_{ij}) - \nabla \ell(\theta_j^{(t)}; Z_{ij}) \right) \right\|^2 &\leq (1 + \frac{p}{n})(1 + \eta_t \beta)^2 \left\| \theta_j^{(t)} - \tilde{\theta}_j^{(t)} \right\|^2 \\ &\quad + \frac{2\eta_t^2(1 + p^{-1})}{n} \|\nabla \ell(\tilde{\theta}_j^{(t)}; Z'_{ij})\|^2 + \frac{2\eta_t^2(1 + p^{-1})}{n} \|\nabla \ell(\theta_j^{(t)}; Z_{ij})\|^2 \end{aligned} \quad (\text{B.11})$$

Substituting equation (B.11) into equation (B.8), we obtain:

$$\begin{aligned} \|\theta_k^{(t+1)} - \tilde{\theta}_k^{(t+1)}\|_2^2 &\leq (1 + \frac{p}{n})(1 + \eta_t \beta)^2 \sum_{l=1}^m W_{kl} \|\theta_k^{(t)} - \tilde{\theta}_k^{(t)}\|_2^2 \\ &\quad + \frac{2\eta_t^2(1 + p^{-1})}{n} W_{kj} \left(\|\nabla \ell(\tilde{\theta}_j^{(t)}; Z'_{ij})\|^2 + \|\nabla \ell(\theta_j^{(t)}; Z_{ij})\|^2 \right) \end{aligned} \quad (\text{B.12})$$

Given that $\mathbb{E}_{S, \tilde{S}, A} [\|\nabla \ell(\theta_j^{(t)}; Z_{ij})\|^2] = \mathbb{E}_{S, \tilde{S}, A} [\|\nabla \ell(\tilde{\theta}_j^{(t)}; Z'_{ij})\|^2]$, and to simplify the notation, we denote $\mathbb{E}_{S, \tilde{S}, A}[\cdot] = \mathbb{E}[\cdot]$, we then obtain the following:

$$\begin{aligned} \mathbb{E}[\|\theta_k^{(t+1)} - \tilde{\theta}_k^{(t+1)}\|_2^2] &\leq (1 + \frac{p}{n})(1 + \eta_t \beta)^2 \sum_{l=1}^m W_{kl} \mathbb{E}[\|\theta_k^{(t)} - \tilde{\theta}_k^{(t)}\|_2^2] \\ &\quad + \frac{4\eta_t^2(1 + p^{-1})}{n} W_{kj} \mathbb{E}[\|\nabla \ell(\theta_j^{(t)}; Z_{ij})\|^2] \end{aligned} \quad (\text{B.13})$$

From the previous equations and let the vector $G^{(t)} \in \mathbb{R}^m$ be defined such that its j -th component is $G_j^{(t)} = \mathbb{E}[\|\nabla \ell(\theta_j^{(t)}; Z_{ij})\|^2]$, we get that $\Delta^{(t+1)}(i, j) \leq (1 + \frac{p}{n})(1 + \eta_t \beta)^2 W \Delta^{(t)}(i, j) + \frac{4\eta_t^2(1 + p^{-1})}{n} W_j \circ G^{(t)}$ (the inequality, and the following ones are meant coordinate-wise), where W_j denotes the j -th column of matrix W , and \circ represents the Hadamard product. Let $\Delta^{(t)} = \frac{1}{mn} \sum_{i,j} \Delta^{(t)}(i, j)$, then using the fact that $\eta_t \leq \frac{c}{t+1}$, $c > 0$, we have $\forall t \geq t_0$:

$$\begin{aligned} \Delta^{(t+1)} &\leq (1 + \eta_t \beta)^2 (1 + \frac{p}{n}) W \Delta^{(t)} + \frac{4\eta_t^2(1 + p^{-1})}{nm} \sum_{j=1}^m W_j \circ G^{(t)} \\ &= (1 + \eta_t \beta)^2 (1 + \frac{p}{n}) W \Delta^{(t)} + \frac{4\eta_t^2(1 + p^{-1})}{nm} G^{(t)} \\ &\leq (1 + \frac{c\beta}{t+1})^2 (1 + \frac{p}{n}) W \Delta^{(t)} + \frac{4(1 + p^{-1})}{nm} \frac{c^2}{(t+1)^2} G^{(t)} \end{aligned} \quad (\text{B.14})$$

Using Lemma 3 and the assumption that $\ell \in [0, 1]$, we can derive $G^{(t)} \leq \sqrt{2\beta} \mathbf{1}$, where $\mathbf{1}$ is the all-ones vector. Then

$$\Delta^{(t+1)} \leq (1 + \frac{c\beta}{t+1})^2 (1 + \frac{p}{n}) W \Delta^{(t)} + \frac{4\sqrt{2\beta}(1 + p^{-1})}{nm} \frac{c^2}{(t+1)^2} \mathbf{1} \quad (\text{B.15})$$

Since $\Delta^{(t_0)} = \mathbf{0}$, we can unroll the previous recursion from T to $t_0 + 1$ and get:

$$\Delta^{(T)} \leq \sum_{s=t_0}^{T-1} \left(\prod_{k=s+1}^{T-1} \left(1 + \frac{c\beta}{k+1} \right)^2 \left(1 + \frac{p}{n} \right) W \right) \cdot \frac{4\sqrt{2}\beta c^2(1+p^{-1})}{nm(s+1)^2} \mathbf{1}. \quad (\text{B.16})$$

Then, we focus on the coordinate of interest k and using the fact that $1+x \leq \exp(x)$, we have:

$$\begin{aligned} \Delta_k^{(T)} &\leq \sum_{s=t_0}^{T-1} \left(\prod_{k=s+1}^{T-1} \exp\left(\frac{2c\beta}{k+1}\right) \left(1 + \frac{p}{n} \right) \right) \cdot \frac{4\sqrt{2}\beta c^2(1+p^{-1})}{nm(s+1)^2} \\ &\leq \sum_{s=t_0}^{T-1} \left(\left(1 + \frac{p}{n} \right)^{T-s-1} W^{T-s-1} \exp\left(2c\beta \sum_{k=s+1}^{T-1} \frac{1}{k+1}\right) \right) \cdot \frac{4\sqrt{2}\beta c^2(1+p^{-1})}{nm(s+1)^2} \\ &\leq \sum_{s=t_0}^{T-1} \left(\left(1 + \frac{p}{n} \right)^{T-s-1} W^{T-s-1} \exp\left(2c\beta \log\left(\frac{T}{s+1}\right)\right) \right) \cdot \frac{4\sqrt{2}\beta c^2(1+p^{-1})}{nm(s+1)^2} \quad (\text{B.17}) \\ &= \sum_{s=t_0}^{T-1} \left(\left(1 + \frac{p}{n} \right)^{T-s-1} W^{T-s-1} \left(\frac{T}{s+1} \right)^{2c\beta} \right) \cdot \frac{4\sqrt{2}\beta c^2(1+p^{-1})}{nm(s+1)^2} \\ &= \sum_{s=t_0}^{T-1} \left(\left(1 + \frac{p}{n} \right)^{T-s-1} \left(\frac{T}{s+1} \right)^{2c\beta} \right) \cdot \frac{4\sqrt{2}\beta c^2(1+p^{-1})}{nm(s+1)^2}. \end{aligned}$$

Let $p = \frac{n}{T-t_0-1} > 1$, then for $s \geq t_0$, we have $\left(1 + \frac{p}{n} \right)^{T-s-1} < \left(1 + \frac{1}{T-t_0-1} \right)^{T-t_0-1} < e$, and also $1 + p^{-1} < 2$, where e is euler's number. Then, we have

$$\begin{aligned} \Delta_k^{(T)} &\leq \sum_{s=t_0}^{T-1} \left(\frac{T}{s+1} \right)^{2c\beta} \cdot \frac{8e\sqrt{2}\beta c^2}{nm(s+1)^2} \quad (\text{B.18}) \\ &\leq \frac{8e\sqrt{2}\beta c^2 T^{2c\beta}}{nm} \cdot \int_{t_0}^{T-1} s^{-2c\beta-2} ds \\ &\leq \frac{8e\sqrt{2}\beta c^2}{(1+2c\beta)nmt_0} \left(\frac{T}{t_0} \right)^{2c\beta} \end{aligned}$$

We then derive the component-wise form of inequality (B.18).

$$\frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m \mathbb{E}[\delta_k^{(T)}(i, j) | \delta^{(t_0)}(i, j) = \mathbf{0}] \leq \frac{8e\sqrt{2}\beta c^2}{(1+2c\beta)nmt_0} \left(\frac{T}{t_0} \right)^{2c\beta} \quad (\text{B.19})$$

B.3 Proof of generalization of DSGD-MGS

By substituting (B.19) into (B.7), we obtain the following.

$$\begin{aligned} |\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| &\leq \frac{t_0}{n} + \frac{1}{2mn\gamma} \sum_{i,j} \mathbb{E}[\|\nabla \ell(A_k(S); Z_{ij})\|^2] \\ &\quad + \frac{\gamma + \beta}{2} \frac{8e\sqrt{2}\beta c^2}{(1+2c\beta)nmt_0} \left(\frac{T}{t_0} \right)^{2c\beta} \quad (\text{B.20}) \end{aligned}$$

Treating equation (B.20) as a function of t_0 , and noting that the left-hand side is independent of t_0 , equation (B.20) holds for any t_0 . Without loss of generality, let $t_0 = \left(\frac{4(\gamma+\beta)e\sqrt{2}\beta c^2 T^{2c\beta}}{m} \right)^{\frac{1}{2c\beta+2}}$, yielding the following expression.

$$\begin{aligned} |\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| &\leq \frac{1}{2mn\gamma} \sum_{i,j} \mathbb{E}[\|\nabla \ell(A_k(S); Z_{ij})\|^2] \\ &\quad + \frac{2(c\beta+1)}{n(2c\beta+1)} \cdot \left(\frac{4(\gamma+\beta)e\sqrt{2}\beta c^2 T^{2c\beta}}{m} \right)^{\frac{1}{2c\beta+2}} \quad (\text{B.21}) \end{aligned}$$

Let $f(\gamma)$ denote the right-hand side of equation (B.21). We proceed to analyze the approximate minimum of $f(\gamma)$. Let $C_1 = \frac{1}{2mn} \sum_{i,j} \mathbb{E}[\|\nabla \ell(A_k(S); Z_{ij})\|^2]$, $C_2 = \frac{2(c\beta+1)}{n(2c\beta+1)} \left(\frac{4e\sqrt{2}\beta c^2 T^{2c\beta}}{m} \right)^{\frac{1}{2c\beta+2}}$, and $\alpha = \frac{1}{2c\beta+2}$. We aim to find an upper bound for the minimum value of the function $f(\gamma)$ defined as:

$$f(\gamma) = C_1\gamma^{-1} + C_2(\gamma + \beta)^\alpha$$

where $\gamma > 0$, $\beta > 0$. We assume $c \geq 1$ and $\beta > 0$, which implies $2c\beta + 2 > 2$, and thus $0 < \alpha < 1/2$.

Finding the exact minimum of $f(\gamma)$ requires solving $f'(\gamma) = -C_1\gamma^{-2} + \alpha C_2(\gamma + \beta)^{\alpha-1} = 0$, which yields the equation $\frac{\gamma^2}{(\gamma + \beta)^{1-\alpha}} = \frac{C_1}{\alpha C_2}$. This equation is generally intractable to solve analytically for γ . Therefore, it is not amenable to analysis, and we need to approximate $f(\gamma)$ to enable an explicit analysis of the upper bound on the generalization error. Next, we employ inequalities to derive an analytically tractable approximation of the generalization bound.

Seeking an analytically tractable approximation of the generalization bound:

We seek an analytically tractable upper bound for the minimum value, $\min_{\gamma>0} f(\gamma)$. We utilize the standard inequality $(x + y)^p \leq x^p + y^p$ which holds for $x, y > 0$ and $0 < p < 1$. Since $0 < \alpha < 1$, we can apply this inequality to the term $(\gamma + \beta)^\alpha$:

$$(\gamma + \beta)^\alpha \leq \gamma^\alpha + \beta^\alpha$$

Substituting this into the expression for $f(\gamma)$ yields an upper bound:

$$f(\gamma) \leq C_1\gamma^{-1} + C_2(\gamma^\alpha + \beta^\alpha)$$

Let $g(\gamma) = C_1\gamma^{-1} + C_2\gamma^\alpha + C_2\beta^\alpha$. The minimum of $f(\gamma)$ is bounded by the minimum of $g(\gamma)$:

$$\min_{\gamma>0} f(\gamma) \leq \min_{\gamma>0} g(\gamma)$$

We find the minimum of $g(\gamma)$ by setting its derivative with respect to γ to zero:

$$g'(\gamma) = \frac{d}{d\gamma}(C_1\gamma^{-1} + C_2\gamma^\alpha + C_2\beta^\alpha) = -C_1\gamma^{-2} + \alpha C_2\gamma^{\alpha-1}$$

Setting $g'(\gamma) = 0$:

$$\begin{aligned} C_1\gamma^{-2} &= \alpha C_2\gamma^{\alpha-1} \\ \gamma^{\alpha+1} &= \frac{C_1}{\alpha C_2} \end{aligned}$$

The minimizer $\tilde{\gamma}^*$ for $g(\gamma)$ is:

$$\tilde{\gamma}^* = \left(\frac{C_1}{\alpha C_2} \right)^{\frac{1}{\alpha+1}}$$

Substituting $\tilde{\gamma}^*$ back into $g(\gamma)$ gives the minimum value of $g(\gamma)$:

$$\begin{aligned} \min_{\gamma>0} g(\gamma) &= g(\tilde{\gamma}^*) = C_1(\tilde{\gamma}^*)^{-1} + C_2(\tilde{\gamma}^*)^\alpha + C_2\beta^\alpha \\ &= C_1 \left(\frac{C_1}{\alpha C_2} \right)^{\frac{-1}{\alpha+1}} + C_2 \left(\frac{C_1}{\alpha C_2} \right)^{\frac{\alpha}{\alpha+1}} + C_2\beta^\alpha \\ &= C_1^{1-\frac{1}{\alpha+1}} (\alpha C_2)^{\frac{1}{\alpha+1}} + C_2^{1-\frac{\alpha}{\alpha+1}} \left(\frac{C_1}{\alpha} \right)^{\frac{\alpha}{\alpha+1}} + C_2\beta^\alpha \\ &= C_1^{\frac{\alpha}{\alpha+1}} (\alpha C_2)^{\frac{1}{\alpha+1}} + C_2^{\frac{1}{\alpha+1}} C_1^{\frac{\alpha}{\alpha+1}} \alpha^{\frac{-\alpha}{\alpha+1}} + C_2\beta^\alpha \\ &= (C_1^\alpha C_2)^{\frac{1}{\alpha+1}} \left(\alpha^{\frac{1}{\alpha+1}} + \alpha^{\frac{-\alpha}{\alpha+1}} \right) + C_2\beta^\alpha \\ &= (C_1^\alpha C_2)^{\frac{1}{\alpha+1}} \alpha^{\frac{-\alpha}{\alpha+1}} \left(\alpha^{\frac{1+\alpha}{\alpha+1}} + 1 \right) + C_2\beta^\alpha \\ &= (C_1^\alpha C_2)^{\frac{1}{\alpha+1}} \alpha^{\frac{-\alpha}{\alpha+1}} (\alpha + 1) + C_2\beta^\alpha \end{aligned}$$

Thus, we have the upper bound:

$$\min_{\gamma>0} f(\gamma) \leq (\alpha + 1) \alpha^{\frac{-\alpha}{\alpha+1}} (C_1^\alpha C_2)^{\frac{1}{\alpha+1}} + C_2 \beta^\alpha \quad (\text{B.22})$$

Now, we substitute the definitions of C_1 , C_2 , and α . Let $\bar{G} = \frac{1}{mn} \sum_{i,j} \mathbb{E}[\|\nabla \ell(A_k(S); Z_{ij})\|^2]$ denote the average expected squared norm of the gradient. Then $C_1 = \bar{G}/2$. Let $H = \frac{e\sqrt{2\beta}c^2T^{2c\beta}}{m}$. Then $C_2 = \frac{2c\beta+2}{n(2c\beta+1)}(4H)^\alpha$.

We also need the following exponent relations based on $\alpha = \frac{1}{2c\beta+2}$:

$$\begin{aligned} \alpha + 1 &= \frac{1}{2c\beta+2} + 1 = \frac{2c\beta+3}{2c\beta+2} \\ \frac{1}{\alpha+1} &= \frac{2c\beta+2}{2c\beta+3} \\ \frac{\alpha}{\alpha+1} &= \frac{1/(2c\beta+2)}{(2c\beta+3)/(2c\beta+2)} = \frac{1}{2c\beta+3} \\ \frac{-\alpha}{\alpha+1} &= -\frac{1}{2c\beta+3} \end{aligned}$$

Let's evaluate the two terms in the bound (B.22).

First Term: $(\alpha + 1) \alpha^{\frac{-\alpha}{\alpha+1}} (C_1^\alpha C_2)^{\frac{1}{\alpha+1}}$

$$\begin{aligned} C_1^\alpha C_2 &= \left(\frac{\bar{G}}{2}\right)^\alpha \frac{2c\beta+2}{n(2c\beta+1)} (4H)^\alpha = \frac{2c\beta+2}{n(2c\beta+1)} \left(\frac{\bar{G}}{2} \cdot 4H\right)^\alpha \\ &= \frac{2c\beta+2}{n(2c\beta+1)} (2\bar{G}H)^\alpha \\ (C_1^\alpha C_2)^{\frac{1}{\alpha+1}} &= \left(\frac{2c\beta+2}{n(2c\beta+1)}\right)^{\frac{1}{\alpha+1}} (2\bar{G}H)^{\frac{\alpha}{\alpha+1}} \\ &= \left(\frac{2c\beta+2}{n(2c\beta+1)}\right)^{\frac{2c\beta+2}{2c\beta+3}} (2\bar{G}H)^{\frac{1}{2c\beta+3}} \end{aligned}$$

The coefficient is:

$$(\alpha + 1) \alpha^{\frac{-\alpha}{\alpha+1}} = \frac{2c\beta+3}{2c\beta+2} \left(\frac{1}{2c\beta+2}\right)^{-\frac{1}{2c\beta+3}} = \frac{2c\beta+3}{2c\beta+2} (2c\beta+2)^{\frac{1}{2c\beta+3}}$$

Combining these parts for the first term:

$$\begin{aligned} \text{First Term} &= \left(\frac{2c\beta+3}{2c\beta+2} (2c\beta+2)^{\frac{1}{2c\beta+3}}\right) \left(\frac{2c\beta+2}{n(2c\beta+1)}\right)^{\frac{2c\beta+2}{2c\beta+3}} (2\bar{G}H)^{\frac{1}{2c\beta+3}} \\ &= \frac{2c\beta+3}{2c\beta+2} (2c\beta+2)^{\frac{1}{2c\beta+3}} \frac{(2c\beta+2)^{\frac{2c\beta+2}{2c\beta+3}}}{(n(2c\beta+1))^{\frac{2c\beta+2}{2c\beta+3}}} (2\bar{G}H)^{\frac{1}{2c\beta+3}} \\ &= (2c\beta+3) (2c\beta+2)^{-1+\frac{1}{2c\beta+3}+\frac{2c\beta+2}{2c\beta+3}} (n(2c\beta+1))^{-\frac{2c\beta+2}{2c\beta+3}} (2\bar{G}H)^{\frac{1}{2c\beta+3}} \\ &= (2c\beta+3) (n(2c\beta+1))^{-\frac{2c\beta+2}{2c\beta+3}} (2\bar{G}H)^{\frac{1}{2c\beta+3}} \\ &= (2c\beta+3) (n(2c\beta+1))^{-\frac{2c\beta+2}{2c\beta+3}} \left(\frac{2\bar{G}e\sqrt{2\beta}c^2T^{2c\beta}}{m}\right)^{\frac{1}{2c\beta+3}} \end{aligned}$$

Second Term: $C_2 \beta^\alpha$

$$C_2 \beta^\alpha = \left(\frac{2c\beta+2}{n(2c\beta+1)} (4H)^\alpha\right) \beta^\alpha = \frac{2c\beta+2}{n(2c\beta+1)} (4\beta H)^\alpha$$

$$= \frac{2c\beta + 2}{n(2c\beta + 1)} \left(\frac{4\beta e\sqrt{2\beta}c^2T^{2c\beta}}{m} \right)^{\frac{1}{2c\beta+2}}$$

Final Upper Bound: Combining the two terms, we obtain the final upper bound for the minimum value of $f(\gamma)$:

$$\min_{\gamma>0} f(\gamma) \leq \frac{2c\beta + 3}{(n(2c\beta + 1))^{\frac{2c\beta+2}{2c\beta+3}}} \left(\frac{2\bar{G}e\sqrt{2\beta}c^2T^{2c\beta}}{m} \right)^{\frac{1}{2c\beta+3}} + \frac{2c\beta + 2}{n(2c\beta + 1)} \left(\frac{4\beta e\sqrt{2\beta}c^2T^{2c\beta}}{m} \right)^{\frac{1}{2c\beta+2}} \quad (\text{B.23})$$

where $\bar{G} = \frac{1}{mn} \sum_{i,j} \mathbb{E}[\|\nabla \ell(A_k(S); Z_{ij})\|^2]$, where $A_k(S) = \theta_k^{(T)}$.

B.4 Proof of optimization error of DSGD-MGS

Next, we will analyze the expression $\bar{G} = \frac{1}{mn} \sum_{i,j} \mathbb{E}[\|\nabla \ell(\theta_k^{(T)}; Z_{ij})\|^2]$ in detail to further understand the impact of algorithmic parameters in DSGD-MGS on the generalization error bound. Prior to this, since the gradient in equation (D.1) corresponds to the gradient of the final iteration, and current research in the academic community has not yet thoroughly investigated the gradient of the final iteration in non-convex settings for DSGD, we need to clarify an assumption that is widely used in non-convex optimization. This assumption establishes a connection between the gradient and the function value, enabling us to analyze the specific upper bound of the gradient in equation (D.1).

Assumption 4. (Polyak-Łojasiewicz Condition) Under the condition that $R_{S_k}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\theta; Z_{ik})$ also satisfies the β -smoothness property, the objective function $R_S(\theta) = \frac{1}{m} \sum_{k=1}^m R_{S_k}(\theta)$ satisfies the Polyak-Łojasiewicz Condition (PLC) with parameter μ , i.e., for all $\forall \theta \in \mathbb{R}^d$.

$$\|\nabla R_S(\theta)\|^2 \geq 2\mu(R_S(\theta) - R_S^*), \quad \mu > 0, \quad R_S^* = \min_{\theta} R_S(\theta).$$

Next, we proceed to estimate the upper bound of \bar{G} . According to the Bounded Stochastic Gradient Noise assumption (Assumption 2) and the Bounded Stochastic Gradient Noise assumption (Assumption 3), the following inequality holds:

$$\begin{aligned} \bar{G} &= \frac{1}{mn} \sum_{i,j} \mathbb{E}[\|\nabla \ell(\theta_k^{(T)}; Z_{ij})\|^2] = \frac{1}{mn} \sum_{i,j} \mathbb{E}[\|\nabla \ell(\theta_k^{(T)}; Z_{ij}) \pm \nabla R_{S_k}(\theta_k^{(T)}) \pm \nabla R_S(\theta_k^{(T)})\|^2] \\ &\leq 3\sigma^2 + 3\xi^2 + 3\mathbb{E}[\|\nabla R_S(\theta_k^{(T)})\|^2] \end{aligned}$$

Since ℓ satisfies the β -smoothness property, it is straightforward to show that $R_S(\theta_k^{(T)})$ also satisfies the β -smoothness property. Consequently, $R_S(\theta)$ also satisfies the self-bounding property in Lemma 3, i.e., $\|\nabla R_S(\theta)\| \leq 2\beta R_S(\theta)$. Then, we have

$$\bar{G} \leq 3\sigma^2 + 3\xi^2 + 6\beta \mathbb{E}_S[R_S(\theta_k^{(T)})] \quad (\text{B.24})$$

Next, we will focus on bounding $\mathbb{E}_S[R_S(\theta)]$. According to the results from [18] [Theorem 1], we have the following lemma:

Lemma 5. Let $\Delta^2 := \max_{\theta^* \in \mathcal{X}^*} \sum_{k=1}^m \|\nabla R_{S_k}(\theta^*)\|^2$, $R_0 := R_S(\theta^{(0)}) - R_S^*$, where $\mathcal{X}^* = \arg \min_{\theta} R_S(\theta)$ and $R_S^* = R_S(\hat{\theta}_{ERM})$. Suppose Assumptions 1 and 4 hold. Define

$$\begin{aligned} Q_0 &:= \log(\bar{\rho}/46) / \log\left(1 - \frac{\delta\tilde{\gamma}}{2}\right), \quad \bar{\rho} := 1 - \frac{\mu}{m\beta}, \\ \tilde{\gamma} &= \frac{\delta}{\delta^2 + 8\delta + (4 + 2\delta)\lambda_{\max}^2(I - W)}. \end{aligned}$$

Then, if the nodes are initialized such that $\theta_k^Q = 0$, for any $Q > Q_0$ after T iterations the iterates of DSGD-MGS with $\eta_t = \frac{1}{\beta}$ satisfy

$$\mathbb{E}_S[R_S(\theta_k^{(T)})] - R_S^* = \mathcal{O}\left(\frac{\Delta^2 e^{-\frac{\delta\tilde{\gamma}Q}{4}}}{1 - \bar{\rho}} + \left[1 + \frac{\beta}{\mu\bar{\rho}}\left(1 + e^{-\frac{\delta\tilde{\gamma}Q}{4}}\right)\right] R_0 \rho^T\right).$$

Here, δ represents the spectral gap of W , and $\rho \triangleq 1 - \delta = |\lambda_2(W)|$, both of which are defined in detail in Definition 2.

By combining Equation (B.24) with Lemma 2, we obtain the upper bound for \bar{G} .

$$\bar{G} = \mathcal{O}(\sigma^2 + \xi^2 + R_S^*) + \mathcal{O}\left(\frac{\Delta^2 e^{-\frac{\delta\bar{\gamma}Q}{4}}}{1 - \bar{\rho}} + \left[1 + \frac{\beta}{\mu\bar{\rho}} \left(1 + e^{-\frac{\delta\bar{\gamma}Q}{4}}\right)\right] R_0 \rho^T\right).$$

C Consensus Error Analysis

Lemma 6 (Consensus Error Recursion for DSGD-MGS). *Consider the DSGD-MGS algorithm (Algorithm 1) under Assumptions 1 (β -smoothness, with $\ell(\theta; z) \in [0, 1]$ implying gradient bound via Lemma 3), 2 (bounded stochastic gradient noise σ^2), and 3 (bounded heterogeneity δ^2), using a symmetric doubly stochastic communication matrix W with $\rho = |\lambda_2(W)| < 1$ (Definition 2). Let $x_t = \mathbb{E}[\frac{1}{m} \sum_{k=1}^m \|\theta_k^{(t)} - \bar{\theta}^{(t)}\|^2]$ be the average consensus error at the start of iteration t , where $\bar{\theta}^{(t)} = \frac{1}{m} \sum_{k=1}^m \theta_k^{(t)}$. Then, for any iteration $t \geq 0$ and number of gossip steps $Q \geq 1$, the consensus error satisfies the following recursion:*

$$x_{t+1} \leq \rho^{2Q} (2 + 24\beta^2 \eta_t^2) x_t + 24\rho^{2Q} (\sigma^2 + \delta^2) \eta_t^2$$

Proof. The proof proceeds in three steps. Let $\mathbb{E}[\cdot]$ denote expectation conditional on the history \mathcal{F}_t .

Step 1: Bounding the consensus error after local updates. Let $\theta_k^{(t,0)} = \theta_k^{(t)} - \eta_t g_k^{(t)}$ and $\bar{\theta}^{(t,0)} = \bar{\theta}^{(t)} - \eta_t \bar{g}^{(t)}$. The consensus error after the local update is $x_{t,0} = \mathbb{E}[\frac{1}{m} \sum_{k=1}^m \|\theta_k^{(t,0)} - \bar{\theta}^{(t,0)}\|^2]$. We have $\theta_k^{(t,0)} - \bar{\theta}^{(t,0)} = (\theta_k^{(t)} - \bar{\theta}^{(t)}) - \eta_t (g_k^{(t)} - \bar{g}^{(t)})$.

$$\begin{aligned} x_{t,0} &= \mathbb{E} \left[\frac{1}{m} \sum_{k=1}^m \|(\theta_k^{(t)} - \bar{\theta}^{(t)}) - \eta_t (g_k^{(t)} - \bar{g}^{(t)})\|^2 \right] \\ &\leq \mathbb{E} \left[\frac{1}{m} \sum_{k=1}^m \left(2\|\theta_k^{(t)} - \bar{\theta}^{(t)}\|^2 + 2\eta_t^2 \|g_k^{(t)} - \bar{g}^{(t)}\|^2 \right) \right] \\ &= 2x_t + 2\eta_t^2 \mathbb{E} \left[\frac{1}{m} \sum_{k=1}^m \|g_k^{(t)} - \bar{g}^{(t)}\|^2 \right]. \end{aligned} \quad (\text{C.1})$$

where the inequality follows from $\|a - b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Next, we bound the gradient difference term. Let $c = \nabla R_S(\bar{\theta}^{(t)})$. Using $\|x - y\|^2 \leq 2\|x - z\|^2 + 2\|y - z\|^2$ and Jensen's inequality:

$$\begin{aligned} \mathbb{E} \left[\frac{1}{m} \sum_k \|g_k^{(t)} - \bar{g}^{(t)}\|^2 \right] &\leq \mathbb{E} \left[\frac{1}{m} \sum_k (2\|g_k^{(t)} - c\|^2 + 2\|\bar{g}^{(t)} - c\|^2) \right] \\ &= 2\mathbb{E} \left[\frac{1}{m} \sum_k \|g_k^{(t)} - c\|^2 \right] + 2\mathbb{E}[\|\bar{g}^{(t)} - c\|^2] \\ &\leq 2\mathbb{E} \left[\frac{1}{m} \sum_k \|g_k^{(t)} - c\|^2 \right] + 2\mathbb{E} \left[\frac{1}{m} \sum_k \|g_k^{(t)} - c\|^2 \right] \\ &= 4\mathbb{E} \left[\frac{1}{m} \sum_k \|g_k^{(t)} - \nabla R_S(\bar{\theta}^{(t)})\|^2 \right]. \end{aligned} \quad (\text{C.2})$$

Now, we bound the term $\mathbb{E}[\frac{1}{m} \sum_k \|g_k^{(t)} - \nabla R_S(\bar{\theta}^{(t)})\|^2]$ by decomposing it into three parts using the triangle inequality:

$$\begin{aligned} &\mathbb{E} \left[\frac{1}{m} \sum_k \|g_k^{(t)} - \nabla R_S(\bar{\theta}^{(t)})\|^2 \right] \\ &\leq \mathbb{E} \left[\frac{1}{m} \sum_k 3 \left(\|g_k^{(t)} - \nabla R_{S_k}(\theta_k^{(t)})\|^2 + \|\nabla R_{S_k}(\theta_k^{(t)}) - \nabla R_S(\theta_k^{(t)})\|^2 \right) \right] \end{aligned}$$

$$\begin{aligned}
& + \|\nabla R_S(\theta_k^{(t)}) - \nabla R_S(\bar{\theta}^{(t)})\|^2 \Big] \\
& \leq 3(\sigma^2 + \delta^2) + 3\beta^2 x_t.
\end{aligned} \tag{C.3}$$

Here, the first inequality uses $\|a + b + c\|^2 \leq 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$. The second inequality applies Assumption 2, Assumption 3, and Assumption 1 (for the β -smoothness of R_S , which follows from the smoothness of ℓ).

Substituting (C.3) and (C.2) into (C.1):

$$x_{t,0} \leq (2 + 24\beta^2\eta_t^2)x_t + 24(\sigma^2 + \delta^2)\eta_t^2. \tag{C.4}$$

Step 2: Analyzing the effect of Q gossip steps. This step analyzes how the consensus error $x_{t,0} = \mathbb{E}[\frac{1}{m} \sum_{k=1}^m \|\theta_k^{(t,0)} - \bar{\theta}^{(t,0)}\|^2]$ evolves during the Q gossip steps defined in Algorithm 1, line 9-11, resulting in the state $\theta_k^{(t+1)} = \theta_k^{(t,Q)}$ with consensus error $x_{t+1} = \mathbb{E}[\frac{1}{m} \sum_{k=1}^m \|\theta_k^{(t+1)} - \bar{\theta}^{(t+1)}\|^2]$.

First, we establish that the average model parameter is invariant under the gossip updates because W is doubly stochastic (Definition 2). Let $\bar{\theta}^{(t,q)} = \frac{1}{m} \sum_k \theta_k^{(t,q)}$. Then,

$$\begin{aligned}
\bar{\theta}^{(t,q+1)} &= \frac{1}{m} \sum_{k=1}^m \theta_k^{(t,q+1)} = \frac{1}{m} \sum_{k=1}^m \sum_{l=1}^m W_{kl} \theta_l^{(t,q)} \\
&= \frac{1}{m} \sum_{l=1}^m \left(\sum_{k=1}^m W_{kl} \right) \theta_l^{(t,q)}.
\end{aligned}$$

Since W is doubly stochastic, its column sums are equal to 1, i.e., $\sum_{k=1}^m W_{kl} = 1$ for all l . Thus,

$$\bar{\theta}^{(t,q+1)} = \frac{1}{m} \sum_{l=1}^m (1) \theta_l^{(t,q)} = \bar{\theta}^{(t,q)}.$$

By induction, $\bar{\theta}^{(t,Q)} = \bar{\theta}^{(t,Q-1)} = \dots = \bar{\theta}^{(t,0)}$. Therefore, the average model after Q steps is the same as before gossip: $\bar{\theta}^{(t+1)} = \bar{\theta}^{(t,0)}$.

Now, let's analyze the evolution of the deviations from the average. Define the deviation for agent k at gossip step q as $\delta_k^{(t,q)} = \theta_k^{(t,q)} - \bar{\theta}^{(t,0)}$ (note we use the constant average $\bar{\theta}^{(t,0)}$). The initial deviation is $\delta_k^{(t,0)} = \theta_k^{(t,0)} - \bar{\theta}^{(t,0)}$ and the final deviation is $\delta_k^{(t+1)} = \theta_k^{(t+1)} - \bar{\theta}^{(t+1)} = \theta_k^{(t,Q)} - \bar{\theta}^{(t,0)}$. The update rule for the deviations is:

$$\begin{aligned}
\delta_k^{(t,q+1)} &= \theta_k^{(t,q+1)} - \bar{\theta}^{(t,0)} = \sum_{l=1}^m W_{kl} \theta_l^{(t,q)} - \bar{\theta}^{(t,0)} \\
&= \sum_{l=1}^m W_{kl} (\delta_l^{(t,q)} + \bar{\theta}^{(t,0)}) - \bar{\theta}^{(t,0)} \\
&= \sum_{l=1}^m W_{kl} \delta_l^{(t,q)} + \left(\sum_{l=1}^m W_{kl} \right) \bar{\theta}^{(t,0)} - \bar{\theta}^{(t,0)}.
\end{aligned}$$

Since W is doubly stochastic, its row sums are also 1, i.e., $\sum_{l=1}^m W_{kl} = 1$. Therefore,

$$\delta_k^{(t,q+1)} = \sum_{l=1}^m W_{kl} \delta_l^{(t,q)}.$$

Stacking the deviations into a large vector $\delta^{(t,q)} = [\delta_1^{(t,q)\top}, \dots, \delta_m^{(t,q)\top}]^\top \in \mathbb{R}^{md}$, the update becomes $\delta^{(t,q+1)} = (W \otimes I_d) \delta^{(t,q)}$, where I_d is the $d \times d$ identity matrix and \otimes denotes the Kronecker product. After Q steps, we have:

$$\delta^{(t+1)} = (W \otimes I_d)^Q \delta^{(t,0)} = (W^Q \otimes I_d) \delta^{(t,0)}.$$

The consensus error after Q steps is $x_{t+1} = \mathbb{E}[\frac{1}{m} \sum_{k=1}^m \|\delta_k^{(t+1)}\|^2] = \frac{1}{m} \mathbb{E}[\|\delta^{(t+1)}\|^2]$. We bound the squared norm:

$$\|\delta^{(t+1)}\|^2 = \|(W^Q \otimes I_d) \delta^{(t,0)}\|^2$$

$$\leq \|W^Q \otimes I_d\|_2^2 \|\delta^{(t,0)}\|^2.$$

Using the property of the spectral norm for Kronecker products, $\|A \otimes B\|_2 = \|A\|_2 \|B\|_2$, we have:

$$\|W^Q \otimes I_d\|_2 = \|W^Q\|_2 \|I_d\|_2 = \|W^Q\|_2.$$

Since $\delta^{(t,0)}$ represents deviations from the mean, it holds that $\sum_{k=1}^m \delta_k^{(t,0)} = \mathbf{0}_d$. This means $\delta^{(t,0)}$ lies in the subspace orthogonal to the consensus subspace (vectors of the form $\mathbf{1}_m \otimes v$ for $v \in \mathbb{R}^d$). Let $J = \frac{1}{m} \mathbf{1}\mathbf{1}^T$ be the projection onto the consensus subspace in \mathbb{R}^m . The action of W^Q on vectors orthogonal to $\mathbf{1}_m$ is equivalent to the action of $(W - J)^Q$. Therefore, when acting on $\delta^{(t,0)}$, the operator $W^Q \otimes I_d$ acts identically to $(W - J)^Q \otimes I_d$. The spectral norm $\|(W - J)^Q\|_2$ corresponds to the largest magnitude eigenvalue of $(W - J)^Q$ acting on the orthogonal subspace. Since W is symmetric, the eigenvalues of $W - J$ are 0 (corresponding to eigenvector $\mathbf{1}$) and $\lambda_i(W)$ for $i = 2, \dots, m$. The eigenvalues of $(W - J)^Q$ are 0 and $\lambda_i(W)^Q$ for $i = 2, \dots, m$. Thus,

$$\|(W - J)^Q\|_2 = \max_{i=2, \dots, m} |\lambda_i(W)^Q| = \left(\max_{i=2, \dots, m} |\lambda_i(W)| \right)^Q = \rho^Q.$$

where $\rho = |\lambda_2(W)|$ by Definition 2. Therefore, $\|W^Q\|_2$ restricted to the relevant subspace is ρ^Q . It follows that:

$$\|\delta^{(t+1)}\|^2 \leq (\rho^Q)^2 \|\delta^{(t,0)}\|^2 = \rho^{2Q} \|\delta^{(t,0)}\|^2.$$

Taking the expectation and dividing by m :

$$x_{t+1} = \frac{1}{m} \mathbb{E}[\|\delta^{(t+1)}\|^2] \leq \frac{1}{m} \mathbb{E}[\rho^{2Q} \|\delta^{(t,0)}\|^2] = \rho^{2Q} \left(\frac{1}{m} \mathbb{E}[\|\delta^{(t,0)}\|^2] \right) = \rho^{2Q} x_{t,0}. \quad (\text{C.5})$$

This concludes the analysis of the gossip steps.

Step 3: Combining the results. Substituting the bound for $x_{t,0}$ from (C.4) into (C.5) yields the final result:

$$\begin{aligned} x_{t+1} &\leq \rho^{2Q} [(2 + 24\beta^2 \eta_t^2) x_t + 24(\sigma^2 + \delta^2) \eta_t^2] \\ &= \rho^{2Q} (2 + 24\beta^2 \eta_t^2) x_t + 24\rho^{2Q} (\sigma^2 + \delta^2) \eta_t^2. \end{aligned}$$

This concludes the proof. \square

Remark 10 (Implications of Lemma 6). *Lemma 6 establishes a recursive bound for the average consensus error $x_t = \mathbb{E}[\frac{1}{m} \sum_k \|\theta_k^{(t)} - \bar{\theta}^{(t)}\|^2]$. This inequality leads to several key insights regarding the behavior of DSGD-MGS (Algorithm 1):*

1. **Exponential Error Reduction via MGS:** The recursion $x_{t+1} \leq C_t x_t + D_t$ involves coefficients $C_t = \rho^{2Q} (2 + 24\beta^2 \eta_t^2)$ and $D_t = 24\rho^{2Q} (\sigma^2 + \delta^2) \eta_t^2$. Both coefficients are scaled by ρ^{2Q} . Since $\rho = |\lambda_2(W)| < 1$ (Definition 2), increasing the number of gossip steps Q causes ρ^{2Q} to decrease exponentially. Consequently, the influence of past consensus error (x_t) and the injection of new error per iteration (D_t) are exponentially suppressed as Q increases.
2. **Sources of Disagreement:** The term $D_t = 24\rho^{2Q} (\sigma^2 + \delta^2) \eta_t^2$ arises from the local updates. It explicitly depends on the variance of stochastic gradients (σ^2 , Assumption 2) and the variance due to data heterogeneity across agents (δ^2 , Assumption 3). Multiple gossip steps mitigate the impact of these factors by the exponential factor ρ^{2Q} .
3. **Convergence of Consensus Error:** The asymptotic behavior of x_t depends on the step size η_t :
 - Decreasing step size: If $\{\eta_t\}$ satisfies $\sum_{t=0}^{\infty} \eta_t = \infty$ and $\sum_{t=0}^{\infty} \eta_t^2 < \infty$, and if the network connectivity and Q are sufficient such that $2\rho^{2Q} < 1$ (i.e., $\rho^Q < 1/\sqrt{2}$), then the contraction factor $C_t \approx 2\rho^{2Q} < 1$ for large t . Since the noise term D_t is proportional to η_t^2 , we have $\sum_{t=0}^{\infty} D_t < \infty$. Under these conditions, standard results for stochastic approximation (e.g., Robbins-Siegmund lemma) imply that $x_t \rightarrow 0$ as $t \rightarrow \infty$. The models across agents asymptotically reach consensus.

- Constant step size: If $\eta_t = \eta$ is constant, convergence to a steady state requires the contraction factor $C = \rho^{2Q}(2 + 24\beta^2\eta^2)$ to be strictly less than 1. This stability condition, $C < 1$, again necessitates $\rho^Q < 1/\sqrt{2}$ and potentially a small enough step size η . If $C < 1$, iterating the recursion $x_{t+1} \leq Cx_t + D$ (where $D = 24\rho^{2Q}(\sigma^2 + \delta^2)\eta^2$) leads to $\limsup_{t \rightarrow \infty} x_t \leq \frac{D}{1-C} = \frac{24\rho^{2Q}(\sigma^2 + \delta^2)\eta^2}{1 - \rho^{2Q}(2 + 24\beta^2\eta^2)}$. This residual consensus error bound decreases exponentially as Q increases.

4. Approximation of Mini-batch SGD: The lemma shows that x_t can be made arbitrarily small by choosing a sufficiently large Q . When $x_t \approx 0$, all local models are close to the average, i.e., $\theta_k^{(t)} \approx \bar{\theta}^{(t)}$ for all k . The effective gradient used to update the average model $\bar{\theta}^{(t+1)}$ is approximately $\frac{1}{m} \sum_k g_k^{(t)} = \frac{1}{m} \sum_k \nabla \ell(\theta_k^{(t)}; Z_{I_k^t}) \approx \frac{1}{m} \sum_k \nabla \ell(\bar{\theta}^{(t)}; Z_{I_k^t})$. This is precisely the stochastic gradient estimate used by Mini-batch SGD with a batch size of m . Therefore, increasing Q makes DSGD-MGS behave increasingly like Mini-batch SGD, with the deviation (characterized by x_t) decaying exponentially with Q . However, this relationship holds only for the iterative updates and not for the final generalization error bound.

D Additional results and discussions

D.1 On the generalization of $A(S) = \bar{\theta}^{(T)}$

Our generalization bound also holds for the average of the final iterates $A(S) = \bar{\theta}^{(T)} \triangleq \frac{1}{m} \sum_{k=1}^m \theta_k^{(T)}$. We proceed to prove this result.

Proposition 1. Let $A(S) = \bar{\theta}^{(T)}$. Under the same set of hypotheses, except for the form of the gradient expression, the upper-bounds derived in Equation (B.23) also valid upper-bounds on $|\mathbb{E}_{A,S}[R(A(S)) - R_S(A(S))]|$.

Proof. By replacing A_k with A in the proof of Lemma 4 and using the fact that $\ell \in [0, 1]$, we obtain:

$$\begin{aligned} & |\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \\ & \leq \frac{t_0}{n} + \frac{1}{2mn\gamma} \sum_{i,j} \mathbb{E}[\|\nabla \ell(\bar{\theta}^{(T)}; Z_{ij})\|^2] + \frac{\gamma + \beta}{2mn} \sum_{i,j} \mathbb{E}[\|\frac{1}{m} \sum_{k=1}^m (\theta_k^{(T)} - \bar{\theta}_k^{(T)}(i, j))\|_2^2 | \mathcal{E}(i, j)] \\ & \leq \frac{t_0}{n} + \frac{1}{2mn\gamma} \sum_{i,j} \mathbb{E}[\|\nabla \ell(\bar{\theta}^{(T)}; Z_{ij})\|^2] + \frac{1}{m} \sum_{k=1}^m \frac{\gamma + \beta}{2mn} \sum_{i,j} \mathbb{E}[\|\theta_k^{(T)} - \bar{\theta}_k^{(T)}(i, j)\|_2^2 | \mathcal{E}(i, j)] \end{aligned}$$

According to equation (B.19), the upper bound of the third term on the right-hand side is independent of the index k . Moreover, the subsequent estimation of the generalization error bound does not rely on the specific form of the gradient but treats it as a constant. Therefore, following the same derivation as for $A(S) = \theta_k^{(T)}$, we obtain the following inequality.

$$\begin{aligned} & |\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \tag{D.1} \\ & \leq \frac{2c\beta + 3}{(n(2c\beta + 1))^{\frac{2c\beta + 2}{2c\beta + 3}}} \left(\frac{2\bar{G}e\sqrt{2\beta}c^2T^{2c\beta}}{m} \right)^{\frac{1}{2c\beta + 3}} + \frac{2c\beta + 2}{n(2c\beta + 1)} \left(\frac{4\beta e\sqrt{2\beta}c^2T^{2c\beta}}{m} \right)^{\frac{1}{2c\beta + 2}} \end{aligned}$$

where $\bar{G} = \frac{1}{mn} \sum_{i,j} \mathbb{E}[\|\nabla \ell(\bar{\theta}^{(T)}; Z_{ij})\|^2]$. This completes the proof. \square

D.2 Theoretical proof extended to the mini-batch setting.

To make our theory more general, in this section we extend the previous results by incorporating the mini-batch parameter (b). Since most of the proof process remains consistent with the earlier analysis, we present only the key modifications and the final conclusions.

First, we modify Assumption 2 to incorporate the mini-batch parameter (b). This assumption is quite intuitive, since as the batch size increases, the variance of each gradient estimate decreases.

Assumption 5. (Bounded Stochastic Gradient Noise with mini-batchsize b) There exists $\sigma^2 > 0$ such that $\mathbb{E}[\|\frac{1}{b} \sum_{i=1}^b \nabla \ell(\theta; Z_{i,j}) - \nabla R_{S_j}(\theta)\|^2] \leq \frac{\sigma^2}{b}$, for any agent $j \in [m]$ and $\theta \in \mathbb{R}^d$.

Secondly, Algorithm line 6 updated to mini-batch gradient:

$$\theta_k^{(t,0)} = \theta_k^{(t)} - \eta_t \frac{1}{b} \sum_{i=1}^b \nabla \ell(\theta_k^{(t)}; Z_{ik})$$

Thirdly, in the probability calculation below Equation B.6, modify it to:

$$\mathbb{P}(\mathcal{E}(i, j)^c) \leq \mathbb{P}(T_0 \leq t_0) = \sum_{t=1}^{t_0} \mathbb{P}(T_0 = t) \leq \sum_{t=1}^{t_0} \frac{b}{n} = \frac{bt_0}{n}$$

With these modifications, we obtain the following generalization bound for decentralized mini-batch SGD with batch size b :

$$\begin{aligned} & |\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \\ & \leq \frac{(2c\beta + 3)b^{\frac{2c\beta+2}{2c\beta+3}}}{(n(2c\beta + 1))^{\frac{2c\beta+2}{2c\beta+3}}} \left(\frac{2\bar{G}e\sqrt{2\beta}c^2T^{2c\beta}}{m} \right)^{\frac{1}{2c\beta+3}} + \frac{b(2c\beta + 2)}{n(2c\beta + 1)} \left(\frac{4\beta e\sqrt{2\beta}c^2T^{2c\beta}}{m} \right)^{\frac{1}{2c\beta+2}} \quad (\text{D.2}) \end{aligned}$$

It is important to note that the term \bar{G} includes a variance-related component of order $\mathcal{O}(\sigma^2/b)$. Combining this with Equation (D.2) and comparing to the single-sample case, we observe that increasing the batch size b actually increases the generalization bound. This implies that larger batches degrade the generalization ability of the algorithm.

From a stability perspective, when drawing a single sample, the probability of selecting the perturbed sample is $\frac{1}{n}$, whereas for batch size b , it increases to $\frac{b}{n}$. This leads to earlier and larger accumulation of deviation in the stability term $\delta_k^{(t)}(i, j) = \|\theta_k^{(t)} - \tilde{\theta}_k^{(t)}(i, j)\|_2^2$, and ultimately results in a looser stability and generalization bound.

In addition, the work [46] provides a complementary explanation: large batch sizes reduce gradient noise, which increases the likelihood of convergence to sharp minima, known to have poor generalization. In contrast, smaller batches introduce more noise, which helps the model find flatter minima with better generalization. This empirical observation aligns well with our theoretical findings.

D.3 Experimental validation of the relationship between (b) and (Q)

As shown in our newly introduced theory on the mini-batch parameter (b), the batch size represents a trade-off: increasing (b) stabilizes gradient estimates but may compromise stability in other aspects. To further validate this conclusion, we conducted experiments on CIFAR-100 using ResNet-18. The results strongly support our claim regarding the interaction between batch size (b) and the number of MGS steps (Q). Below are the test accuracies (%) after 300 communication rounds, which clearly reveal the complex interplay between (b) and (Q).

Table 1: Test accuracy (%) on CIFAR-100 after 300 communication rounds.

Mini-batch size (b)	Number of MGS steps (Q)			
	1	3	5	10
16	16.08	17.98	18.75	18.95
32	21.75	24.00	24.72	24.95
64	28.38	27.21	27.80	27.39
96	30.39	30.50	30.01	29.66

From Table 1, we obtain the following key observations:

- **Effectiveness of MGS is Conditional on Batch Size:** For smaller batch sizes ($b = 16, b = 32$), increasing the number of MGS steps (Q) consistently and significantly improves performance. For instance, with $b = 32$, increasing Q from 1 to 10 boosts accuracy by over 3 percentage points. This aligns with our theory that frequent communication helps mitigate model divergence when local updates are noisy (due to small b).

- **Diminishing or Negative Returns of MGS with Large Batches:** Conversely, for larger batch sizes ($b = 64, b = 96$), the benefit of increasing Q diminishes or even becomes negative. With $b = 64$, the best performance is achieved with $Q = 1$, and further increasing Q harms performance. Similarly, for $b = 96$, the peak is at $Q = 3$, after which accuracy declines. This suggests that when local gradient estimates are already of high quality (due to large b), excessive communication may introduce unnecessary overhead or other negative effects without providing significant consensus benefits.
- **Non-trivial Trade-off and Optimal Configuration:** The results clearly demonstrate that there is no single optimal value for Q that works across all batch sizes. The optimal configuration (b, Q) is a result of a complex trade-off. For instance, the overall best performance in this early stage of training is achieved at $b = 96, Q = 3$, not at the highest Q or largest b . This empirically validates our argument that local computation and communication are not independent in practice but are linked through a resource and performance trade-off.

Overall, these experiments reveal that the optimal configuration of Q and b is the result of a complex trade-off. From this, we can derive empirical guidelines that balance communication efficiency with model performance:

- **When the batch size (b) is small (e.g., $b = 16, 32$):** In this regime, local gradient updates are subject to significant stochasticity (i.e., high gradient noise). Under these conditions, increasing the number of MGS steps (Q) yields consistent and substantial performance gains. For instance, raising Q from 1 to 10 effectively promotes model consensus across nodes, mitigating the model divergence caused by gradient noise and thereby enhancing final generalization. This suggests that in scenarios with limited computational resources or where rapid iterations are desired, investing in a moderate increase in communication overhead is highly beneficial.
- **When the batch size (b) is large (e.g., $b = 64, 96$):** In this case, local gradient estimates are already more accurate, and the impact of gradient noise is reduced. Consequently, the benefits of increasing Q diminish or can even become detrimental. Our results show that the optimal Q is small ($Q = 1$ or $Q = 3$) in this setting. A possible explanation is that when local updates are of high quality, the marginal gains from intensive communication (high Q) do not outweigh the associated communication costs and potential synchronization overhead. It might even disrupt well-trained local features. Therefore, in scenarios where computational power is ample enough to support large-batch training, priority should be given to ensuring sufficient local computation, complemented by a more economical communication strategy.

In summary, this experiment provides valuable insights for hyperparameter selection in practical applications: b and Q are not independently tunable but must be co-designed based on available computational and communication resources to strike the optimal balance between performance and cost.

D.4 Appendix X: On the Technical Necessity and Role of the Polyak-Łojasiewicz Condition

In this section, we provide a detailed discussion on the technical role of the Polyak-Łojasiewicz (PL) condition within our generalization analysis. We elucidate why this assumption is instrumental for bounding the final iterate’s gradient in the complex setting of non-convex decentralized optimization with Multiple Gossip Steps (MGS), and how it enables the derivation of our main results.

D.4.1 The Core Challenge: Bounding the Final Iterate’s Gradient Norm

Our main generalization bound in Theorem 3 is derived from the stability analysis in Lemma 2. A critical component of this bound is the term G , which represents the expected squared norm of the stochastic gradient at the **final iterate** of the algorithm, averaged over all clients:

$$G = \frac{1}{mn} \sum_{i,j} \mathbb{E} \left[\|\nabla \ell(\theta_k^{(T)}; Z_{ij})\|^2 \right]$$

To make this bound useful, we further bound G by a term related to the expected squared norm of the full gradient, $\bar{G} = \mathbb{E} [\|\nabla R_S(\theta^{(T)})\|^2]$ (as shown in Equation 4.1 and the subsequent analysis). Therefore, the tightness and applicability of our final generalization error bound are directly contingent on our ability to establish a rigorous upper bound for the gradient norm of the **final iterate**, $\theta^{(T)}$.

However, providing such a bound is a notoriously difficult problem in optimization theory, especially under the confluence of three challenging conditions present in our work: (1) a non-convex objective function, (2) a decentralized training paradigm, and (3) the inclusion of the MGS mechanism. In general non-convex optimization, most convergence guarantees are for the minimum gradient norm over all iterations (i.e., $\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\nabla R_S(\theta^{(t)})\|^2]$), as convergence of the final iterate’s gradient is a much stronger and harder-to-prove property.

D.4.2 Limitations of Existing Last-Iterate Convergence Analyses

The analysis of last-iterate convergence in non-convex decentralized settings is an active and challenging research frontier. While significant progress has been made, existing theoretical frameworks are not directly applicable to our specific setting.

For instance, the seminal work by Yuan et al. [44] provides a last-iterate convergence analysis for D-SGD under non-convexity. However, their analysis is tailored to the standard D-SGD algorithm (equivalent to MGS with $Q = 1$) and does not account for the accelerated consensus dynamics introduced by multiple gossip steps ($Q > 1$). The MGS mechanism fundamentally alters the interplay between local computation and inter-node communication, rendering direct application of their bounds unsuitable. Other contemporary works on last-iterate convergence often provide bounds on the **function value gap** (i.e., $\mathbb{E}[\ell(\theta^{(T)})] - R_S^*$) rather than the gradient norm. In a general non-convex landscape, a small function value gap does not necessarily imply a small gradient norm, making these results insufficient for our purpose of bounding \bar{G} .

D.4.3 The PL Condition as a Principled Bridge

To overcome this theoretical impasse, we adopt the Polyak-Łojasiewicz (PL) condition. The PL condition, defined as $\|\nabla R_S(\theta)\|^2 \geq 2\mu(R_S(\theta) - R_S^*)$, establishes a direct relationship between the squared gradient norm and the function value gap. This is not an ad-hoc choice, but rather a standard and widely accepted technique in the optimization literature when a direct analysis of the gradient norm is intractable. For example, Sun et al. [34] also employed the PL condition in their analysis of decentralized learning to derive tighter theoretical bounds.

The strategic advantage of this approach lies in the fact that a tight, MGS-aware upper bound on the function value gap does exist in the literature, as established by the analysis in Hashemi et al. [18]. By leveraging the PL condition, we can translate this existing, powerful result on the function value into a rigorous upper bound on the final iterate’s gradient norm, \bar{G} , which is precisely what our generalization framework requires.

D.4.4 The Benefit: Enabling Fine-Grained, Interpretable Generalization Bounds

This technical choice is what enables us to move beyond high-level, generic bounds and derive some of the **first fine-grained, MGS-aware generalization guarantees**. By connecting the gradient norm to the MGS-sensitive function value gap, our final bounds in Theorem 3 and its subsequent remarks explicitly and quantitatively capture the impact of key algorithmic and architectural hyperparameters. These include:

- The number of MGS steps (Q), showing an exponential reduction in error.
- The communication topology, via the spectral properties of the gossip matrix (ρ).
- The learning rate (c) and total number of iterations (T).
- The number of clients (m) and per-client data size (n)

This level of detail provides concrete, actionable insights for practitioners and stands in sharp contrast to classic stability analyses (e.g., the L2-stability analysis in [33]), which typically yield more abstract bounds, such as a high-level $\mathcal{O}(1/T)$ rate for the optimization error, without explicitly showing the influence of network structure or MGS.

D.4.5 Modularity and Extensibility of Our Framework

Finally, it is crucial to recognize that the use of the PL condition is a component of our optimization error analysis (Theorem 2), not a fundamental limitation of our stability framework itself. Our overall analytical framework is modular.

This modularity implies that our contribution is extensible. Should future research in optimization theory provide a direct, assumption-free upper bound for the final iterate's gradient norm (\bar{G}) in the DSGD-MGS setting, that result could be seamlessly "plugged into" our framework. The stability-derived components of our generalization bound would remain valid, and the overall result would be immediately strengthened and generalized. This highlights that while our work relies on the current state-of-the-art in optimization theory, it is also designed to incorporate future advances.

In summary, our adoption of the PL condition is a deliberate and well-justified technical decision that addresses a significant challenge in current theory, enabling us to provide novel, detailed insights into the generalization behavior of MGS.