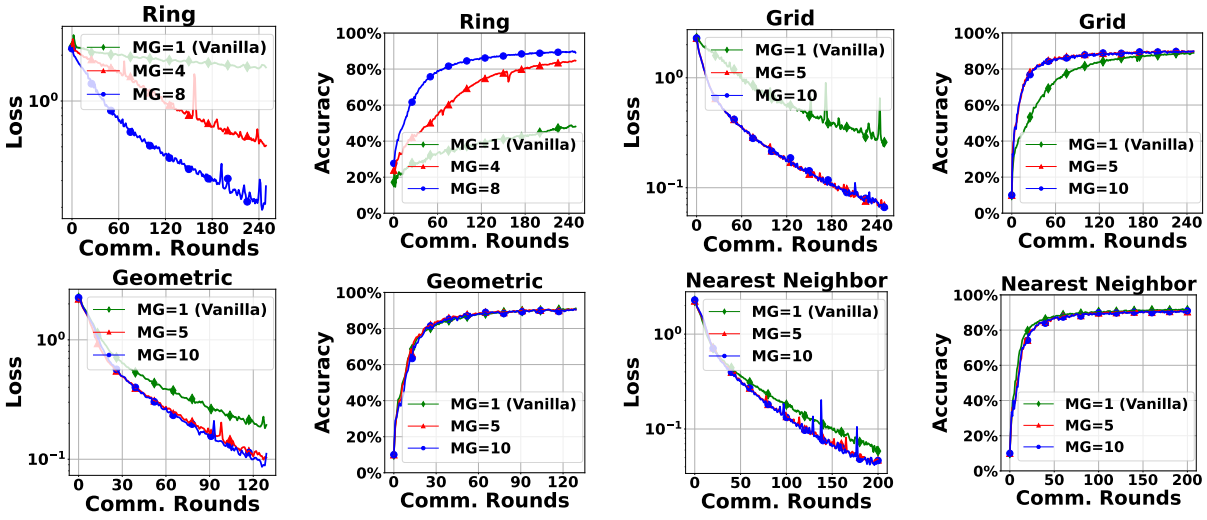

Figure 1: Training loss and test accuracy of ResNet-18 trained using MG-PullDiag-GT and vanilla PullDiag-GT on the CIFAR-10 dataset. MG-PullDiag-GT consistently outperforms the baseline across all network topologies. The detailed experiment setting will be included in Appendix E.

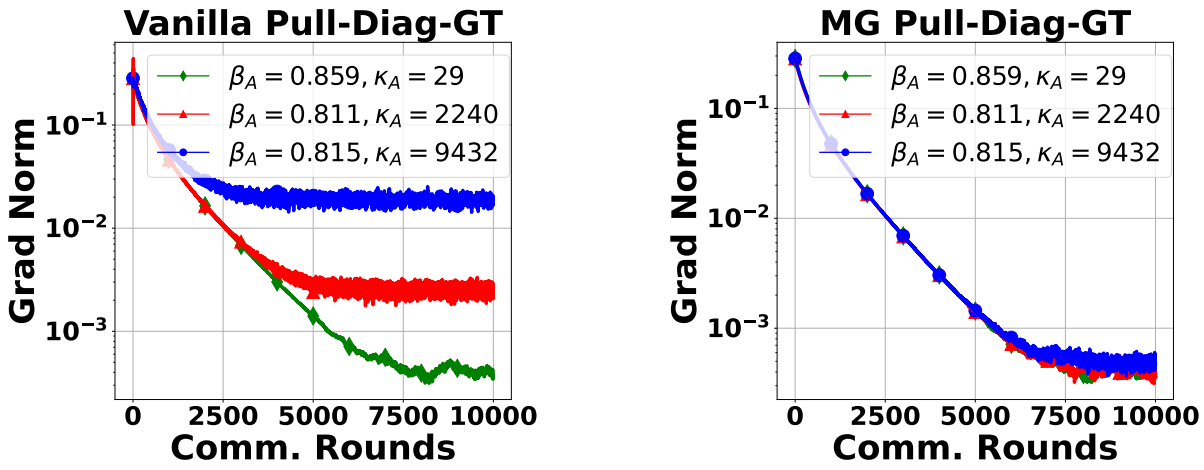

Figure 2: Training nonconvex logistic model on the synthetic dataset. Left: Performance of Pull-Diag-GT on different mixing matrices with similar values of β_A but different values of κ_A. Right: Performance if MG-Pull-Diag-GT (Mg=5) using the same mixing matrices with the left plot. The mixing matrices are generated based on the topology of success run chain with 10 nodes. This proves that κ_A has a significant influence on Pull-Diag-GT, while MG can eliminate this influence. Learning rates are set as 0.001 across all experiments.

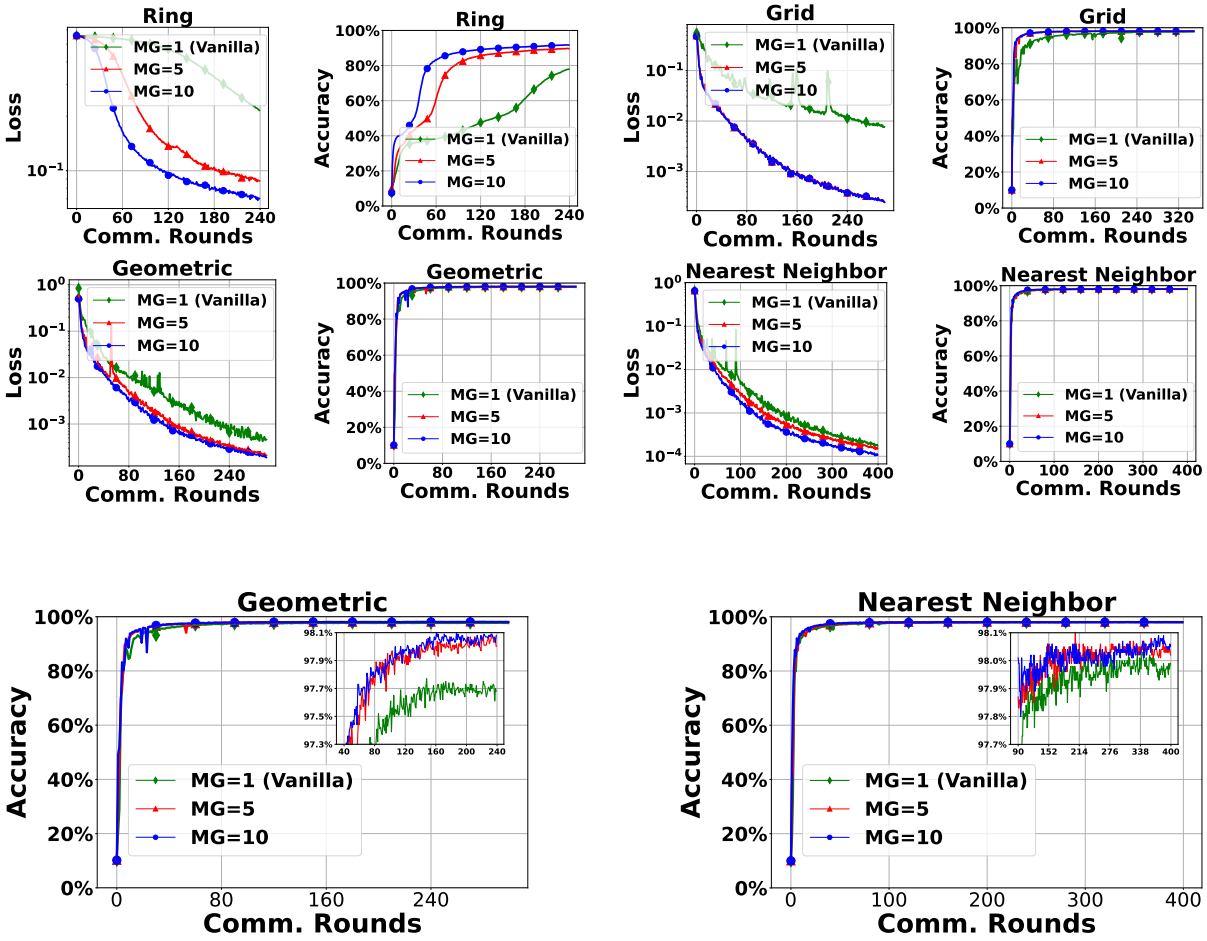

Figure 3: Training loss and test accuracy of a 4-layer neural network trained using MG-Pull-Diag-GT and vanilla Pull-Diag-GT on the MNIST dataset. The data is distributed in a heterogeneous manner (see Figure 4).

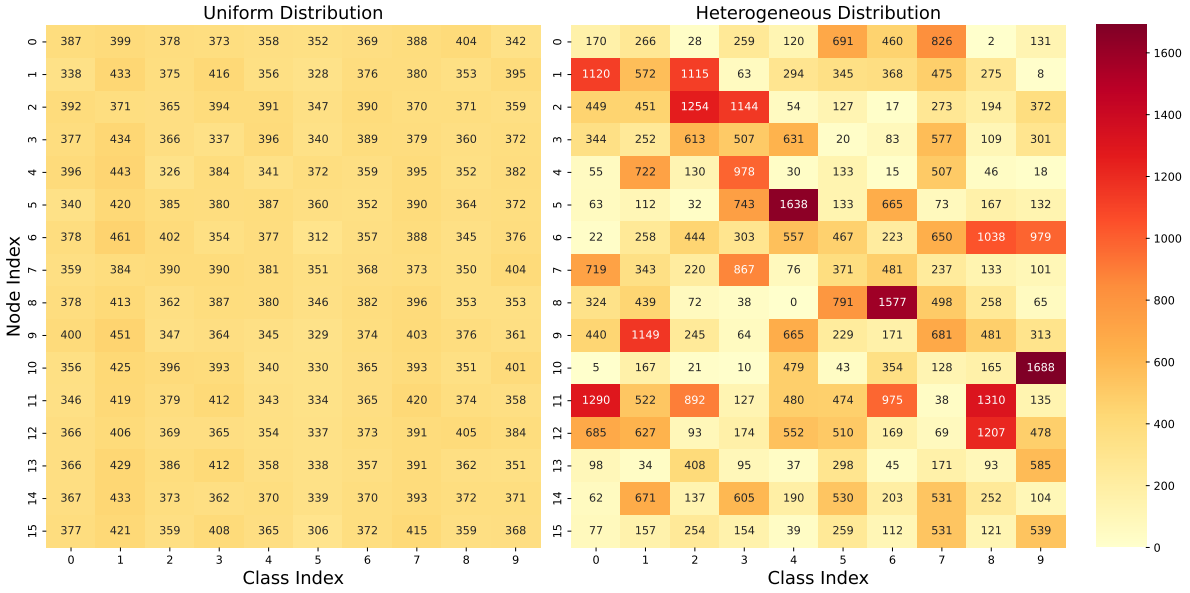

Figure 4: Left: Uniform MNIST data distribution used in the experiments in Section 6. Right: Heterogeneous MNIST data distribution used in the additional experiment shown in Figure 3.