A Supplemental Figures and Tables

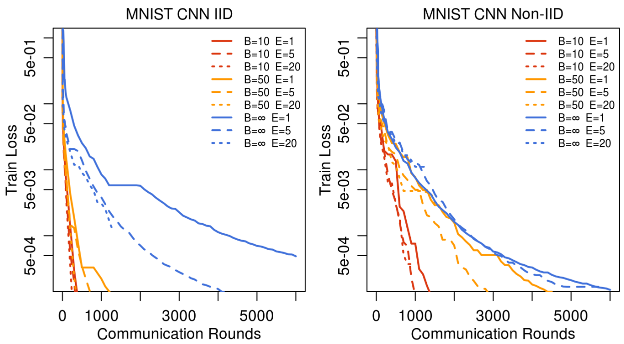

Figure 6: Training set convergence for the MNIST CNN. Note the y-axis is on a log scale, and the x-axis covers more training than Figure 2. These plots fix $C = 0.1$.

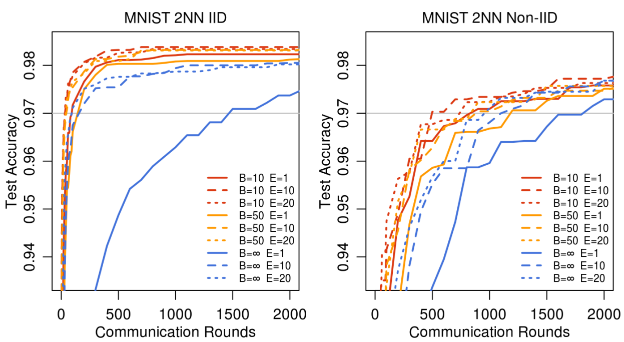

Figure 7: Test set accuracy vs. communication rounds for MNIST 2NN with $C = 0.1$ and optimized η. The left column is the IID dataset, and right is the pathological 2-digits-per-client non-IID data.

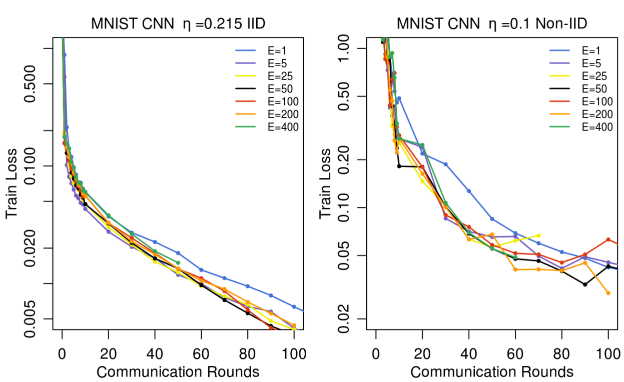

Figure 8: The effect of training for many local epochs (large E) between averaging steps, fixing $B = 10$ and $C = 0.1$. Training loss for the MNIST CNN. Note different learning rates and y-axis scales are used due to the difficulty of our pathological non-IID MNIST dataset.

Table 4: Speedups in the number of communication rounds to reach a target accuracy of 97% for FedAvg, versus FedSGD (first row) on the MNIST 2NN model.

MNIST 2NN	E	B	u	IID	NON-IID
FEDSGD	1	∞	1	1468	1817
FEDAVG	10	∞	10	156 (9.4×)	1100 (1.7×)
FEDAVG	1	50	12	144 (10.2×)	1183 (1.5×)
FEDAVG	20	∞	20	92 (16.0×)	957 (1.9×)
FEDAVG	1	10	60	92 (16.0×)	831 (2.2×)
FEDAVG	10	50	120	45 (32.6×)	881 (2.1×)
FEDAVG	20	50	240	39 (37.6×)	835 (2.2×)
FEDAVG	10	10	600	34 (43.2×)	497 (3.7×)
FEDAVG	20	10	1200	32 (45.9×)	738 (2.5×)

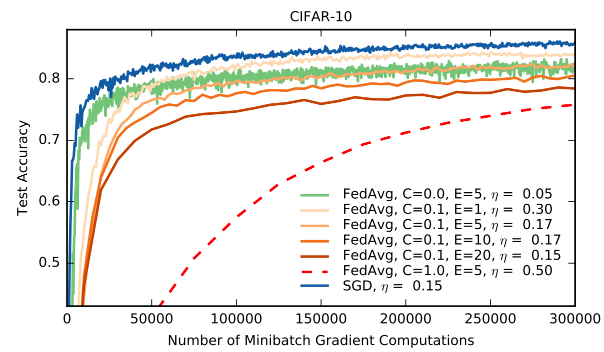

Figure 9: Test accuracy versus number of minibatch gradient computations ($B = 50$). The baseline is standard sequential SGD, as compared to FedAvg with different client fractions C (recall $C = 0$ means one client per round), and different numbers of local epochs E.

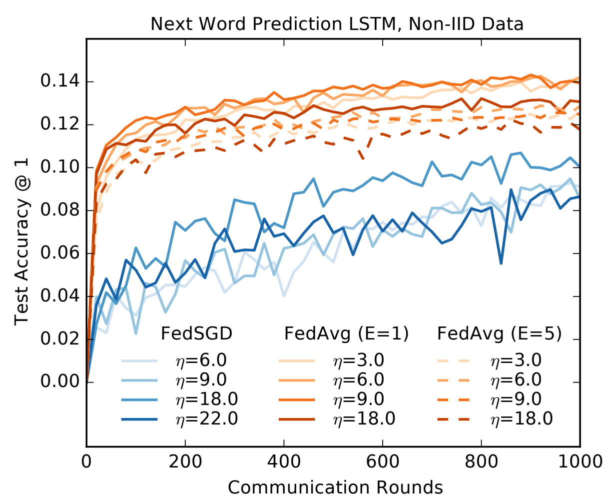

Figure 10: Learning curves for the large-scale language model word LSTM, with evaluation computed every 20 rounds. FedAvg actually performs better with fewer local epochs E (1 vs 5), and also has lower variance in accuracy across evaluation rounds compared to FedSGD.