Supporting Information - Mixing Gradients in Neural Networks to Enhance Privacy in Federated Learning

Convergence Analysis

We carried out analysis to verify that the number of iterations and the learning rate of the LBFGS optimiser used are sufficient to determine if an optimisation attempt to recover input data is successful.

We run representative tests on MNIST to support our experimental setup for convergence conditions of attack and show recovery rates from 600 experiments of linear dense layer attacks with batch sizes of $1, 2, 4$ in Table 1.

Table 1: Representative recovery rate results are shown for various LBFGS parameters: number of iteration (i), and learning rate (lr). We compare the recovery rates between 1200 iterations and 550 iterations, and rate=0.05, and 0.025.

		Recovery Rates		
Experiment Settings	Batch Size	i=1200, lr=0.025	i=1200, lr=0.05	i=550 , lr=0.05
CEL + Batch with Random Labels	1,2,4	100, 87.71, 44.6	100, 87.93, 60.17	92.30, 66.66, 42
CEL + Batch with Equal Labels	1,2,4	98.113, 12.74, 5.45	98.03, 23.07, 4.38	92., 18.96, 6.89
MSE	1,2,4	100, 23.33, 0.0	90, 26.66, 0.0	76.33, 21.0, 0.0

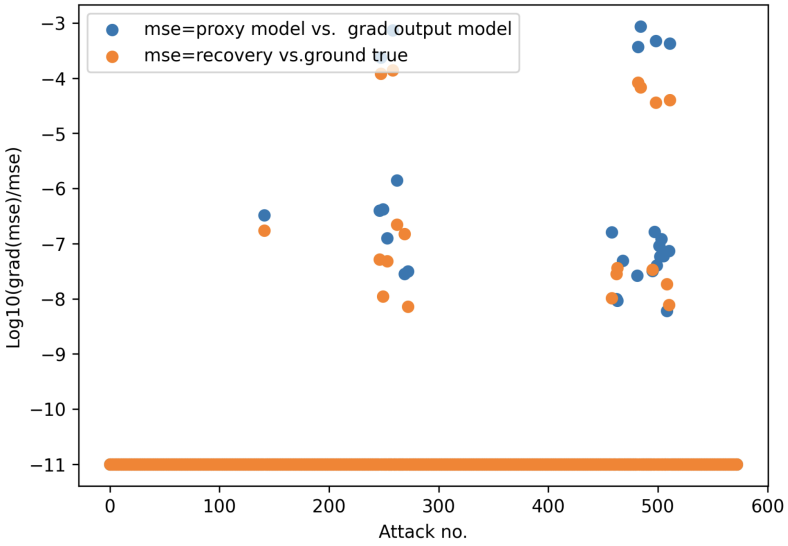

Figure 1: The logarithm of the ratio between MSE gradients and MSE as a convergence estimate calculated from the final 10 iterations are below 10^{-3} for all cases and zero in most cases. This indicates that experiments have achieved steady MSEs values or sufficient stedy MSEs values. The results shown here are obtained from a sample of 600 experiments on linear dense layer in 1,2,4 batch sizes, and different defense strategies; CEL equal labels, CEL random labels, and MSE.

We observe that the recovery rates for batch sizes 2 and 4, when comparing 1200 versus 550 iterations allow us to use 0.05 rate and 550 as the rate and limit on the number of iterations for our attack study. Table 1 shows that the deviations between results are relatively small despite changes

in type of loss function and batching strategies, which therefore supports our experimental results in the main paper and conclusion regarding the difficulty in recovering data from different strategies.

Additionally, over iterations, the gradients in the convergence test are calculated to analyse the changes in mean square errors (MSE) between the recovered data and the ground truth input, as well as the MSE loss between gradients of the proxy model and the output gradients. The gradients as a fraction of the absolute MSE values are shown in logarithm scale in figure 1 for the linear layer and figure 2 for LeNET. It shows that for most of the trials, attack values in the final iterations are steady MSE values, a few percentage of the attack trials are sufficiently steady MSEs values where most of them are less than 10^{-3} fractional changes in the gradients at the end of the attack. This indicates that our experimental setup is sufficient to use as a measure for recovery success rates. Note that we can conveniently ignore MSEs that are well below the threshold of successful attack, to avoid magnifying the fractional estimates due to small absolute MSEs values.

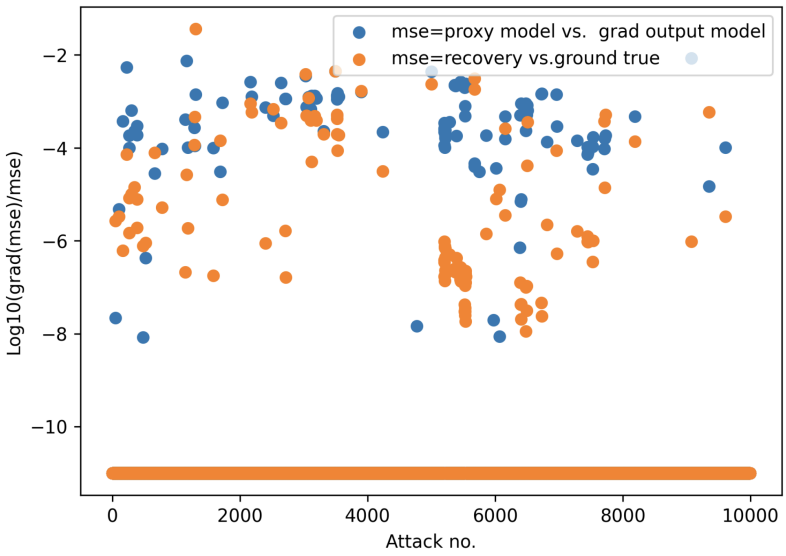

Figure 2: The logarithm of the ratio between MSE gradients and MSE as convergence estimates in LeNET (CNN) calculated from the final 10 iterations, with a setup of 550 iterations and 0.05 rate. The results shown here are obtained from a sample of 1040 experiments and total of 10000 trials, in 1,2,4 batch sizes, and different defense strategies; CEL equal labels, CEL random labels, and MSE loss. The convergence estimates shows more than 96% of points attained to steady MSE values. Less than 4% shows close to steady MSE values and mostly below a sufficient estimate of 10^{-3} values and with only few trials exceeding 10^{-3} gradient fractions.