Fed-EF: Compressed Federated Learning with Error Feedback for Non-Convex Optimization

Anonymous Author(s)
Affiliation
Address
email

1 An Illustrative Example on the Effect of Stale Error Compensation

In this section, we further provide an illustrative numerical example to demonstrate the influence of stale error compensation in Fed-EF, which means that under partial participation (PP), at round t, the local update is compensated by the error that was last accumulated in round $t-s$, where s might be large. This "lagged" information could intuitively explain the slower convergence (Theorem 3) of Fed-EF under PP.

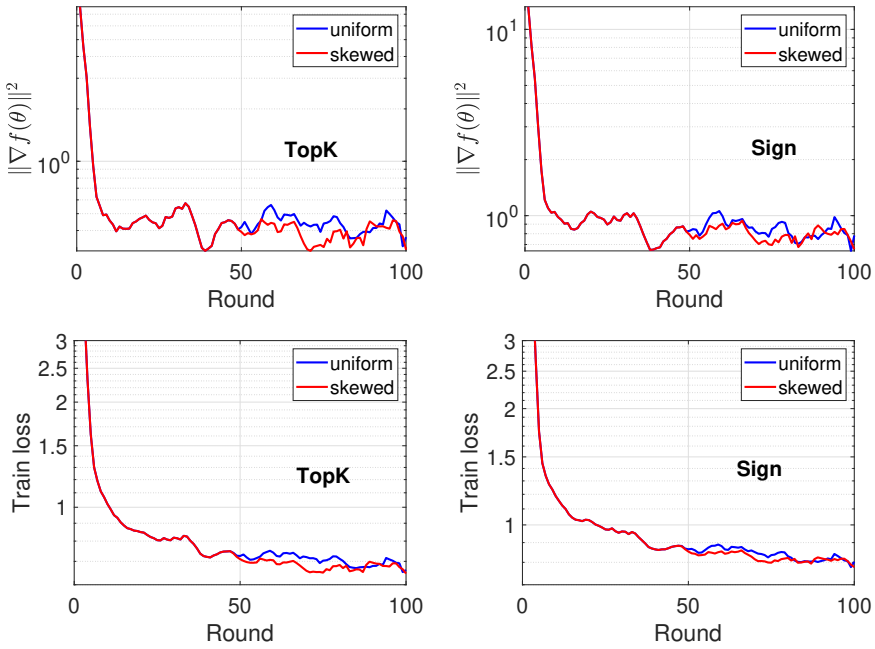

Figure 1: The squared gradient norm and training loss of training a logistic regression on MNIST, with learning rate $\eta = 1$, $\eta_l = 0.001$. The system has 200 clients with participation rate $p = 0.1$. For the first 50 rounds, fully random sampling is applied. For the last 50 rounds, we compare fully random sampling versus skewed sampling.

Specifically, we train a logistic regression model on MNIST using Fed-EF-SGD for 100 rounds. Following the theory, for the first 50 rounds, we apply fully random sampling. At this point, ideally the model is close to the unique stationary point (since the model is convex). Then, for the last 50 rounds, we compare random sampling with the following skewed sampling procedure.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

- In each round t, we assign higher probability to the clients that are picked in the previous iteration, i.e., clients in $\mathcal{M}_{t-1}$. In particular, the selection probability of each client in $\mathcal{M}_{t-1}$ is 5 times greater than that of clients not in $\mathcal{M}_{t-1}$.

This approach leads to smaller r_t compared with random sampling, effectively reducing the number of lagged errors in local updates. We report the $\|\nabla f(\theta)\|^2$ and training loss as the metrics to measure convergence. All results are averaged over 5 independent runs. As we can see from the figures, the gradient norm of the skewed sampling procedure tends to be smaller than uniform random sampling starting from the 50th round. This suggests that such skewed sampling is likely to make the model stay closer to the global minima. This can also be seen from the training loss plots where skewed sampling achieves smaller train loss faster. This example demonstrates how too much stale error compensation in fully random sampling slows down the convergence.

While the case of training deep neural nets in practice is much more complicated than this example, we hope that it provides sufficient intuition on the effect of delayed error compensation. We hope this numerical example could better illustrate the intuition on the effect of "lagged" errors in EF under partial participation, which might be helpful for FL system design.