Defending Large-Scale Distributed Machine Learning Against Adversarial Attacks

Mid-year Report

Lili Su, Vidyasagar Sadhu, Seyyed Fatemi, Rehana Mahfuz

The team has been corresponding via email, sharing updates in the form of simulation observations and reciprocal discussions/suggestions.

We ran some simulations in Python to visualize our abstract model of a distributed system where agents communicate via local estimates calculated as:

$$x_i[t+1] = \frac{\sum_{i=1}^{n} x_i[t]}{n} - \lambda[t] \cdot f_i'\left(\frac{\sum_{i=1}^{n} x_i[t]}{n}\right).$$

(for the i^{th} agent at time t)

to converge to a single estimate. The existence of a 'selfish' agent, which calculates its local estimate as:

$$x_i[t+1] = x_i[t] - \lambda[t] \cdot f_i'(x_i[t]).$$

endangers the converged estimate to deviate from the true value to reflect the selfish agent's interests more closely.

When the local estimates of the agents were plotted across time, it was found that the selfish agent's local estimate is visibly different, as shown in Figure 1.

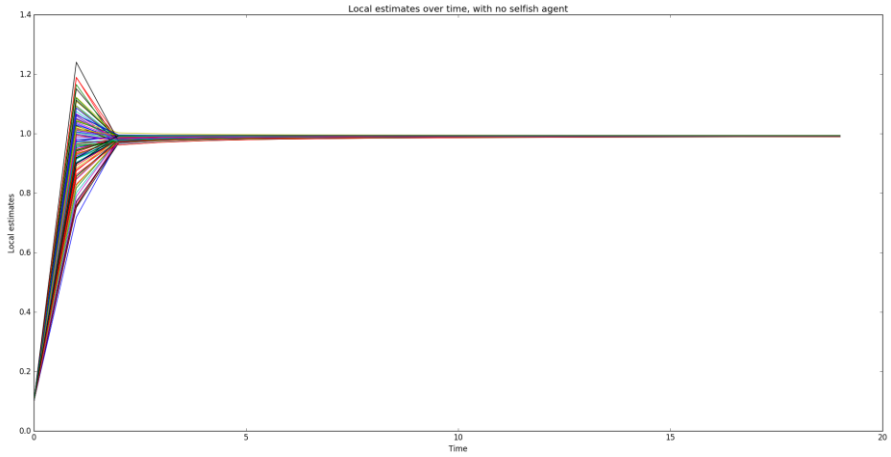

Figure 1(a): Local estimates of a distributed system vs. time, when there is no selfish agent

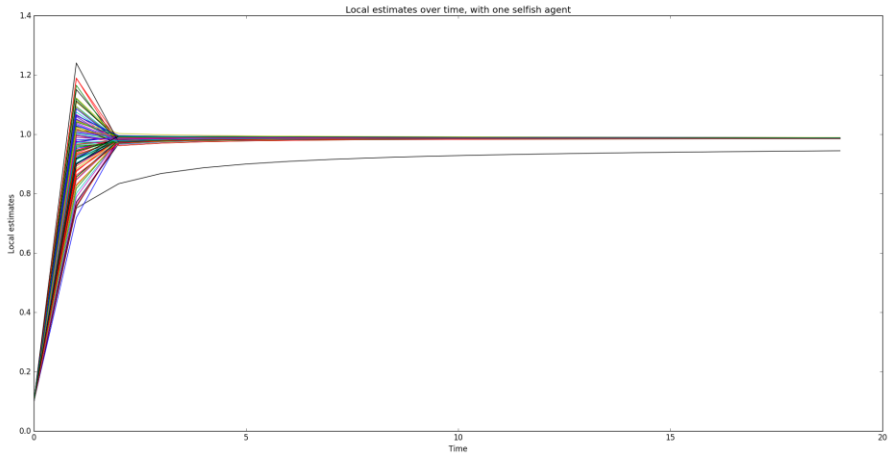

Figure 1(b): Local estimates of a distributed system vs. time, where the selfish agent's local estimate can be distinguished

This method can be used to detect the selfish agent, and other agents can avoid using this selfish agent's local estimates in their calculations to prevent the global converging estimate from being biased.

This work was presented as a poster during the NSF Annual Site Visit in December at Purdue University.

The team plans to continue working on this project for the next few months, and possibly submit a paper at a machine learning or distributed computing conference (example: COLT, NIPS, PODC, DISC).