EP3260: Fundamentals of Machine Learning over Networks
Computer Assignment 5, Due Date: April 25, 2020
Group 4: H. Hellström, M. Tavana, F. Vannella, and M. Zeng

Computer Assignment 5. *ADMM*
Split the "MNIST" dataset to 10 random disjoint subsets, each for one worker, and consider SVM classifier in the form of $\min \frac{1}{N}\sum_{i\in[N]} f_i(\boldsymbol{w})$ with $N = 10$.

(a) Run decentralized GD (from Lecture 6) with 10 workers. Characterize the convergence against the total number of signaling exchanges among all nodes, denoted by T.

Solution: For this part, we follow the work of CA4, with the only difference here being that we add the signaling exchanges among all nodes. For each iteration, a signaling exchanges of 20 is needed, since we have 10 nodes an each needs to send and receive from the master node. The convergence rate of the training process is shown in Fig. 1.

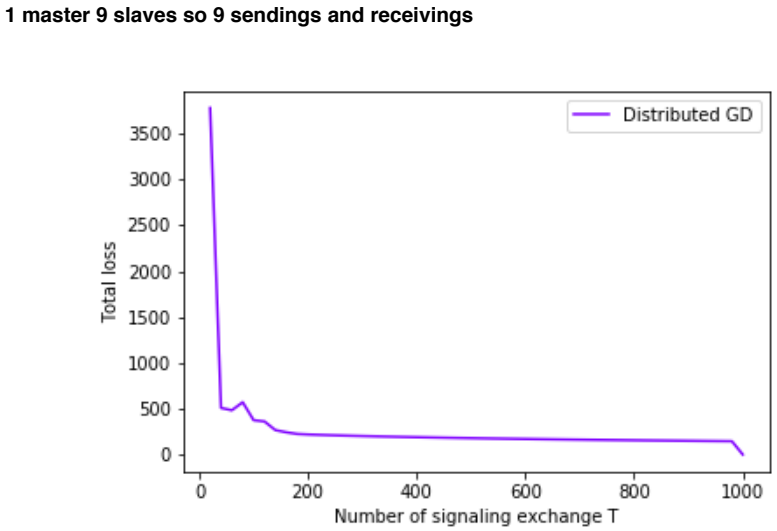

Figure 1: Convergence rate of GD over signaling exchange T.

(b) Consider a two-star topology with communication graph $(1, 2, 3, 4) - 5 - 6 - (7, 8, 9, 10)$ and run decentralized subgradient method (from Lecture 6) and ADMM over the network (from Lecture 7). Characterize the convergence against T. Tune hyper-parameters to improve the convergence rate.

Solution:

- For decentralized subgradient method, we follow the work of CA4, with the only difference here being that we add the signaling exchanges among all nodes. For each iteration, a signaling exchange of 18 is needed, since we have 9 edges whose connecting nodes need to communicate with each other. The convergence rate of the training process is shown in Fig. 2.
- For ADMM, we follow the formulas on page 30 of Lecture 7 to update the variables. For each iteration, a signaling exchange of 54 is needed, since we have 9 edges whose connecting nodes need to send and receive from its neighbours the three variables, i.e., $x_i, y_{i,j}$ and $z_{i,j}$. Fig. 3 shows the total loss vs. T when the hyperparameter λ is set to 1. Then, we further tune the hyperparameter λ to improve the convergence rate. The corresponding result is shown in Fig. 4.

(c) Propose an approach to reduce T with a marginal impact on the convergence. Do not limit your imaginations and feel free to propose any solution. While being nonsense in some applications, your

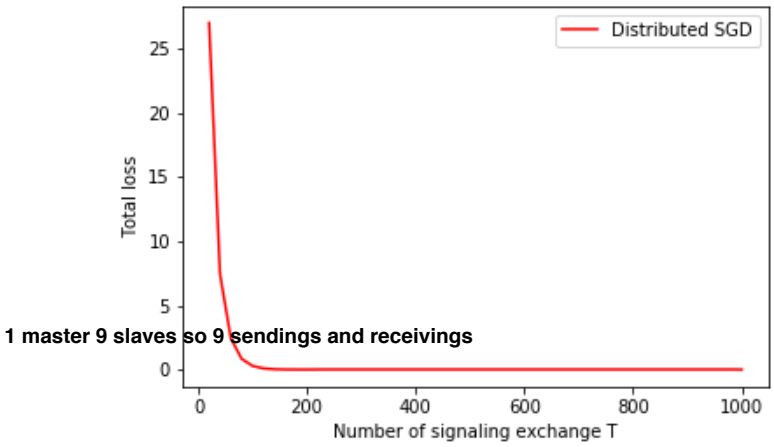

Figure 2: Convergence rate of decentralized SGD over signaling exchange T.

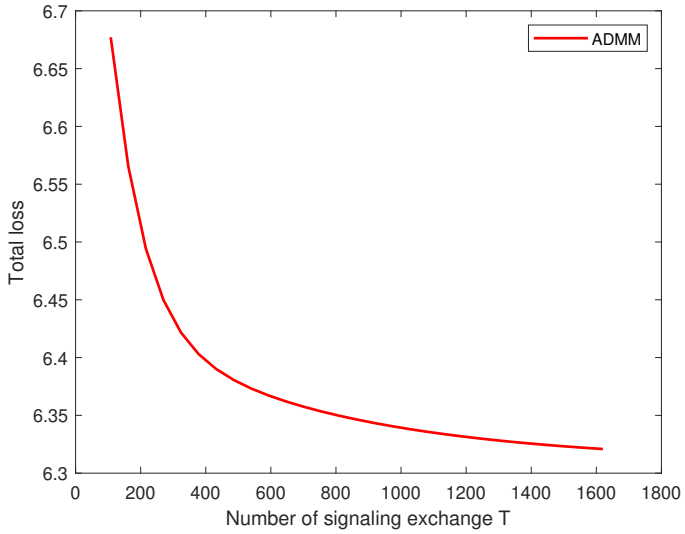

Figure 3: Convergence rate of ADMM over signaling exchange T.

solution may actually make sense in some other applications. Discuss pros and cons of your solution and possibly provide numerical evidence that it reduces T.

Solution: According to existing literature, there are some approaches which can be deployed to reduce the communication overhead among agents/workers.

First, we could try to reduce the size of information required for transmission at each agent. This can be done using source coding. In particular, it is found that the vectors $x_i, y_{i,j}$ and $z_{i,j}$ contain many zero values elements, which clearly can be compressed. However, this increases the complexity at the agents, since encoding and decoding need to be performed. Also, we can avoid unnecessary information transmission. When calculating the signaling exchange in (b) for ADMM, we assume that all three vectors $x_i, y_{i,j}$ and $z_{i,j}$ should be exchanged between its neighbours. However, after giving a closer look, it can be seen that the vector $y_{i,j}$ does not need to transmitted. This could reduce the signaling exchange by one third in each iteration. **good to reduce**

Second, we can use quantization to reduce the size of data to be transmitted. This, however, comes at

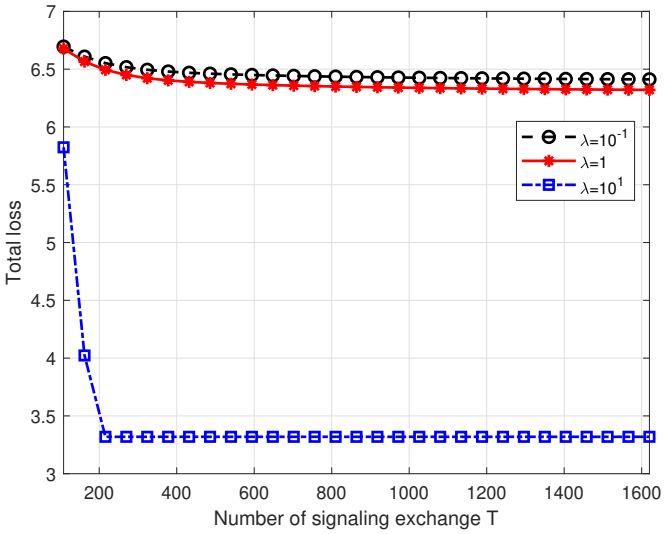

Figure 4: Convergence rate of ADMM over signaling exchange T under different λ.

a price of the convergence rate. To illustrate this, we consider the 10 workers network with a master node. The GD algorithm is used. Then, we assume that each agent randomly (it can follow a certain rule so that all agents know the position of the selected elements) assigns 300 elements from the total 784 value of the weight vector to zero. Then, these zeros values do not require to be exchanged. This can reduce the signaling overhead by $\frac{300}{784} \approx 40\%$. On the other hand, the algorithm still converges although some fluctuation exists due to the information compression, as shown in Fig. 5.

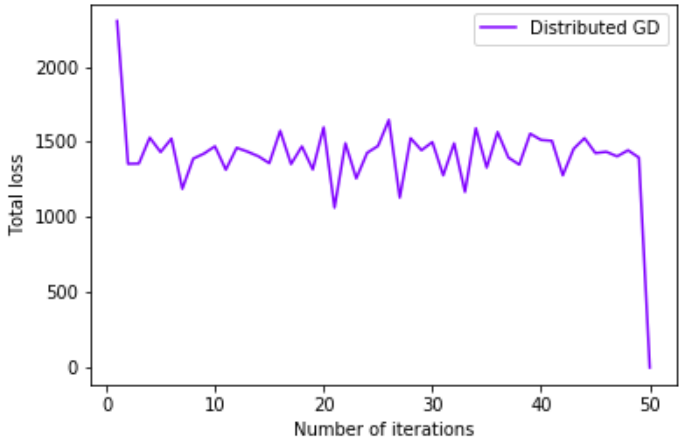

Figure 5: Convergence rate of GD over iterations with compression.

Last, during signal exchange, we may use broadcasting to replace peer-to-peer transmission. This works because the information each node sends to its neighbours is the same. By doing this, we may also reduce the signaling overhead. This approach can be applied to wireless communication networks, where broadcasting is more efficient since the transmission media is shared by all agents.

network topology could be discussed more here