MLoN Computer Assignment 6

Group 1

Split "MNIST" dataset to 10 random disjoint subsets, each for one worker, and consider SVM classifier in the form of $min_w \frac{1}{N} \sum_{i \in [N]} f_i(w)$ with $N = 10$. An alternative approach to improve communication-efficiency is to compress the information message to be exchanged (usually gradients – either in primal or dual forms).
Consider two compression/quantization methods for a vector: (Q1) keep only K values of a vector and set the rest to zero and (Q2) represent every element with fewer bits (e.g. 4 bits instead of 32 bits).

1 a

Problem

Repeat parts a-b from CA5 using Q1 and Q2. Can you integrate Q1/Q2 to your solution in part c of CA5? Discuss.

Solution

For Q1 compression, we simply downscale the weight to half of origin. A more complex protocal shall be designed easily e.g. 4x downscaling and the initial non-zero dimension loops through 4 positions.
For Q2 compression, we simply reduces the precision.
We shall run DSM and ADMM with Q1/Q2 compression. For DSM, the w and converge are shown in Fig. 1 and Fig. 2. Q1 compression can cause loss increase at the beginning. And both compression methods converge in the end.

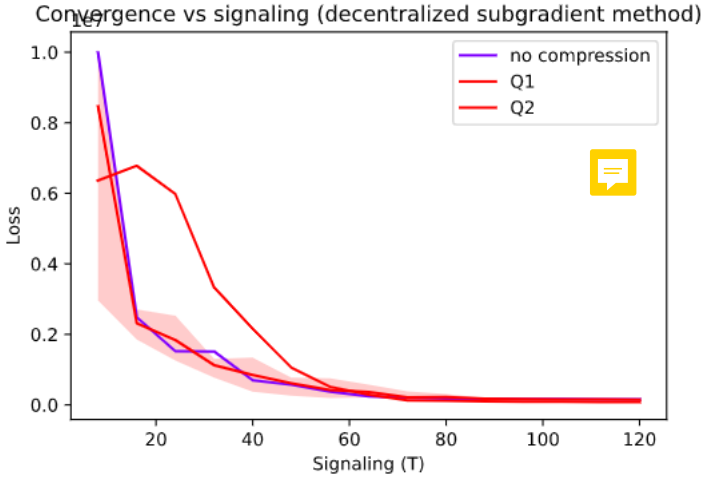
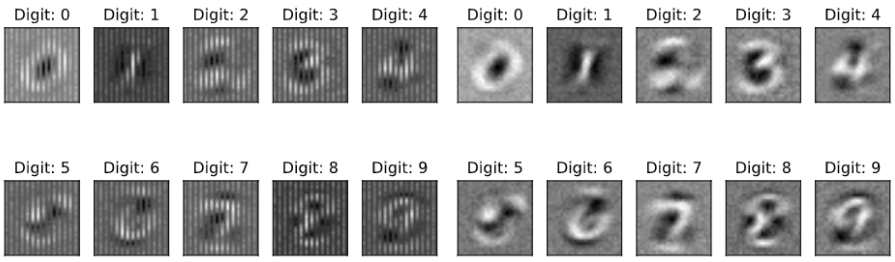


Figure 1: convergence vs signaling (DSM with compression)



(a) Q1 (b) Q2

Figure 2: grey-scale model (DSM with compression)

For ADMM, the w and converge are shown in Fig. 3 and Fig. 4. The robustness of ADMM removes the loss increase caused by compression.
The communication reduction method in part (c) of CA5 is obviously compatible with the compression/quantization method thus we shall not run it here.

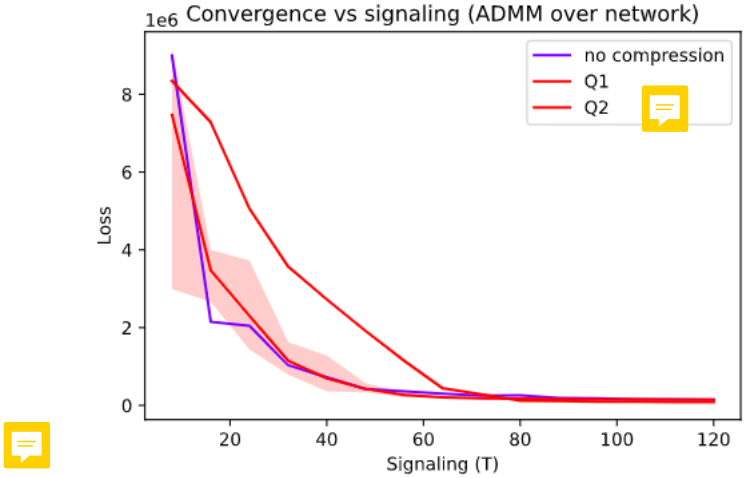
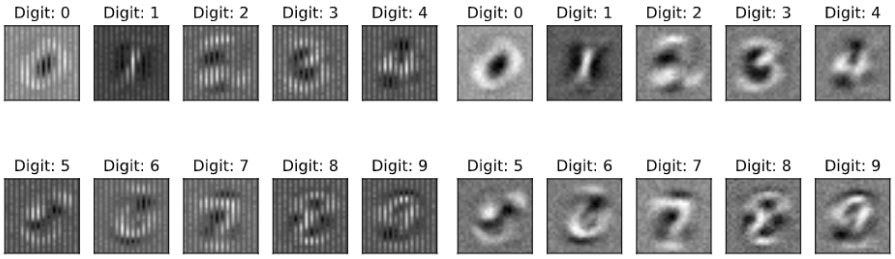


Figure 3: convergence vs signaling (ADMM with compression)



(a) Q1 (b) Q2

Figure 4: grey-scale model (ADMM with compression)

2 b

Problem

How do you make SVRG and SAG communication efficient for large-scale ML?

Solution

Although topology change, hierarchical update and compression can always be applied to decentralized gradient methods. We shall take a closer look at stochastic average gradient (SAG) and stochastic variance reduced gradient (SVRG) which modifies the gradient update.
Naturally, the gradient sampling can be integrated in the communication network. The communication protocol is straight-forward as each node just sends

a flag to other nodes whether it will accept new gradient or not. As result, gradient not sampled to update will not be shared thus reducing communication.