Sparse-SignSGD Optimizer for Communication-Efficient Distributed Learning

Chanho Park
Electrical Engineering
POSTECH
Pohang, South Korea
chanho26@postech.ac.kr

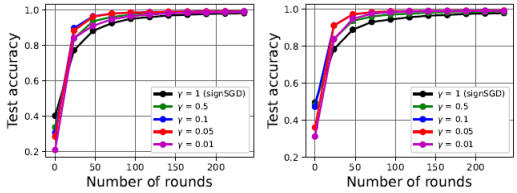
Namyoon Lee
School of Electrical Engineering
Korea University
Seoul, South Korea
namyoon@korea.ac.kr

Abstract—As the deep learning technique grows up, the size of deep neural network increases exponentially. This induces a large amount of communication cost between workers and a parameter server during iterations in the distributed learning system. To address this communication bottleneck, in this work, we present a novel communication-efficient algorithm which utilizes the synergistic benefits of sign quantization and top-K sparsification to the gradient components, called sparse-signSGD. Each worker in sparse-signSGD select the top-K magnitude components of its local gradient vector and only send the signs of its components to the server. Then, the server aggregates the signs for all the gradient components, and returns the results via a majority vote rule. We demonstrate that sparse-signSGD can converge almost at the same rate as signSGD with some certain mild conditions, while tremendously reducing the communication costs. Even, if the sparsification parameter K is chosen properly based on the number of workers and the size of the neural network model, sparse-signSGD can achieve a convergence rate faster than the conventional method, signSGD-MV with a theoretical approach. Experimental results using both independent and identically distributed (IID) and non-IID datasets verify that sparse-signSGD can attain higher test accuracy than signSGD-MV, and reduce the communication costs 300x compared with the representative optimizer, stochastic gradient descent. These findings highlight the potential of sparse-signSGD as a promising solution for communication-efficient distributed optimization in deep learning.

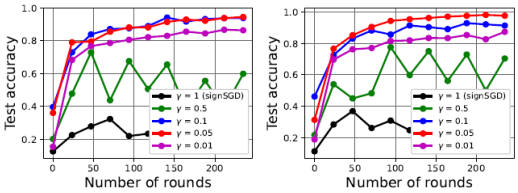
Keywords—Distributed optimization, gradient compression, convergence rate

Algorithm 1 S^{3}GD-MV

Input: initial model $\mathbf{x}^0$, learning rate δ^t, gradient sparsity K, the number of workers M, initial accumulated error $\mathbf{e}_m^0 = \mathbf{0}$, error weight η, total iteration T
for $t = 0 : T-1$ **do**
  for worker $m = 1 : M$ **do**
    compute $\tilde{\mathbf{g}}_m^t$ (local gradient)
    $\mathbf{g}_m^t \leftarrow \tilde{\mathbf{g}}_m^t + \eta \mathbf{e}_m^t$
    $\mathbf{e}_m^{t+1} \leftarrow \mathbf{g}_m^t - \mathsf{TopK}(\mathbf{g}_m^t)$
    send $\mathsf{TopKSign}(\mathbf{g}_m^t)$ to **server**
    receive $\mathrm{sgn}\left[\sum_{m=1}^M \mathsf{TopKSign}(\mathbf{g}_m^t)\right]$ from **server**
    $\mathbf{x}^{t+1} \leftarrow \mathbf{x}^t - \delta^t \mathrm{sgn}\left[\sum_{m=1}^M \mathsf{TopKSign}(\mathbf{g}_m^t)\right]$
  end for
end for



(a) $M = 10$ and MNIST dataset (b) $M = 100$ and MNIST dataset



(a) $M = 10$ and MNIST dataset (b) $M = 100$ and MNIST dataset

Our novel distributed learning algorithm, sparse-signSGD, can optimize the trade-off between the learning performance and the communication costs. Based on signSGD-MV optimizer, which only uses the sign quantizer to the gradient components, sparse-signSGD achieves two synergistic benefits from sign quantization and top-K sparsification under the majority vote principle in which the procedures are described in the left figure. By these benefits, sparse-signSGD diminishes the communication costs considerably than the state-of-the-art algorithms (e.g., SGD, sparsified-SGD). Sparse-signSGD can also achieve a higher convergence rate than the conventional method, signSGD-MV, by using a proper sparsification parameter K in terms of model size N and the number of workers M. We theoretically demonstrate these tendencies through the convergence rate analysis in L1-geometry. In addition, the experimental results in the right 4 figures represent that sparse-signSGD can attain higher test accuracy compared with signSGD, even with non-IID data distribution for the workers.

[1] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. Int. Conf. Comput. Statist. (COMPSTAT)*, Sep. 2010, pp. 177-186.
[2] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2018, pp. 560-569.
[3] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 5973-5983.