THC: Accelerating Distributed Deep Learning Using Tensor Homomorphic Compression
Minghao Li, et al.

This paper introduces THC, a novel bidirectional gradient compression framework that enables direct aggregation of compressed values, eliminating computational overheads. The motivation of the paper is to mitigate the growing communication bottleneck in large-scale distributed deep learning, where synchronization costs account for up to 49% of training time for models such as BERT. Existing gradient compression methods require decompressing gradients before aggregation, leading to inefficiencies. THC allows for in-network aggregation on programmable switches, reducing communication overhead. The authors propose an encoding scheme leveraging uniform homomorphic compression to preserve mathematical properties required for gradient updates. The authors then evaluate THC against state-of-the-art distributed deep learning frameworks such as BytePS and Horovod, demonstrating a 1.4–1.47x speed improvement with in-network aggregation and a 1.28–1.33x speed increase using a software parameter server. The authors assume that structured gradient compression does not significantly degrade model accuracy, which ends up holding true under their tested conditions, but this may not be the case one extrapolated to other training setups. One limitation of THC is that it is optimized for parameter server architectures, meaning its improvements may not extend to decentralized training schemes. Another limitation is that THC's efficiency depends on maintaining a fixed bit budget per worker which may introduce challenges in training environments where this is not possible. Overall, the paper presents a compelling solution for large-scale training applicable to cloud-based ML workloads with substantial improvements in computation and network overhead.

Distributed Mean Estimation with Limited Communication
Amanda Theertha Suresh, et al.

This paper focuses on the problem of distributed mean estimation which is fundamental to distributed optimization and federated learning. The goal is to efficiently estimate the mean of high-dimensional data distributed across multiple clients while minimizing the communication costs. Prior work assumes independently and identically distributed data distributions, however this paper considers arbitrary data distributions, making it much more generalizable. The authors first analyze a naive stochastic rounding approach which achieves an MSE of d/n while transmitting a constant number of bits per dimension. To improve this, the authors propose structured random rotation, reducing MSE to (log d)/n and variable-length coding which reduces the MSE down to 1/n. The authors validate the proposed algorithms through applications in distributed k-means clustering and PCA, showing significant communication savings without a tremendous compromise in accuracy. Their assumptions include random rotations preserving the statistical properties of the data and that clients can independently encode their data without coordination , which may not be the case in, for example, adversarial contexts. One limitation is that these methods do not address privacy concerns, which is a critical problem in federated learning. One extension of the paper could address this limitation by introducing some type of secure aggregation into the quantization scheme. The paper provides a solid theoretical foundation for optimal mean estimation considering communication constraints, making it more practical, in distributed learning, making it very plausible to be deployed in edge computing environments where bandwidth is limited.