Sparsified Distributed Adaptive Learning with Error Feedback

Anonymous Author(s)
Affiliation
Address
email

Abstract

To be completed...

1 Introduction

Some related work:

[2] develops variant of signSGD (as a biased compression schemes) for distributed optimization. Contributions are mainly on this error feedback variant. In [3], the authors provide theoretical results on the convergence of sparse Gradient SGD for distributed optimization (we want that for AMS here). [4] develops a variant of distributed SGD with sparse gradients too. Contributions include a memory term used while compressing the gradient (using top k for instance). Speeding up the convergence in $\frac{1}{T^3}$.

2 Preliminaries

Distributed Learning.

Sparse Optimization.

Sketch and Quantization based FL.

3 Method

Consider standard synchronous distributed optimization setting. AMSGrad is used as the prototype, and the local workers is only in charge of gradient computation.

3.1 TopK AMSGrad with Error Feedback

The key difference (and interesting part) of our TopK AMSGrad comprared with the following arxiv paper "Quantized Adam"`https://arxiv.org/pdf/2004.14180.pdf` is that, in our model only gradients are transmitted. In "QAdam", each local worker keeps a local copy of moment estimator m and v, and compresses and transmits m/v as a whole. Thus, that method is very much like the sparsified distributed SGD, except that g is changed into m/v. In our model, the moment estimates m and v are computed only at the central server, with the compressed gradients instead of the full gradient. This would be the key (and difficulty) in convergence analysis.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Algorithm 1 SPARS-AMS for Federated Learning

Input: parameter β_1, β_2, learning rate η_t.
Initialize: central server parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^d$; $e_{t,i} = 0$ the error accumulator for each worker; sparsity parameter k; N local workers; $m_0 = 0$, $v_0 = 0$, $\hat{v}_0 = 0$
for $t = 1$ to T **do**
    parallel for worker $i \in [n]$ **do**:
        Receive model parameter θ_{t-1} from central server
        Compute stochastic gradient $g_{t,i}$ at θ_t
        Compute $\tilde{g}_{t,i} = TopK(g_{t,i} + e_{t,i}, k)$
        Update the error $e_{t+1,i} = e_{t,i} + g_{t,i} - \tilde{g}_{t,i}$
        Send $\tilde{g}_{t,i}$ back to central server
    end parallel
    Central server do:
    $\bar{g}_t = \frac{1}{N}\sum_{i=1}^{N} \tilde{g}_{t,i}$
    $m_t = \beta_1 m_{t-1} + (1-\beta_1)\bar{g}_t$
    $v_t = \beta_2 v_{t-1} + (1-\beta_2)\bar{g}_t^2$
    $\hat{v}_t = \max(v_t, \hat{v}_{t-1})$
    Update global model $\theta_t = \theta_{t-1} - \eta_t \frac{m_t}{\sqrt{\hat{v}_t}}$
end for

3.2 Convergence Analysis

Several mild assumptions to make: Nonconvex and smooth loss function, unbiased stochastic gradient, bounded variance of the gradient, bounded norm of the gradient, control of the distance between the true gradient and its sparse variant.

Check [1] for proofs starting with single machine and extending to distributed settings (several machines).

3.2.1 Single machine

Under the centralized setting, the goal is to derive an upper bound to the second order moment of the gradient of the objective function at some iteration $T_f \in [1, T]$.

We first define multiple auxiliary sequences. For the first moment, define

$$\begin{aligned}
&\bar{m}_t = m_t + \mathcal{E}_t, \\
&\mathcal{E}_t = \beta_1 \mathcal{E}_{t-1} + (1-\beta_1)(e_{t+1} - e_t),
\end{aligned}$$

such that

$$\begin{aligned}
\bar{m}_t &= \bar{m}_t + \mathcal{E}_t \\
&= \beta_1(m_t + \mathcal{E}_t) + (1-\beta_1)(\bar{g}_t + e_{t+1} - e_1) \\
&= \beta_1 \bar{m}_{t-1} + (1-\beta_1) g_t.
\end{aligned}$$

TBD...

3.2.2 Multiple machine

4 Experiments

Our proposed TopK-EF with AMSGrad matches that of full AMSGrad, in distributed learning. Number of local workers is 20. Error feedback fixes the convergence issue of using solely the TopK gradient.

5 Conclusion

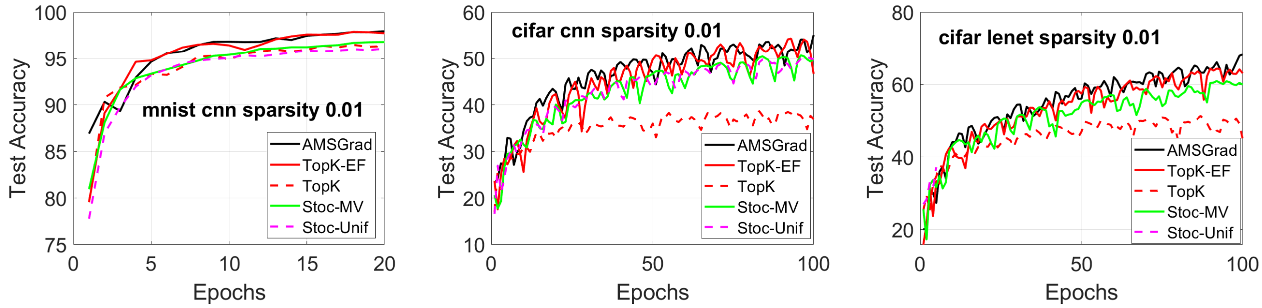

Figure 1: Test accuracy.

References

[1] Congliang Chen, Li Shen, Haozhi Huang, Qi Wu, and Wei Liu. Quantized adam with error feedback. *arXiv preprint arXiv:2004.14180*, 2020.

[2] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. *arXiv preprint arXiv:1901.09847*, 2019.

[3] Shaohuai Shi, Kaiyong Zhao, Qiang Wang, Zhenheng Tang, and Xiaowen Chu. A convergence analysis of distributed sgd with communication-efficient gradient sparsification. In *IJCAI*, pages 3411–3417, 2019.

[4] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems*, pages 4447–4458, 2018.

A Appendix