Sparsified Distributed Adaptive Learning with Error Feedback

Anonymous Author(s)
Affiliation
Address
email

Abstract

To be completed...

1 Method

Most modern machine learning tasks can be casted as a large finite-sum optimization problem written as:

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \tag{1}$$

where n denotes the number of workers, f_i represents the average loss for worker i and θ the global model parameter taking value in Θ, a subset of $\mathbb{R}^d$.

Some related work:

[**?**] develops variant of signSGD (as a biased compression schemes) for distributed optimization. Contributions are mainly on this error feedback variant. In [**?**], the authors provide theoretical results on the convergence of sparse Gradient SGD for distributed optimization (we want that for AMS here). [**?**] 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 $\frac{1}{T^3}$.

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

1.1 TopK AMSGrad with Error Feedback

The key difference (and interesting part) of our TopK AMSGrad compared 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 Distributed Learning

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

1.2 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 [?] starting with single machine and extending to distributed settings (several machines).

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

1.3 Mild Assumptions

We begin by making the following assumptions.

A 1. *(Smoothness) For $i \in [\![n]\!]$, f_i is L-smooth: $\|\nabla f_i(\theta) - \nabla f_i(\vartheta)\| \leq L\|\theta - \vartheta\|$.*

A 2. *(Unbiased and Bounded gradient **per worker**) For any iteration index $t > 0$ and worker index $i \in [\![n]\!]$, the stochastic gradient is unbiased and bounded from above: $\mathbb{E}[g_{t,i}] = \nabla f_i(\theta_t)$ and $\|g_{t,i}\| \leq G_i$.*

A 3. *(Bounded variance **per worker**) For any iteration index $t > 0$ and worker index $i \in [\![n]\!]$, the variance of the noisy gradient is bounded: $\mathbb{E}[|g_{t,i} - \nabla f_i(\theta_t)|^2] < \sigma_i^2$.*

Denote by $Q(\cdot)$ the quantization operator Line 7 of Algorithm 1, which takes as input a gradient vector and returns a quantized version of it, and note $\tilde{g} := Q(g)$. Assume that

A 4. *(Bounded Quantization) For any iteration $t > 0$, there exists a constant $q > 0$ such that $\|g_{t,i} - \tilde{g}_{t,i}\| \leq q\|g_{t,i}\|$, where $g_{t,i}$ is the stochastic gradient computed at iteration t for worker i. (high q means large quantization so loss of precision on the true gradient)*

Denote for all $\theta \in \Theta$:

$$f(\theta) := \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)\,, \tag{2}$$

where n denotes the number of workers.

2 Single Machine

Single machine method

Algorithm 2 SPARS-AMS : Single machine setting

Input: parameter β_1, β_2, learning rate η_t.
Initialize: central server parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^d$; $e_0 = 0$ the error accumulator; sparsity parameter k; $m_0 = 0$, $v_0 = 0$, $\hat{v}_0 = 0$
for $t = 1$ to T **do**
    Compute stochastic gradient $g_t = g_{t,i_t}$ at θ_t for randomly sampled index i_t
    Compute $\tilde{g}_t = TopK(g_t + e_t, k)$
    Update the error $e_{t+1} = e_t + g_t - \tilde{g}_t$
    $m_t = \beta_1 m_{t-1} + (1 - \beta_1)\tilde{g}_t$
    $v_t = \beta_2 v_{t-1} + (1 - \beta_2)\tilde{g}_t^2$
    $\hat{v}_t = \max(v_t, \hat{v}_{t-1})$
    Update global model $\theta_t = \theta_{t-1} - \eta_t \frac{m_t}{\sqrt{\hat{v}_t + \epsilon}}$
end for

Belhal Try for Single Machine Setting:

Define the auxiliary model

$$\begin{aligned}
\theta'_{t+1} &:= \theta_{t+1} - e_{t+1} \\
&= \theta_t - \eta a_t - e_{t+1} \\
&= \theta_t - \eta a_t - e_t - g_t + \tilde{g}_t \\
&= \theta_t - \eta a_t - e_t - \Delta_t \\
&= \theta'_t - \eta a_t - \Delta_t
\end{aligned}$$

where $a_t := \frac{m_t}{\sqrt{\hat{v}_t + \epsilon}}$ and $\Delta_t := g_t - \tilde{g}_t$. By smoothness assumption we have

$$f(\theta'_{t+1}) \leq f(\theta'_t) - \langle \nabla f(\theta'_t), \eta a_t + \Delta_t \rangle + \frac{L}{2} \|\theta'_{t+1} - \theta'_t\|^2.$$

Thus,

$$\begin{aligned}
\mathbb{E}[f(\theta'_{t+1}) - f(\theta'_t)] &\leq -\mathbb{E}[\langle \nabla f(\theta'_t), \eta a_t + \Delta_t \rangle] + \frac{L}{2}\mathbb{E}[\|\eta a_t + \Delta_t\|^2] \\
&\leq \eta \mathbb{E}[\langle \nabla f(\theta_t) - \nabla f(\theta'_t), \eta a_t + \Delta_t \rangle] - \mathbb{E}[\langle \nabla f(\theta_t), \eta a_t + \Delta_t \rangle] + \frac{L}{2}\mathbb{E}[\|\eta a_t + \Delta_t\|^2]
\end{aligned}$$

Using the smoothness assumption A1 we have

$$\mathbb{E}[\langle \nabla f(\theta_t) - \nabla f(\theta'_t), \eta a_t + \Delta_t \rangle] \leq L\mathbb{E}[\|\theta_t - \theta'_t\|]E[\|\eta a_t + \Delta_t\|]$$

Hence,

$$\begin{aligned}
\mathbb{E}[f(\theta'_{t+1}) - f(\theta'_t)] &\leq -\mathbb{E}[\langle \nabla f(\theta'_t), \eta a_t + \Delta_t \rangle] + \frac{L}{2}\mathbb{E}[\|\eta a_t + \Delta_t\|^2] \\
&\leq -\left(\eta \frac{1}{\sqrt{G^2 + \epsilon}} + q\right) \mathbb{E}\|\nabla f(\theta_t)\|^2 + L\mathbb{E}[\|\theta_t - \theta'_t\|]E[\|\eta a_t + \Delta_t\|] + \frac{L}{2}\mathbb{E}[\|\eta a_t + \Delta_t\|^2] \\
&\leq -\left(\eta \frac{1}{\sqrt{G^2 + \epsilon}} + q\right) \mathbb{E}\|\nabla f(\theta_t)\|^2 + L\mathbb{E}[\|e_t\| \, \|\eta a_t + \Delta_t\|] + \frac{L}{2}\mathbb{E}[\|\eta a_t + \Delta_t\|^2]
\end{aligned}$$

Summing from $t = 0$ to $t = T_{\text{m}} - 1$ and divide it by T_{m} yields:

$$\begin{aligned}
&\left(\eta \frac{1}{\sqrt{G^2 + \epsilon}} + q\right) \frac{1}{T_{\text{m}}} \sum_{t=0}^{T_{\text{m}}-1} \mathbb{E}[\|\nabla f(\theta_t)\|^2] \\
&\leq \sum_{t=0}^{T_{\text{m}}-1} \frac{\mathbb{E}[f(\theta'_t) - f(\theta'_{t+1})]}{T_{\text{m}}} + \frac{1}{T_{\text{m}}} \sum_{t=0}^{T_{\text{m}}-1} \mathbb{E}[\|e_t\| \, \|\eta a_t + \Delta_t\|] + \frac{L}{2T_{\text{m}}} \sum_{t=0}^{T_{\text{m}}-1} \mathbb{E}[\|\eta a_t + \Delta_t\|^2]
\end{aligned}$$

3 Conclusion

A Appendix