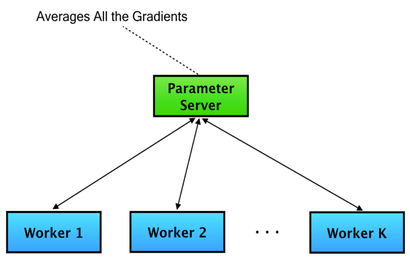
Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification, and Local Computations

Debraj Basu, Deepesh Data, Can Karakus, Suhas Diggavi

Relevant to IOBT CRA Research Area / Task: Research Area 1, Task 1.3

Motivation



Communication bottlenecks in exchange of stochastic gradients for distributed training of high dimensional models over bandlimited networks

State of the Art:

1) **Quantization** (Q): Stochastic such as QSGD, TernGrad, etc. or deterministic such as signSGD

2) **Sparsification** (Comp_k): Selecting TopK or RandomK elements

3) Increased **Local Computations**: Mini batch and local iterations

Composition of Quantizer and Sparsifier

- Step 1. Select the Top k or Random k elements of the d dimensional
- vector v to be transmitted
- Step 2. Quantize the sparse vector using aforementioned quantizers

Ex. 1: $\frac{\text{QSGD}(\text{Top}_k(v))}{1+\beta_{k,s}}$ has a compression factor of $\gamma = \frac{k}{d(1+\beta_{k,s})} \in (0,1]$

Ex. 2: $\frac{||\text{Top}_k(v)||\text{Sign}(\text{Top}_k(v))}{k}$ has comp. factor $\max\left\{\frac{1}{d}, \frac{k}{d}\left(\frac{||\text{Top}_k(v)||_1}{\sqrt{d}\,||\text{Top}_{k(v)}||_2}\right)^2\right\}$

Qsparse-SGD

At time t on worker r,

- Use a composed operator $Q\ Comp_k$
- Compress the previous error to be compensated and the new mini batch update together

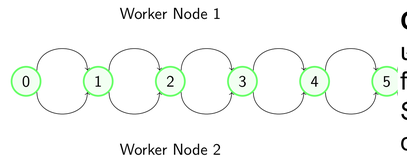
$$g_t^{(r)} \leftarrow Q\ Comp_k\left(m_t^{(r)} + \eta_t \nabla^{(r)} f_{i_t}(\mathbf{x}_t)\right)$$

Send $g_t^{(r)}$ to the master

- Master will compute $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \frac{1}{R}\sum_{r=1}^{R} g_t^{(r)}$ and broadcast to the workers.
- Store the compression error in memory[1]

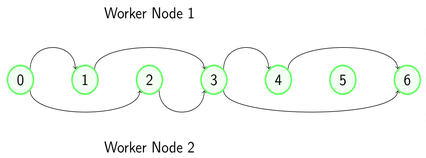
$$m_{t+1}^{(r)} \leftarrow m_t^{(r)} + \eta_t \nabla^{(r)} f_{i_t}(\mathbf{x}_t) - g_t^{(r)}$$

Synchronous Operation

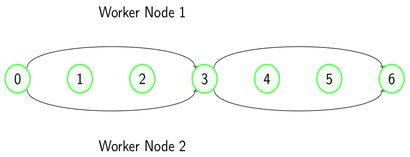


Qsparse-SGD: Sparse Quantized updates together with error feedback at each worker locally. Synchronization schedule same as distributed vanilla SGD

Qsparse-local-SGD: *Synchronous* operation with infrequent communication where the sum of net local update since last synchronization and the historical error is sparsified and quantized before transmission



Asynchronous Operation



Qsparse-local-SGD: Asynchronous operation where only a subset of workers communicates with the master in each iteration

Error Compensation

- The transmitted updates in Qsparse-local-SGD are highly biased due to the combination of different forms of compression with infrequent communication
- The savings in communication is almost for free and is a consequence of the error compensation technique [Seide et al. 2014, Stich et al. 2018]

Synchronous Operation

For batch size b, number of workers R, iteration T, compression $\gamma \in (0,1]$ and synchronization period H, we have the following:

Theorem 1 (Non-convex). H must be $O(\gamma T^{1/4}/(bR)^{3/4})$ for convergence at a rate of $O(1/\sqrt{(bRT)})$

$$\frac{1}{RT}\sum_{t=0}^{T-1}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\hat{\mathbf{x}}_t^{(r)})\|^2 \le \mathcal{O}\left(\frac{J\sigma_{max}}{\sqrt{bRT}}\right) + \mathcal{O}\left(\frac{J^2 bRH^2}{\gamma^2 T}\right)$$

Theorem 2 (Strongly-convex). H must be $O(\gamma(T/bR)^{1/2})$ for convergence at a rate of $O(1/bRT)$

$$\mathbb{E}\left[f\left(\frac{\sum_{t=0}^{T-1} w_t \hat{\mathbf{x}}_t}{\sum_{t=0}^{T-1} w_t}\right)\right] - f^* \le \mathcal{O}\left(\frac{\sigma_{max}^2 \kappa}{\mu bRT}\right) + \mathcal{O}\left(\frac{G^2 H^2 \kappa}{\mu^2 \gamma^2 T^2}\right) + \mathcal{O}\left(\frac{\kappa H^3 G^2}{\mu \gamma^3 T^3}\right)$$

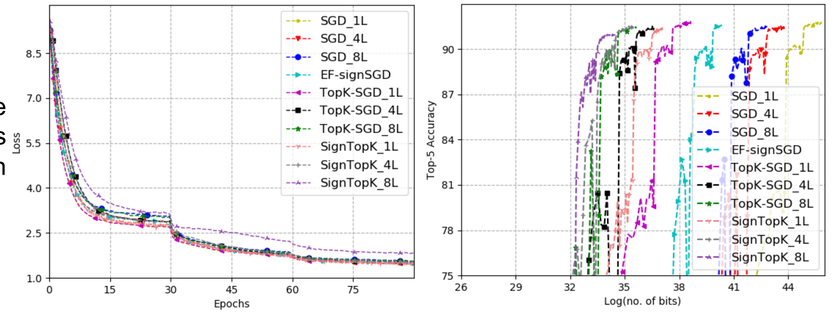
Asynchronous Operation

Theorem 3 (Non -convex). H must be $O(\sqrt{\gamma}T^{1/8}/(bR)^{3/8})$ for convergence at a rate of $O(1/\sqrt{(bRT)})$

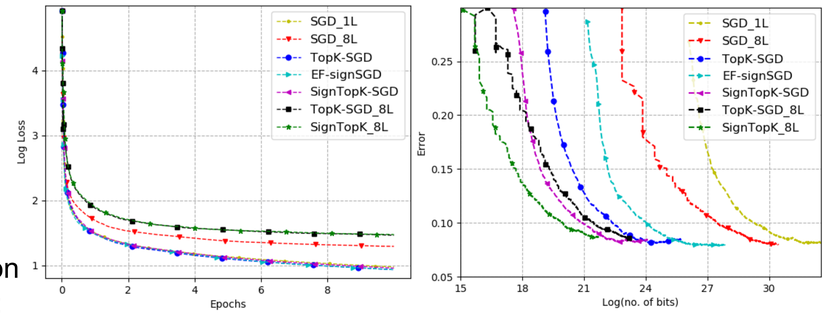
Theorem 4 (Strongly-convex). H must be $O(\sqrt{\gamma}(T/bR)^{1/4})$ for convergence at a rate of $O(1/bRT)$

Experiments

ResNet-50 trained on ImageNet:



Multiclass logistic classifier trained on MNIST:



We recover convergence at rates matching distributed vanilla SGD for non-convex and convex objectives with 15-20x savings in communicated bits over state-of-the-art and 1000x savings over full precision SGD

Publications

Debraj Basu, **Deepesh Data**, Can Karakus, Suhas Diggavi, "Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification, and Local Computations", NeurIPS, 2019.

POINT OF CONTACT:

Deepesh Data, Suhas Diggavi
{deepeshdata, suhasdiggavi}@ucla.edu