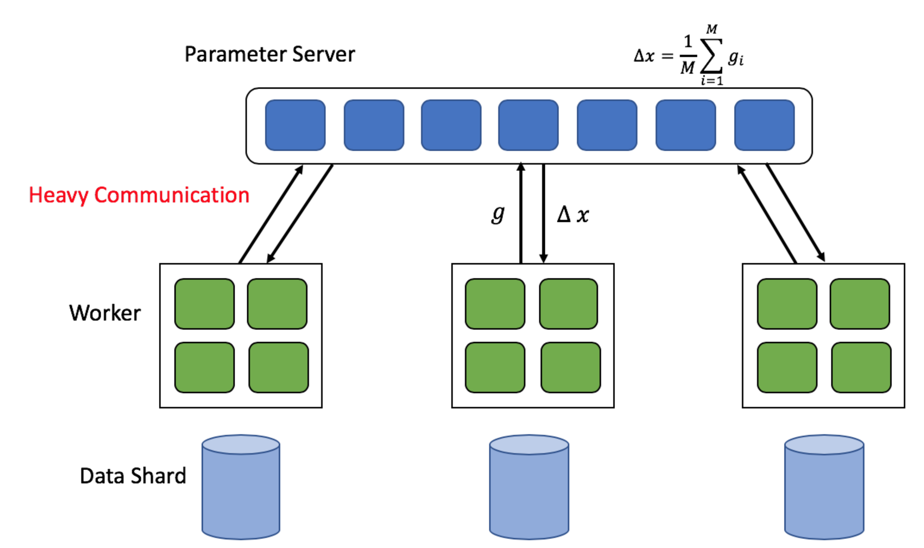
Communication-Efficient Distributed Blockwise Momentum SGD with Error-Feedback

Shuai Zheng, Ziyue Huang, James T. Kwok

Communication is a Bottleneck



To mitigate the communication bottleneck, two common approaches are

- **gradient sparsification**, which sends the most significant, information preserving gradient entries.
- **gradient quantization**, which lowers the gradient's floating-point precision with a smaller bit width.

Our Contributions

We propose a general distributed compressed SGD with Nesterov's momentum, with the following properties

- **two-way compression**, which compresses the gradients both to and from workers.
- **same convergence rates** as full-precision distributed SGD/Momentum SGD (SGDM) for general nonconvex objectives under common assumptions.
- compatible with **general stepsize schedule** for a class of compressors (including the commonly used sign-operator and top-k sparsification).
- **a blockwise compressor** which partitions the gradients into blocks and compresses each block using 1-bit quantization with a scaling factor.

Main Techniques

- **δ-approximate compressor**: an operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ which satisfies

$$\|\mathcal{C}(x) - x\|_2^2 \leq (1-\delta)\|x\|_2^2$$

  e.g., $\mathcal{C}(x) = \|x\|_1/d \cdot \text{sign}(x)$ or $\mathcal{C}(x) = \text{topk}(x)$

- **error feedback**:
  - on each machine i, we keep the difference between the stochastic gradient and the compressed gradient in each iteration as $e_{t,i}$, which is used to correct the compressed gradient in the next iteration.
  - on the server, we maintain another $\tilde{e}_t$ to correct for the compressed aggregated gradient.

Distributed SGD with Error-Feedback

Algorithm 2 Distributed SGD with Error-Feedback (dist-EF-SGD)

Input: stepsize sequence $\{\eta_t\}$ with $\eta_{-1} = 0$; number of workers M; compressor $\mathcal{C}(\cdot)$.
Initialize: $x_0 \in \mathbb{R}^d$; $e_{0,i} = 0 \in \mathbb{R}^d$ on each worker i; $\tilde{e}_0 = 0 \in \mathbb{R}^d$ on server.
for $t = 0, \ldots, T-1$ **do**
  on each worker i
    $p_{t,i} = g_{t,i} + \frac{\eta_{t-1}}{\eta_t} e_{t,i}$ {stochastic gradient $g_{t,i} = \nabla f(x_t, \xi_{t,i})$}
    push $\Delta_{t,i} = \mathcal{C}(p_{t,i})$ **to server** and **pull** $\tilde{\Delta}_t$ **from server**
    $x_{t+1} = x_t - \eta_t \tilde{\Delta}_t$
    $e_{t+1,i} = p_{t,i} - \Delta_{t,i}$
  on server
    pull $\Delta_{t,i}$ **from each worker** i and $\tilde{p}_t = \frac{1}{M}\sum_{i=1}^{M} \Delta_{t,i} + \frac{\eta_{t-1}}{\eta_t}\tilde{e}_t$
    push $\tilde{\Delta}_t = \mathcal{C}(\tilde{p}_t)$ **to each worker**
    $\tilde{e}_{t+1} = \tilde{p}_t - \tilde{\Delta}_t$
end for

Convergence Analysis

- **error-corrected iterate**: $\tilde{x}_t = x_t - \eta_{t-1}(\tilde{e}_t + \frac{1}{M}\sum_{i=1}^{M} e_{t,i})$
- **virtual recurrence**: $\tilde{x}_{t+1} = \tilde{x}_t - \frac{\eta_t}{M}\sum_{i=1}^{M} g_{t,i}$
- **bounded error**: $\mathbb{E}[\|\tilde{e}_t + \frac{1}{M}\sum_{i=1}^{M} e_{t,i}\|_2^2] \leq \frac{8(1-\delta)G^2}{\delta^2}[1 + \frac{16}{\delta^2}]$, which implies that $\nabla F(\tilde{x}_t) \approx \nabla F(x_t)$.

Then we can utilize the tools used on the full-precision distributed SGD and show dist-EF-SGD has a convergence rate of $\mathcal{O}(1/\sqrt{MT})$.

Nesterov's Momentum

Algorithm 4 Distributed Blockwise Momentum SGD with Error-Feedback (dist-EF-blockSGDM)

Input: stepsize sequence $\{\eta_t\}$ with $\eta_{-1} = 0$; momentum parameter $0 \leq \mu < 1$; number of workers M; block partition $\{\mathcal{G}_1, \ldots, \mathcal{G}_B\}$.
Initialize: $x_0 \in \mathbb{R}^d$; $m_{-1,i} = e_{0,i} = 0 \in \mathbb{R}^d$ on each worker i; $\tilde{e}_0 = 0 \in \mathbb{R}^d$ on server
for $t = 0, \ldots, T-1$ **do**
  on each worker i
    $m_{t,i} = \mu m_{t-1,i} + g_{t,i}$ {stochastic gradient $g_{t,i} = \nabla f(x_t, \xi_{t,i})$}
    $p_{t,i} = \mu m_{t,i} + g_{t,i} + \frac{\eta_{t-1}}{\eta_t} e_{t,i}$
    push $\Delta_{t,i} = \left[\frac{\|p_{t,i,\mathcal{G}_1}\|_1}{d_1}\text{sign}(p_{t,i,\mathcal{G}_1}), \ldots, \frac{\|p_{t,i,\mathcal{G}_B}\|_1}{d_B}\text{sign}(p_{t,i,\mathcal{G}_B})\right]$ **to server**
    $x_{t+1} = x_t - \eta_t \tilde{\Delta}_t$ {$\tilde{\Delta}_t$ **is pulled from server**}
    $e_{t+1,i} = p_{t,i} - \Delta_{t,i}$
  on server
    pull $\Delta_{t,i}$ **from each worker** i and $\tilde{p}_t = \frac{1}{M}\sum_{i=1}^{M} \Delta_{t,i} + \frac{\eta_{t-1}}{\eta_t}\tilde{e}_t$
    push $\tilde{\Delta}_t = \left[\frac{\|\tilde{p}_{t,\mathcal{G}_1}\|_1}{d_1}\text{sign}(\tilde{p}_{t,\mathcal{G}_1}), \ldots, \frac{\|\tilde{p}_{t,\mathcal{G}_B}\|_1}{d_B}\text{sign}(\tilde{p}_{t,\mathcal{G}_B})\right]$ **to each worker**
    $\tilde{e}_{t+1} = \tilde{p}_t - \tilde{\Delta}_t$
end for

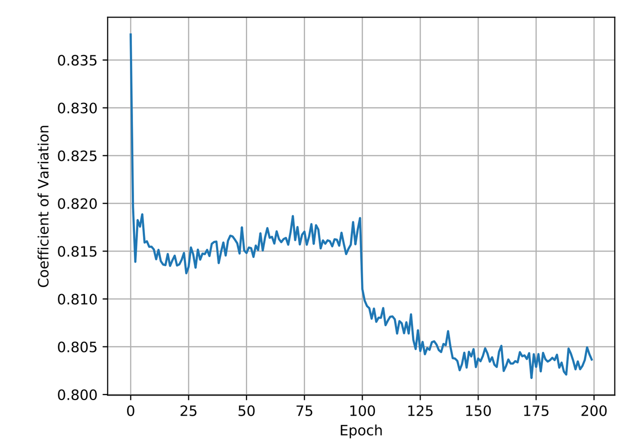
Similar with the analysis for dist-EF-SGD, we can show dist-EF-SGDM has a convergence rate of

$$\mathcal{O}([(1-\mu)(F(x_0) - F_*) + \sigma^2/(1-\mu)]/\sqrt{MT})$$

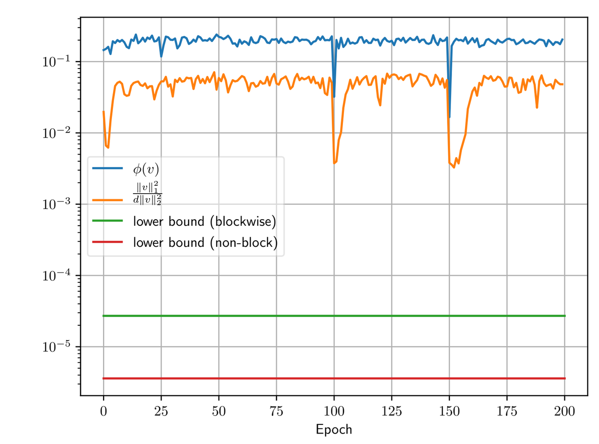
- μ balances between initial optimality gap $F(x_0) - F_*$ and variance σ^2

Experiments

Blockwise/Non-Block Compressor



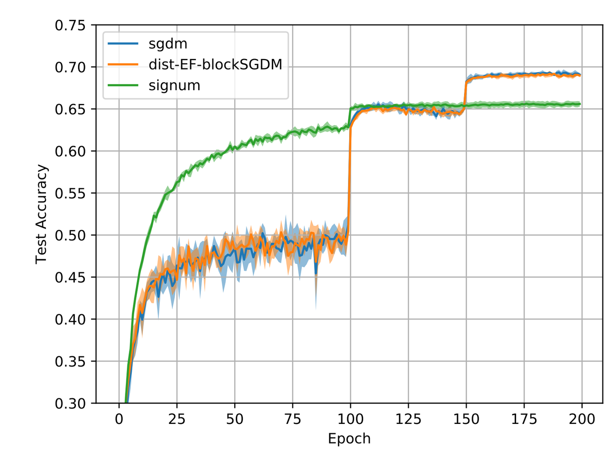
(a) Coefficient of variation (CV) of $\{|g_{t,i}|\}_{i \in \mathcal{G}_b}$.



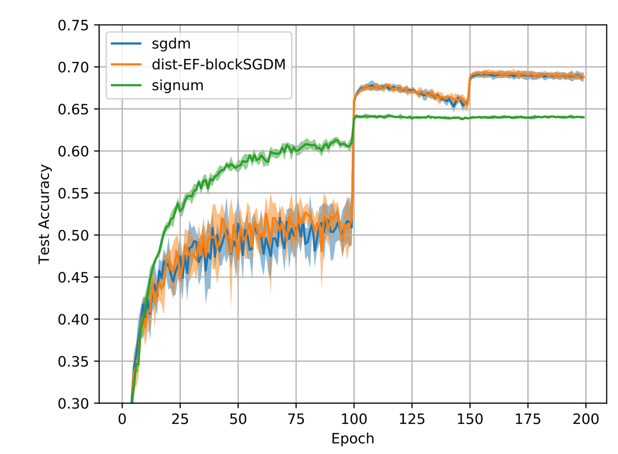
(b) δ for blockwise and non-block versions.

- $CV < 1$ means gradients in each block have low variability.
- blockwise compressor achieves larger δ than non-block compressor.

Multi-GPU Experiment on CIFAR-100



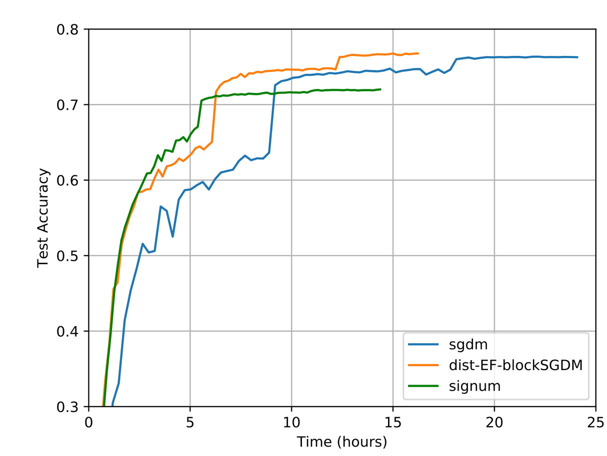
(c) batch size 8 per GPU



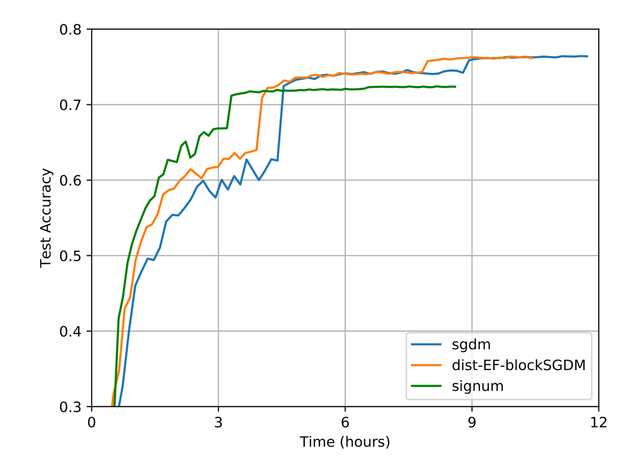
(d) batch size 32 per GPU

- dist-EF-blockSGDM matches the performance of full-precision SGDM, while signum has worse accuracy.

Distributed Training on ImageNet



(e) 7 workers: Test accuracy w.r.t. epoch.



(f) 15 workers: Test accuracy w.r.t. epoch.

- for 7 workers, dist-EF-blockSGDM reaches SGDM's highest accuracy in around 13 hours, while SGDM takes 24 hours, leading to a 46% speedup.
- with 15 workers, we expect we can achieve more speedup by using more parameter servers.