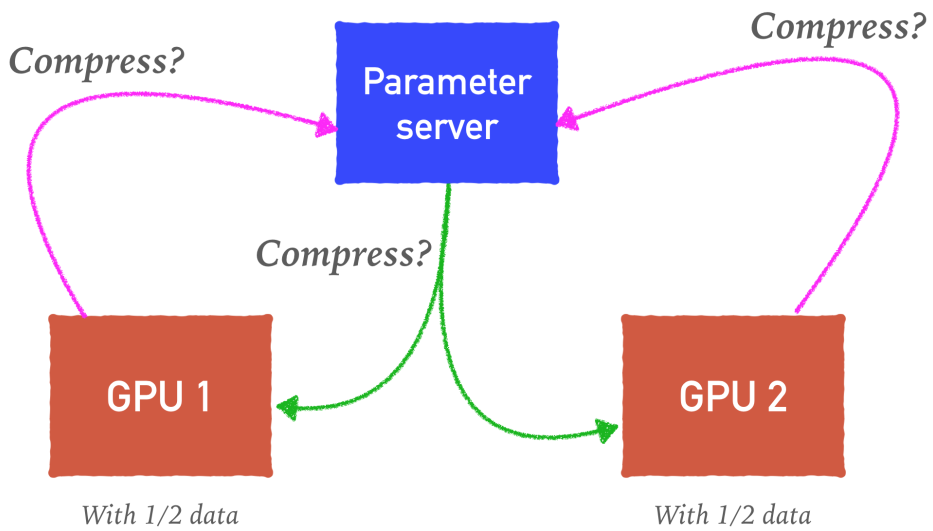
Compression by the signs: distributed learning is a two-way street

Jeremy Bernstein[1,3], Yu-Xiang Wang[3], Kamyar Azizzadenesheli[2], Anima Anandkumar[1,3]

[1]Caltech [2]UCI [3]Amazon AI

Motivation

The sign gradient method performs 1-bit quantisation of gradients, and empirically converges just as fast as SGD for deep networks. Therefore it has potential for **distributed optimisation** where gradient communication across machines is a bottleneck.



Existing gradient quantisation schemes like **QSGD** [1] have good practical performance but weak theoretical foundations. signSGD also resembles **Adam** [2], a popular optimiser which also has weak theoretical foundations.

signSGD takes the sign of the stochastic gradient

$$x_{k+1} = x_k - \eta_k \text{sign}(g_k)$$

The question is, how can we extend signSGD to the multi-worker setting and still have gradient compression benefits?

We propose signSGD with majority vote. The scheme is elegant since **all communication is 1 bit quantised**.

Majority vote lets M workers vote on the true gradient sign

$$x_{k+1} = x_k - \eta_k \,\text{sign}\left[\sum_{i=1}^{M} \text{sign}(g_k)\right]$$

1. each worker sends its stochastic sign gradient to the parameter server
2. the parameter server sums the independent estimates and returns the majority decision

Single worker theory

The first step is to establish the properties of the single worker algorithm, which is just signSGD.

We work in the non-convex setting, under very general assumptions.

Assumptions

1. Objective function has a lower bound f^*
2. Objective function has coordinate-wise Lipschitz smoothness $\vec{L}$
3. Stochastic gradient has a coordinate-wise variance bound $\vec{\sigma}$

We prove the convergence rate of signSGD to first order critical points (either saddles or local minima).

Convergence rate for single-worker signSGD

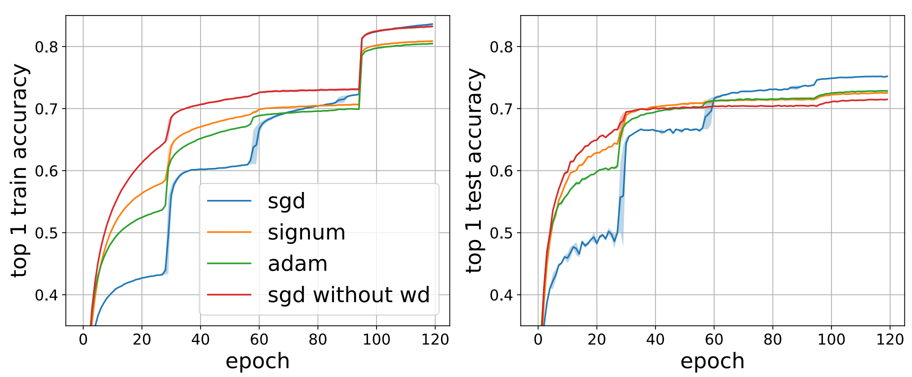
$$\mathbb{E}\left[\min_{0\le k\le K-1} \|g_k\|_1\right]^2 \le \frac{1}{\sqrt{N}}\left[\sqrt{\|\vec{L}\|_1}\left(f_0 - f_* + \frac{1}{2}\right) + 2\|\vec{\sigma}\|_1\right]^2$$

Comparing the rate to SGD

1. N measures the number of stochastic gradient calls up to step K
2. the $1/\sqrt{N}$ rate matches SGD
3. ℓ_1 norms replace the typical SGD-style ℓ_2 norms
4. the theory relies on a large batch size which has systems benefits

Single worker performance on Imagenet

We find that signSGD has **extremely similar Imagenet performance to Adam**.



signSGD performs slightly worse than SGD, but this may be because we used a much smaller batch size than suggested by theory.

Multi worker theory

Remarkably we are able to show that signSGD with majority vote gets the **same theoretical speedup as full-precision distributed SGD**.

The result holds under one additional assumption:

Assumptions

1. Objective function has a lower bound f^*
2. Objective function has coordinate-wise Lipschitz smoothness $\vec{L}$
3. Stochastic gradient has a coordinate-wise variance bound $\vec{\sigma}$
4. **Stochastic gradient noise is symmetric about the mean**

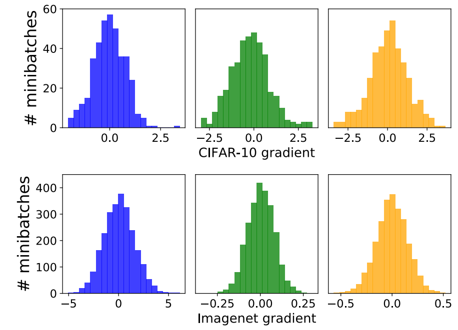
This additional assumption is reasonable by the central limit theorem.

Convergence rate for M-worker majority vote

$$\mathbb{E}\left[\min_{0\le k\le K-1} \|g_k\|_1\right]^2 \le \frac{1}{\sqrt{N}}\left[\sqrt{\|\vec{L}\|_1}\left(f_0 - f_* + \frac{1}{2}\right) + \frac{2}{\sqrt{M}}\|\vec{\sigma}\|_1\right]^2$$

Validating the additional assumption

We plot stochastic gradient histograms for 6 randomly chosen neural network weights. We find that **gradient noise in real problems is approximately symmetric**, even for mini-batches of size O(100).



Next steps

1. single worker signSGD is available in mxnet now!
2. we are currently implementing and testing majority vote

Bibliography

[1] Dan Alistarh et al. QSGD. 2017

[2] Diederik P. Kingma, Jimmy Ba. Adam: A Method for Stochastic Optimization. 2014