signSGD with majority vote is communication efficient and fault tolerant

Jeremy Bernstein[1], Jiawei Zhao[1,2], Kamyar Azizzadenesheli[1,3], Anima Anandkumar[1]

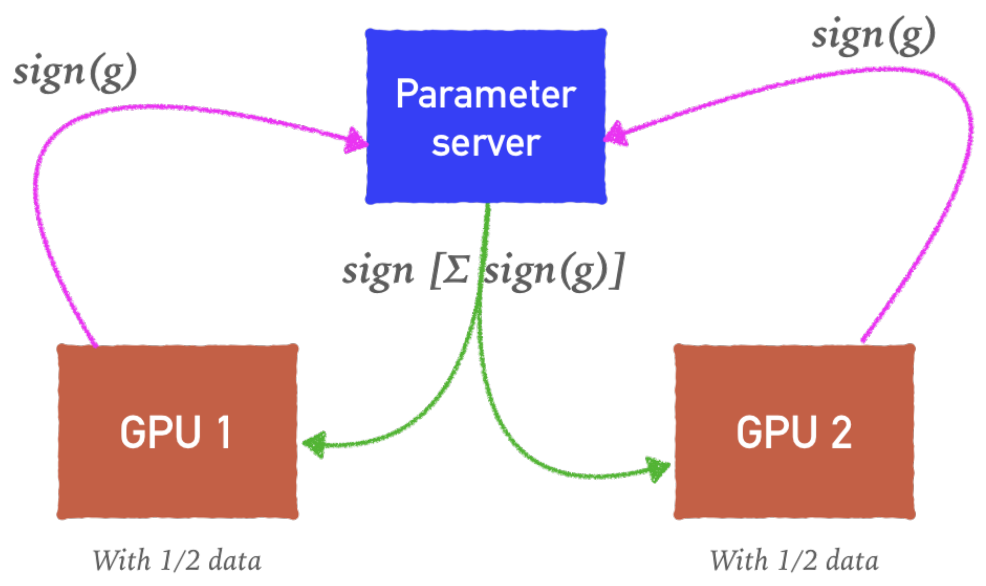
[1]Caltech [2]NUAA [3]UCI

Motivation

We would like a scheme for **distributed optimisation** that satisfies some natural desiderata:

1. fast algorithmic convergence;
2. good generalisation performance;
3. communication efficiency;
4. robustness to network faults.

signSGD was proposed in [1] as a way to accomplish desiderata 1–3.



Overview of signSGD with majority vote

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

Majority vote lets M workers vote on the true gradient sign

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

The algorithm is nice because all communication is 1-bit compressed, and good empirical performance was established in [1].

But prior work was limited since the theory relied on a much larger batch size than needed in practice, and desiderata 4 was not addressed.

Why care about small batch size?

Prior theoretical analysis [1] of signSGD required a large batch size that grew with the total number of iterations: $n \propto K$.

Then why does a small batch (e.g. batch size $= 128$) version of the algorithm work well in practice?

Also, since **signSGD is the $\beta \to 0$ limit of Adam** [2], a small batch theory would improve the theoretical understanding of Adam.

Small batch theory

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 $\vec{\sigma}$
4. **Gradient noise is unimodal & symmetric about the mean**

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

Convergence rate for small batch signSGD

Run single worker signSGD for K iterations under Assumptions 1 to 4. Set the learning rate, η, and mini-batch size, n, as

$$\eta = \sqrt{\frac{f_0 - f_*}{\|\vec{L}\|_1 K}}, \qquad n = 1.$$

Let H_k be the set of gradient components at step k with signal-to-noise ratio $S_i := \frac{|g_{k,i}|}{\sigma_i}$ larger than $\frac{2}{\sqrt{3}}$. Then we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\sum_{i\in H_k}|g_{k,i}| + \sum_{i\notin H_k}\frac{g_{k,i}^2}{\sigma_i}\right] \le 3\sqrt{\frac{\|\vec{L}\|_1(f_0 - f_*)}{K}}.$$

Remarks on the theory

1. the $1/\sqrt{K}$ rate matches SGD
2. the two terms on LHS suggests that convergence goes through two phases as gradients move from high to low SNR
3. assumption 4 is reasonable by the central limit theoerem

Fault tolerance

For extremely large-scale distributed optimisation, it may not be possible to ensure the trustworthiness of all workers or network links.

Naïve SGD offers zero protection since any worker may corrupt the entire model at any time by sending an infinite gradient.

Majority vote protects against **blind multiplicative adversaries** that element-wise multiply their gradient estimate $\tilde{g}$ by any v not conditioned on $\tilde{g}$. This class includes rescalings, randomisations and inversions.

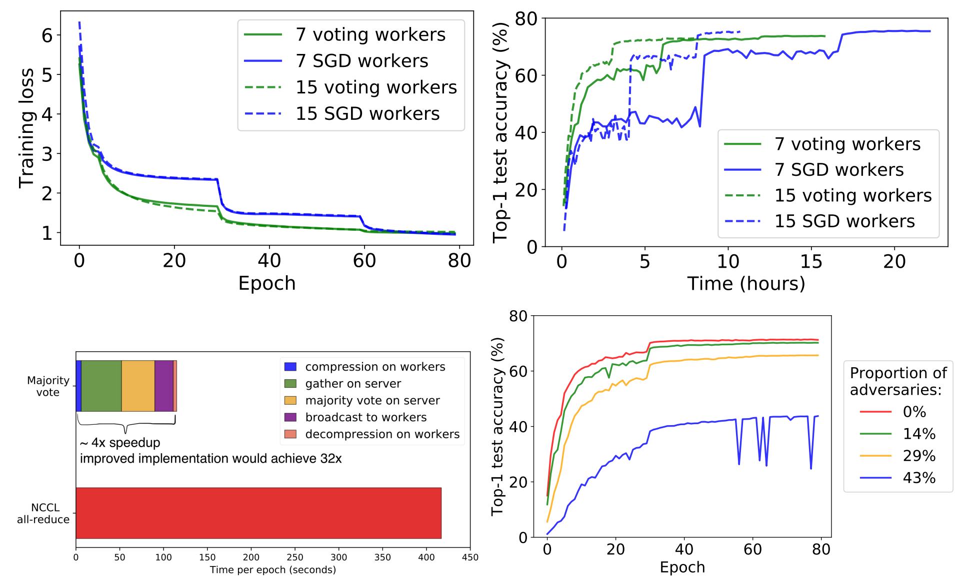
Convergence rate for majority vote with adversaries

Run signSGD with majority vote for K iterations under Assumptions 1 to 4. Set the learning rate as $\eta = \sqrt{\frac{f_0 - f_*}{\|\vec{L}\|_1 K}}$ and mini-batch size per worker as $n = K$.

Assume that a fraction $\alpha < \frac{1}{2}$ of the M workers are blind multiplicative adversaries, and let $N = K^2$ be the total number of stochastic gradient calls per worker up to step K. Then majority vote converges at rate

$$\left[\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\,\|g_k\|_1\right]^2 \le \frac{4}{\sqrt{N}}\left[\frac{1}{1-2\alpha}\frac{\|\vec{\sigma}\|_1}{\sqrt{M}} + \sqrt{\|\vec{L}\|_1(f_0 - f^*)}\right]^2$$

Empirical validation on Imagenet



Bibliography

[1] Jeremy Bernstein, Yu-Xiang Wang et al. signSGD. 2018

[2] Diederik P. Kingma, Jimmy Ba. Adam. 2014