signSGD with majority vote is communication efficient and fault tolerant

Jeremy Bernstein1, Jiawei Zhao1-2, Kamyar Azizzadenesheli1-3, Anima Anandkumar1
1Caltech 2NUAA 3UCI

Motivation
We would like a scheme for distributed optimisation that satisfies some natural desiderata:
- fast algorithmic convergence;
- good generalisation performance;
- communication efficiency;
- robustness to network faults.

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

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 \( \beta \rightarrow 0 \) limit of Adam [2], a small batch theory would improve the theoretical understanding of Adam.

Overview of signSGD with majority vote

Majority vote lets \( M \) workers vote on the true gradient sign
\[ x_{k+1} = x_k - \eta_k \text{ sign} \left( \sum_{i=1}^{M} \text{sign}(g_i) \right) \]

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.

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

Naive 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 \( \hat{g} \) by any \( v \) not conditioned on \( \hat{g} \). 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 \( \eta = \frac{\sqrt{\log M}}{\sqrt{L} \sqrt{n}} \) and mini-batch size per worker as \( n = K \).
Assume that a fraction \( \alpha < \frac{1}{M} \) 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
\[ \left( \frac{1}{K} \sum_{k=0}^{K-1} \|g_k\|_1 \right)^2 \leq \frac{4}{\sqrt{N}} \left( 1 - 2\alpha \sqrt{M} + \sqrt{\|\tilde{L}\|_1 (f_0 - f^*)} \right)^2 \]

Empirical validation on Imagenet

Bibliography