SIGNSGD: COMPRESSED OPTIMIZATION

FOR NON-CONVEX PROBLEMS

JEREMY BERNSTEIN , YU-XIANG WANG, KAMYAR AZIZZADENESHELI, ANIMA ANANDKUMAR

Presentation by

Sai Saran Putta

vputta@buffalo.edu

50483814

University at Buffalo The State University of New York

Structure

- Introduction
- Background
- signSGD in single worker setting
- signSGD in Distributed Setting
- signum
- Experiments
- Conclusion

Introduction

SIGNSGD: Compressed Optimization for Non-Convex Problems

- Considers only the sign of the gradients.
- Compressed optimization technique: reduces the overall training time.
- For non-convex problems (typical in case of DNN).

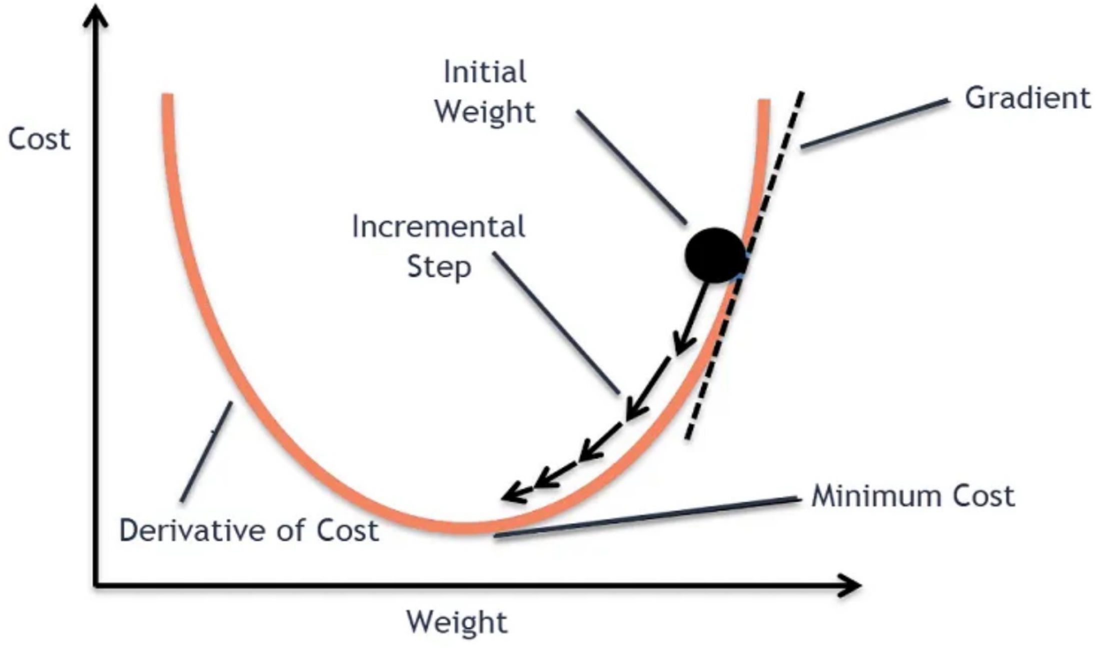
Background

Gradient Descent:

Step 1: Calculate gradient at current point

Step 2: Move in the opposite direction of slope increase by the computed amount

$$x_{k+1} = x_k - \delta * \widetilde{g_k}$$



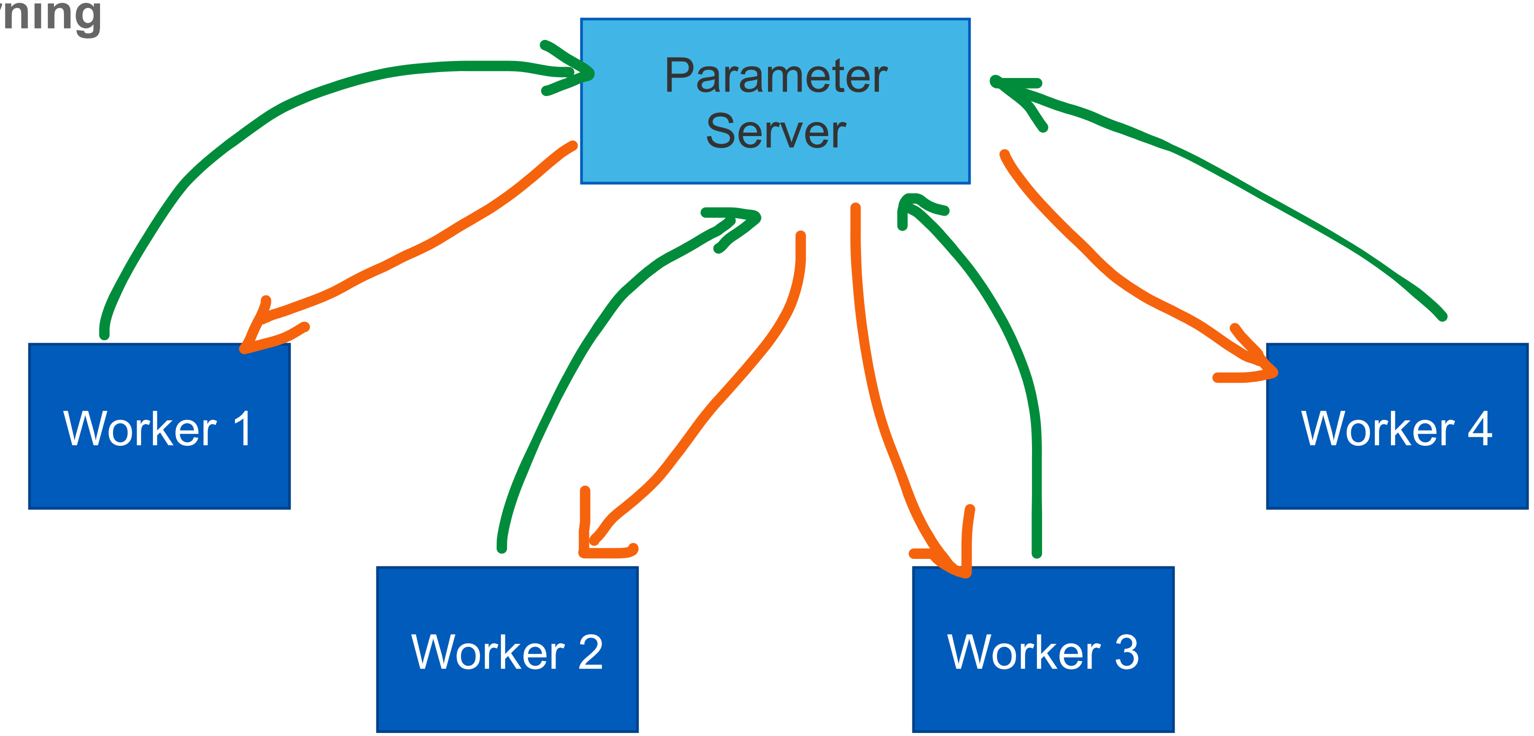
Background

Distributed Learning

- Optimization tasks are computationally resource intensive
- They are not very scalable.
- Can be accelerated by using distributed systems.

Background

Distributed Learning



SignSGD

Algorithm 1 SIGNSGD

Input: learning rate δ, current point x_k

$\tilde{g}_k \leftarrow \text{stochasticGradient}(x_k)$

$x_{k+1} \leftarrow x_k - \delta\,\text{sign}(\tilde{g}_k)$

SignSGD

Convergence Rate

Assumptions:

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

Define

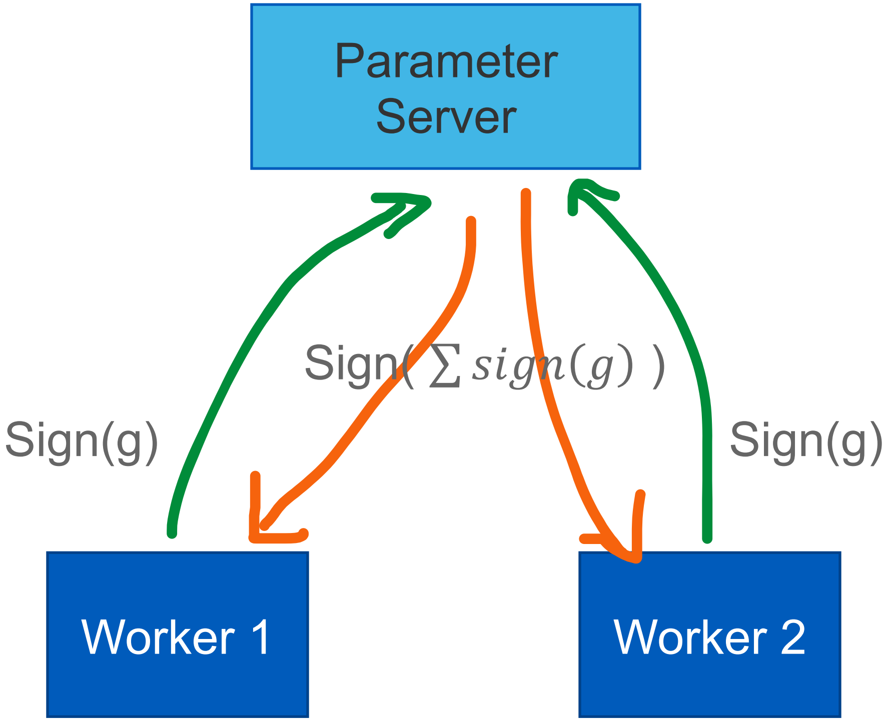
Number of Iterations : K
Number of cumulative gradient calls: N
Learning rate : $\frac{1}{\sqrt{\|\vec{L}\| K}}$

SGD gets rate $$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|g_k\|_2^2\right] \leq \frac{1}{\sqrt{N}}\left[2\|\vec{L}\|_\infty(f_0 - f_*) + \|\vec{\sigma}\|_2^2\right]$$

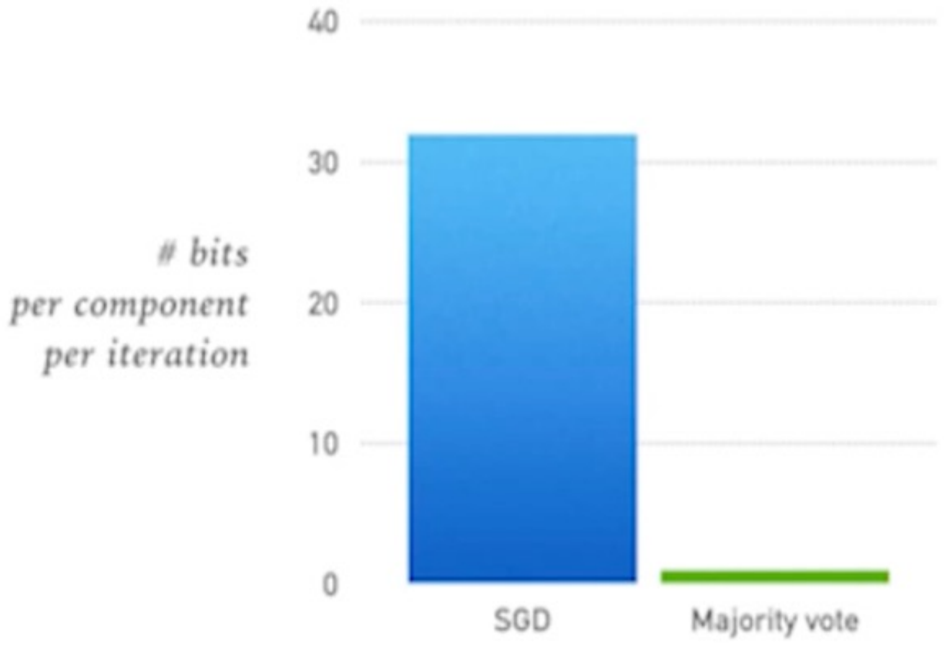
signSGD gets rate $$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|g_k\|_1\right]^2 \leq \frac{1}{\sqrt{N}}\left[\sqrt{\|\vec{L}\|_1}\left(f_0 - f_* + \frac{1}{2}\right) + 2\|\vec{\sigma}\|_1\right]^2$$

SignSGD in Distributed Setting

Majority Voting



Compression Savings



SignSGD with Majority Voting

Algorithm 3 Distributed training by majority vote

Input: learning rate δ, current point x_k, # workers M each with an independent gradient estimate $\tilde{g}_m(x_k)$

on server

- **pull** sign($\tilde{g}_m$) **from** each worker
- **push** sign$\left[\sum_{m=1}^{M} \text{sign}(\tilde{g}_m)\right]$ **to** each worker

on each worker

- $x_{k+1} \leftarrow x_k - \delta \, \text{sign}\left[\sum_{m=1}^{M} \text{sign}(\tilde{g}_m)\right]$

SignSGD with Majority Voting

Convergence Rate

Assumptions:

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

Define

Number of Iterations : K
Number of gradient calls: N

SGD gets rate

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|g_k\|_2^2\right] \le \frac{1}{\sqrt{N}}\left[2\|\vec{L}\|_\infty(f_0 - f_*) + \frac{\|\vec{\sigma}\|_2^2}{\sqrt{M}}\right]$$

if gradient noise is unimodal symmetric majority vote gets

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

Signum

- Momentum can be added to speed up the training
- Instead of taking single gradient, Momentum considers running average of recent gradients
- Take sign of momentum to incorporate momentum into signSGD

Signum

Algorithm 2 SIGNUM

Input: learning rate δ, momentum constant $\beta \in (0, 1)$, current point x_k, current momentum m_k

$\tilde{g}_k \leftarrow \text{stochasticGradient}(x_k)$

$m_{k+1} \leftarrow \beta m_k + (1 - \beta)\tilde{g}_k$

$x_{k+1} \leftarrow x_k - \delta \, \text{sign}(m_{k+1})$

Experiments

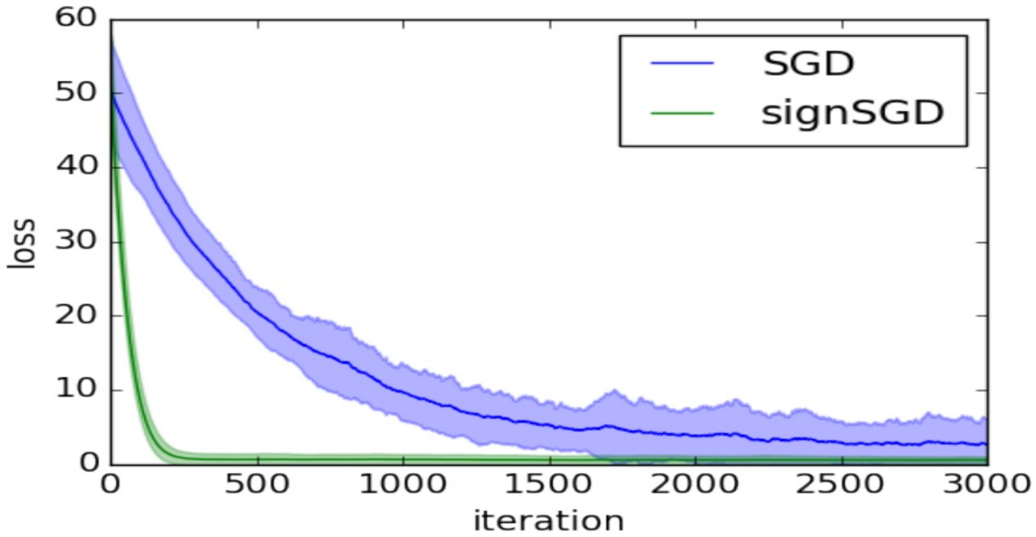
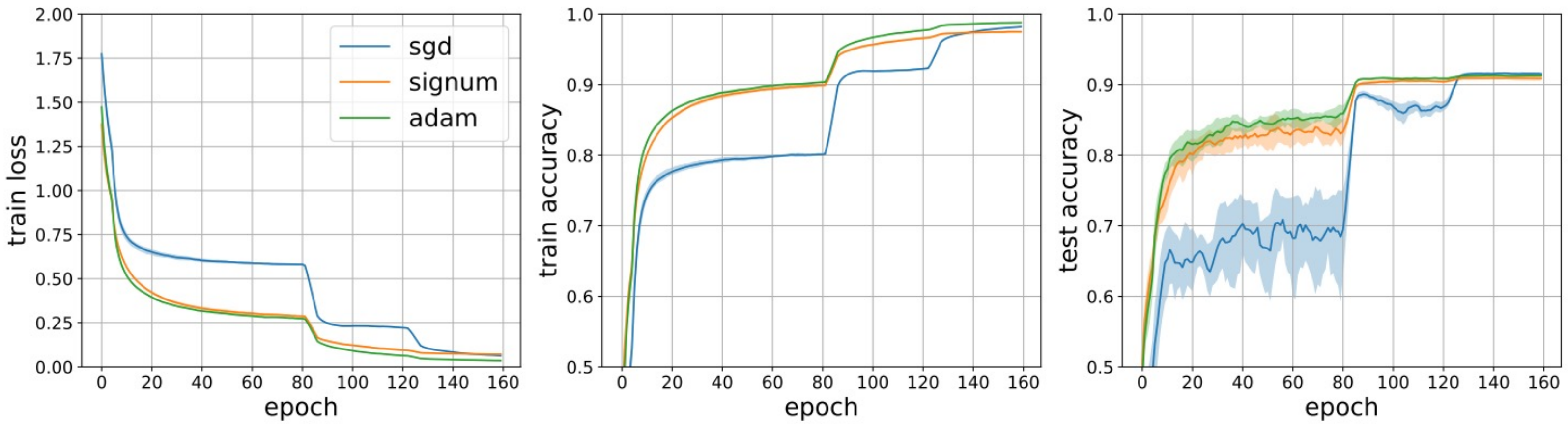


Figure A.1. A simple toy problem where SIGNSGD converges faster than SGD. The objective function is just a quadratic $f(x) = \frac{1}{2}x^2$ for $x \in \mathbb{R}^{100}$. The gradient of this function is just $g(x) = x$. We construct an artificial stochastic gradient by adding Gaussian noise $\mathcal{N}(0, 100^2)$ to only the first component of the gradient. Therefore the noise is extremely sparse. The initial point is sampled from a unit variance spherical Gaussian. For each algorithm we tune a separate, constant learning rate finding 0.001 best for SGD and 0.01 best for SIGNSGD. SIGNSGD appears more robust to the sparse noise in this problem. Results are averaged over 50 repeats with ± 1 standard deviation shaded.

Experiments



CIFAR-10 results using SIGNUM to train a Resnet-20 model.

Conclusion

- A general framework for studying sign-based methods in stochastic non-convex optimization.
- Provides concrete proofs that these algorithms converge under certain assumptions
- Yet to be benchmarked for realistic scenarios on the distributed systems.

THANK YOU!