signSGD

Compressed optimisation for non-convex problems

Jeremy Bernstein
Caltech

Yu-Xiang Wang
UCSB/Amazon

Kamyar Azizzadenesheli
UCI

Anima Anandkumar
Caltech/Amazon
signSGD

Compressed optimisation for non-convex problems

- Snap gradient components to ±1
- Reduces communication time
- Realistic for deep learning
Why care about signSGD?

Theoretical convergence results

Empirical characterisation of neural net landscape

Imagenet results
GRADIENT COMPRESSION . . . . . WHY CARE?

Parameter server

GPU 1
With 1/2 data

GPU 2
With 1/2 data
GRADIENT COMPRESSION . . . . . . WHY CARE?

**COMPRESS?**

- GPU 1
- GPU 2
- GPU 3
- GPU 4
- Parameter server
- GPU 5
- GPU 7
- GPU 7
DISTRIBUTED SGD

Parameter server

\[ \sum g \]

GPU 1

With 1/2 data

GPU 2

With 1/2 data
SIGN SGD WITH MAJORITY VOTE

Parameter server

\[ \text{sign}(g) \quad \text{sign} \left[ \sum \text{sign}(g) \right] \quad \text{sign}(g) \]

GPU 1

GPU 2

With 1/2 data

With 1/2 data
COMPRESSION SAVINGS OF MAJORITY VOTE

# bits per component per iteration

SGD | Majority vote
**signSGD IS A SPECIAL CASE OF ADAM**

$$\text{signSGD} \quad \text{sign}(g_k) = \frac{g_k}{\sqrt{g_k^2}} \quad \text{Adam} \quad \frac{g_k + \beta g_{k-1} + \beta^2 g_{k-2} + \ldots}{\sqrt{g_k^2 + \beta g_{k-1}^2 + \beta^2 g_{k-2}^2 + \ldots}}$$

*Signum*  \[ \text{sign}(g_k + \beta g_{k-1} + \beta^2 g_{k-2} + \ldots) \]

*(Sign momentum)*
ADAM..................WHY CARE?

# of Google scholar citations

SGD
Robbins & Monro
Kingma & Ba
Turing test
Turing

Adam
UNIFYING ADAPTIVE GRADIENT METHODS + COMPRESSION

Sign descent
- weak theoretical foundation
- incredibly popular (e.g. Adam)

Compressed descent
- weak theoretical foundation
- take pains to correct bias
- empirically successful

Sign-based gradient compression?
Why care about signSGD?

Theoretical convergence results

Empirical characterisation of neural net landscape

Imagenet results
DOES $\text{signSGD}$ EVEN CONVERGE?

What might we fear?

➤ Might not converge at all
➤ Might have horrible dimension dependence
➤ Majority vote may give no speedup by adding extra machines

*Compression can be a free lunch*

Our results

➤ It does converge
➤ We characterise functions where $\text{signSGD} \ & \ \text{majority vote}$ are as nice as SGD
➤ Suggest these functions are typical in deep learning
SINGLE WORKER RESULTS

Assumptions

- Objective function lower bound $f_*$
- Coordinate-wise variance bound $\overrightarrow{\sigma}$
- Coordinate-wise gradient Lipschitz $\overrightarrow{L}$

Define

- Number of iterations $K$
- Number of backpropagations $N$

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

\[ \mathbb{E} \left[ \frac{1}{K} \sum_{k=0}^{K-1} ||g_k||_1 \right]^2 \leq \frac{1}{\sqrt{N}} \left[ \sqrt{||\overrightarrow{L}||_1} \left( f_0 - f_* + \frac{1}{2} \right) + 2 ||\overrightarrow{\sigma}||_1 \right]^2 \]
SINGLE WORKER RESULTS

Assumptions

> Objective function lower bound $f^*$
> Coordinate-wise variance bound $\overrightarrow{\sigma}$
> Coordinate-wise gradient Lipschitz $\overrightarrow{L}$

Define

> Number of iterations $K$
> Number of backpropagations $N$

SGD gets rate

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

signSGD gets rate

$$\mathbb{E} \left[ \frac{1}{K} \sum_{k=0}^{K-1} \sqrt{d} \| \overrightarrow{g}_k \|_2 \right]^2 \leq \frac{1}{\sqrt{N}} \left[ \sqrt{d} \| \overrightarrow{L} \|_{\infty} \left( f_0 - f^* + \frac{1}{2} \right) + 2 \sqrt{d} \| \overrightarrow{\sigma} \|_2 \right]^2$$

Define

- $f^*$ ➤ Objective function lower bound
- $\overrightarrow{\sigma}$ ➤ Coordinate-wise variance bound
- $\overrightarrow{L}$ ➤ Coordinate-wise gradient Lipschitz
- $K$ ➤ Number of iterations
- $N$ ➤ Number of backpropagations
**MULTI WORKER RESULTS** *with M workers*

Assumptions

- Objective function lower bound $f_*$
- Coordinate-wise variance bound $\sigma$
- Coordinate-wise gradient Lipschitz $L$

### SGD gets rate

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

**if gradient noise is unimodal symmetric**

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

**majority vote gets**
Why care about signSGD?

Theoretical convergence results

Empirical characterisation of neural net landscape

Imagenet results
CHARACTERISING THE DEEP LEARNING LANDSCAPE EMPIRICALLY

➤ **signSGD** cares about **gradient density**

Natural measure of density

\[
\phi(\vec{v}) = \frac{\|\vec{v}\|_1}{d\|\vec{v}\|_2} = 1 \text{ for fully dense } \vec{v} \\
\approx 0 \text{ for fully sparse } \vec{v}
\]

➤ **majority vote** cares about **noise symmetry**

For large enough mini-batch size, reasonable by Central Limit Theorem.
Why care about signSGD?

Theoretical convergence results

Empirical characterisation of neural net landscape

Imagenet results
SIGNUM IS COMPETITIVE ON IMAGENET

Performance very similar to Adam
May want to switch to SGD towards end?
DOES MAJORITY VOTE WORK?

Cifar-10, Resnet-18

Train Accuracy

Test Accuracy

Jiawei Zhao
NUAA
on server

pull $\text{sign}(\tilde{g}_m)$ from each worker

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

on each worker

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

**Poster tonight!**

6.15—9 PM @ Hall B #72