Jeremy Bernstein
Caltech

signSGD

Yu-Xiang Wang
UCSB/Amazon

Compressed optimisation for non-convex problems

Kamyar Azizzadenesheli
UCI

Anima Anandkumar
Caltech/Amazon

signSGD

Compressed optimisation for non-convex problems

Reduces communication time

Snap gradient components to ±1

Realistic for deep learning

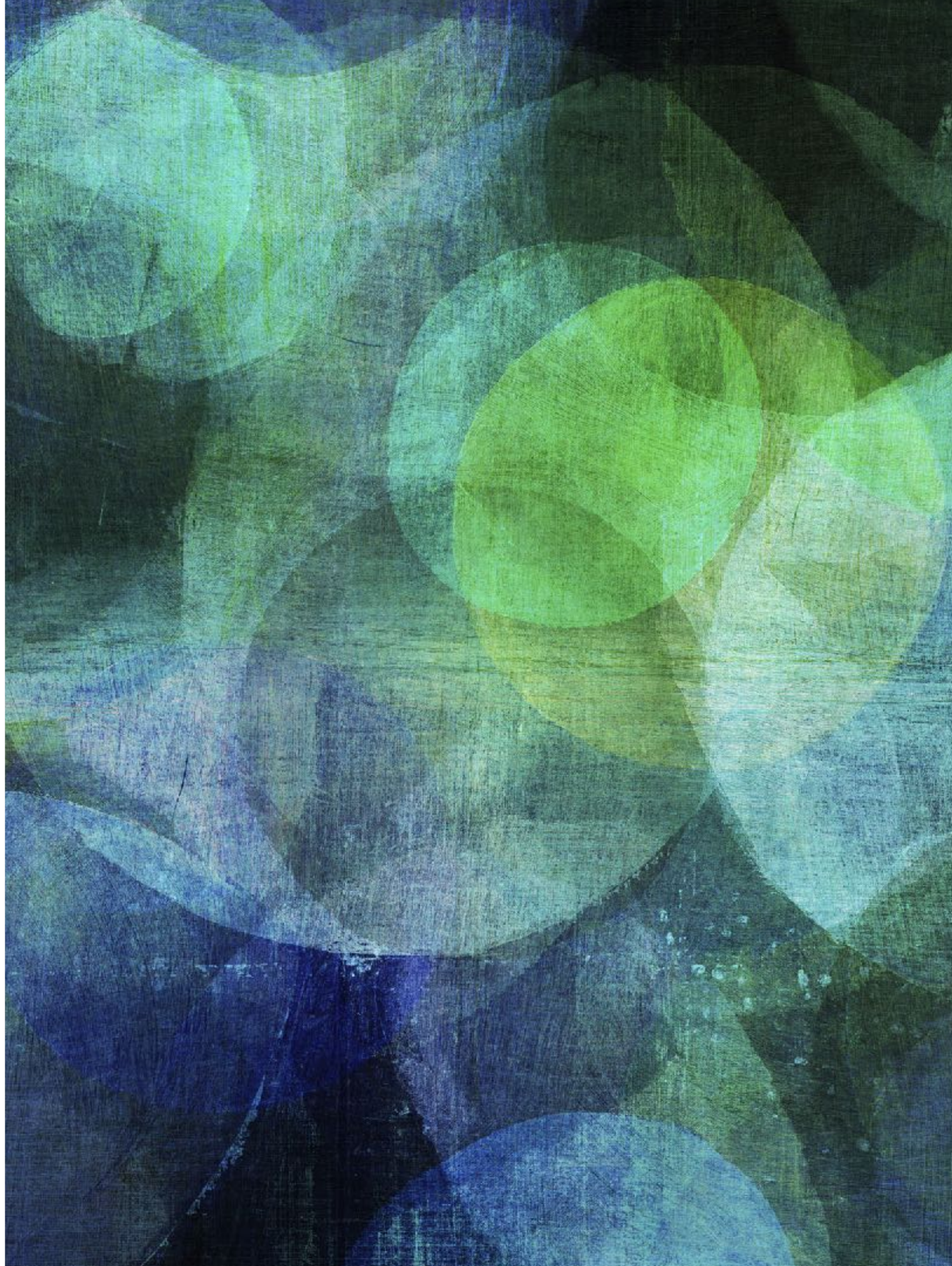
STRUCTURE

Why care about signSGD?

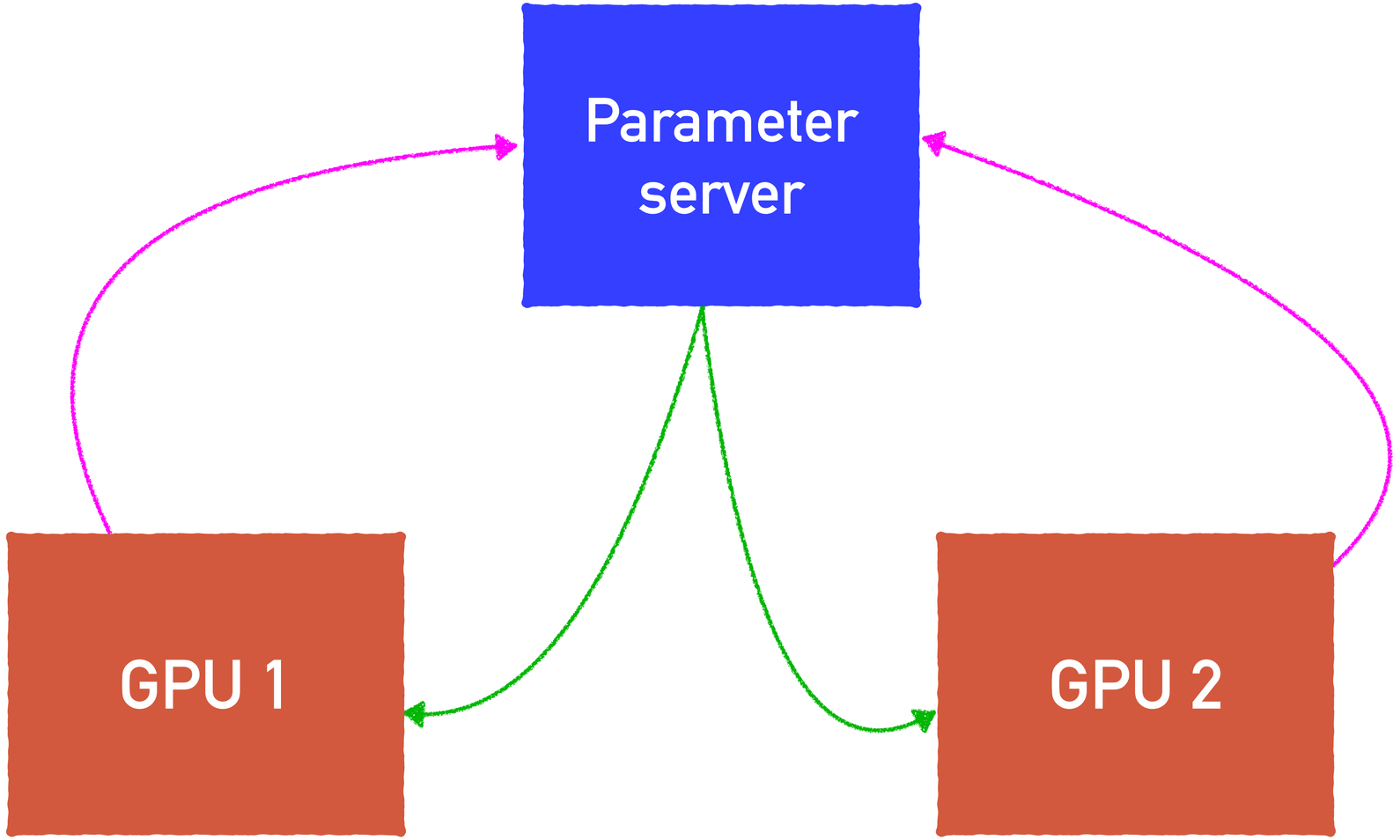
Theoretical convergence results

Empirical characterisation of neural net landscape

Imagenet results



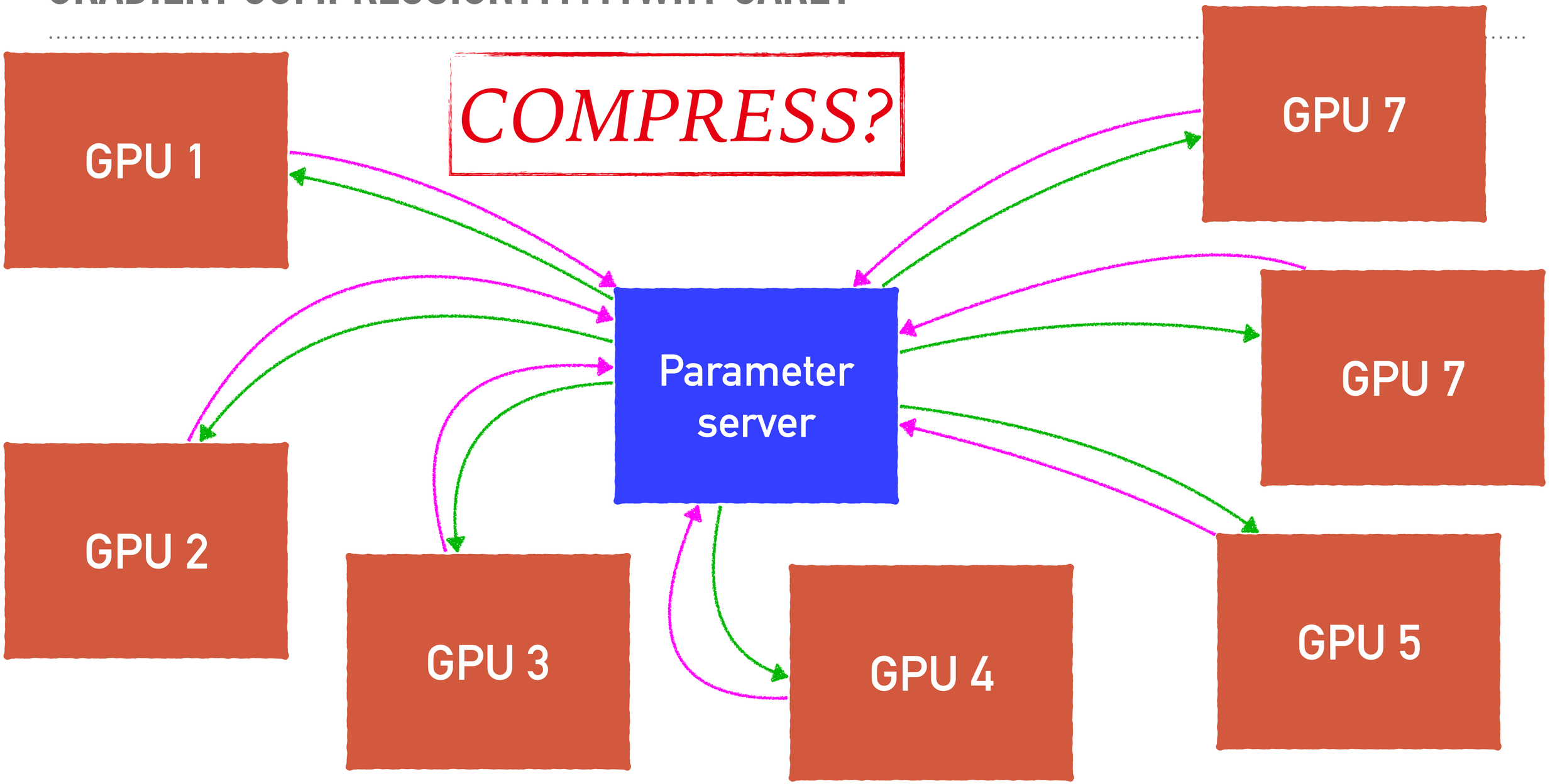
GRADIENT COMPRESSION……WHY CARE?



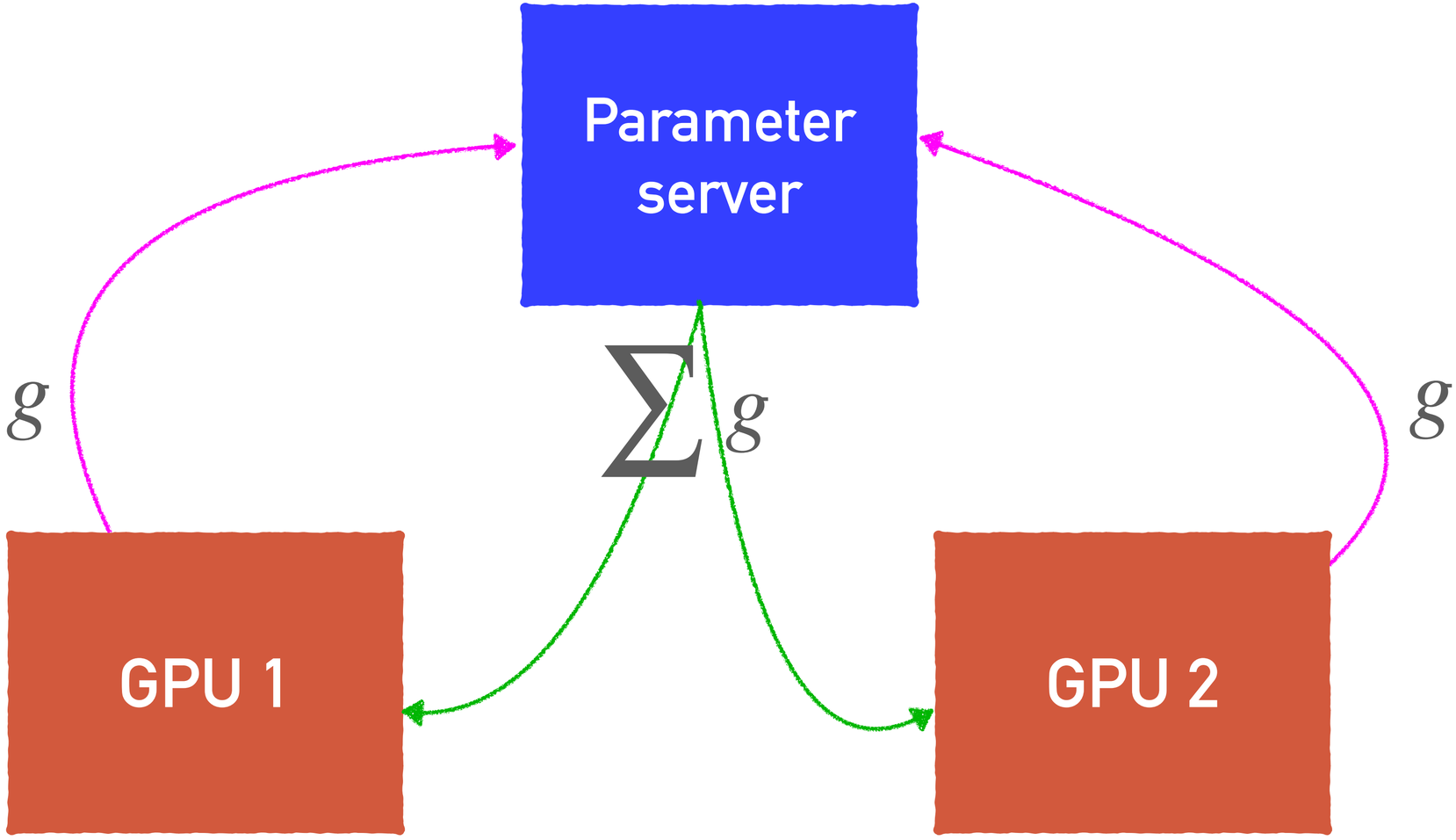
With 1/2 data

With 1/2 data

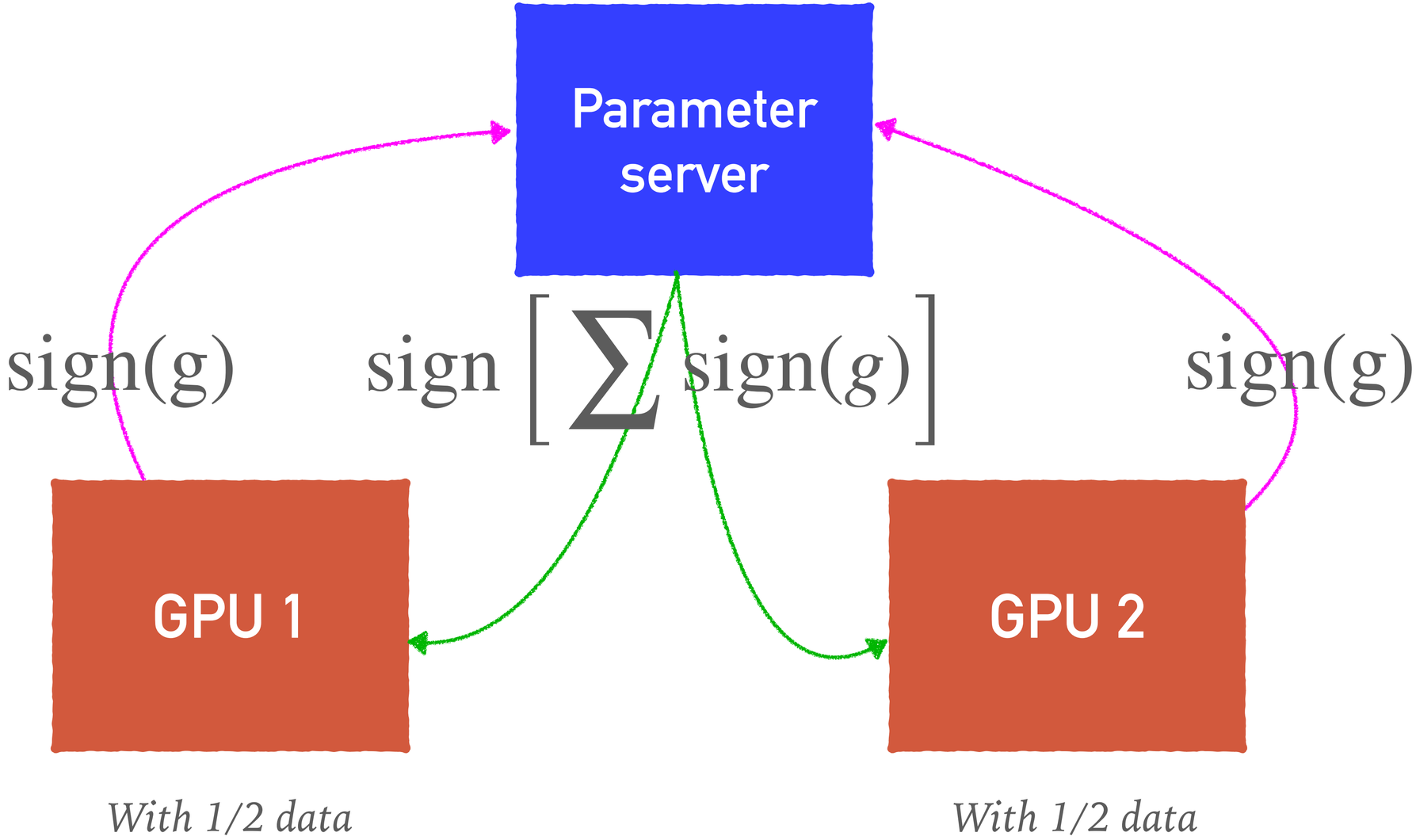
GRADIENT COMPRESSION……WHY CARE?



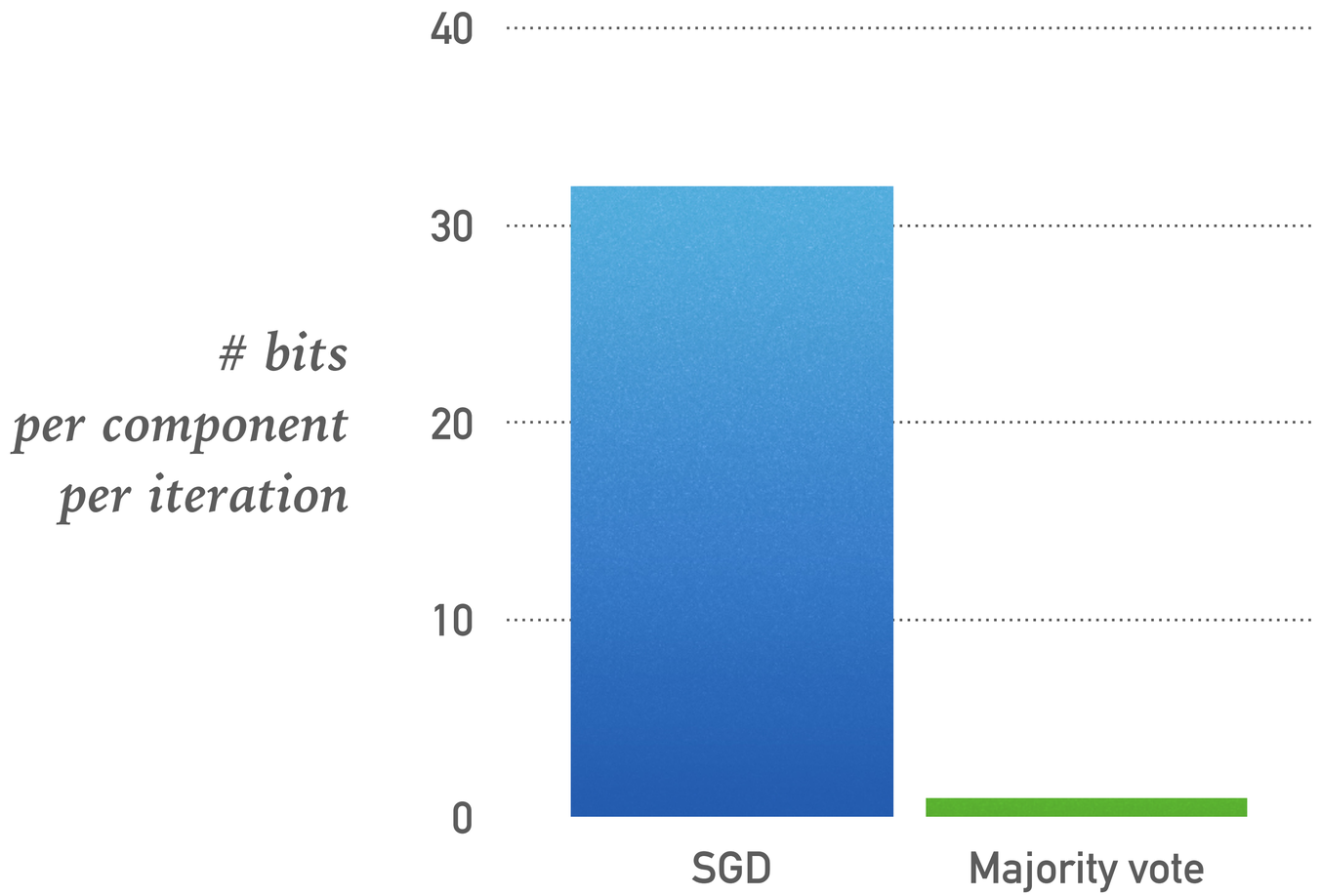
DISTRIBUTED SGD



SIGNSGD WITH MAJORITY VOTE



COMPRESSION SAVINGS OF MAJORITY VOTE



SIGNSGD IS A SPECIAL CASE OF ADAM

signSGD $\operatorname{sign}(g_k) = \dfrac{g_k}{\sqrt{g_k^2}}$

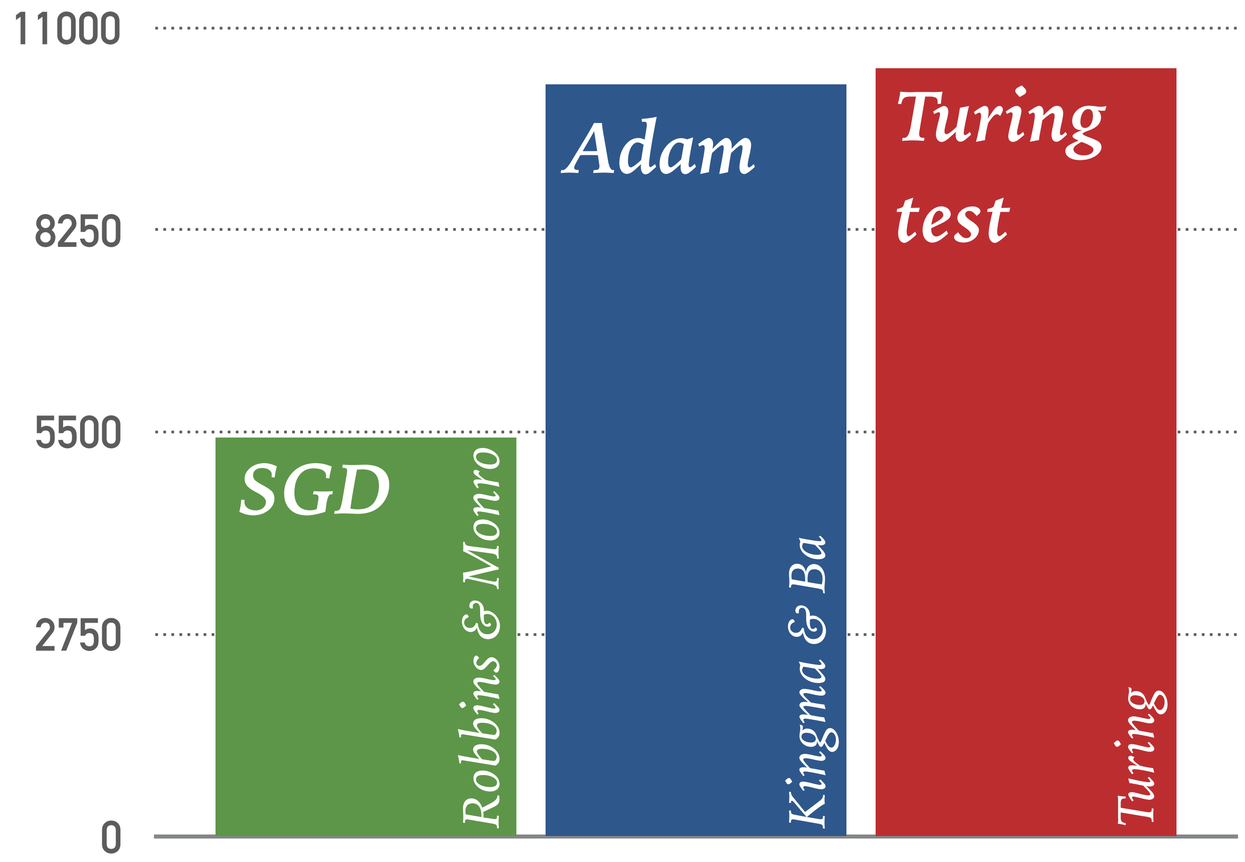
Adam $\dfrac{g_k + \beta g_{k-1} + \beta^2 g_{k-2} + \dots}{\sqrt{g_k^2 + \beta g_{k-1}^2 + \beta^2 g_{k-2}^2 + \dots}}$

Signum $\operatorname{sign}\left(g_k + \beta g_{k-1} + \beta^2 g_{k-2} + \dots\right)$

(*Sign momentum*)

ADAM.....WHY CARE?

# of Google scholar citations



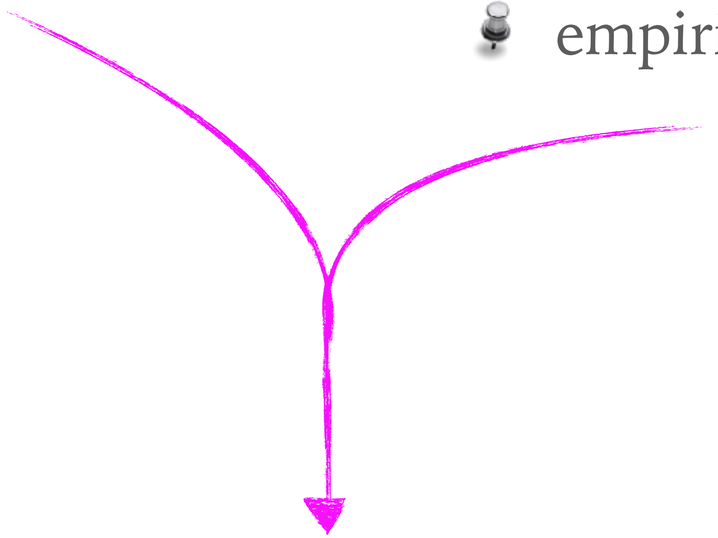
UNIFYING ADAPTIVE GRADIENT METHODS + COMPRESSION

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?

Need to build a theory

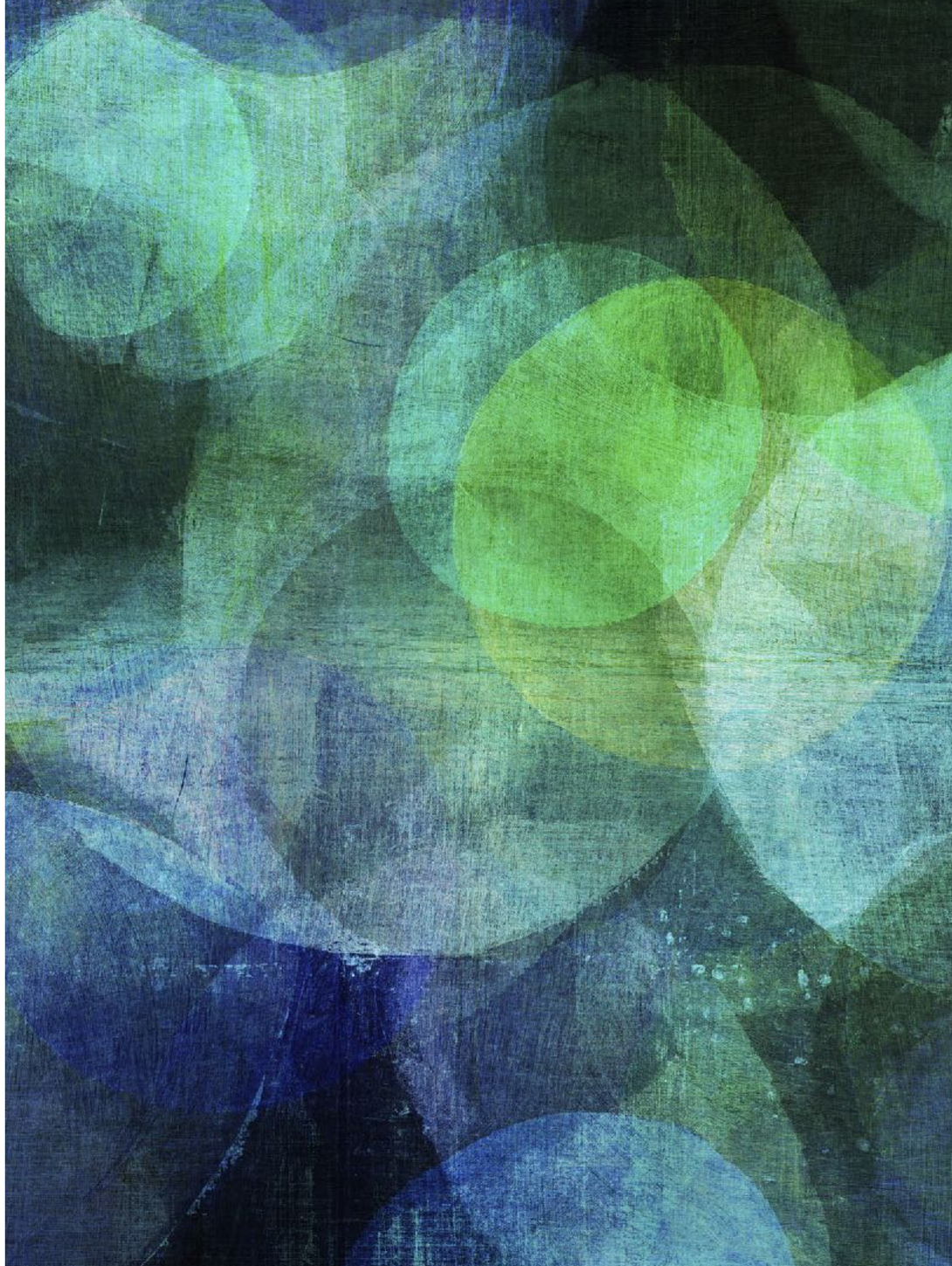
STRUCTURE

Why care about signSGD?

Theoretical convergence results

Empirical characterisation of neural net landscape

Imagenet results



DOES 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 **signSGD & 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

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

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

SINGLE WORKER RESULTS

Assumptions

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

Define

- Number of iterations K
- Number of backpropagations N

SGD gets rate

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

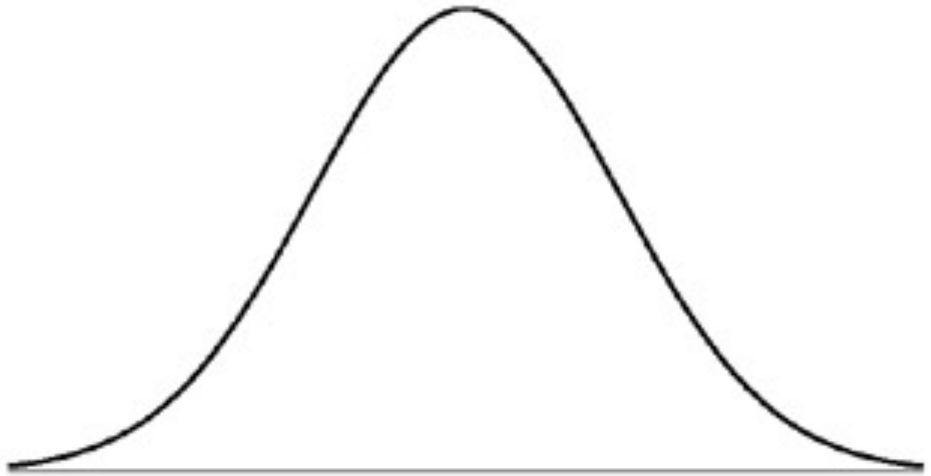
signSGD gets rate

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

MULTI WORKER RESULTS *with M workers*

Assumptions

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



SGD gets rate

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

if gradient noise is unimodal symmetric

majority vote gets

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

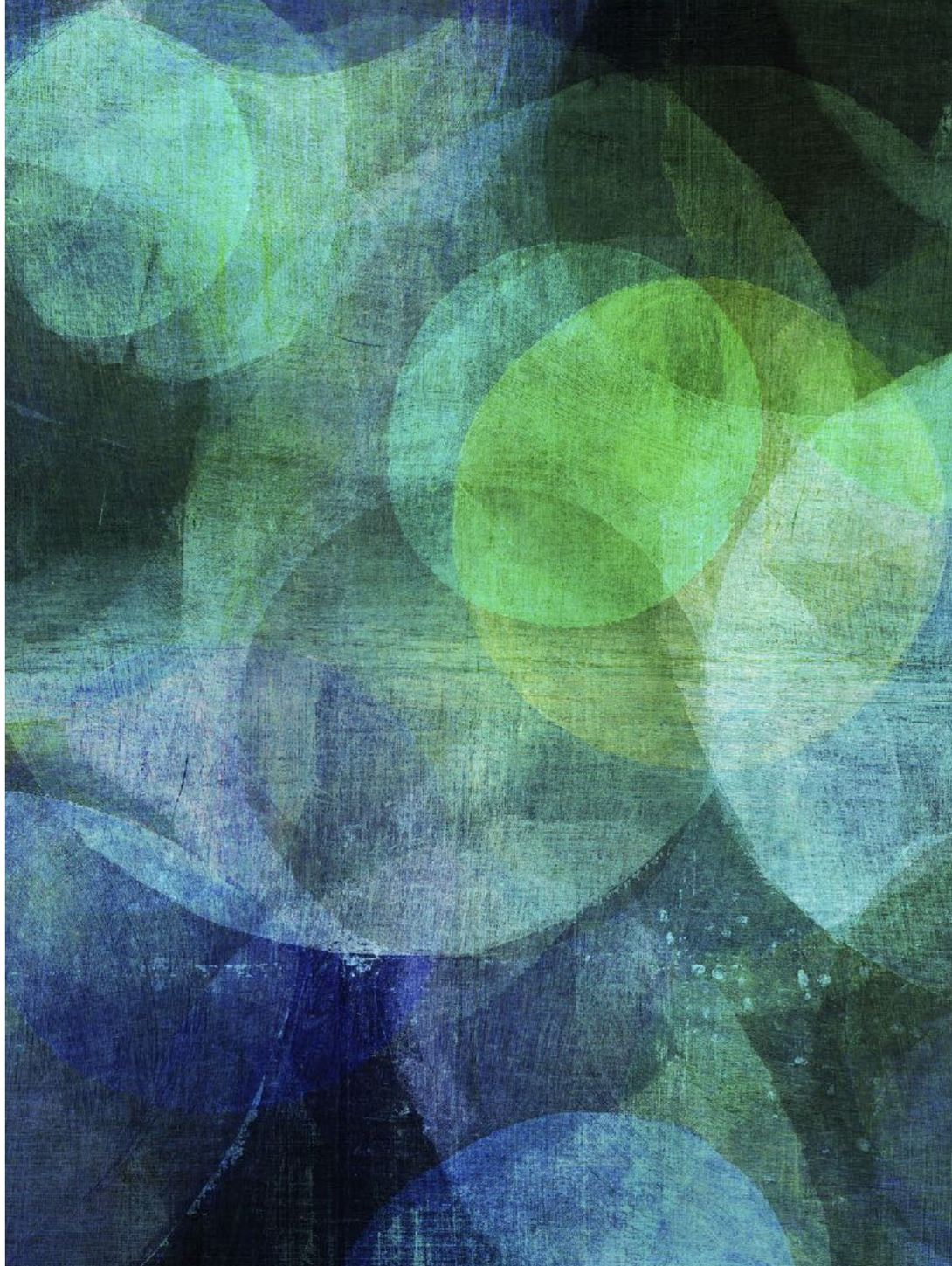
STRUCTURE

Why care about signSGD?

Theoretical convergence results

Empirical characterisation of neural net landscape

Imagenet results



CHARACTERISING THE DEEP LEARNING LANDSCAPE EMPIRICALLY

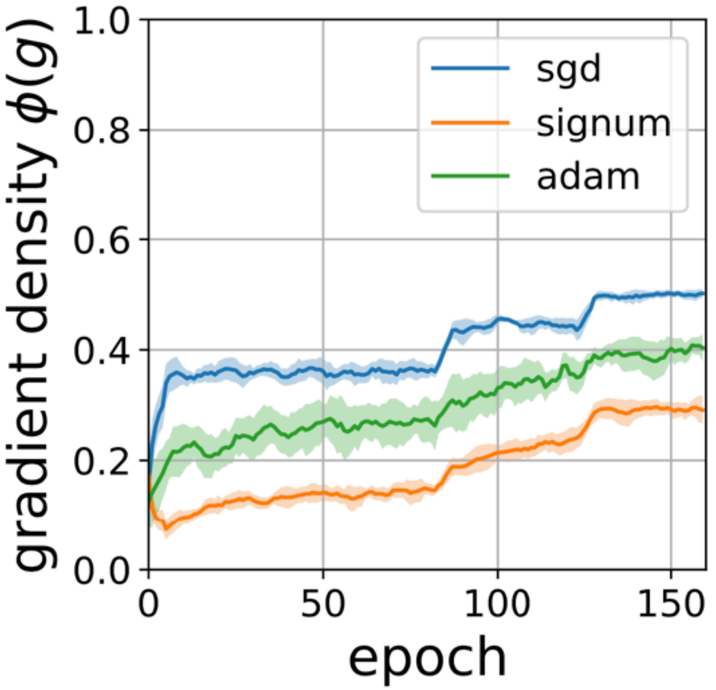
- **signSGD** cares about gradient density
- **majority vote** cares about noise symmetry

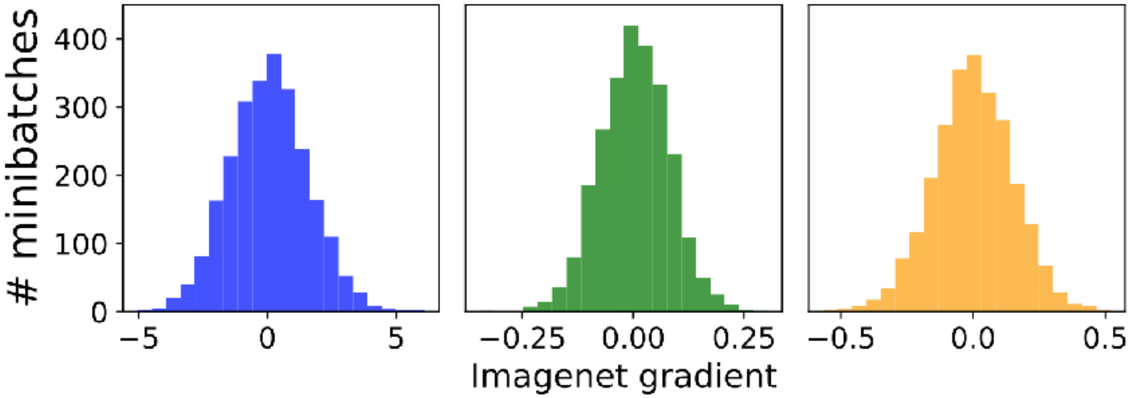
Natural measure of density

$$\phi(\vec{v}) = \frac{\|\vec{v}\|_1^2}{d\|\vec{v}\|_2^2}$$

=1 for fully dense v

≈0 for fully sparse v





For large enough mini-batch size,

reasonable by Central Limit Theorem.

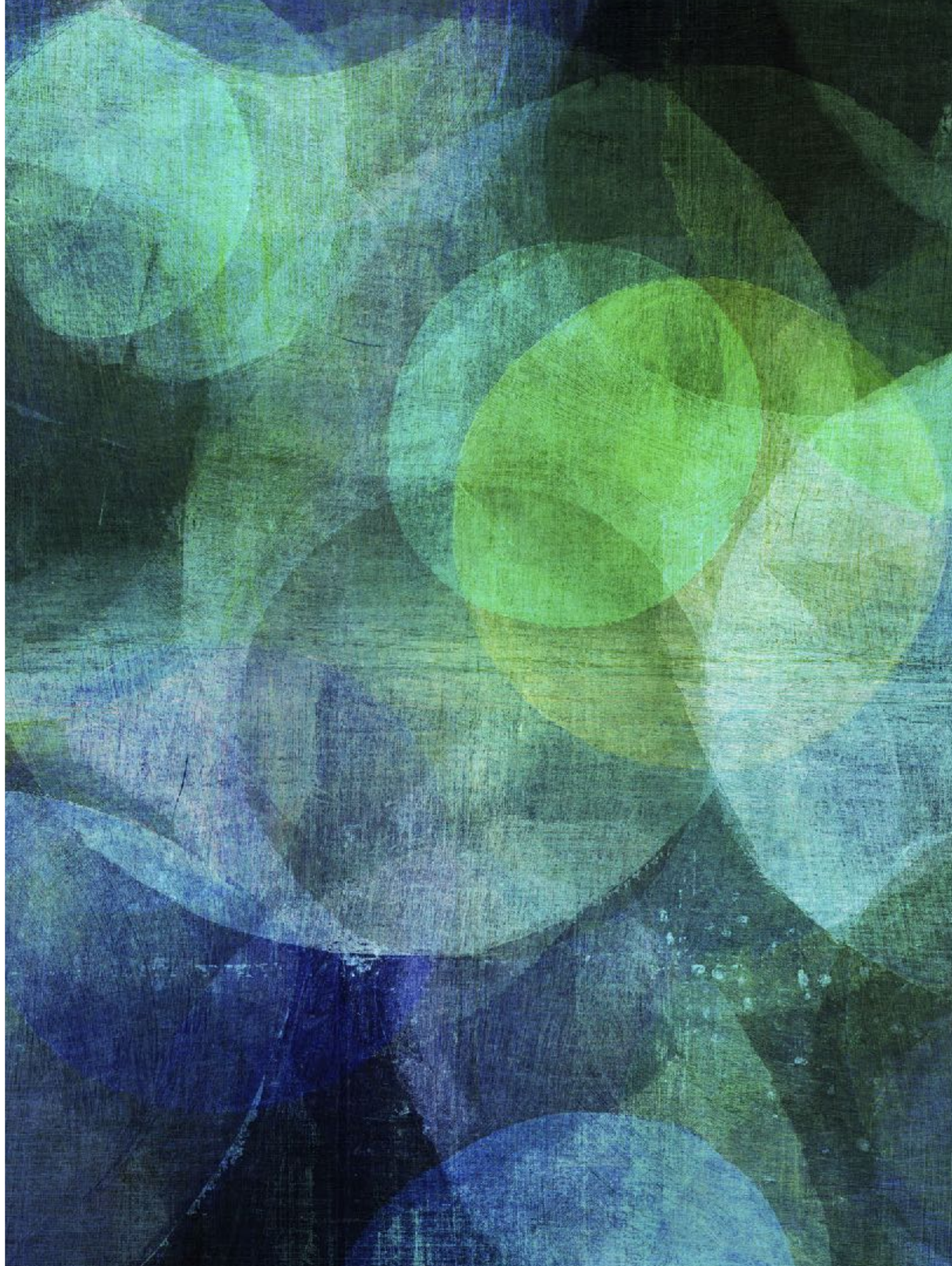
STRUCTURE

Why care about signSGD?

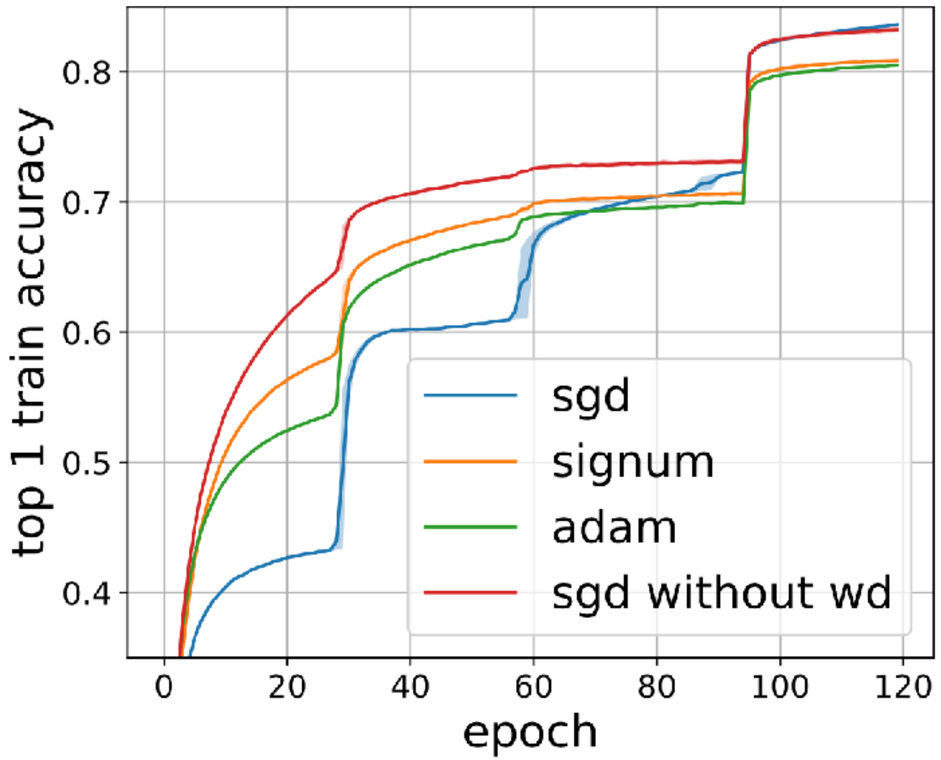
Theoretical convergence results

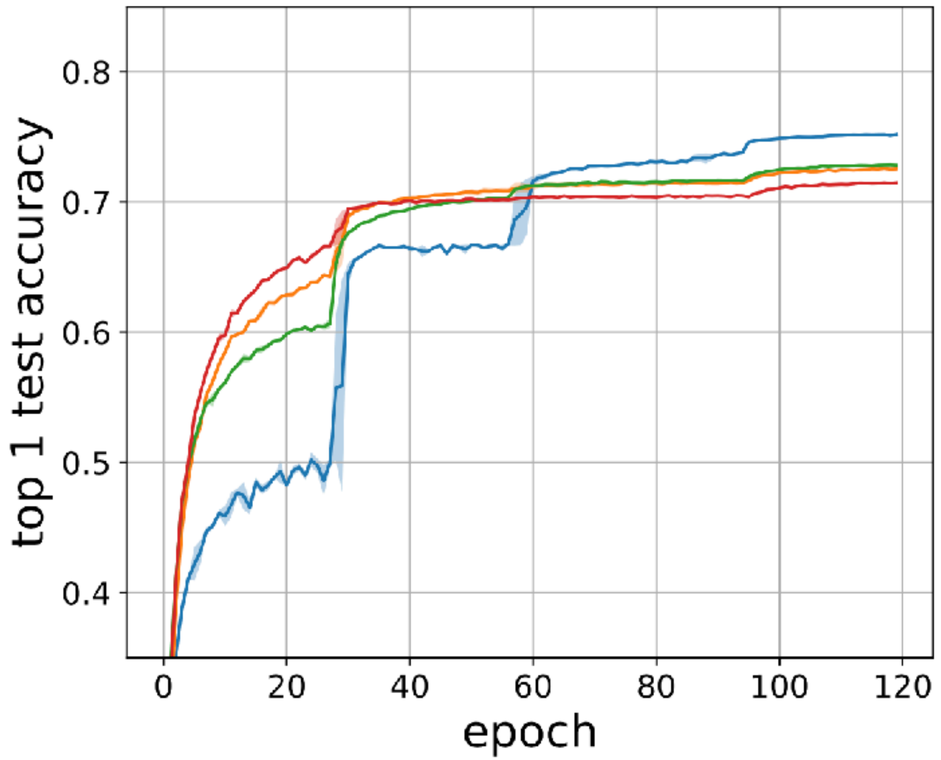
Empirical characterisation of neural net landscape

Imagenet results



SIGNUM IS COMPETITIVE ON IMAGENET





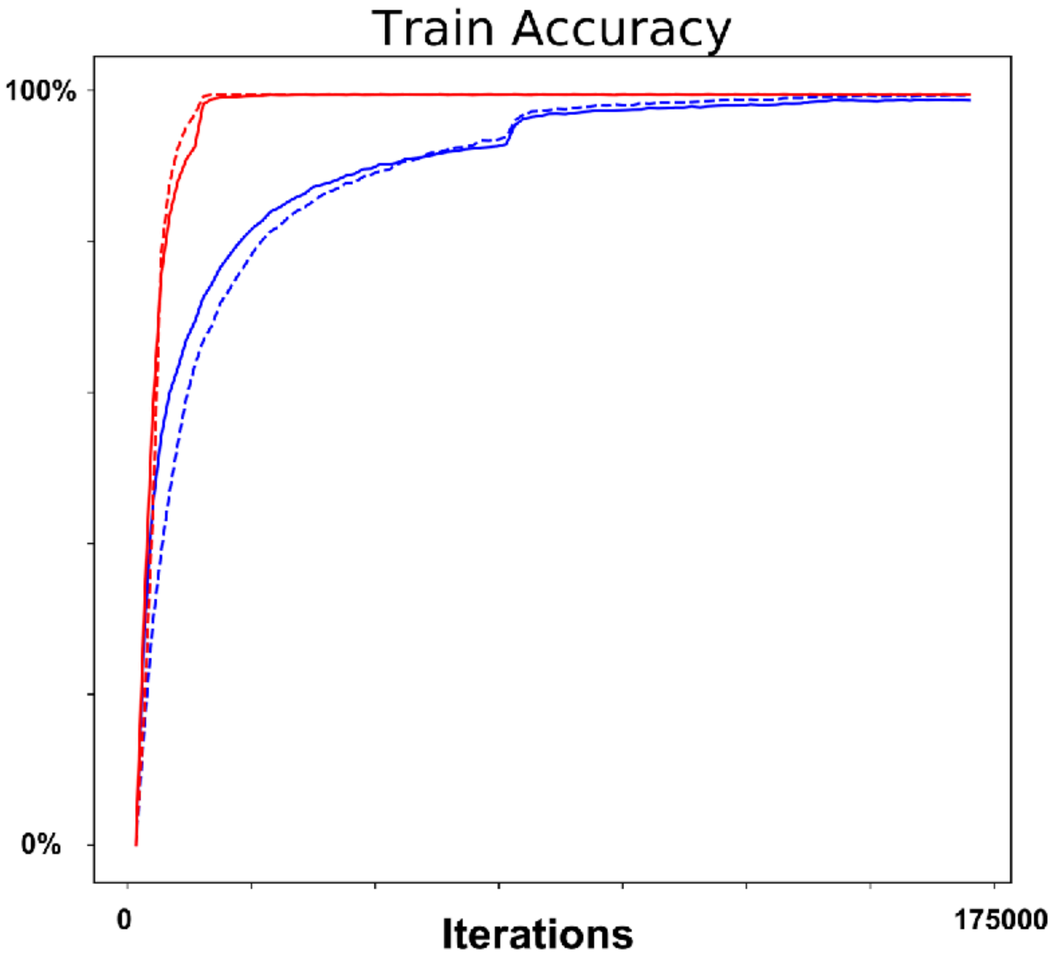
Performance very similar to Adam

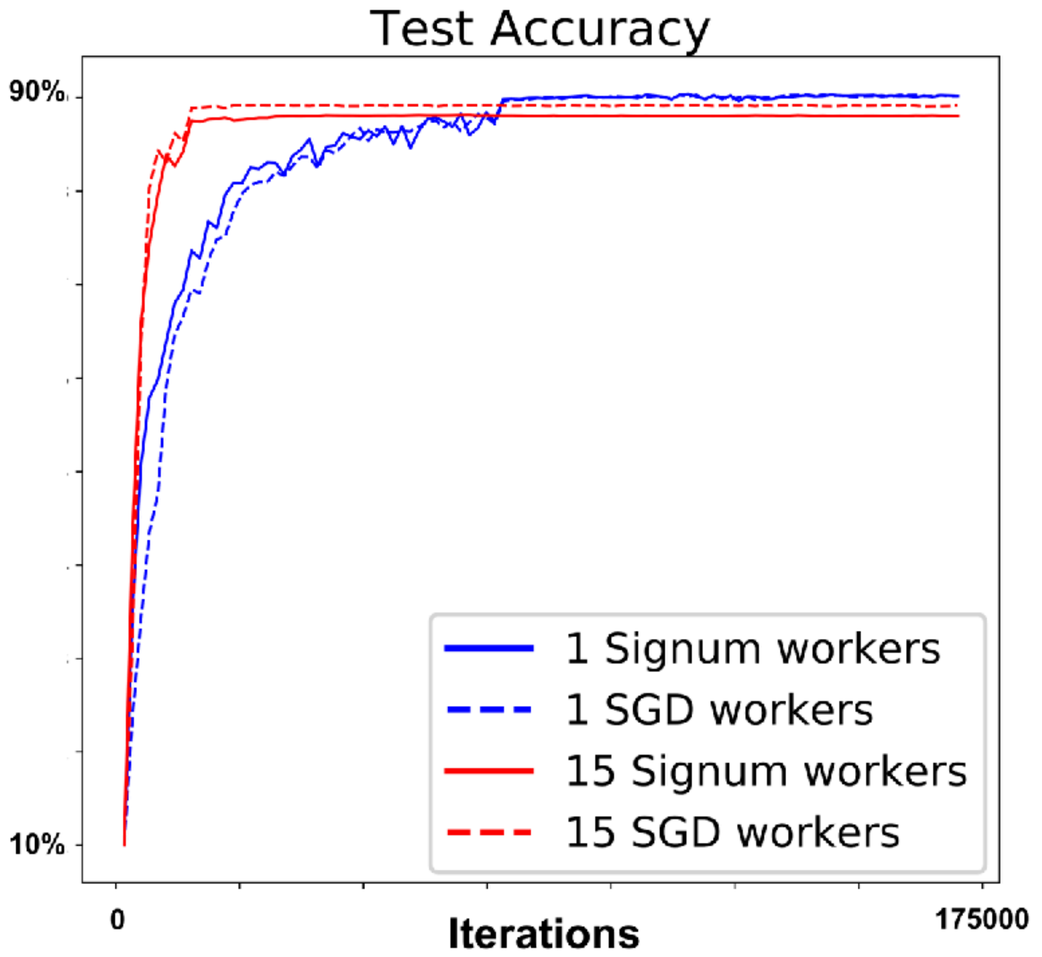
May want to switch to SGD towards end?

DOES MAJORITY VOTE WORK?

Jiawei Zhao
NUAA

Cifar-10, Resnet-18





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