

Some Success Stories in Bridging Theory and Practice In Optimization

Anima Anandkumar

Bren Professor at Caltech

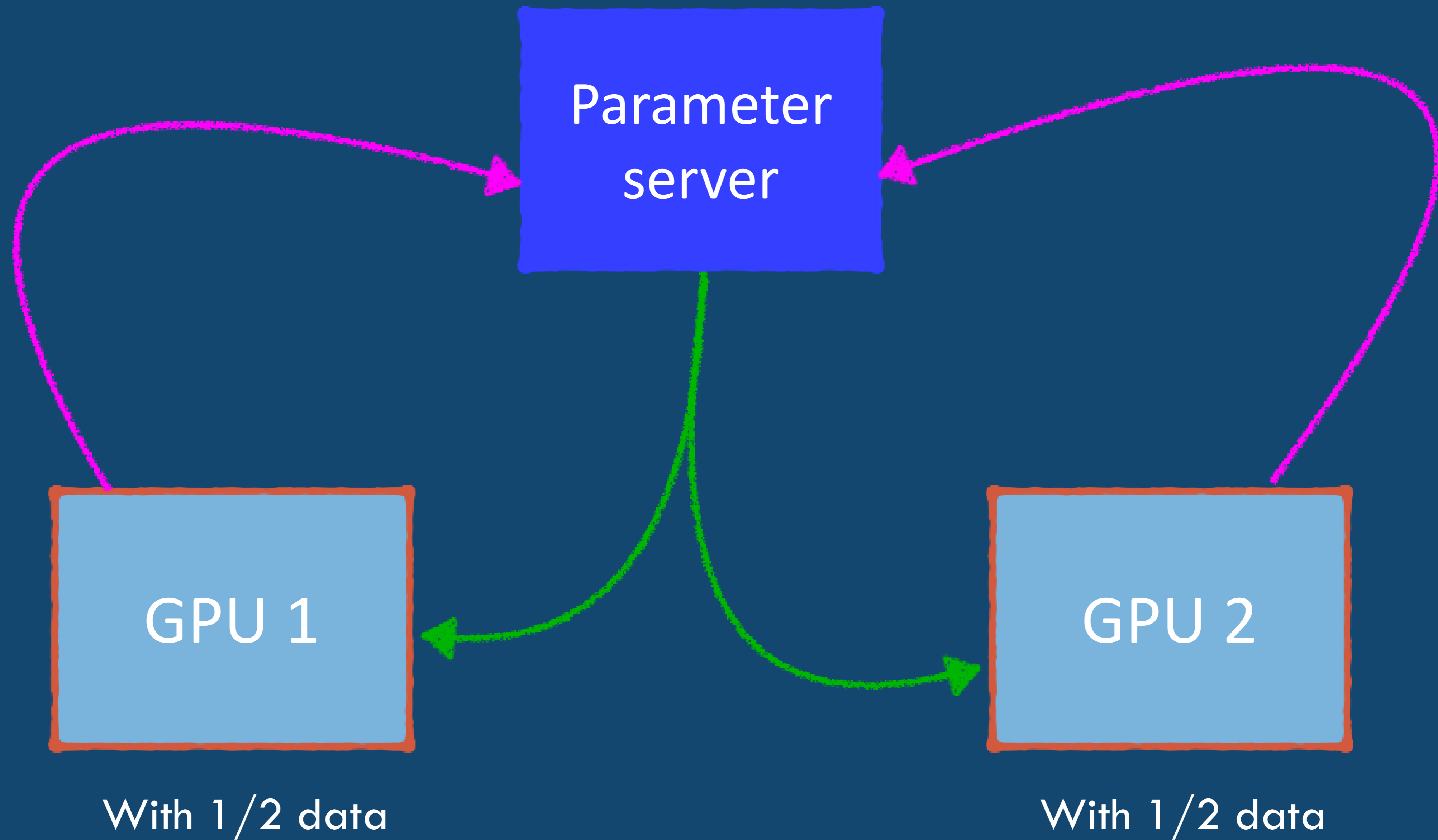
Director of ML Research at NVIDIA

The background is a solid teal color. In the four corners, there are decorative white line-art elements that resemble circuit traces or neural network connections, with small circles at the end of the lines.

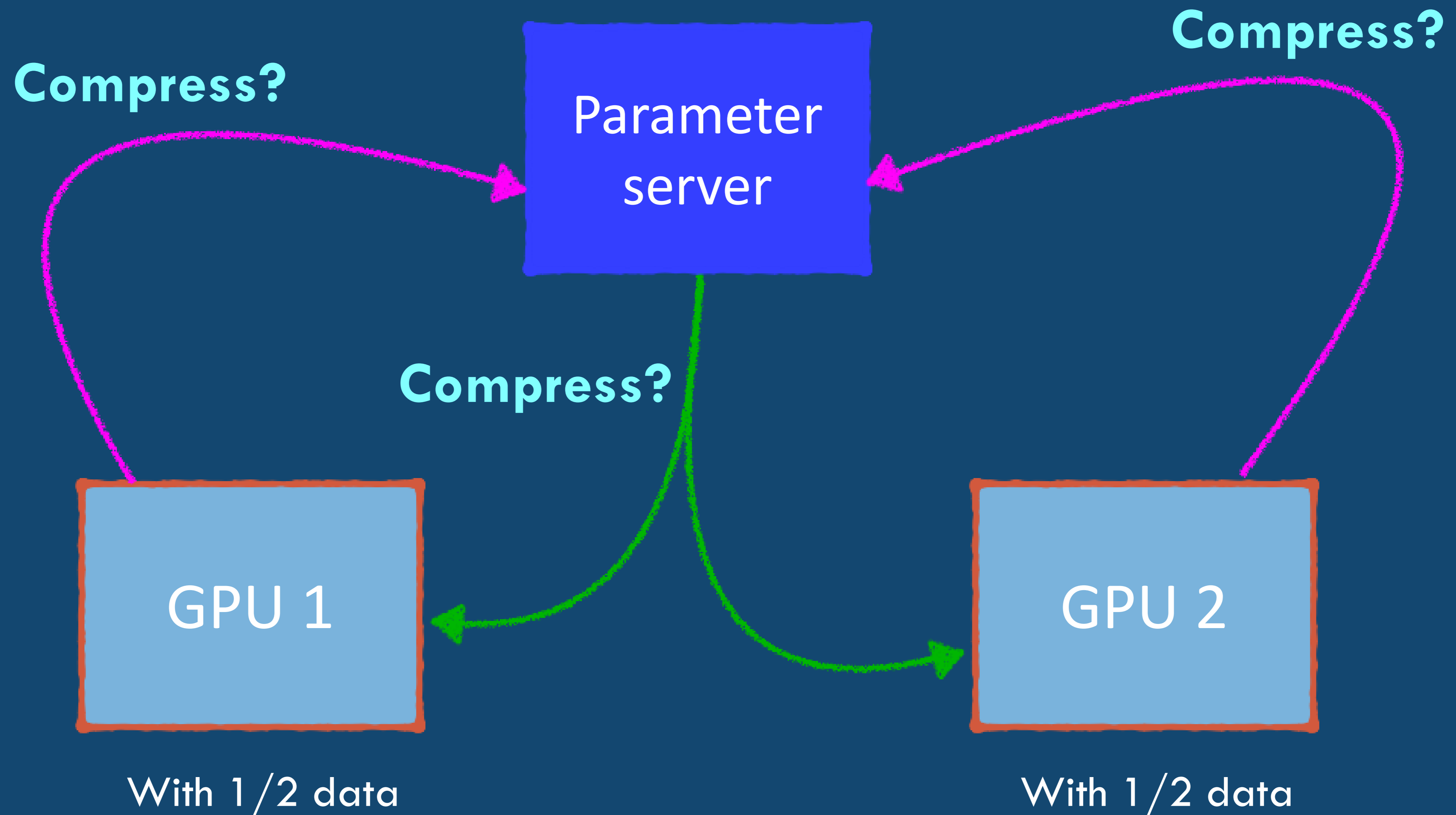
SIGNSGD: COMPRESSED OPTIMIZATION FOR NON-CONVEX PROBLEMS

JEREMY BERNSTEIN, JIAWEI ZHAO, KAMYAR AZZIZADENESHELI, YU-XIANG
WANG, ANIMA ANANDKUMAR

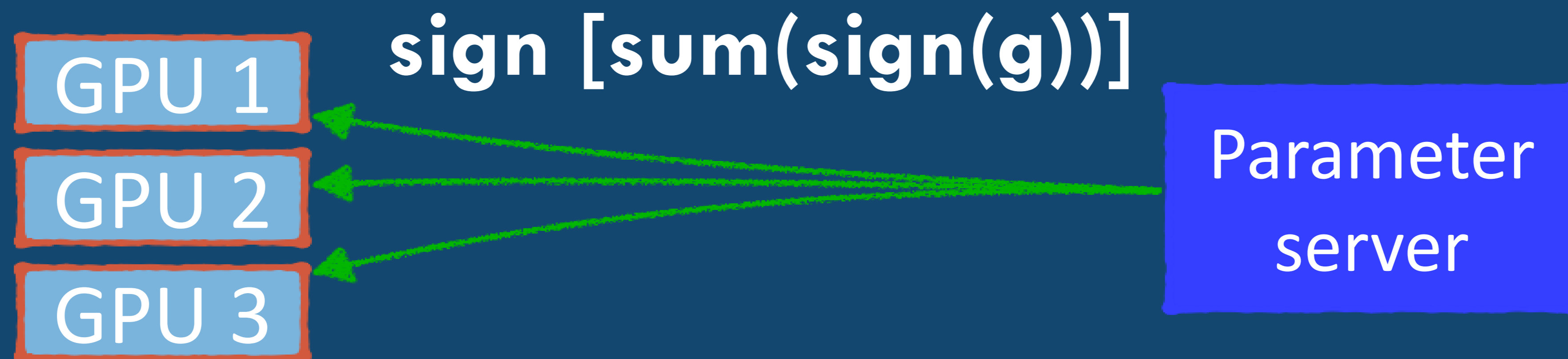
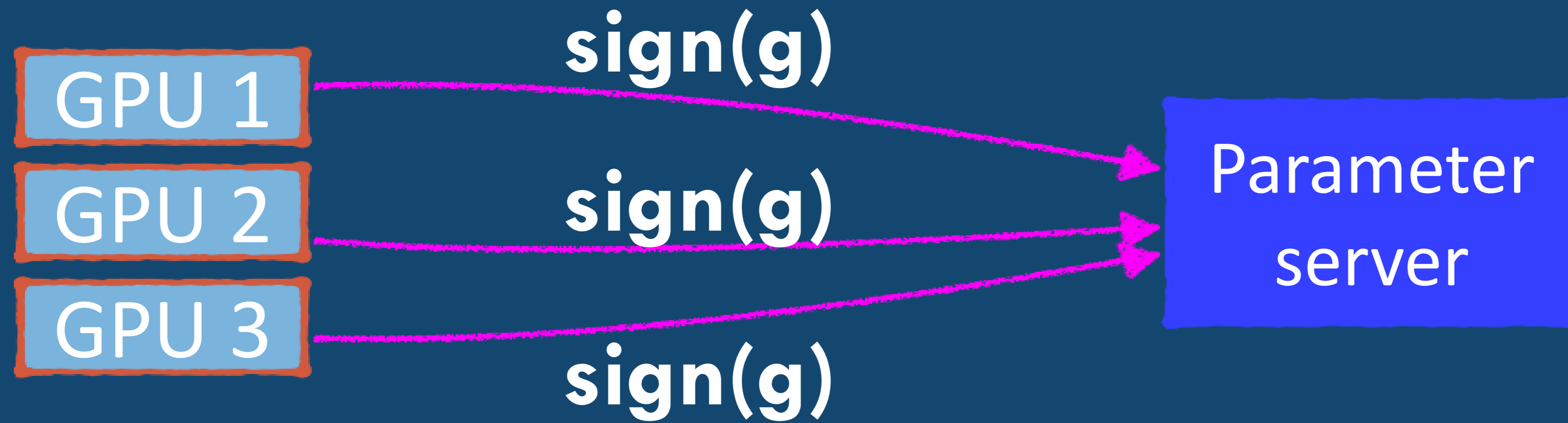
DISTRIBUTED TRAINING INVOLVES COMPUTATION & COMMUNICATION



DISTRIBUTED TRAINING INVOLVES COMPUTATION & COMMUNICATION



DISTRIBUTED TRAINING BY MAJORITY VOTE



VARIANTS OF SGD DISTORT THE GRADIENT IN DIFFERENT WAYS

SGD

$$g_k$$

Signum

$$\text{sign}(g_k + \beta g_{k-1} + \beta^2 g_{k-2} + \dots)$$

Adam

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

LARGE-BATCH ANALYSIS

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

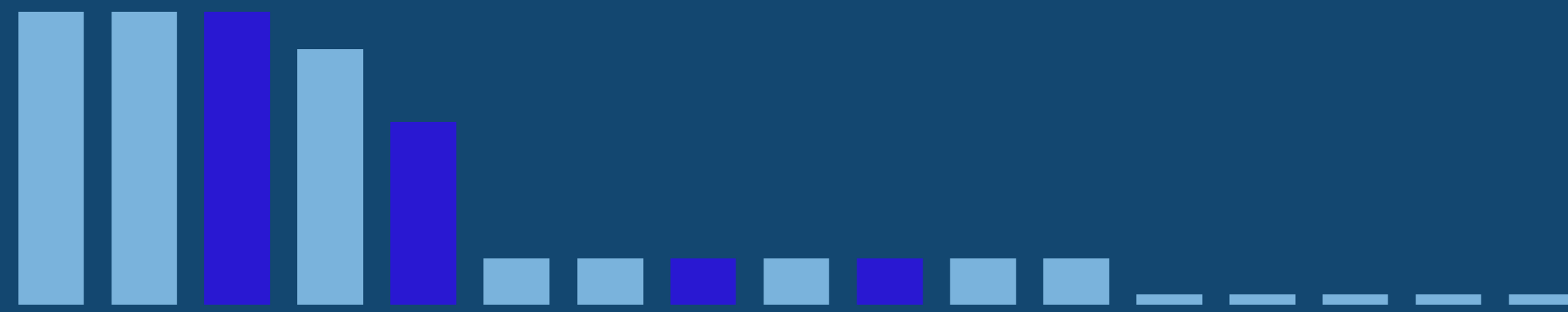
$$\mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \|g_k\|_2^2 \right] \leq \frac{1}{\sqrt{N}} \left[2\|\vec{L}\|_\infty (f_0 - f_*) + \|\vec{\sigma}\|_2^2 \right]$$

signSGD gets rate

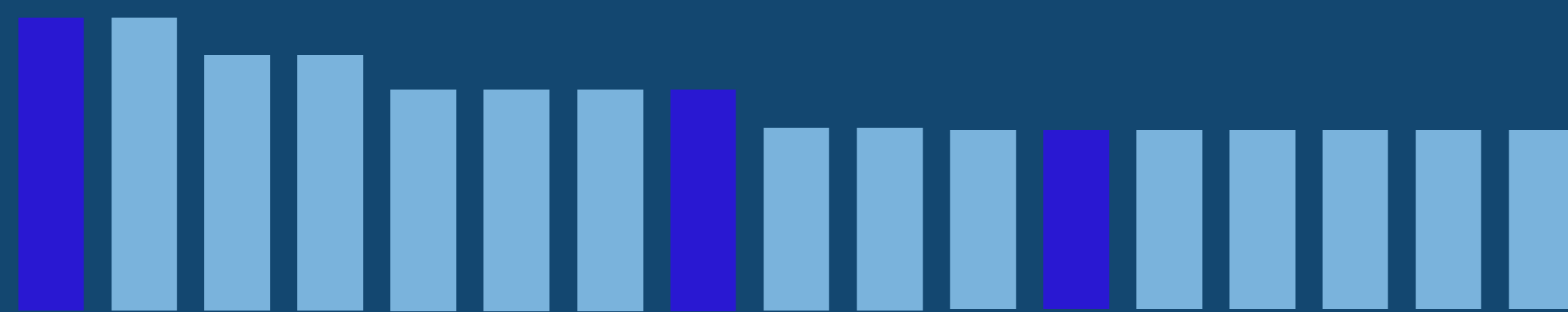
$$\mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \sqrt{d} \|\vec{g}_k\|_2 \right]^2 \leq \frac{1}{\sqrt{N}} \left[\sqrt{d} \sqrt{\|\vec{L}\|_\infty} \left(f_0 - f_* + \frac{1}{2} \right) + 2\sqrt{d} \|\vec{\sigma}\|_2 \right]^2$$

VECTOR DENSITY & ITS RELEVANCE IN DEEP LEARNING

A sparse vector



A dense vector

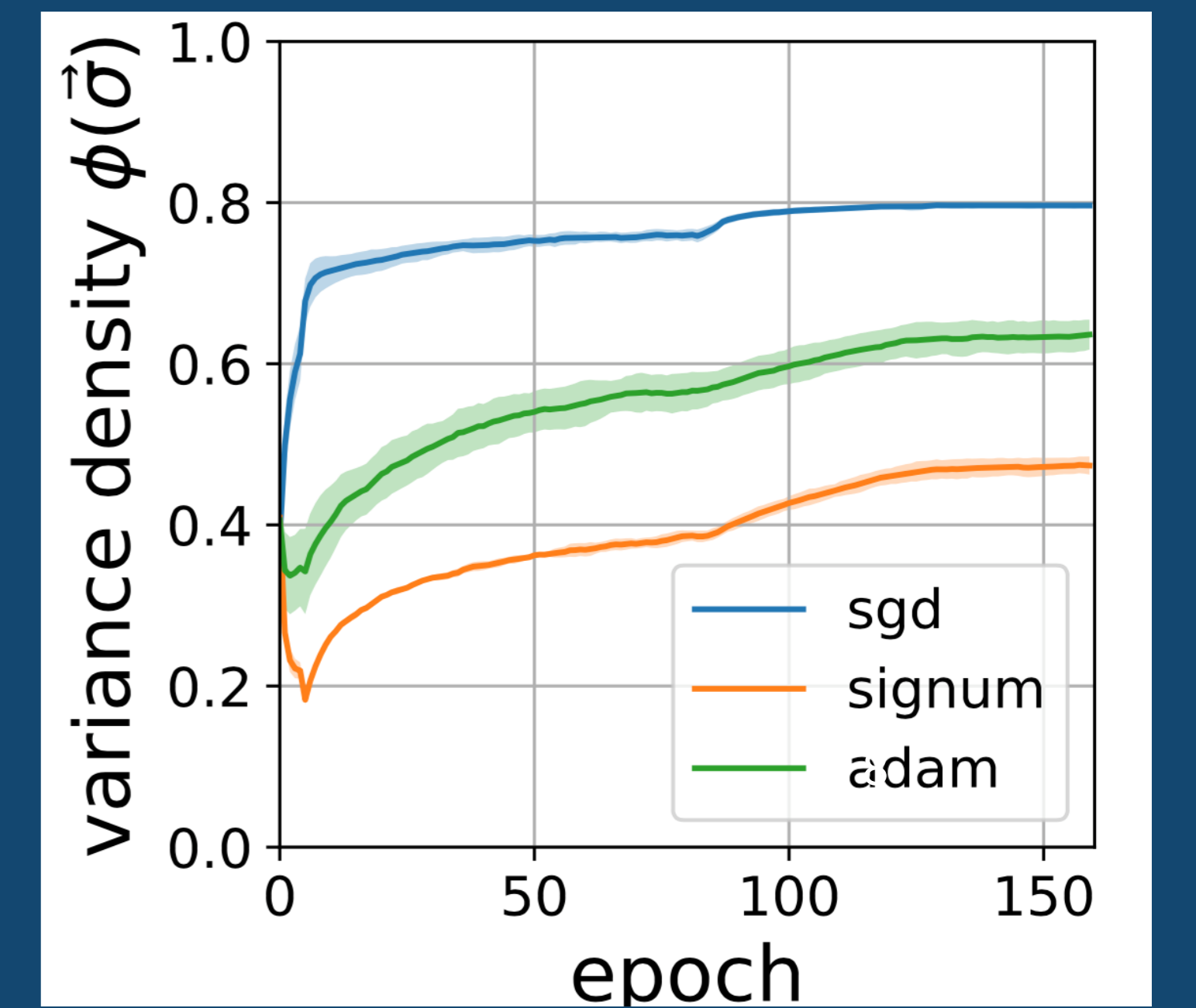
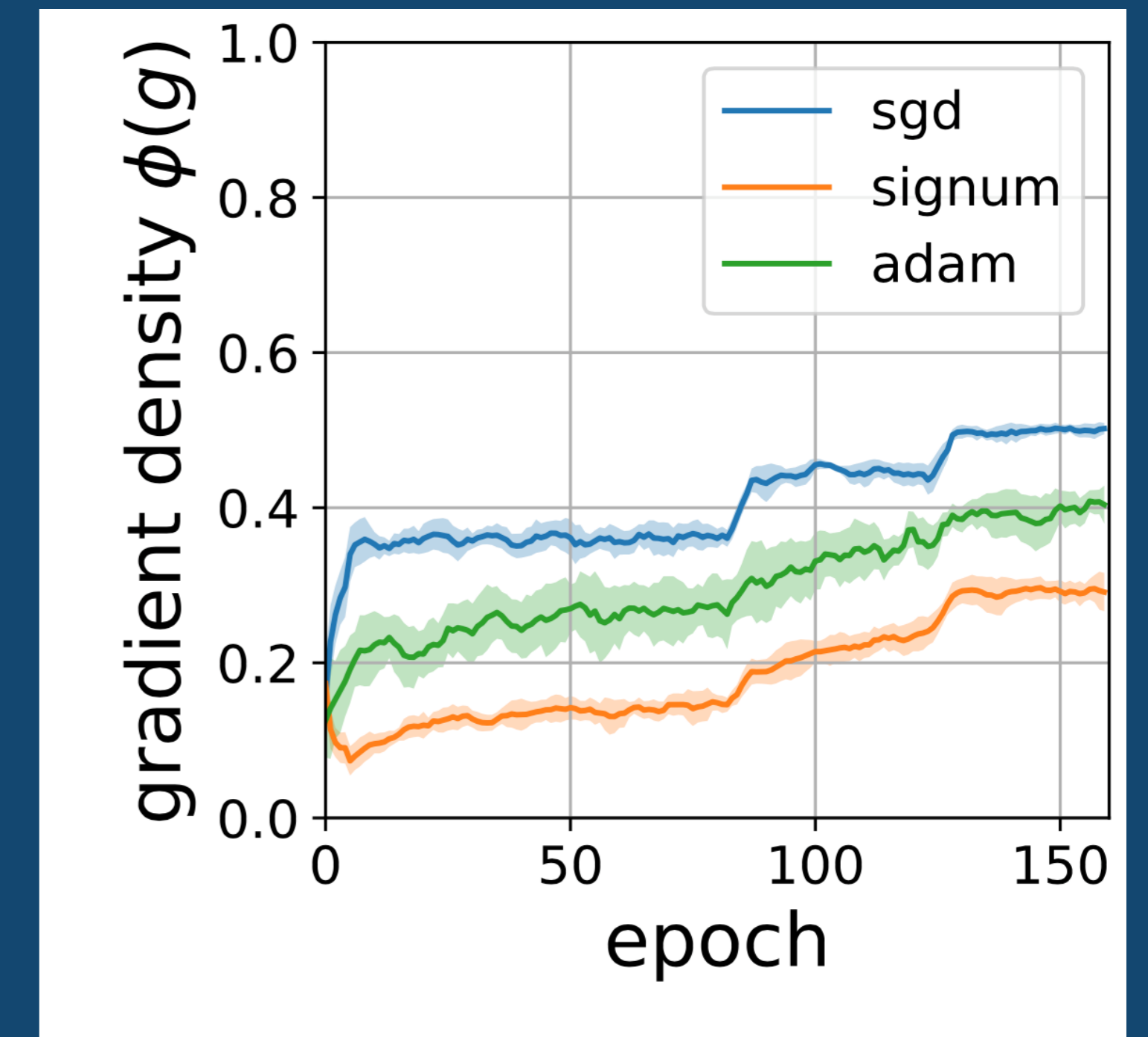
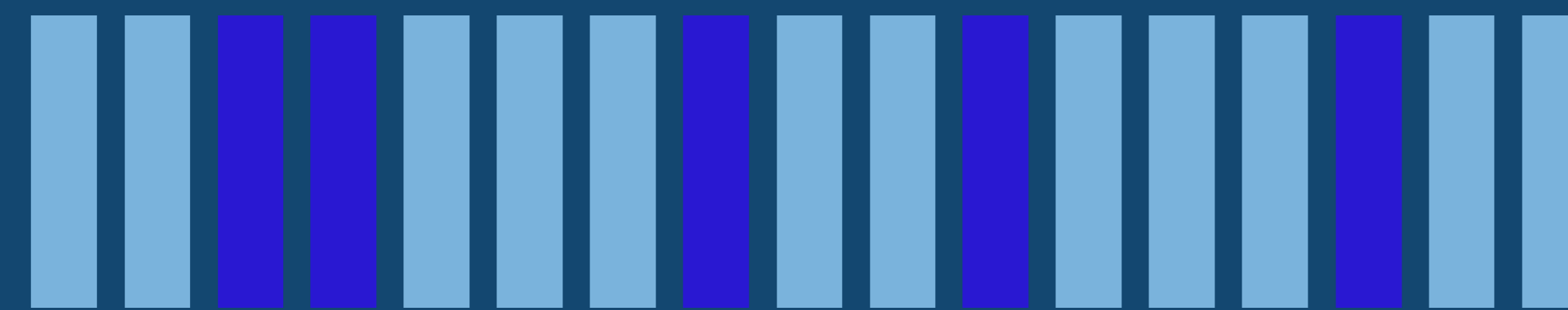


Natural measure of density

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

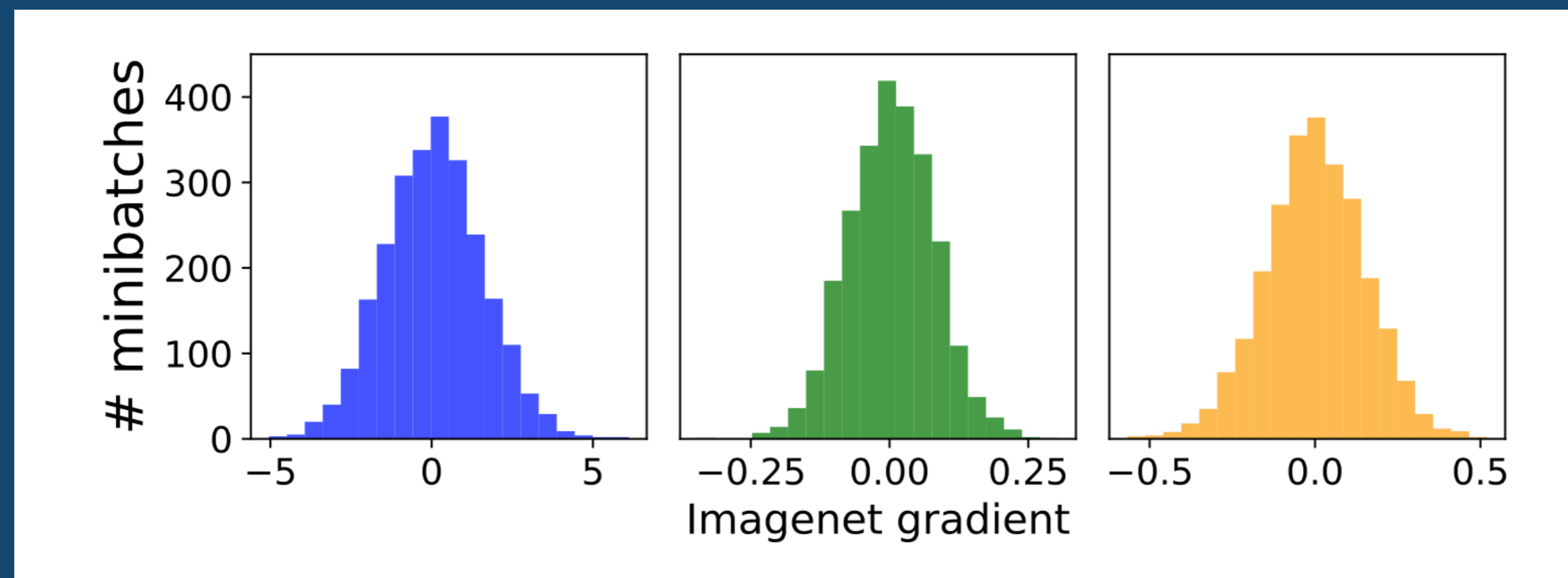
=1 for fully dense v
 ≈ 0 for fully sparse v

Fully dense vector.....a sign vector



DISTRIBUTED SIGNSGD: MAJORITY VOTE THEORY

If gradients are unimodal and symmetric...



...reasonable by central limit theorem...

...majority vote with M
workers converges at rate:

$$\mathbb{E} \left[\min_{0 \leq k \leq K-1} \|g_k\|_1 \right]^2$$

Same variance
reduction as SGD

$$\leq \frac{1}{\sqrt{N}} \left[\sqrt{\|\vec{L}\|_1} \left(f_0 - f_* + \frac{1}{2} \right) + \frac{2}{\sqrt{M}} \|\vec{\sigma}\|_1 \right]^2$$

MINI-BATCH ANALYSIS

Under symmetric noise assumption:

Theorem 1 (Non-convex convergence rate of small-batch SIGNSGD). *Run the following algorithm for K iterations under Assumptions 1 to 4: $x_{k+1} = x_k - \eta \text{sign}(\tilde{g}_k)$. Set the learning rate, η , and mini-batch size, n , as*

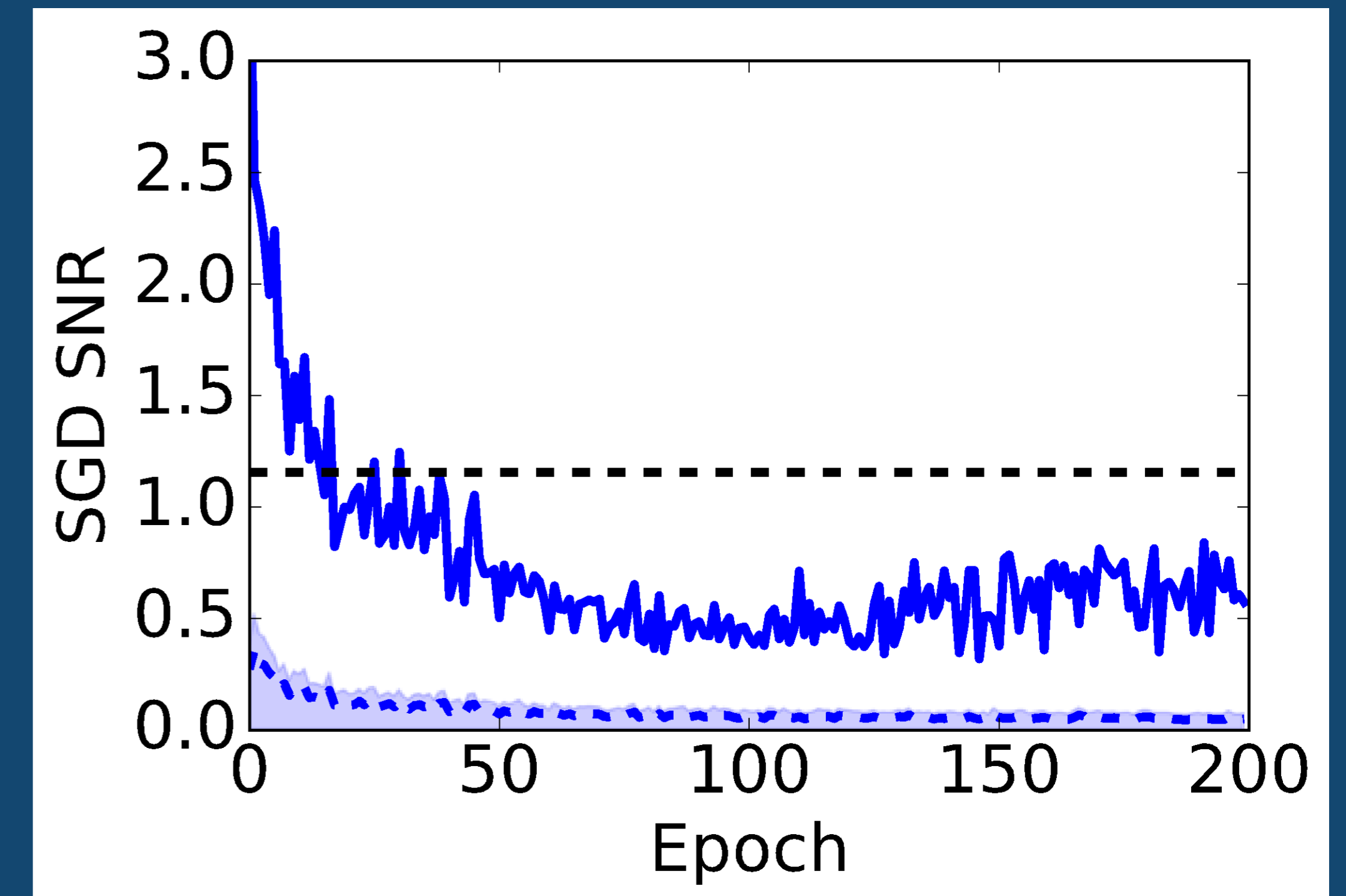
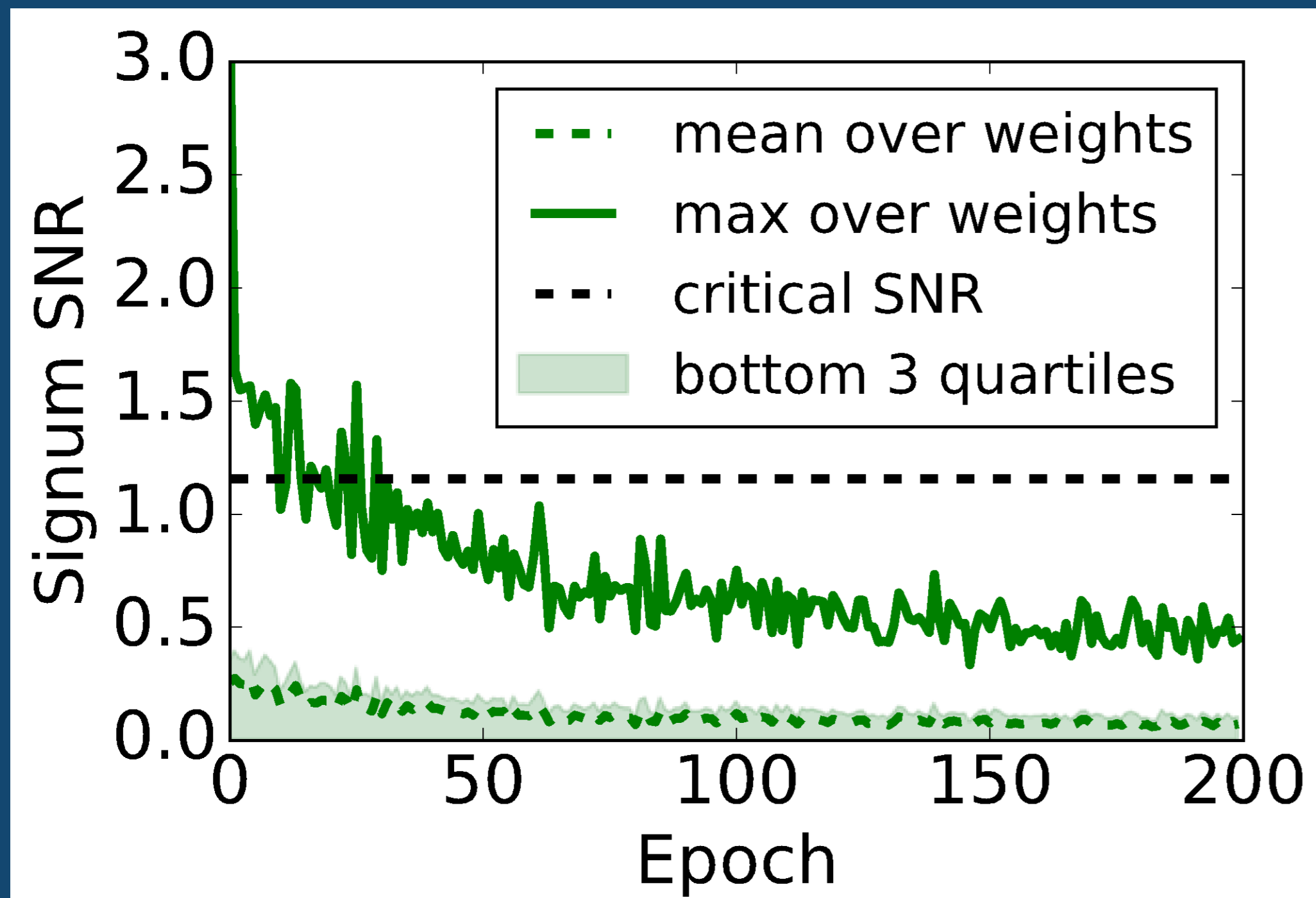
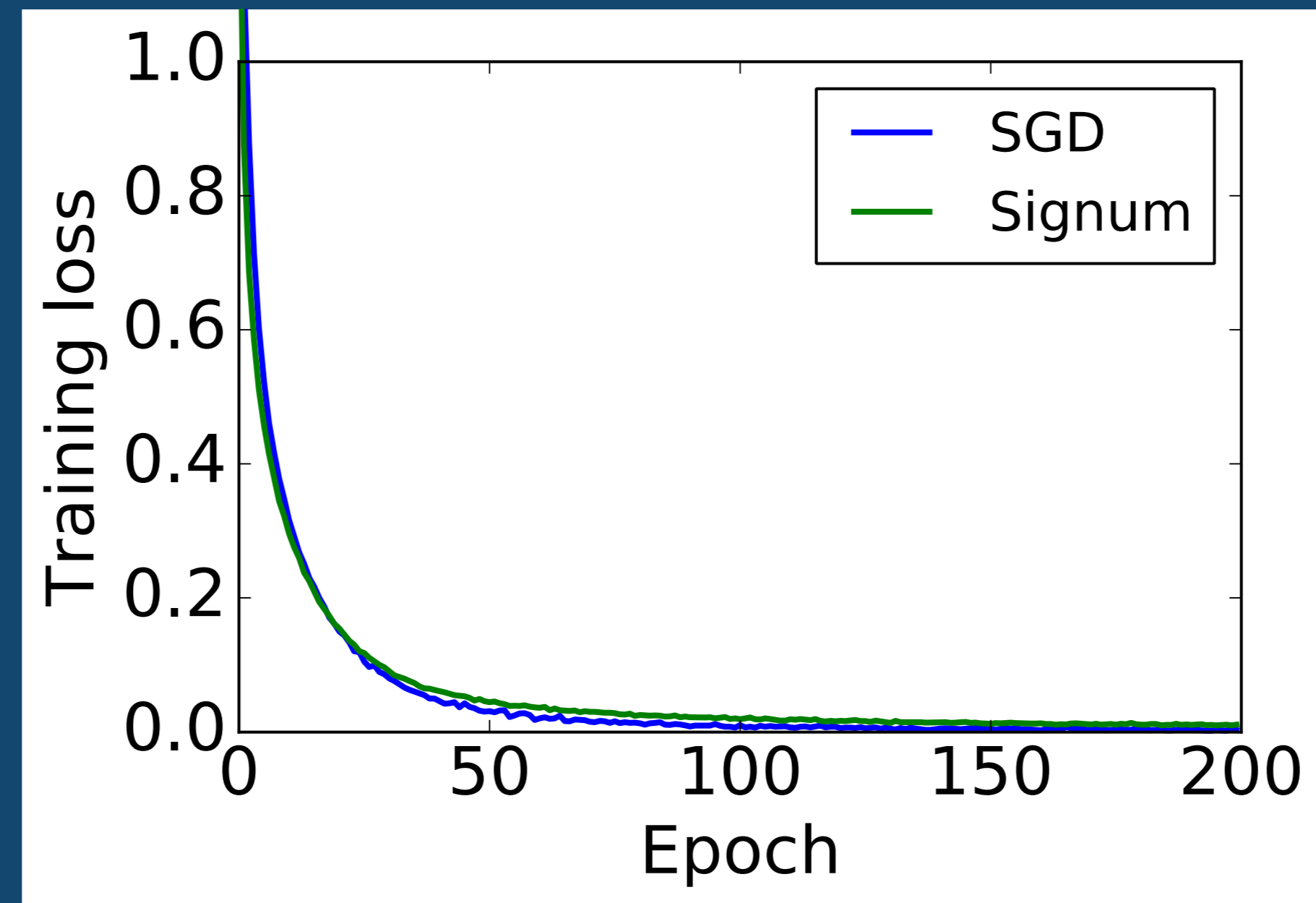
$$\eta = \sqrt{\frac{f_0 - f_*}{\|\vec{L}\|_1 K}}, \quad n = 1.$$

Let H_k be the set of gradient components at step k with large signal-to-noise ratio $S_i := \frac{|g_{k,i}|}{\sigma_i}$, i.e. $H_k := \left\{ i \mid S_i > \frac{2}{\sqrt{3}} \right\}$. We refer to $\frac{2}{\sqrt{3}}$ as the ‘critical SNR’. Then we have

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

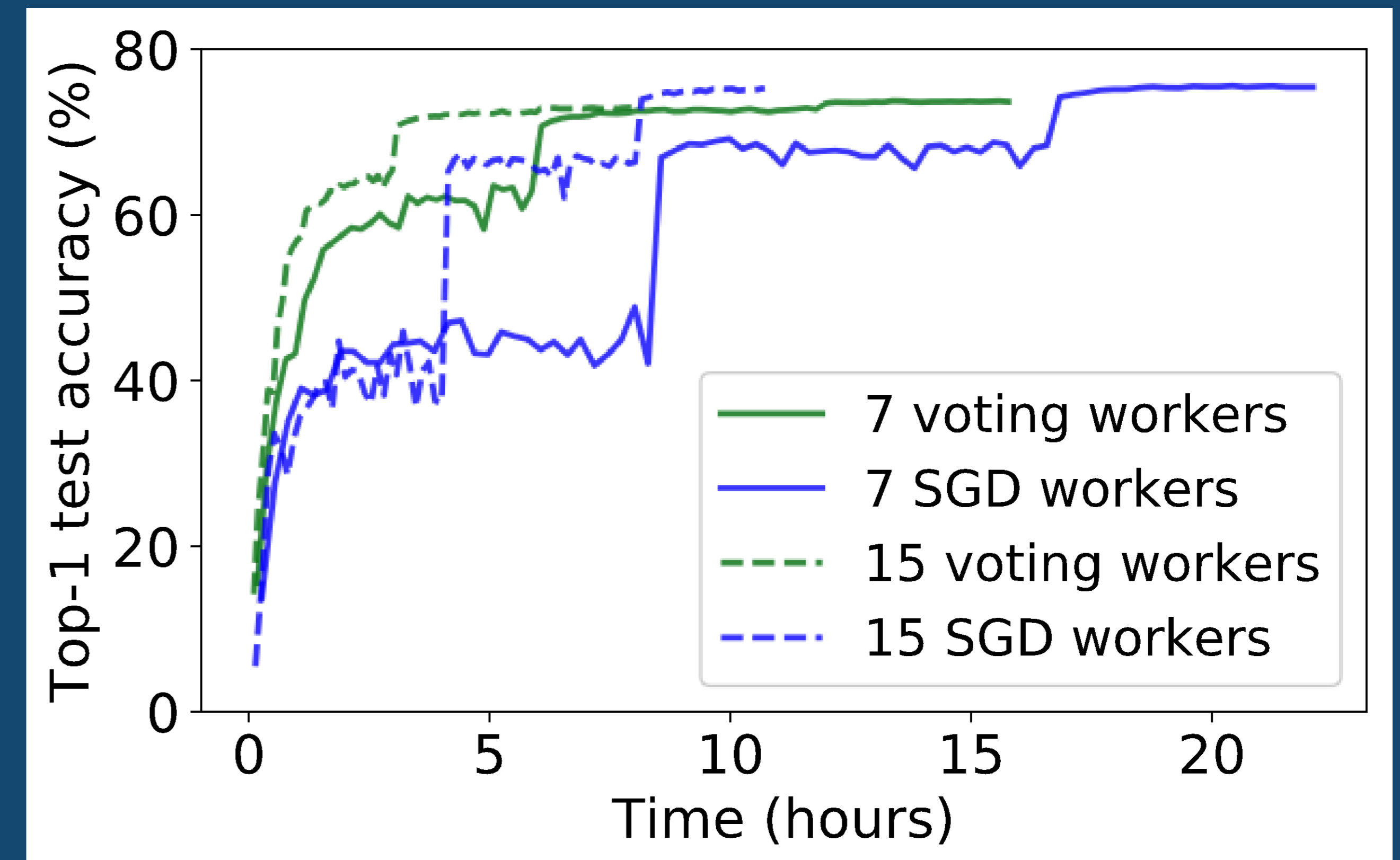
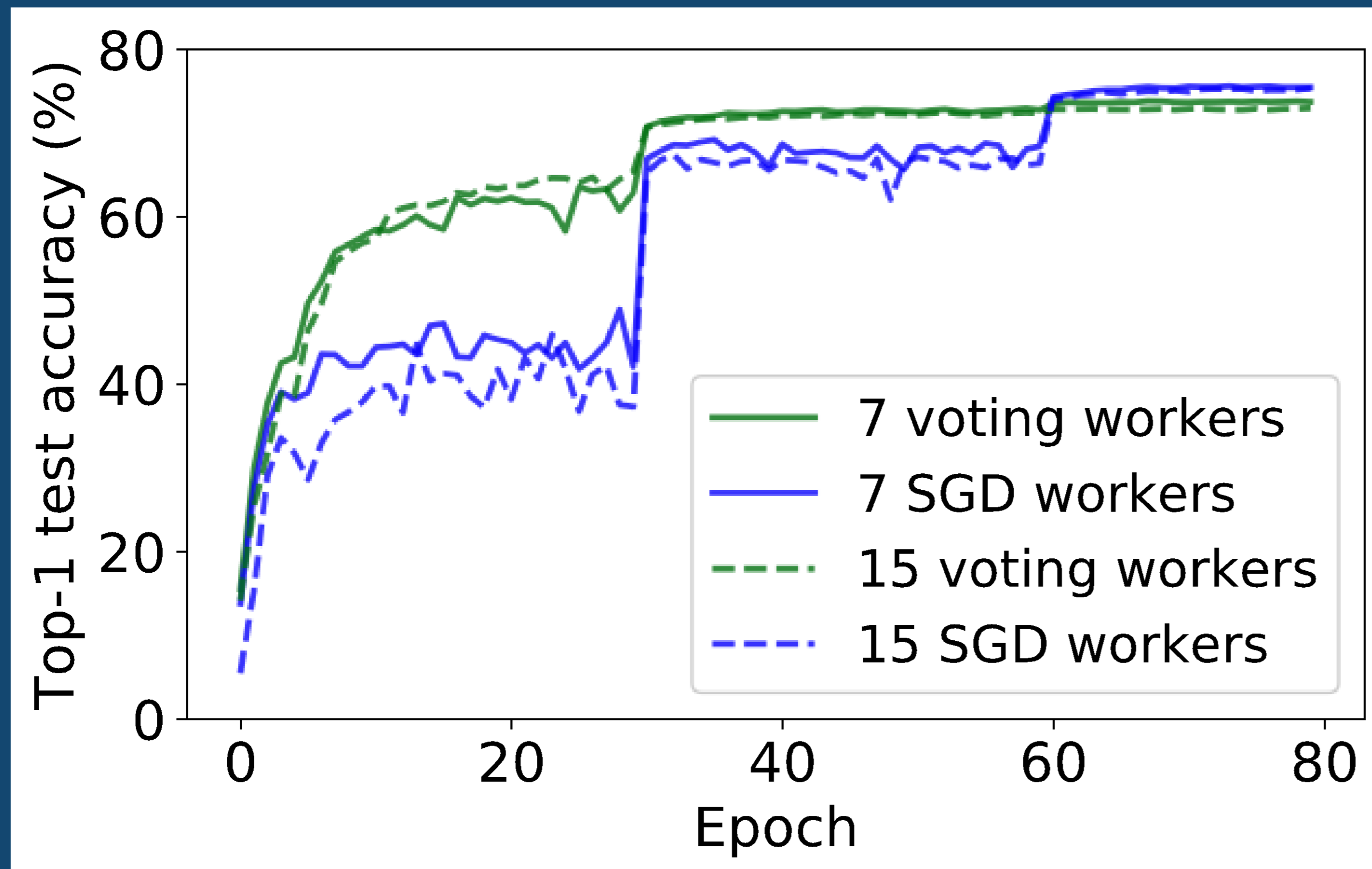
where $N = K$ is the total number of stochastic gradient calls up to step K .

CIFAR-10 SNR



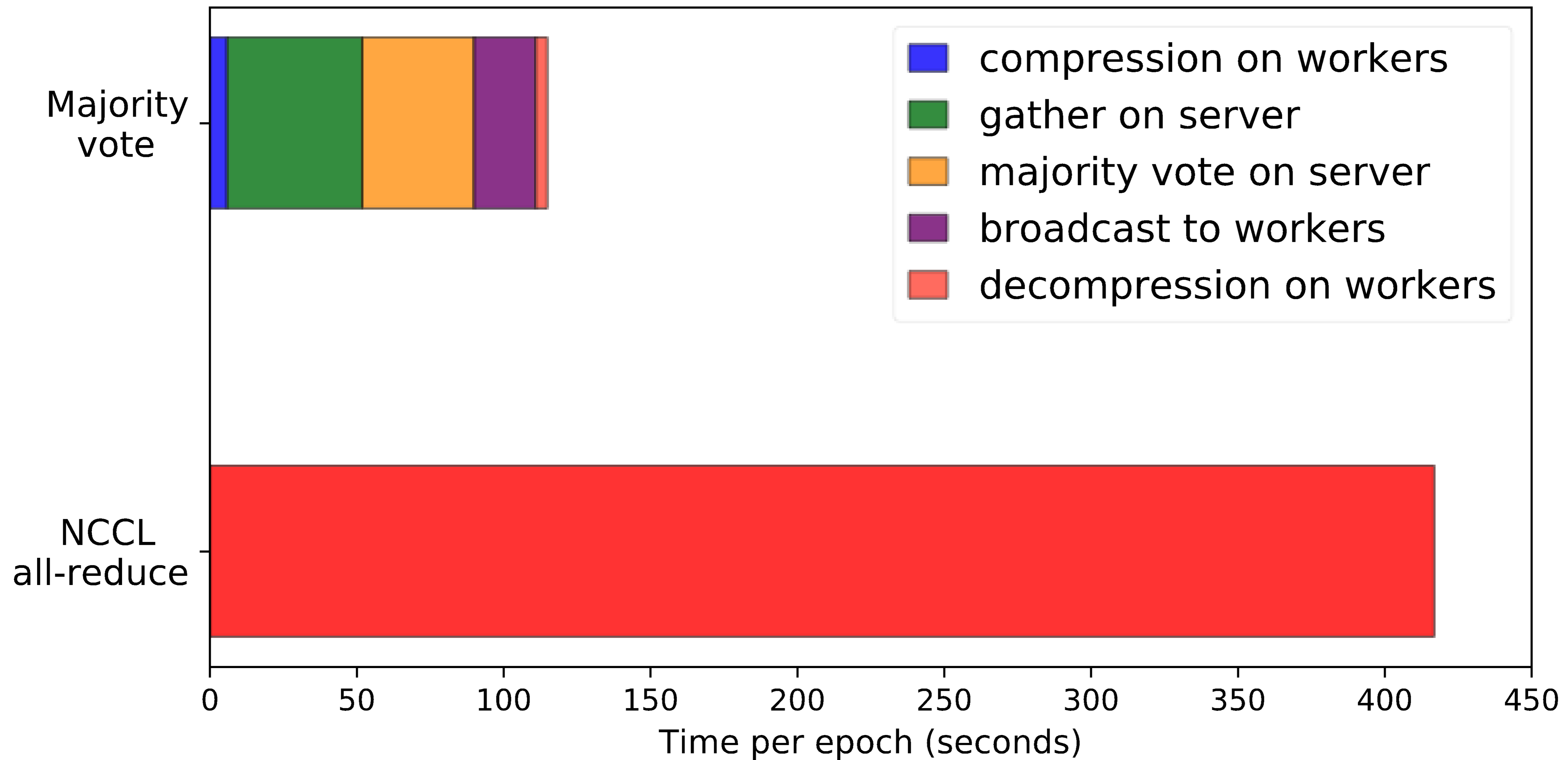
SIGNSGD PROVIDES “FREE LUNCH”

P3.2x machines on AWS, Resnet50 on imagenet

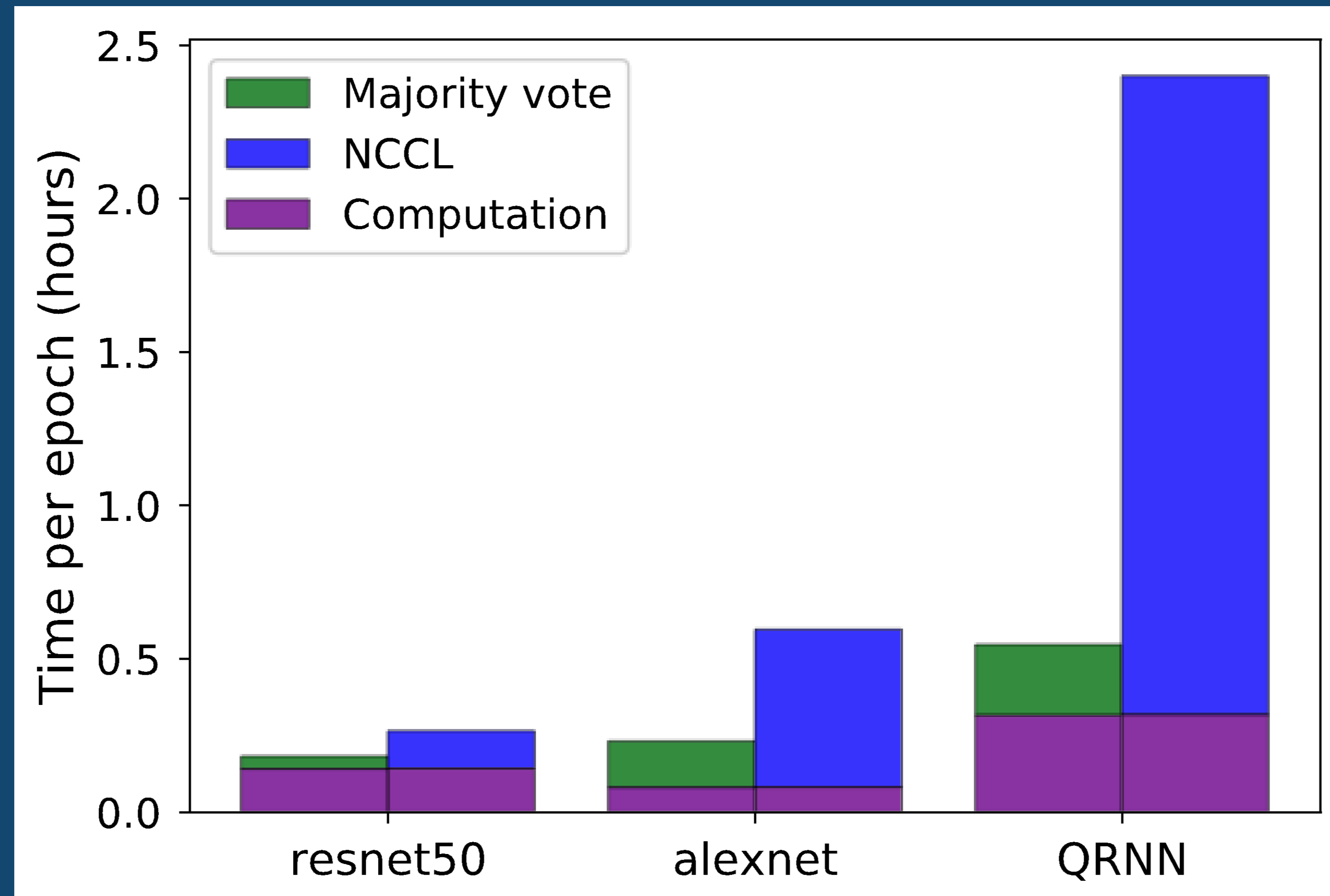


Throughput gain with only tiny accuracy loss

SIGNSGD: TIME PER EPOCH



SIGNSGD ACROSS DOMAINS AND ARCHITECTURES



Huge throughput gain!

BYZANTINE FAULT TOLERANCE

Under symmetric noise assumption:

Theorem 2 (Non-convex convergence rate of majority vote with adversarial workers). *Run algorithm 1 for K iterations under Assumptions 1 to 4. Switch off momentum and weight decay ($\beta = \lambda = 0$). Set the learning rate, η , and mini-batch size, n , for each worker as*

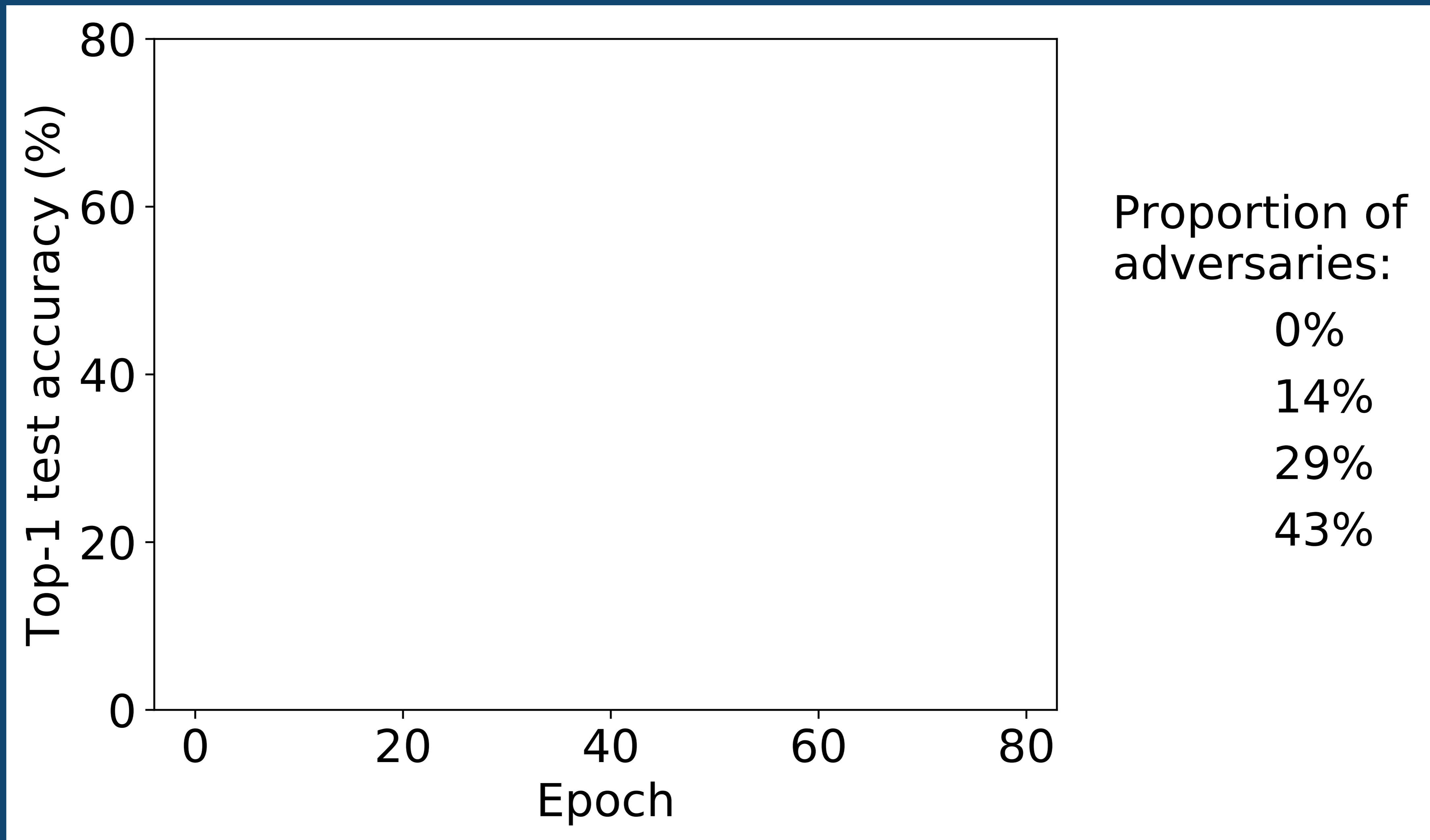
$$\eta = \sqrt{\frac{f_0 - f_*}{\|L\|_1 K}}, \quad n = K.$$

Assume that a fraction $\alpha < \frac{1}{2}$ of the M workers behave adversarially by sending to the server the negation of their sign gradient estimate. Then majority vote converges at rate:

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

where $N = K^2$ is the total number of stochastic gradient calls per worker up to step K .

SIGNSGD IS ALSO BYZANTINE FAULT TOLERANT



TAKE-AWAYS FOR SIGN-SGD

- Convergence even under biased gradients and noise.
- **Faster than SGD** in theory and in practice.
- For distributed training, similar variance reduction as SGD.
- In practice, similar accuracy but with **far less communication**.

The background is a dark teal gradient. In the four corners, there are decorative white line-art elements resembling circuit traces or neural network connections, with small circles at the end of the lines.













LEARNING FROM NOISY SINGLY-LABELED DATA

ASHISH KHETAN, ZACHARY C. LIPTON, ANIMA ANANDKUMAR

CROWDSOURCING: AGGREGATION OF CROWD ANNOTATIONS

Majority rule

- Simple and common.
- Wasteful: ignores **annotator quality** of different workers.

						
	✓		✓		×	
	✓	×			×	
			✓	×		×
		×	✓		×	
	×			×		×
		✓		✓		✓
Majority Voting	✓	×	✓	×	×	×

training data for supervised learning













CROWDSOURCING: AGGREGATION OF CROWD ANNOTATIONS

Majority rule

- Simple and common.
- Wasteful: ignores **annotator quality** of different workers.

Annotator-quality models










- Can improve accuracy.
- Hard: needs to be estimated without ground-truth.

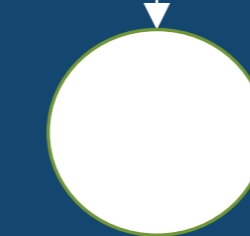
						
	✓		✓		×	
	✓	×			×	
			✓	×		×
		×	✓		×	
	×			×		×
		✓		✓		✓
Majority Voting	✓	×	✓	×	×	×

training data for supervised learning

PROPOSED CROWDSOURCING ALGORITHM

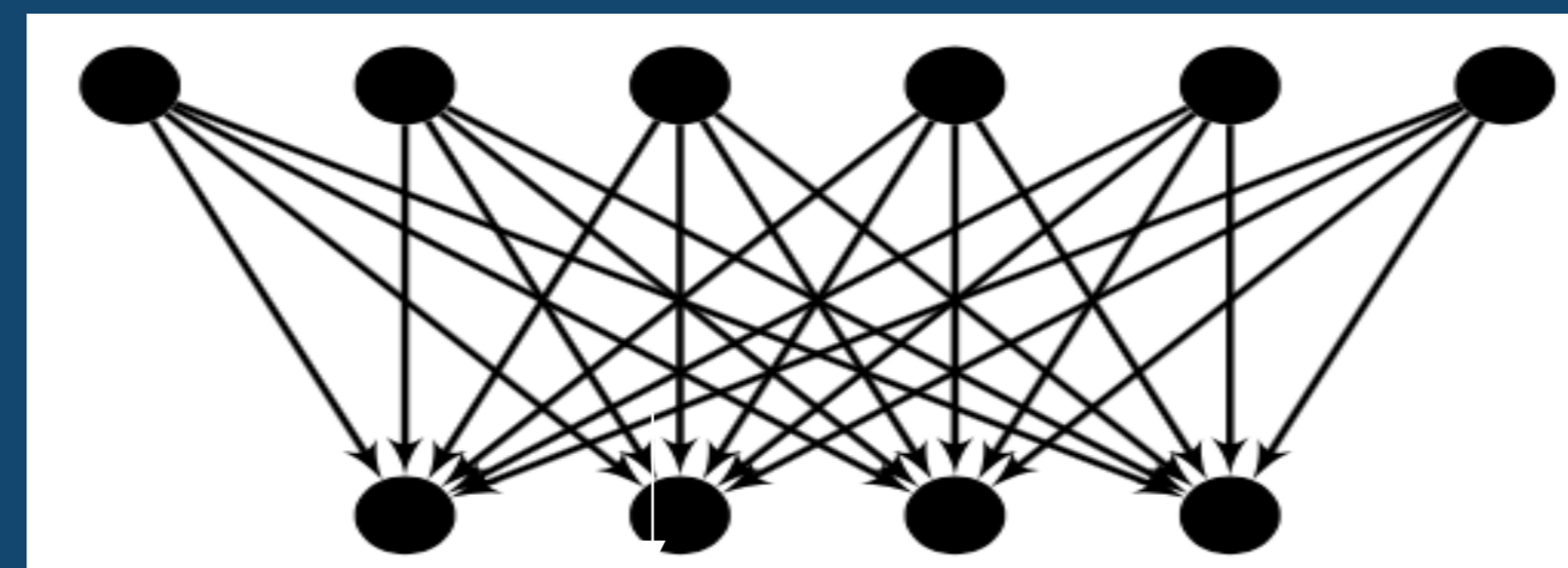
Noisy crowdsourced annotations

						
	×	×	✓	✓	×	✓
	×	✓	×	×	✓	×
	✓	×	×	✓	✓	✓

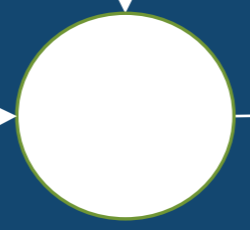





Repeat

cat	1/3	1/3	1/3	2/3	2/3	2/3
not cat	2/3	2/3	2/3	1/3	1/3	1/3



cat	0	0	1	1	1	0
not cat	1	1	0	0	0	1



0.7	0.3	0.5
		

Posterior of ground-truth labels given annotator quality model

Training with weighted loss.
Use posterior as weights

Use trained model to infer ground-truth labels

MLE : update Annotator quality using inferred labels from model

LABELING ONCE IS OPTIMAL: THEORY

Theorem:

Under fixed budget, generalization error minimized with **single annotation** per sample.

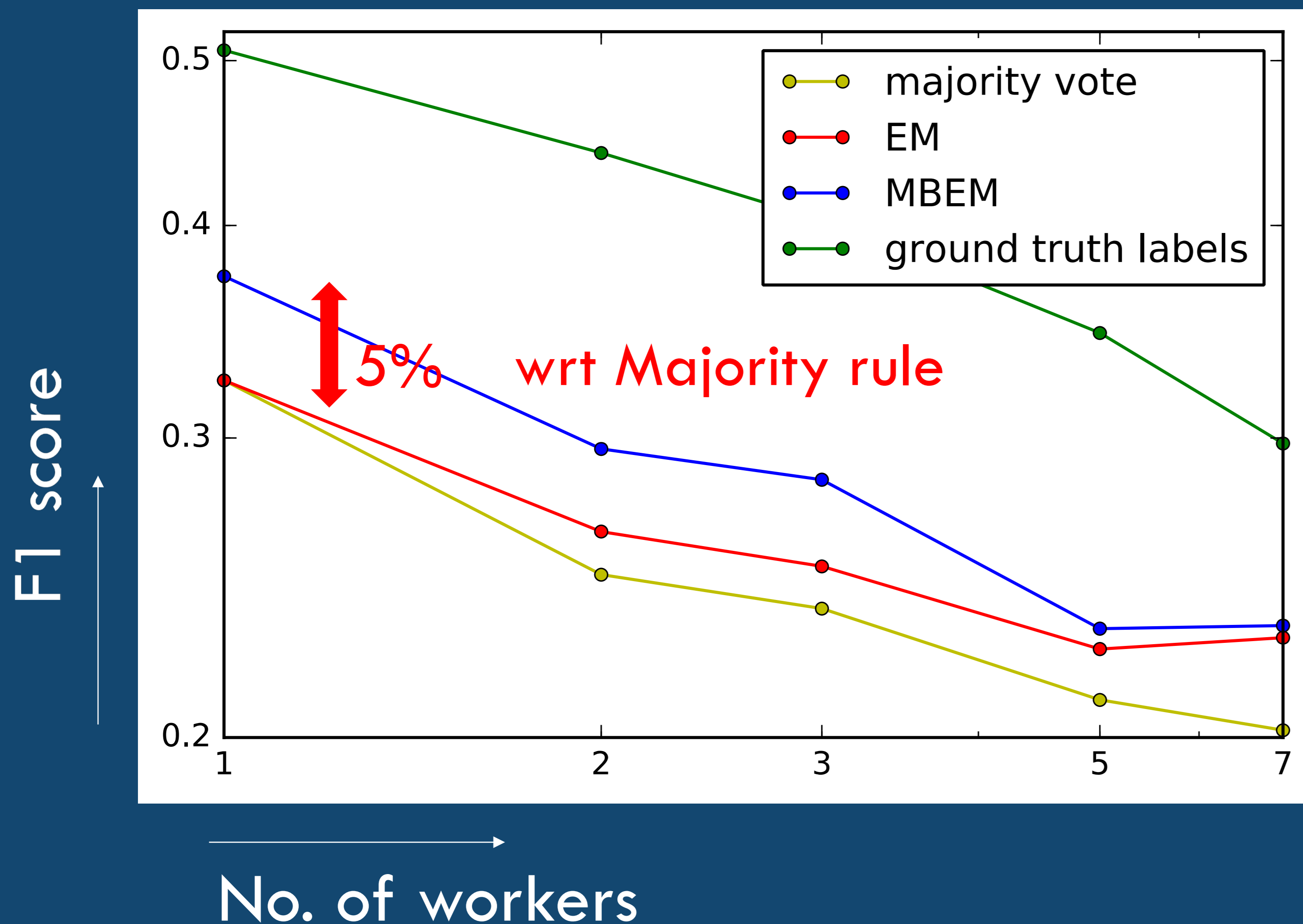
Assumptions:

- Best predictor is accurate enough (under no label noise).
- Simplified case: All workers have same quality.
- Prob. of being correct $> 83\%$

LABELING ONCE IS OPTIMAL: PRACTICE

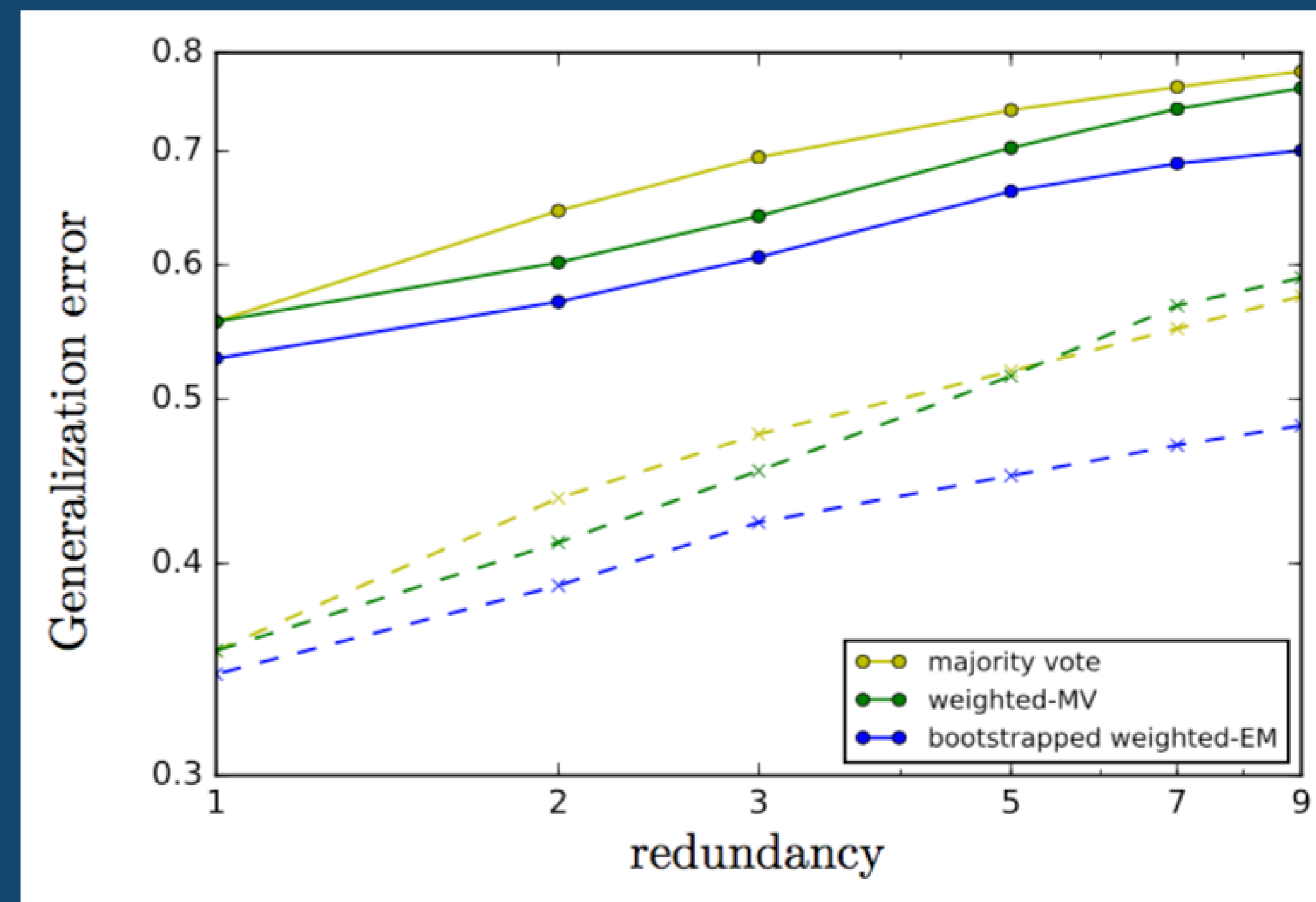
MS-COCO dataset.

Fixed budget: 35k annotations



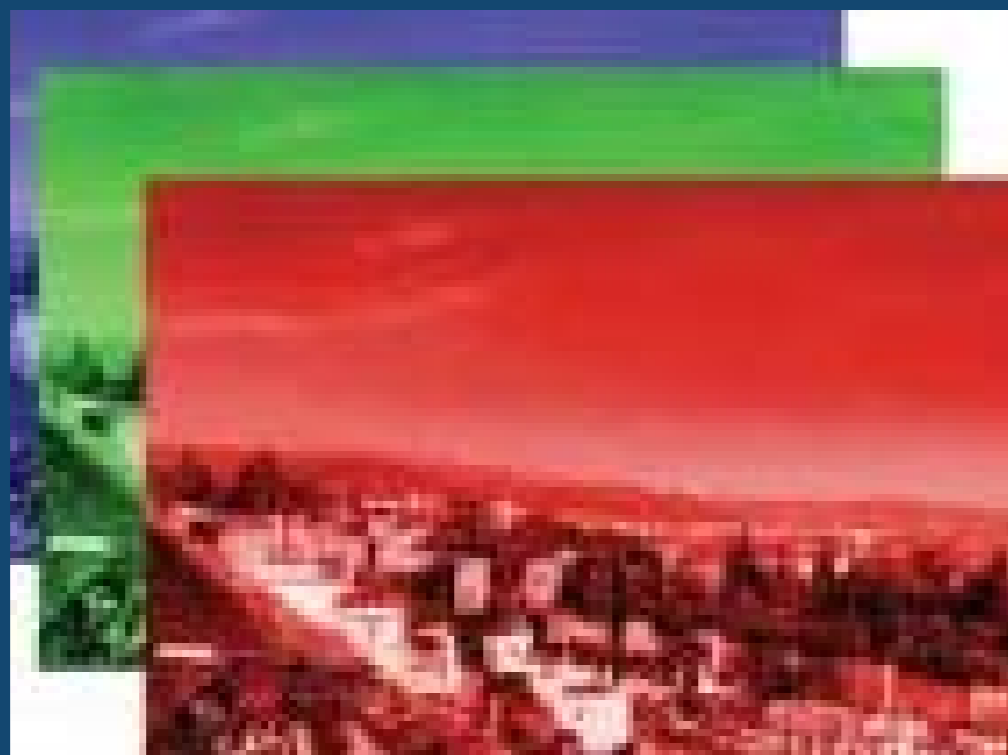
Imagenet dataset.

Simulated workers and fixed budget



TENSOR METHODS

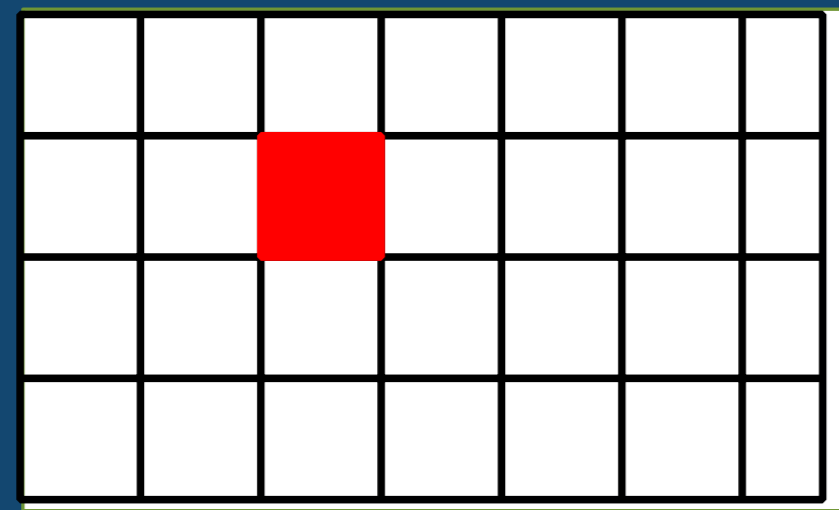
TENSORS FOR MULTI-DIMENSIONAL DATA AND HIGHER ORDER MOMENTS



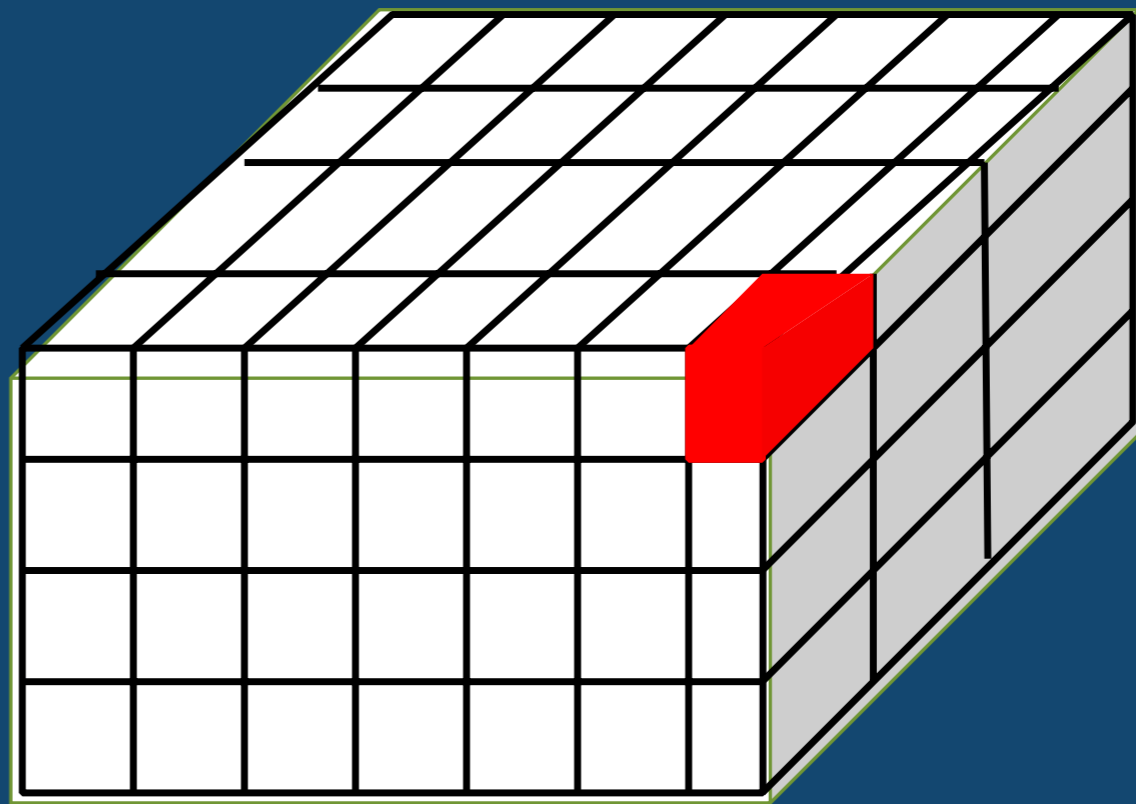
Images: 3 dimensions



Videos: 4 dimensions



Pairwise correlations



Triplet correlations

UNSUPERVISED LEARNING OF TOPIC MODELS THROUGH TENSOR METHODS



SECTIONS HOME SEARCH The New York Times

COLLEGE FOOTBALL

At Florida State, Football Clouds Justice

By MIKE McINTIRE and WALT BOGDANICH OCT. 16, 2014

Now, an examination by The New York Times of **police** and court records, along with interviews with **crime witnesses**, has found that, far from an aberration, the treatment of the Winston complaint was in keeping with the way the **police** on numerous occasions have soft-pedaled allegations of wrongdoing by Seminoles football players. From criminal mischief and motor-vehicle theft to domestic violence, arrests have been avoided, **investigations** have stalled and players have escaped serious consequences.

In a community whose self-image and economic well-being are so tightly bound to the fortunes of the nation's top-ranked college football team, law enforcement officers are finely attuned to a suspect's football connections. Those ties are cited repeatedly in **police** reports examined by The Times. What's more, dozens of officers work second jobs directing traffic and providing security at home football **games**, and many express their devotion to the Seminoles on social media.

On Jan. 10, 2013, a female student at Florida State spotted the man she believed had raped her the previous month. After **learning** his name, Jameis Winston, she reported him to the Tallahassee **police**.

In the 23 months since, Florida State officials have said little about how they handled the case, which is no **investigated** by the federal Department aggressively **investigate** the rape accusation. It did not become public until November, when a Tampa reporter, Matt Baker, acting on a tip, sought records of the **police investigation**.

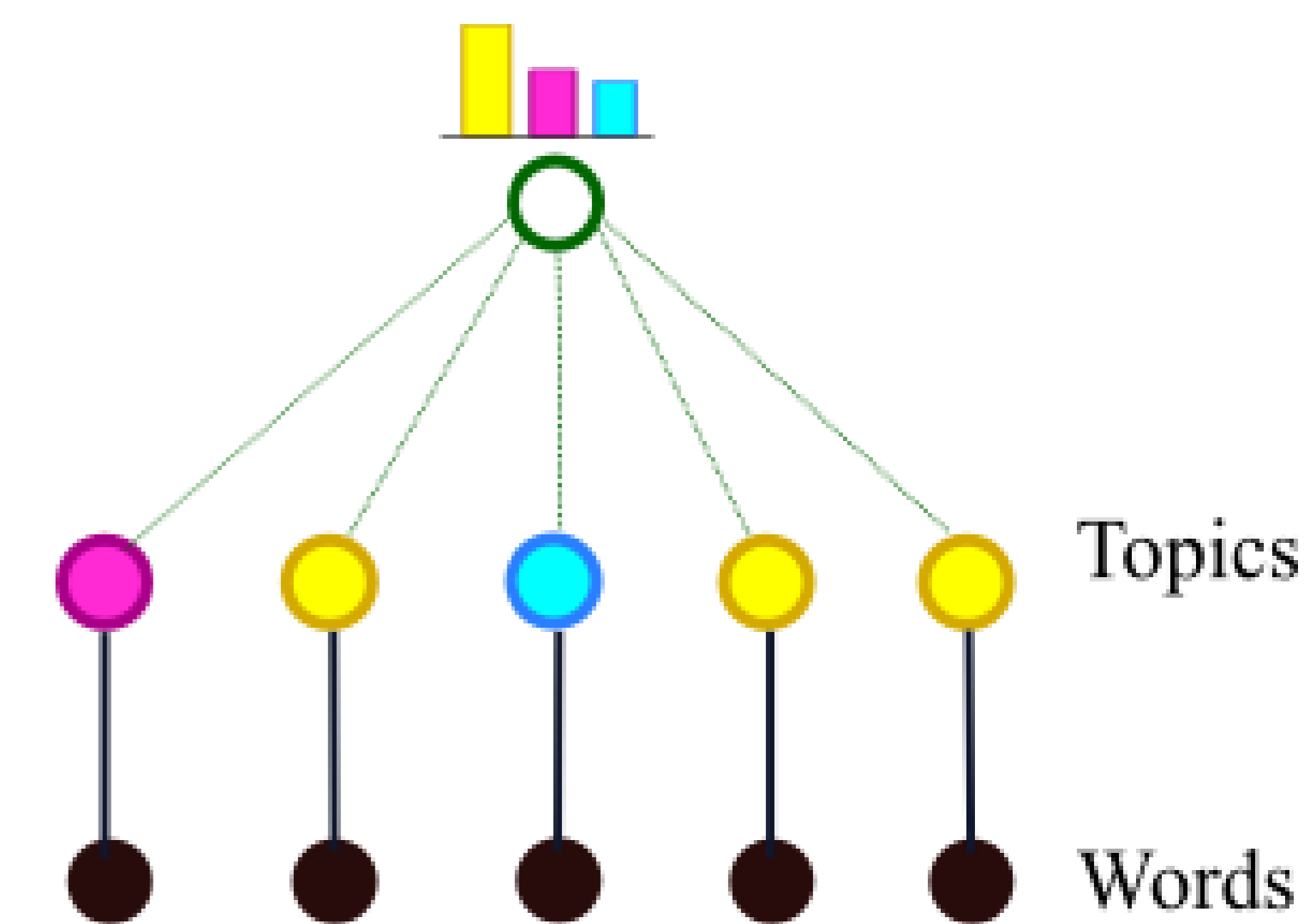
Most recently, university officials suspended Mr. Winston for one **game** after he stood in a public place on **campus** and, playing off a running Internet gag, shouted a crude reference to a sex act. In a news conference afterward, his coach, Jimbo Fisher, said, "Our hope and belief is Jameis will **learn** from this and use better judgment and language and decision-making."

TMZ, the gossip website, also requested the **police** report and later asked the school's deputy **police** chief, Jim L. Russell, if the **campus police** had interviewed Mr. Winston about the rape report. Mr. Russell responded by saying his officers were not **investigating** the case, omitting any reference to the city **police**, even though the **campus police** knew of their involvement. "Thank you for contacting me regarding this rumor — I am glad I can dispel that one!" Mr. Russell told TMZ in an email. The university said Mr. Russell was unaware of any other **police investigation** at the time of the inquiry. Soon after, the Tallahassee **police** belatedly sent their files to the news media and to the **prosecutor**, William N. Meggs. By then critical evidence had been lost and Mr. Meggs, who criticized the **police's** handling of the case, declined to

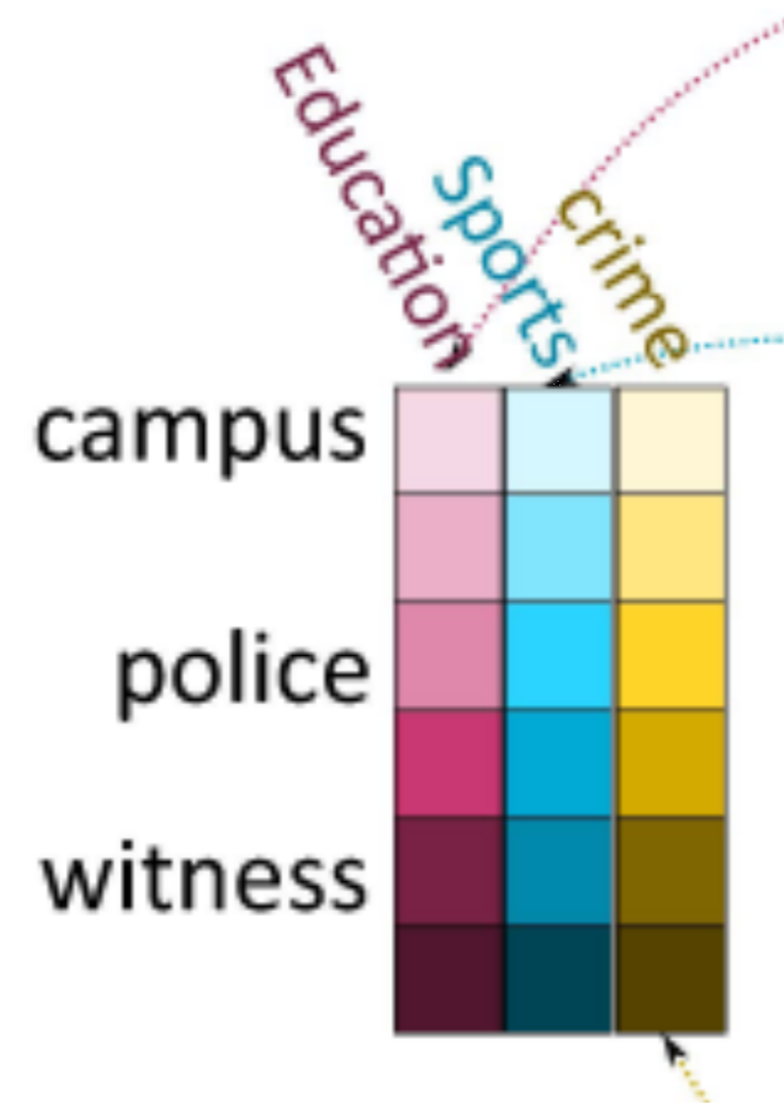
ison after the Seminoles' first **game**, five am's second-leading receiver.

Upon **learning** of Mr. Baker's inquiry, Florida State, having shown little curiosity about the rape accusation, suddenly took a keen interest in the journalist seeking to report it, according to emails obtained by The Times.

"Can you share any details on the requesting source?" David Perry, the university's **police** chief, asked the Tallahassee **police**. Several hours later, Mr.

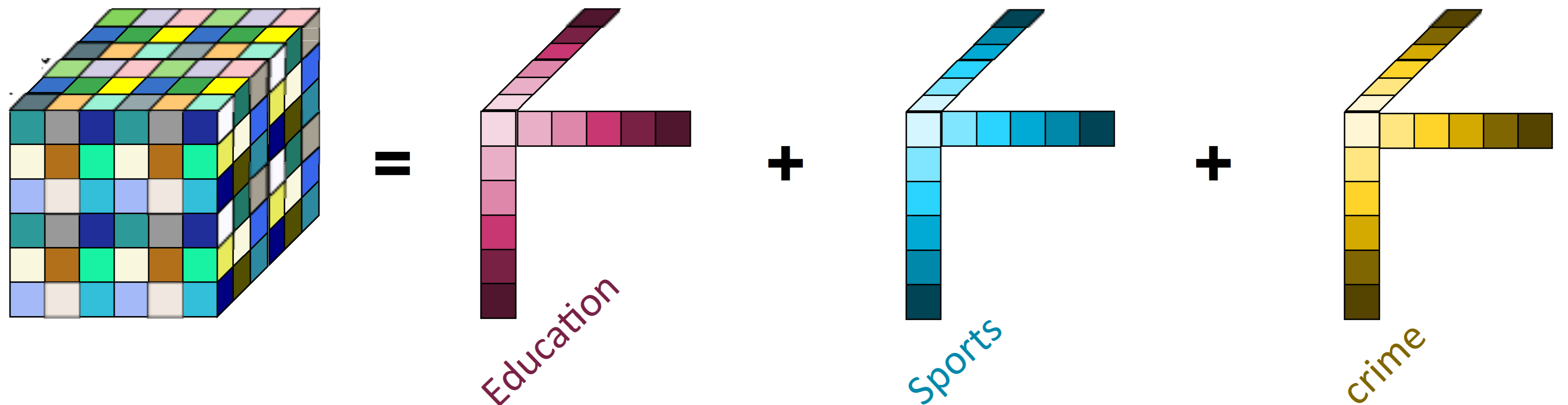


LEARNING LDA MODEL



- Topic-word matrix $P[\text{word} = i | \text{topic} = j]$
- Topic proportions $P[\text{topic} = j | \text{document}]$

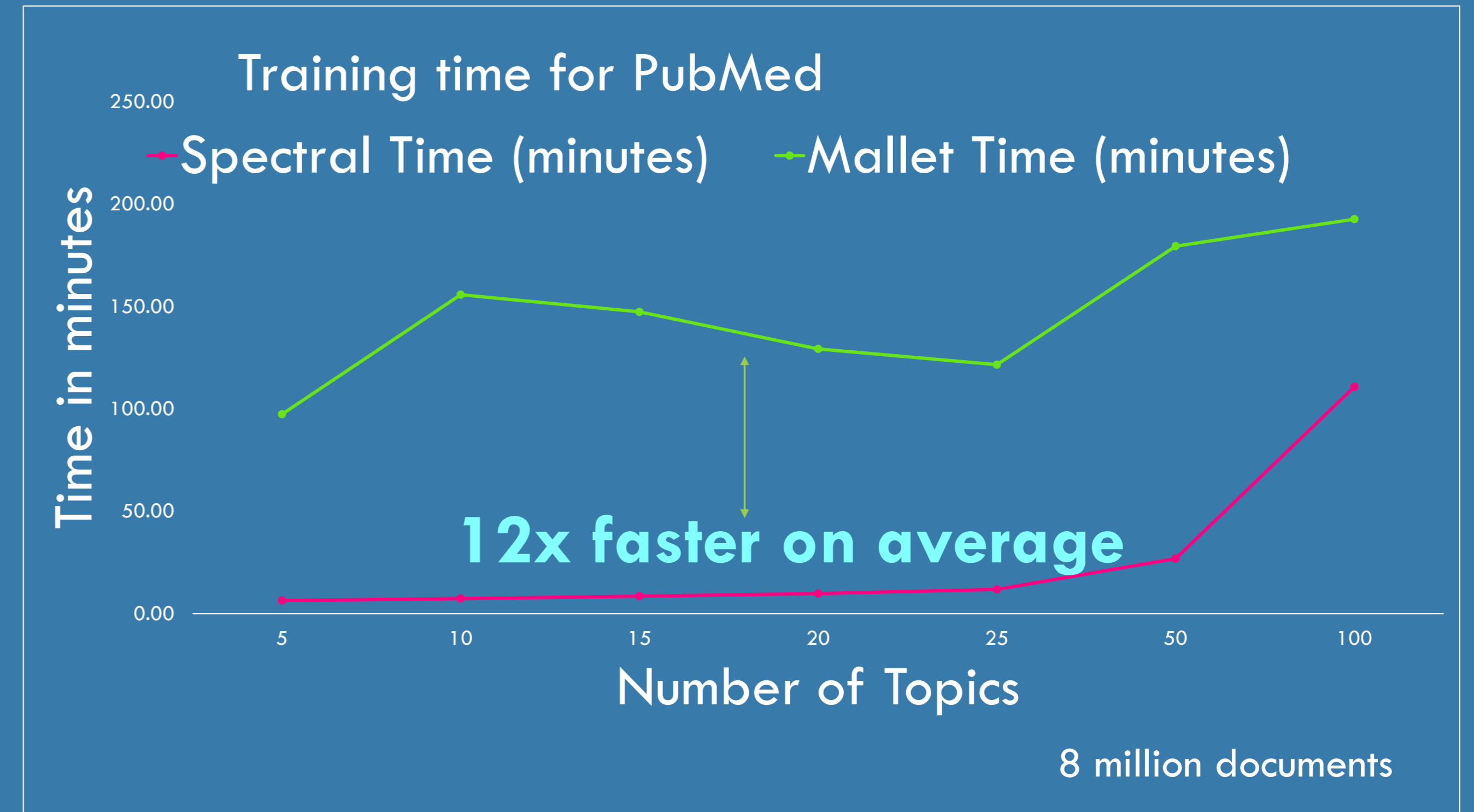
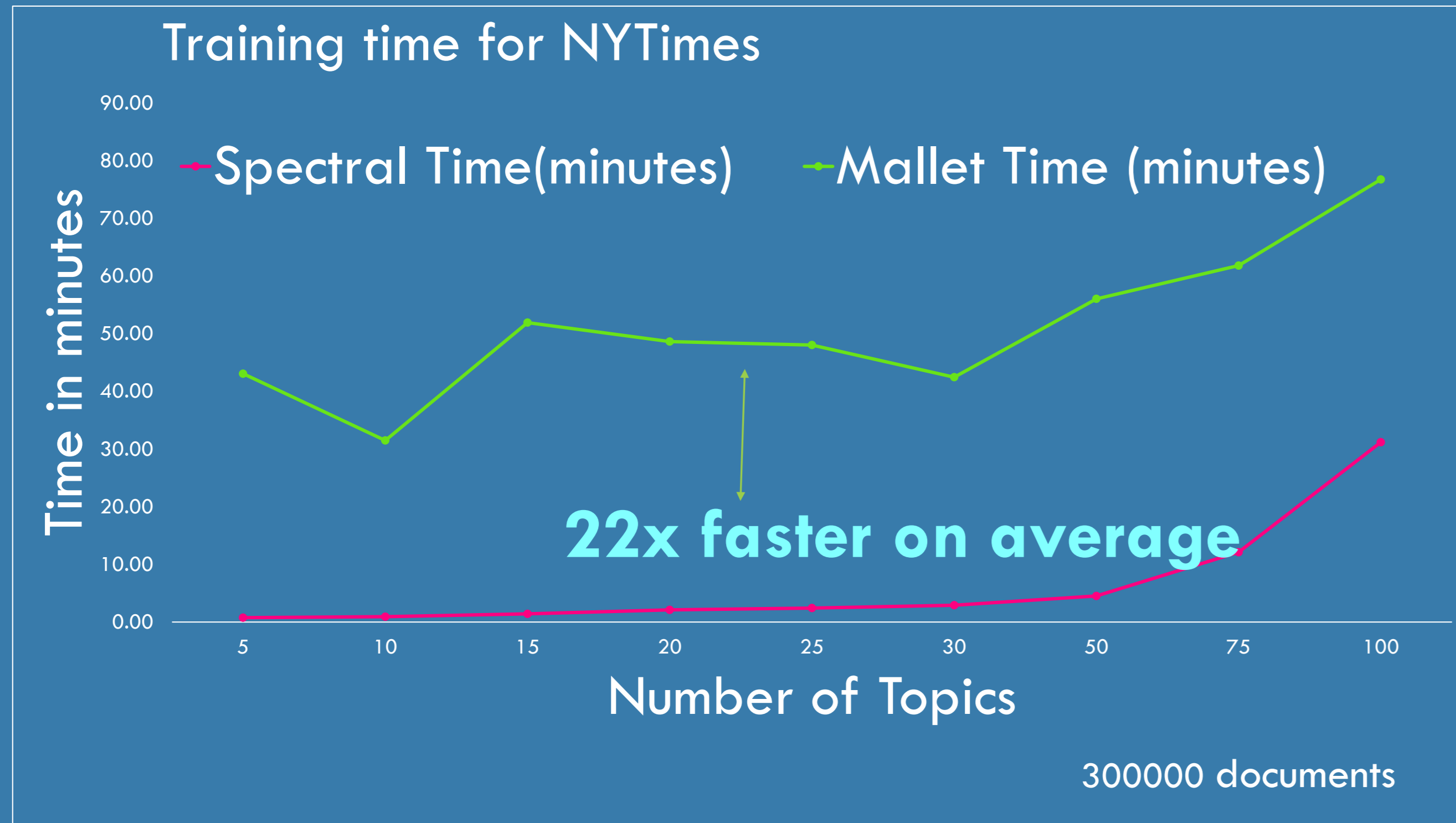
Moment Tensor: Co-occurrence of Word Triplets



WHY TENSORS?

- *Statistical reasons:*
 - Incorporate **higher order** relationships in data
 - Discover hidden topics (not possible with matrix methods)
- *Computational reasons:*
 - Tensor algebra is **parallelizable** like linear algebra.
 - **Faster** than other algorithms for LDA
 - **Flexible:** Training and inference decoupled
 - **Guaranteed** in theory to converge to global optimum

TENSOR-BASED LDA TRAINING IS FASTER



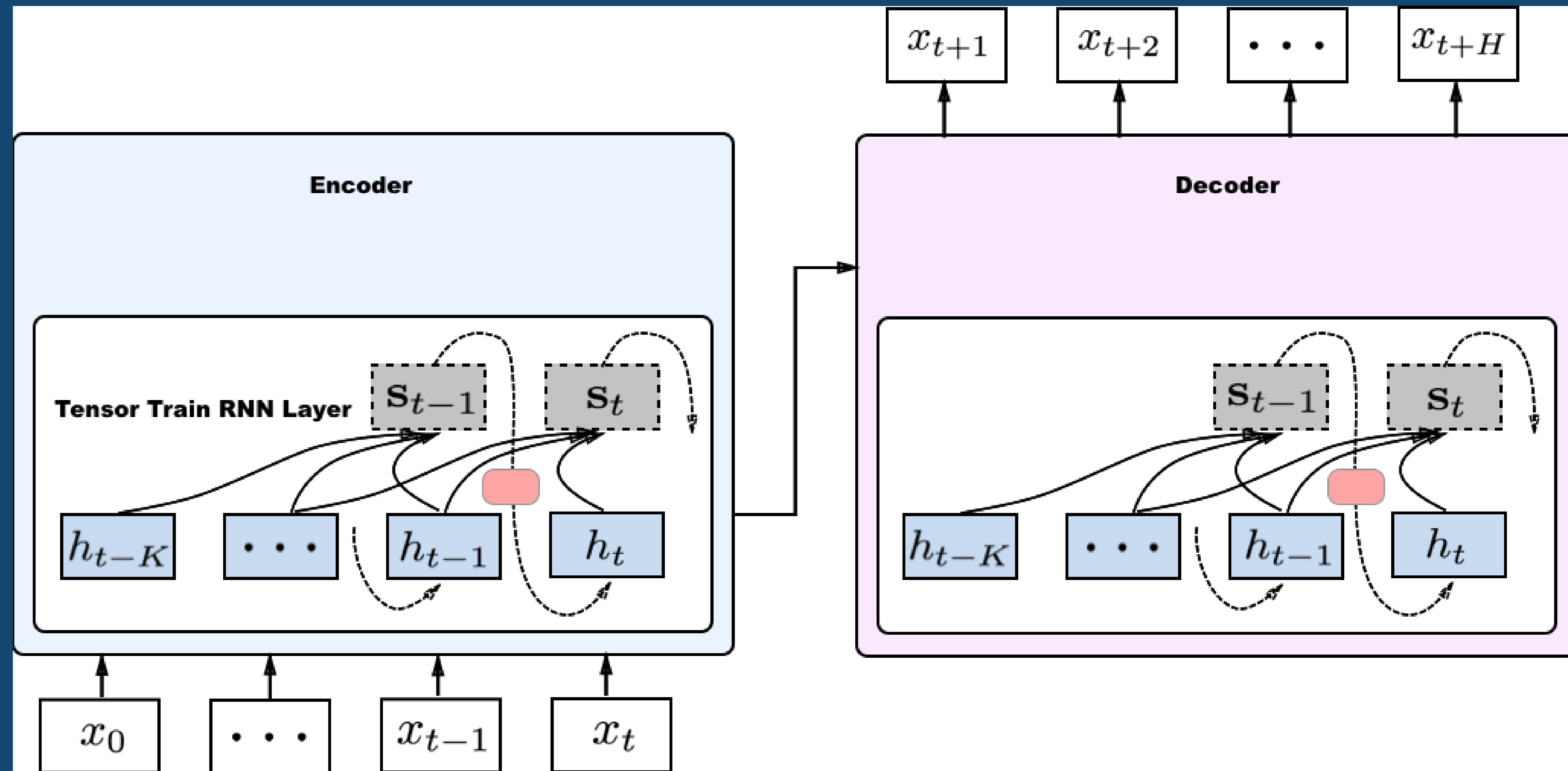
- Mallet is an open-source framework for topic modeling
- Benchmarks on [AWS SageMaker Platform](#)
- Built into [AWS Comprehend NLP service](#).

TENSORS FOR LONG-TERM FORECASTING

Tensor Train RNN and LSTMs

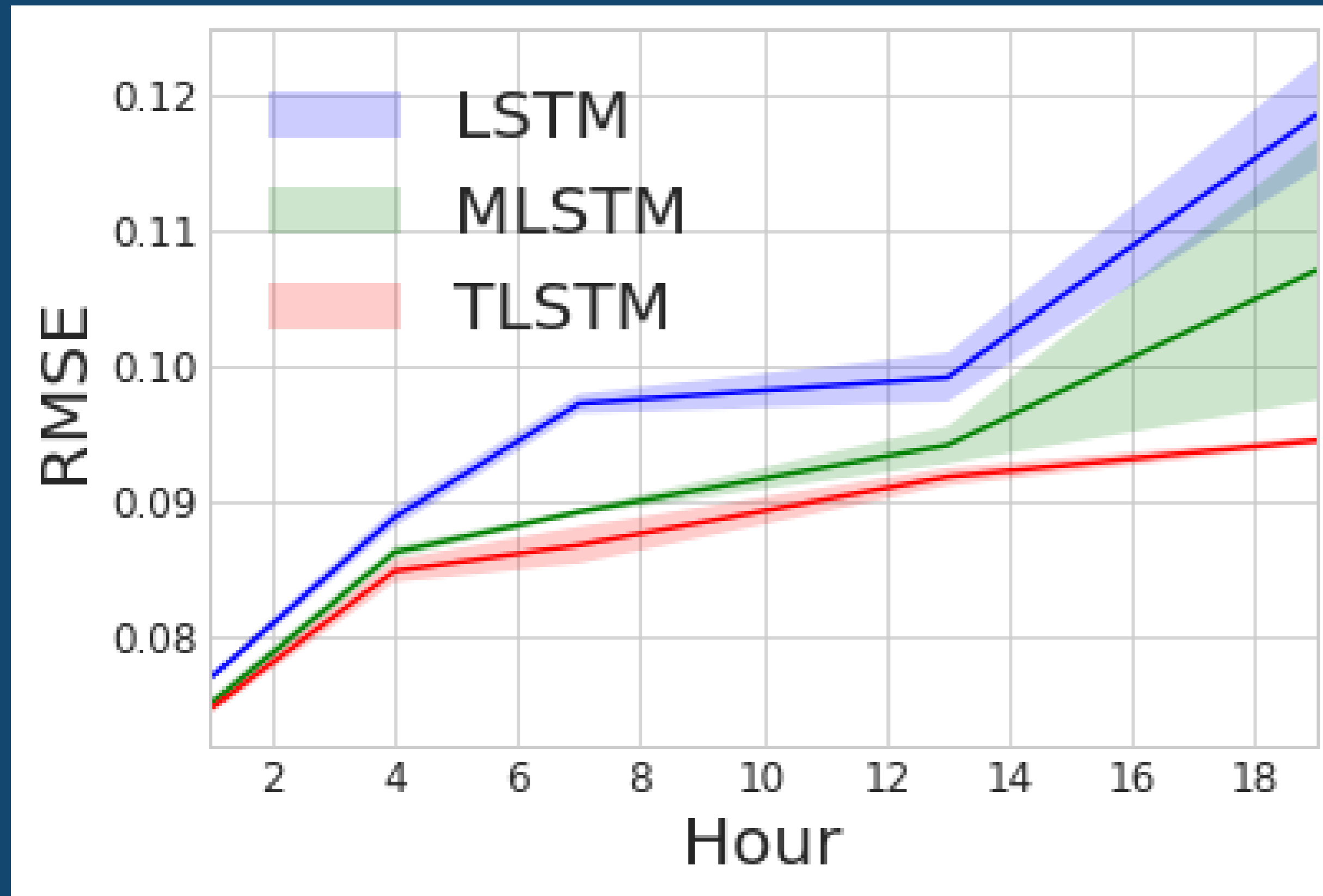
Challenges:

- Long-term dependencies
- High-order correlations
- Error propagation

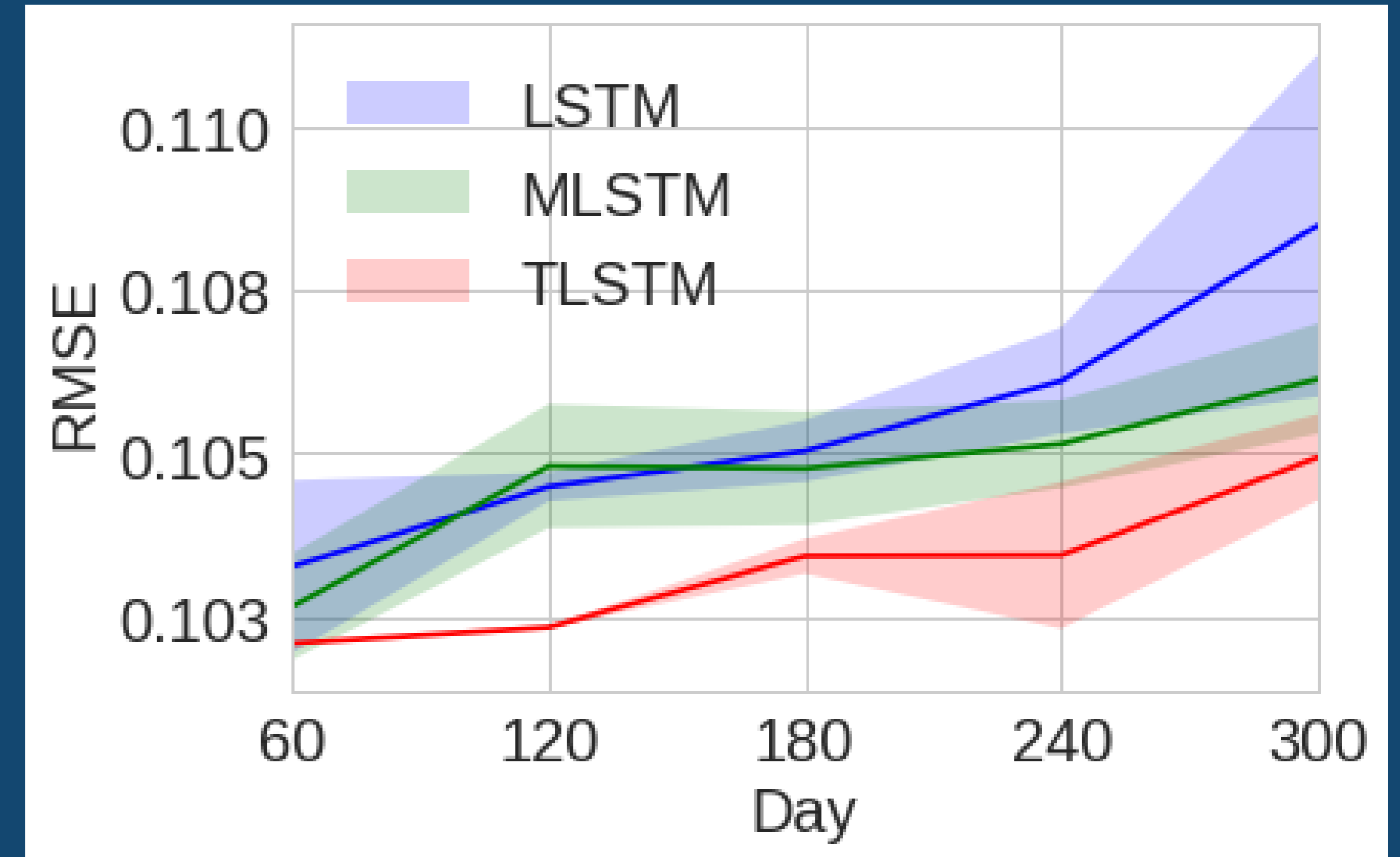


TENSOR LSTM FOR LONG-TERM FORECASTING

Traffic dataset



Climate dataset



APPROXIMATION GUARANTEES FOR TT-RNN

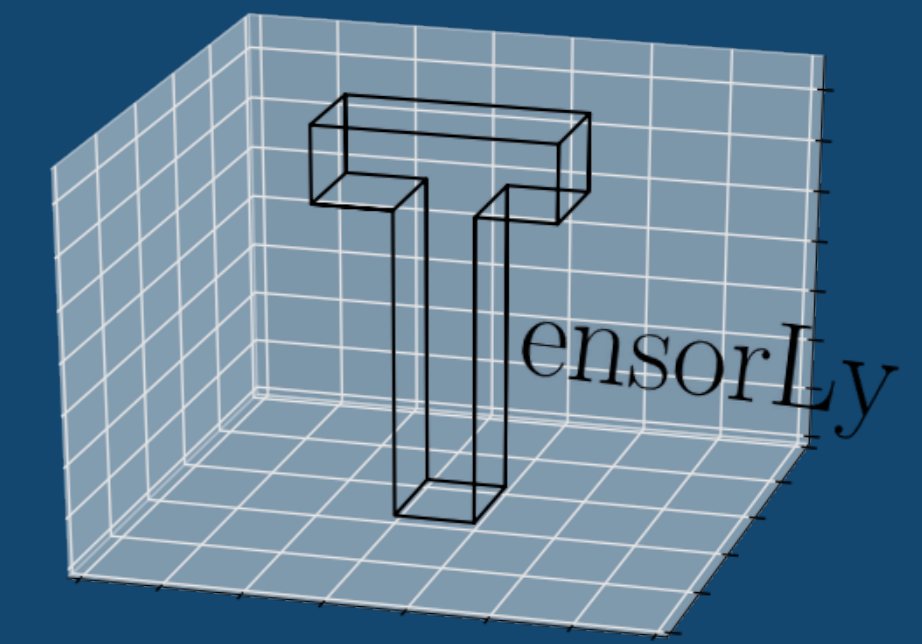
- Approximation error : bias of best model in function class.
- No such guarantees exist for RNNs.

Theorem: TT-RNN with m units approx. with error ϵ

$$m \leq O \left(\frac{C^2}{\epsilon} (dr^{-k} + p^{-k}) \right)$$

- Dimension d , tensor-train rank r . Window p .
- Bounded derivatives order k , smoothness C
- Easier to approximate if function is **smooth and analytic**.
- **Higher rank and bigger window** more efficient.

TENSORLY: HIGH-LEVEL API FOR TENSOR ALGEBRA



Tensor decomposition

Tensor regression

Tensors + Deep

Basic tensor operations

Unified backend



- Python programming
- User-friendly API
- Multiple backends: flexible + scalable
- Example notebooks in repository

NEURAL LANDER

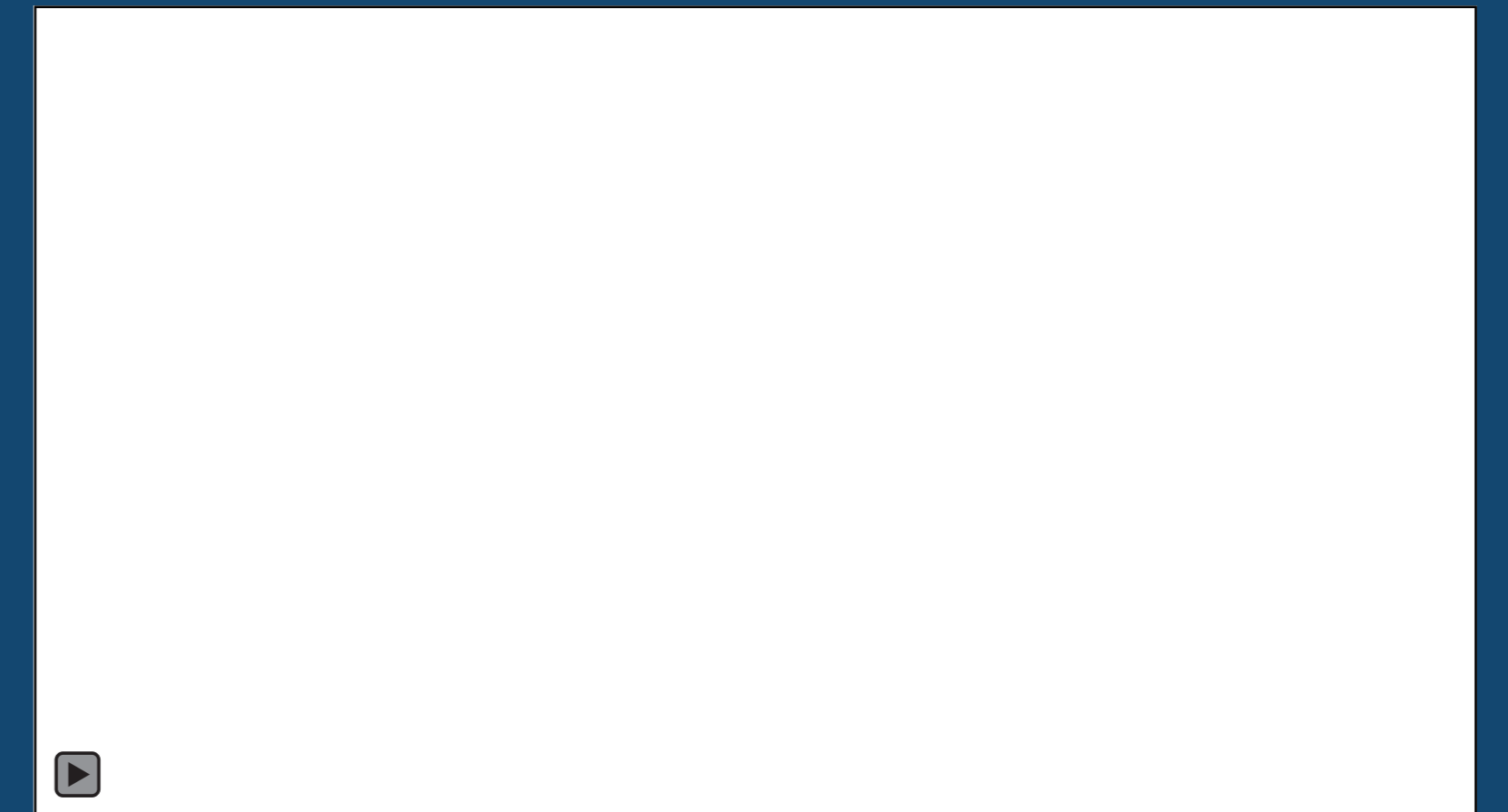
- Key Components

- Dynamics Model Assumption

$$s_{t+1} = f(s_t, a_t) + \hat{f}_a(s_t, a_t) + \epsilon$$

- Using deep learning for \hat{f}_a

- Simple Contraction Mapping Controller, cancelling out \hat{f}_a



NEURAL LANDER

- Key Theoretical Results:
 - Stability

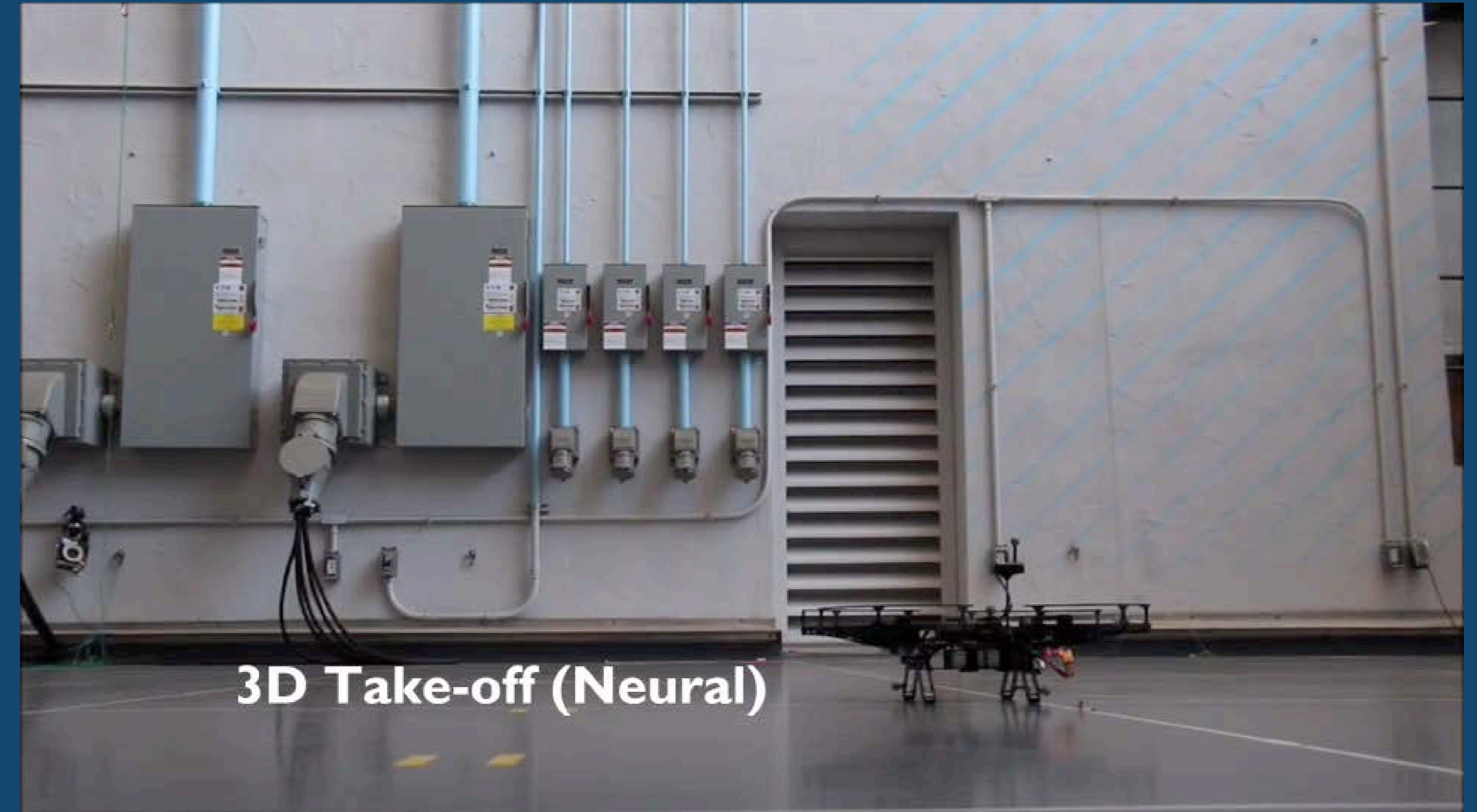
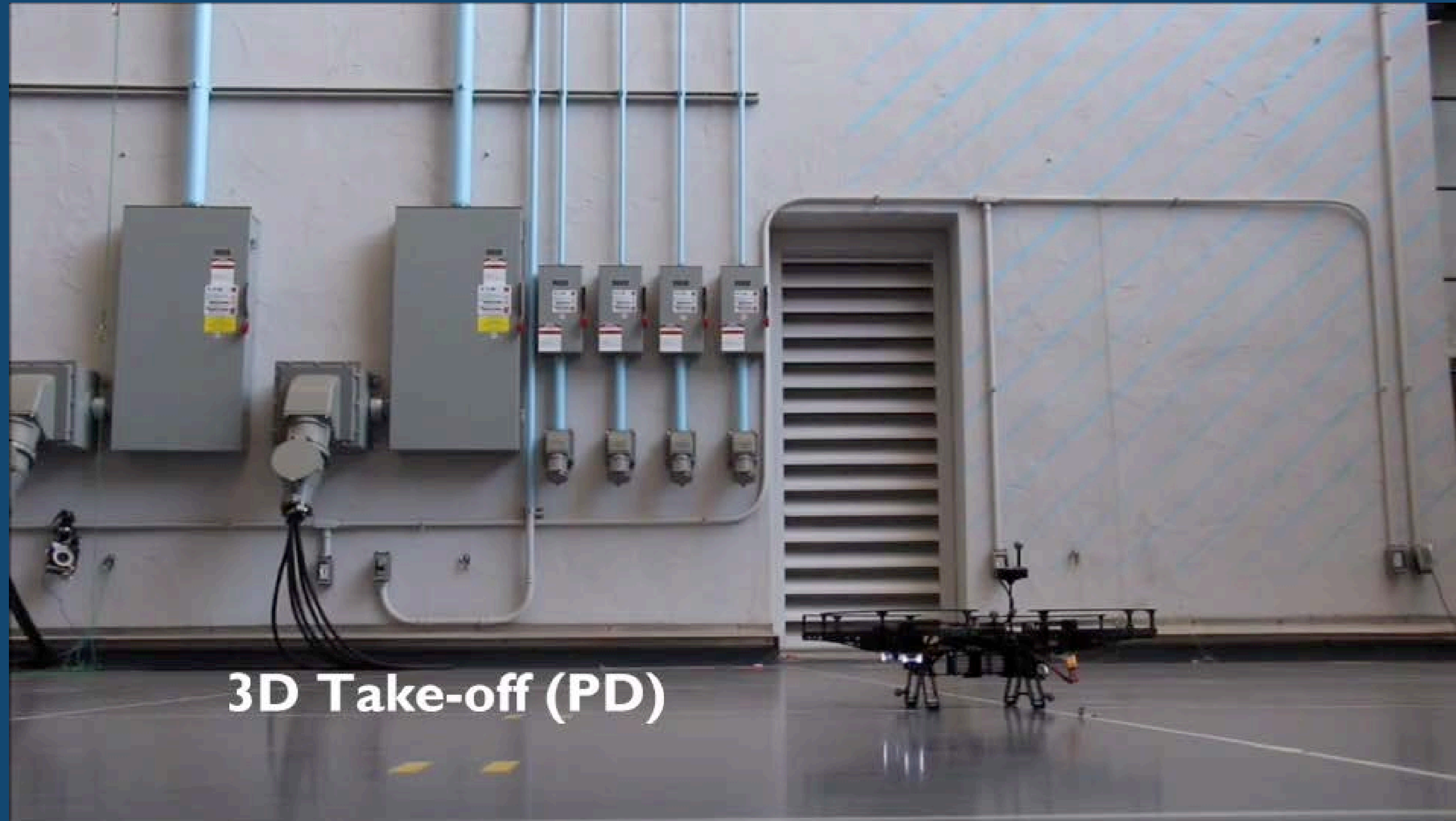
Generalization Error upper bound

$$\|\mathbf{s}(t)\| \leq \|\mathbf{s}(t_0)\| \exp\left(-\frac{\lambda - L_a \rho}{m}(t - t_0)\right) + \frac{\epsilon_m}{\lambda - L_a \rho}$$

Minimum eigenvalue of Gain matrix
- Lipschitz Constant of f_a * time delay

- Requirement:
 - Spectral Normalized ReLU for \hat{f}_a
 - Generalization error of \hat{f}_a is bounded

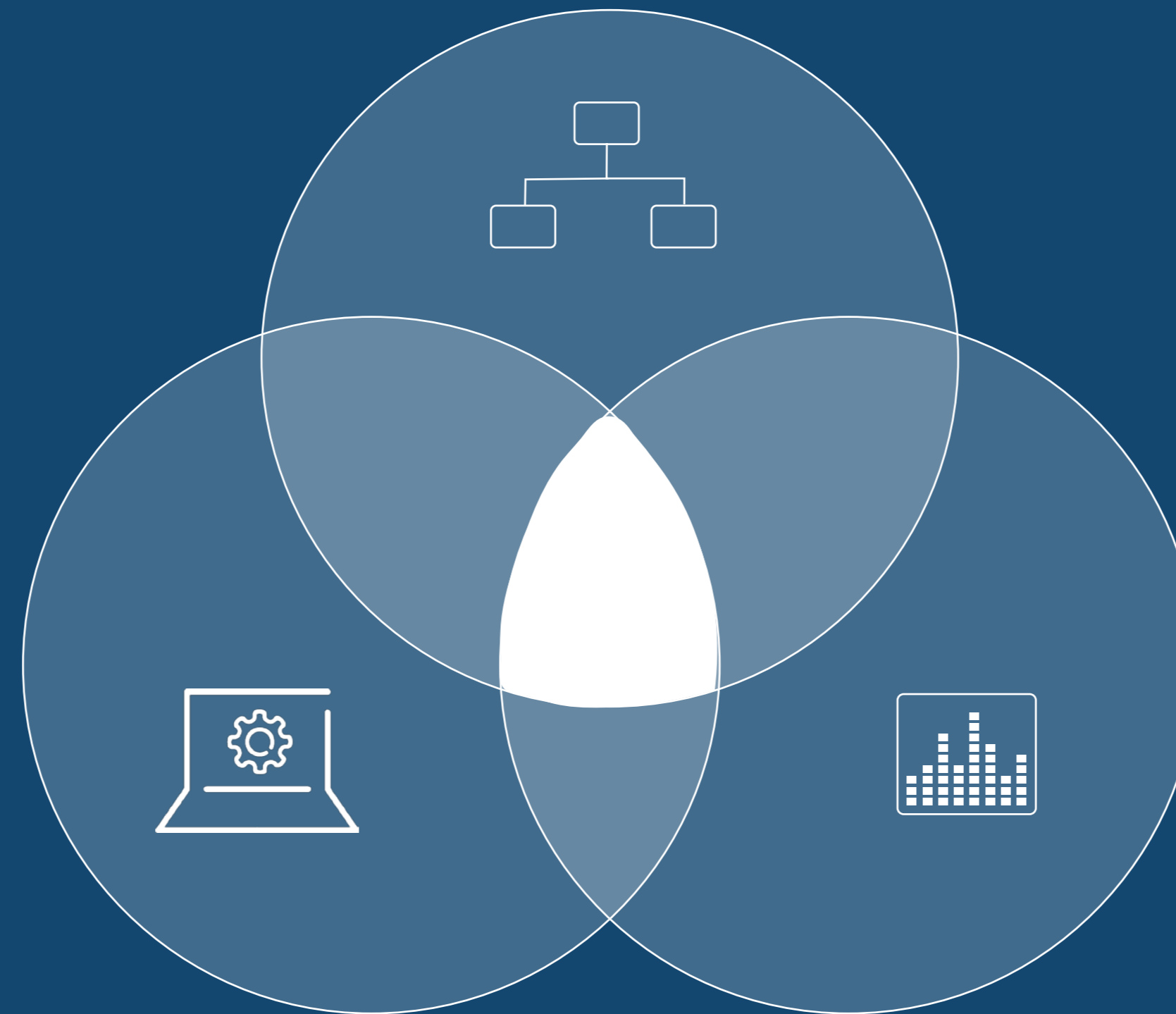
FIRST SET OF RESULTS: LEARNING TO LAND



TRINITY FUELING ARTIFICIAL INTELLIGENCE

ALGORITHMS

- OPTIMIZATION
- SCALABILITY
- MULTI-DIMENSIONALITY



INFRASTRUCTURE

FULL STACK FOR ML

- APPLICATION SERVICES
- ML PLATFORM
- GPUS

DATA

- COLLECTION
- AGGREGATION
- AUGMENTATION

COLLABORATORS (LIMITED LIST)

