Some Success Stories in Bridging Theory and Practice

Anima Anandkumar

Bren Professor at Caltech
Director of ML Research at NVIDIA
SIGNSGD: COMPRESSED OPTIMIZATION FOR NON-CONVEX PROBLEMS

JEREMY BERNSTEIN, JIAWEI ZHAO, KAMYAR AZZIZADENESHELI, YU-XIANG WANG, ANIMA ANANDKUMAR
DISTRIBUTED TRAINING INVOLVES COMPUTATION & COMMUNICATION

Parameter server

GPU 1
With 1/2 data

GPU 2
With 1/2 data
DISTRIBUTED TRAINING INVOLVES COMPUTATION & COMMUNICATION

Parameter server

- GPU 1: With 1/2 data
- GPU 2: With 1/2 data

Compress? Compress? Compress?
DISTRIBUTED TRAINING BY MAJORITY VOTE

Parameter

sign(g)

sign(g)

sign(g)

Parameter server

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

Parameter server

GPU 1

GPU 2

GPU 3

GPU 1

GPU 2

GPU 3
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

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

\( \approx 0 \) for fully sparse \( \vec{v} \)

Fully dense vector....................a sign vector

---

Graphs showing gradient and variance density over epochs for different optimization algorithms (sgd, signum, adam).
DISTRIBUTED SIGNSGD: MAJORITY VOTE THEORY

If gradients are unimodal and symmetric...

...reasonable by central limit theorem...

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

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

Same variance reduction as SGD
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, $\eta$, and mini-batch size, $n$, as

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

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. $H_k := \{i \mid S_i > \frac{2}{\sqrt{3}}\}$. We refer to $\frac{2}{\sqrt{3}}$ as the ‘critical SNR’. Then we have

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[ \sum_{i \in H_k} |g_{k,i}| + \sum_{i \not\in H_k} \frac{g_{k,i}^2}{\sigma_i} \right] \leq 3 \sqrt{\frac{\|L\|_1 (f_0 - f_\ast)}{N}}.$$ 

where $N = K$ is the total number of stochastic gradient calls up to step $K$. 
CIFAR-10 SNR

The diagram shows the training loss over epochs for different SNR settings. The plots illustrate the performance of SGD and Signum methods in terms of training loss. The graphs track the evolution of the loss function across epochs, highlighting the effectiveness of each algorithm in minimizing loss.
SIGNSGD PROVIDES "FREE LUNCH"

P3.2x machines on AWS, Resnet50 on imagenet

Throughput gain with only tiny accuracy loss
SIGNSGD: TIME PER EPOCH

- Majority vote
  - Compression on workers
  - Gather on server
  - Majority vote on server
  - Broadcast to workers
  - Decompression on workers

- NCCL all-reduce

Time per epoch (seconds)

0 50 100 150 200 250 300 350 400 450
SIGNSGD ACROSS DOMAINS AND ARCHITECTURES

Huge throughput gain!
**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 (β = λ = 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:

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

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.
**CROWDSOURCING: AGGREGATION OF CROWD ANNOTATIONS**

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

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>4</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>5</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Majority Voting</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

*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 rule to estimate annotator quality

• **Justification:** Majority rule approaches ground-truth when enough workers.

• **Downside:** Requires large number of annotations for each example for majority rule to be correct.
PROPOSED CROWDSOURCING ALGORITHM

Repeat

Posterior of ground-truth labels given annotator quality model

Noisy crowdsourced annotations

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
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

\[ F1 \text{ score} \]
\[ \text{No. of workers} \]
NEURAL RENDERING MODEL (NRM): JOINT GENERATION AND PREDICTION FOR SEMI-SUPERVISED LEARNING

Nhat Ho, Tan Nguyen, Ankit Patel, A. , Michael Jordan, Richard Baraniuk
SEMI-SUPERVISED LEARNING WITH GENERATIVE MODELS?

**Merits**
- Captures statistics of natural images
- Learnable

**Peril**
- Feedback is real vs. fake: different from prediction.
- Introduces artifacts

Diagram:
- **Generator (G)**:
  - Input: Noise
  - Output: Fake
- **Discriminator (D)**:
  - Input: Real vs Fake, Unlabeled Real
PREDICTIVE VS GENERATIVE MODELS

One model to do both?

- SOTA prediction from CNN models.
- What class of $p(x \mid y)$ yield CNN models for $p(y \mid x)$?
Neural Deep Rendering Model (NRM)

Design joint priors for latent variables based on reverse-engineering CNN predictive architectures.
NEURAL RENDERING MODEL (NRM)

---

CNN: Inference

- Image
- Unpooled feature map
- Max
- Rectified feature map
- Upsampled template

NRM: Generation

- Render
- Upsample, select location
- Choose render or not
- Class template
- Masked template
- Rendered image

Output:
- 0.5 dog
- 0.2 cat
- 0.1 horse
- ...
MAX-MIN CROSS-ENTROPY ➔ MAX-MIN NETWORKS

Cross-Entropy Loss for Training the CNNs with Labeled Data

\[
\min_{\theta \in \Theta} H_{p,q}(y|x, z_1^\text{max}) \geq \min_{(z_i)_{i=1}^n} \frac{1}{n} \sum_{i=1}^{n} -\log p(y_i|x_i, z_i; \theta)
\]

Max-Min Loss for Training the CNNs with Labeled Data

\[
\alpha \max H_{p,q}(y|x, z_1^\text{max}) + \alpha \min H_{p,q}(y|x, z_1^\text{min})
\]

- Max cross-entropy maximizes the posteriors of correct labels.
- Min cross-entropy minimizes the posteriors of incorrect labels.
- Co-learning: Max and Min networks try to learn from each other
Statistical Guarantees for the NRM

Bound on the generalization error

$$\text{Risk} \leq \frac{\text{Number of active rendering paths}}{n^{1/2}}$$

- Rendering path normalization:
- New form of regularization

Max-Min NRM with RPN achieves SOTA on benchmarks
EMPIRICAL RESULTS

Semi-Supervised Learning
• NRM + Max-Min improves SOTA by 0.7%-1.8% on CIFAR10, CIFAR100, SVHN.
• Especially in low labeled data setting.

Supervised Learning
• Max-Min improves SOTA on CIFAR10 by 0.26% and on ImageNet by 0.17% (top 5 error)

Max-Min NRM achieves SOTA on semi-supervised & supervised learning
TENSOR METHODS
Images: 3 dimensions

Pairwise correlations

Videos: 4 dimensions

Triplet correlations
Tensor Contraction

Extends the notion of matrix product

Matrix product

\[ Mv = \sum_j v_j M_j \]

Tensor Contraction

\[ T(u, v, \cdot) = \sum_{i,j} u_i v_j T_{i,j,:} \]
At Florida State, Football Clouds Justice

By MIKE McINTIRE and WALT ROGDANICH OCT. 10, 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 that 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.

Most recently, university officials suspended Mr. Winston for one game after he stood in a public place on campus and, playing off of 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.”

TM2, 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 TM2 in an email. The university said Mr. Russell was unaware of any other police investigations 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 issue a statement after the Seminoles’ first game.

In the 21 months since, Florida State officials have said little about how they handled the case, which is no longer under investigation by the federal Department of Justice. The Tallahassee police also failed to 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.

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.

“One can share any details on the requesting source?” David Perry, the university’s police chief, asked the Tallahassee police. Several hours later, Mr.
Tensor-based LDA Training is faster

- Training time for NYTimes
  - Spectral Time (minutes)
  - Mallet Time (minutes)
  - 22x faster on average

- Training time for PubMed
  - Spectral Time (minutes)
  - Mallet Time (minutes)
  - 12x faster on average

- Mallet is an open-source framework for topic modeling
- Benchmarks on AWS SageMaker Platform
- Built into AWS Comprehend NLP service.
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
Center for Autonomous Systems and Technologies

A New Vision for Autonomy
CHALLENGES IN LANDING A QUADROTOR DRONE

• Unknown aerodynamic forces & moments.
• Example 1: when drone is close to ground.
• Example 2: as velocity goes up, air drag.
• Example 3: external wind conditions.

Wind generation in CALTECH CAST wind tunnel
Our idea: use DNNs to learn unknown aerodynamic forces and then design nonlinear controller to cancel it (unknown moments are very limited in landing)

• **Challenge 1**: DNNs are data-hungry
• **Challenge 2**: DNNs can be unstable and generate unpredictable output
• **Challenge 3**: DNNs are difficult to analyze and design provably stable controller based on them

**Our approach**: using **Spectral Normalization** to control Lipschitz property of DNNs and then design stable nonlinear controller (**Neural-Lander**)
FIRST SET OF RESULTS: LEARNING TO LAND
SOME RESEARCH LEADERS AT NVIDIA

Chief Scientist
Bill Dally

Graphics
Dave Luebke
Alex Keller
Aaron Lefohn

Learning & Perception
Jan Kautz

Robotics
Dieter Fox

Computer vision
Sanja Fidler

Core ML
Me!

Applied research
Bryan Catanzaro

Programming
Michael Garland

Networks
Larry Dennison

Architecture
Steve Keckler
Dave Nellans

VLSI
Mike O’Connor

Circuits
Brucek Khailany

Tom Gray
TRINITY FUELING ARTIFICIAL INTELLIGENCE

ALGORITHMS
- OPTIMIZATION
- SCALABILITY
- MULTI-DIMENSIONALITY

INFRASTRUCTURE
FULL STACK FOR ML
- APPLICATION SERVICES
- ML PLATFORM
- GPUS

DATA
- COLLECTION
- AGGREGATION
- AUGMENTATION
Thank you