Application of Approximate Matrix Multiplication to Neural Networks and Distributed SLAM

Brian Plancher*, Camelia D. Brumar*, Iulian Brumar*, Lillian Pentecost*,

Saketh Rama*, David Brooks

Harvard University

Motivation: Applying Theory

- Linear algebra is compute-intensive
- Mid-1990s and 2000s: Algorithmic analyses of randomized approximations for linear algebra

Motivation: Hardware

- Can this benefit **resource-constrained hardware**?

Motivation: Hardware

- Can this benefit **resource-constrained hardware**?

(Answer: **Maybe.**)

Outline

1. Overview of approximate linear algebra
2. Evaluating some end-to-end sampling strategies
3. Predicting end-to-end error bounds

Outline

1. Overview of **approximate** linear algebra
2. Evaluating some end-to-end sampling strategies
3. Predicting end-to-end error bounds

Outline

1. Overview of approximate linear algebra
2. Evaluating some end-to-end **sampling** strategies
3. Predicting end-to-end error bounds

Outline

1. Overview of approximate linear algebra
2. Evaluating some end-to-end sampling strategies
3. Predicting end-to-end **error bounds**

Outline

1. **Overview of approximate linear algebra**
2. Evaluating some end-to-end sampling strategies
3. Predicting end-to-end error bounds

Randomized Approximations

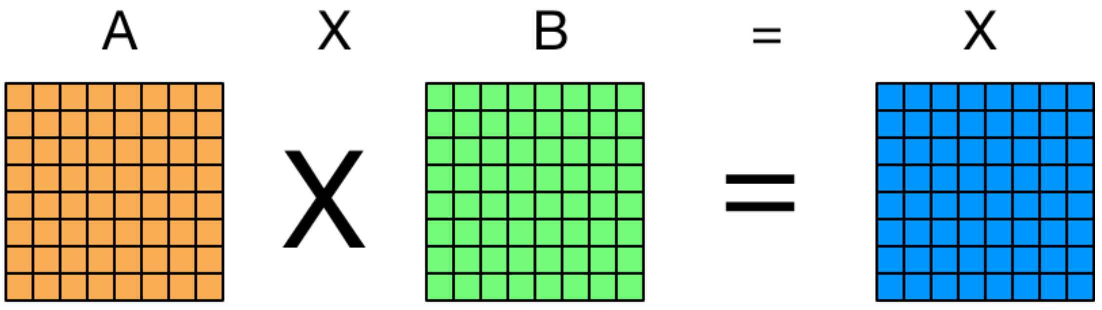
- Low-rank approximations
  - Frieze, Kannan and Vempala (1998, 2004)

Randomized Approximations

- Low-rank approximations
  - Frieze, Kannan and Vempala (1998, 2004)
- Matrix multiplication
- Singular value decomposition (SVD)
- Dimensionality reduction
- Linear regression

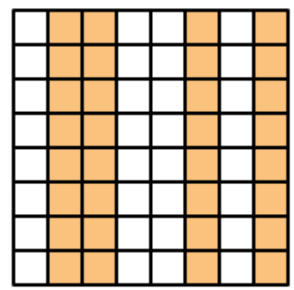
Exact Matrix Multiplication

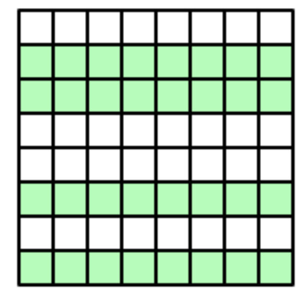
No Approximation



Sampling for Matrix Multiplication

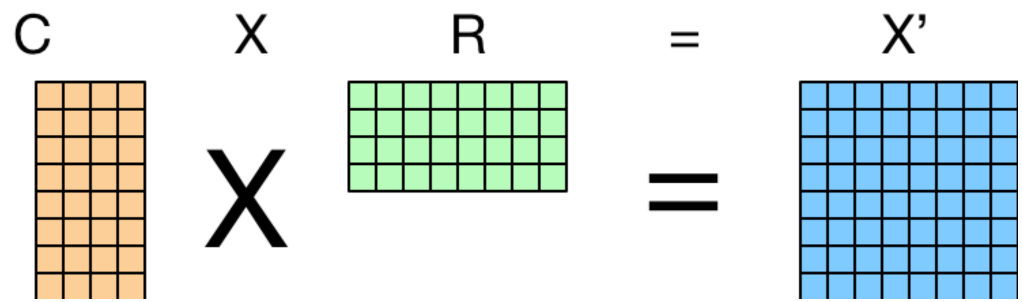
Sampling A and B





Monte Carlo Matrix Multiplication

With Approximation



- In general, for a custom sampling distribution, and c sampled column-row pairs, we construct C and R:

$$C^t = \frac{A^{i_t}}{\sqrt{c * p_{i_t}}} \qquad R_t = \frac{B_{i_t}}{\sqrt{c * p_{i_t}}}$$

Theoretical Bounds

$$\frac{\|AB - CR\|}{\|AB\|} \leq factor * \frac{\|A\| * \|B\|}{\sqrt{c} * \|AB\|}$$

"Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication"
[Drineas et al. 2006]

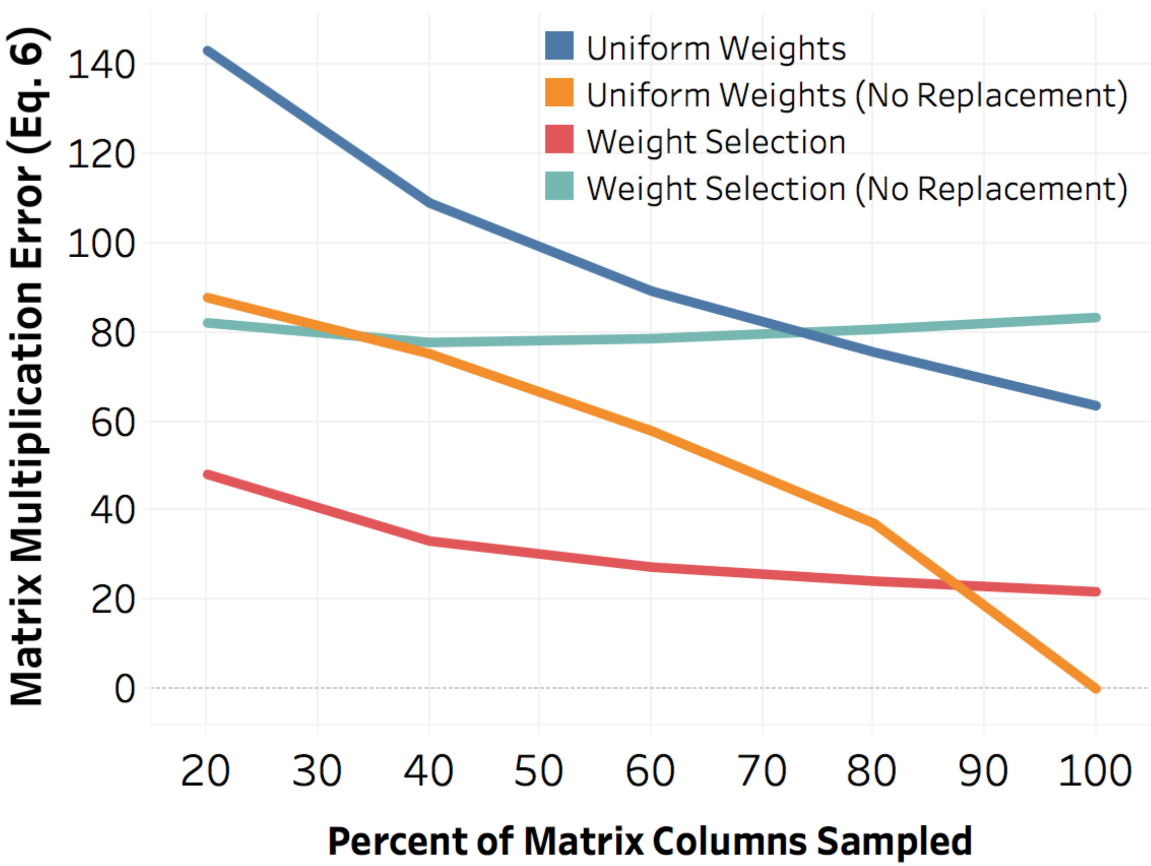
Some steps before application…

- Asymptotic bounds
  - What do the constant factors look like?
- Bounds on relative values

Outline

1. Overview of approximate linear algebra
2. **Evaluating some end-to-end sampling strategies**
3. Predicting end-to-end error bounds

Evaluation of Sampling Strategy



Application: SLAM

- Simultaneous Localization and Mapping

(Image: UPenn, Kumar Lab)

D-SLAM: Most Expensive Step

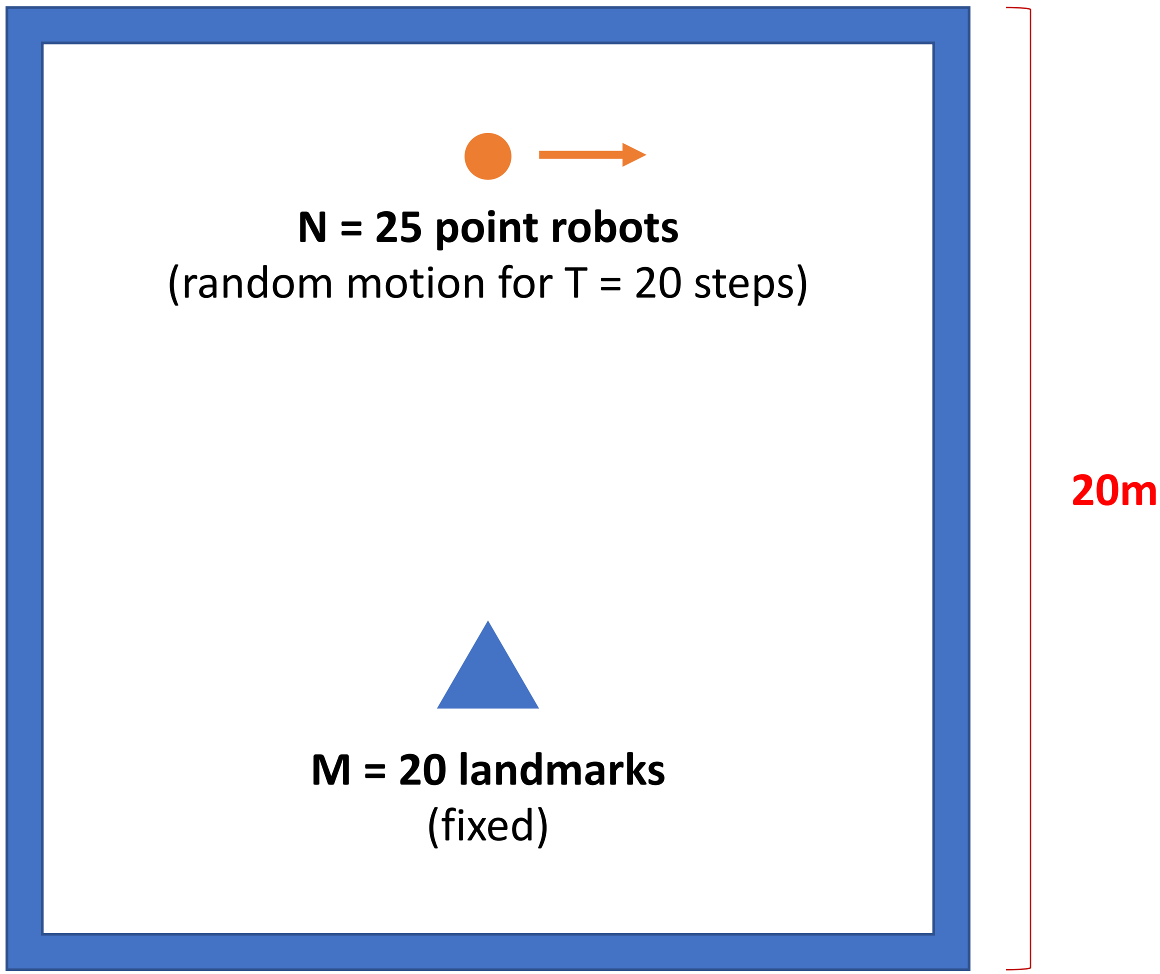
- D-SLAM: Evaluate on the **distributed** case
  - Bottleneck: Computing **covariance matrix** (Σ)
  - More robots = larger covariance matrix

$$\Sigma \in \mathbb{R}^{Nn+Mm}$$

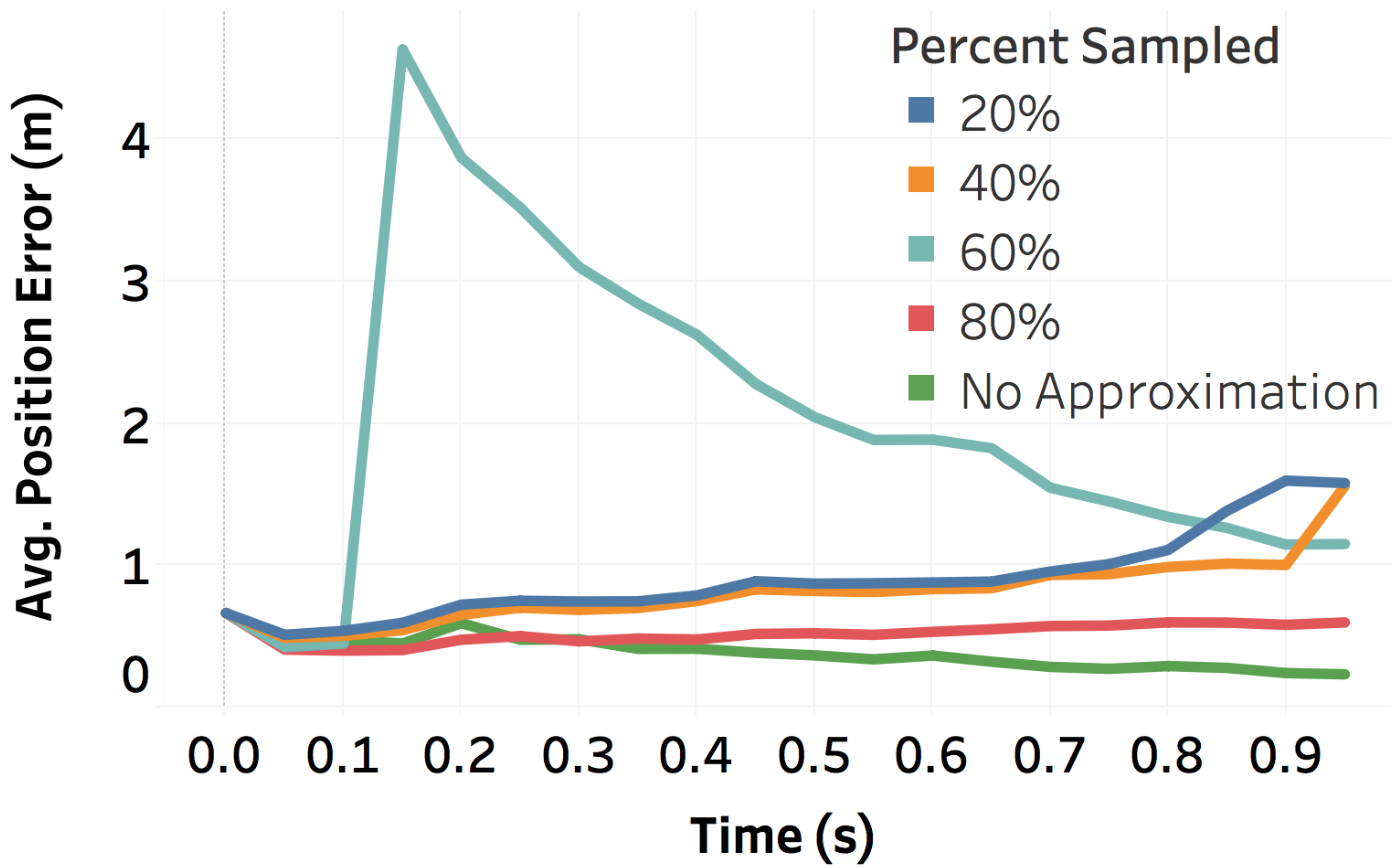
N = # of robots
(D-SLAM)

(M = # of landmarks)

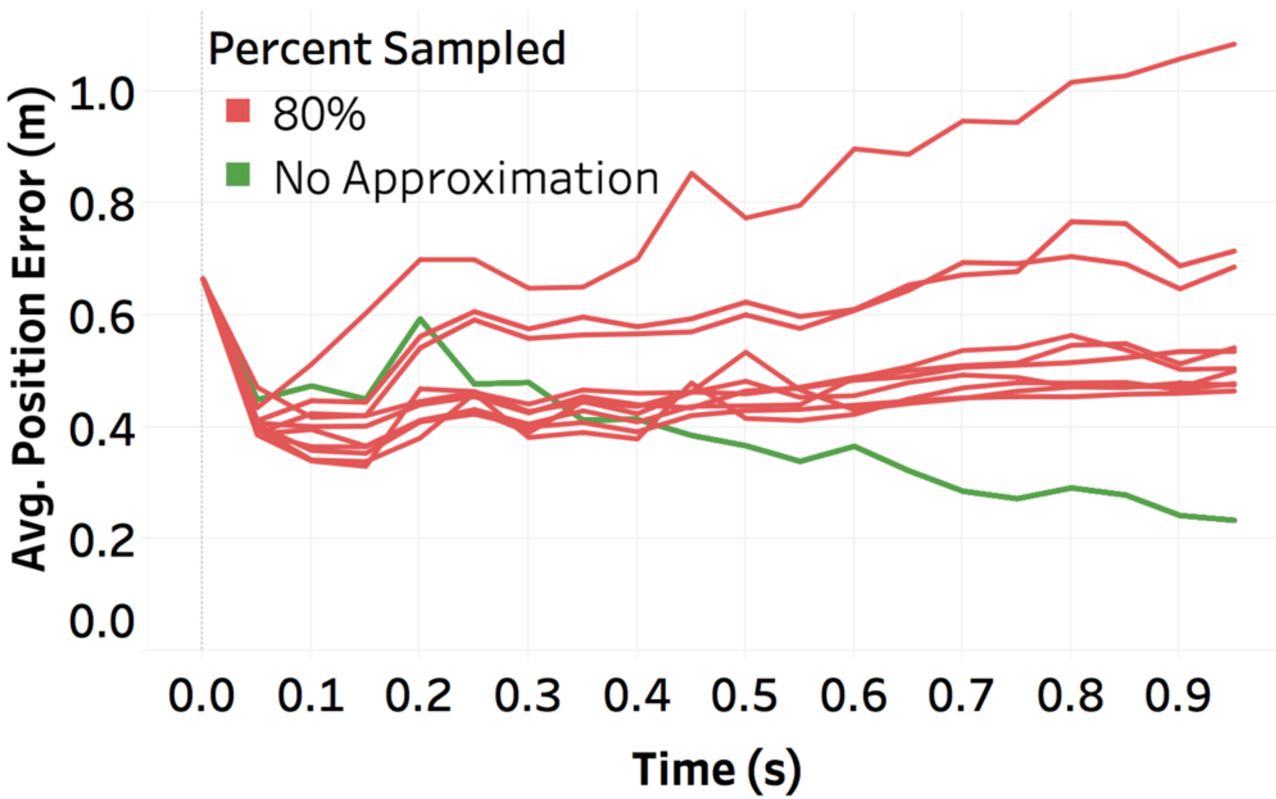
D-SLAM: Position Error over Time



D-SLAM: Position Error over Time



D-SLAM: Per-Trial Position Error



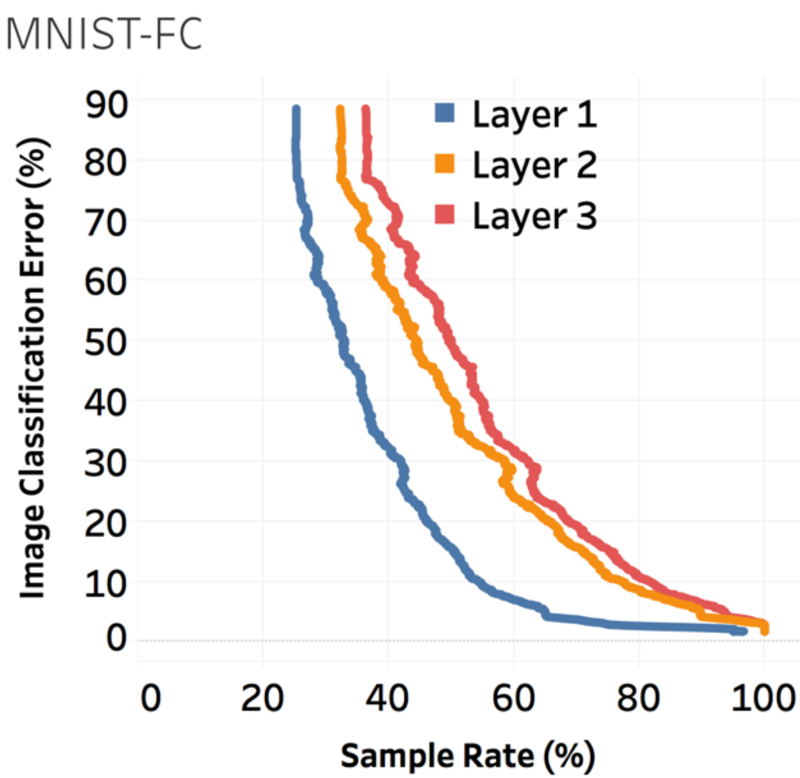
D-SLAM: Results

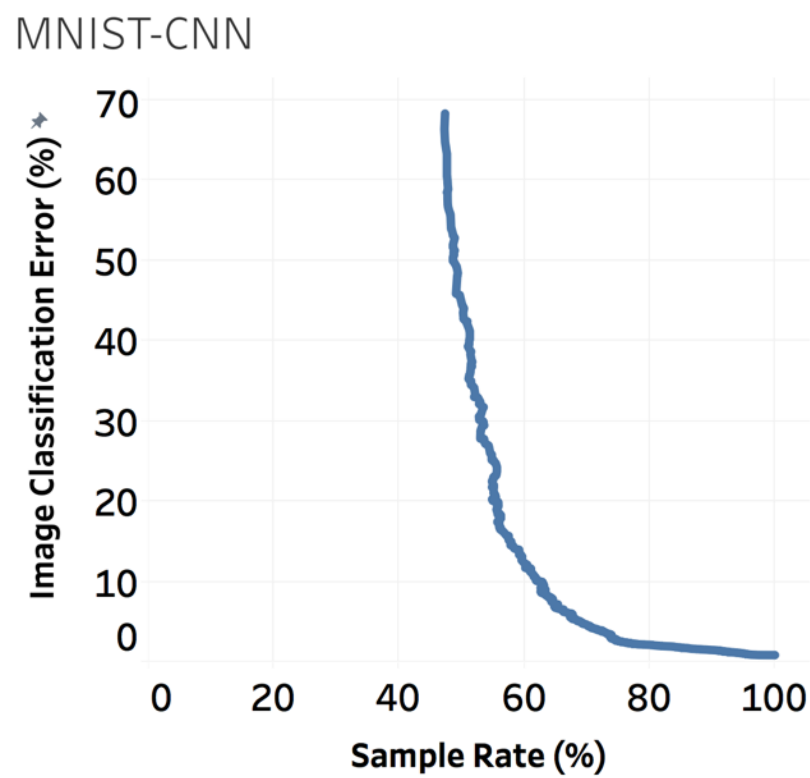
- Variance bad
- But acceptable for some spatial resolutions (~1m)
  - e.g., formation of autonomous drones

Application: Neural Networks

- Known: neural networks are resilient
- Two different networks on MNIST
  - Fully-Connected
  - CNN

Neural Networks: Results





Neural Networks: Results

- Works for certain sampling rates
- Different layers react differently
  - Consistent with reliability studies

Outline

1. Overview of approximate linear algebra
2. Evaluating end-to-end sampling strategies
3. **Predicting end-to-end error bounds**

Why Predict Error Bounds?

- Adaptive runtime control for sampling strategies

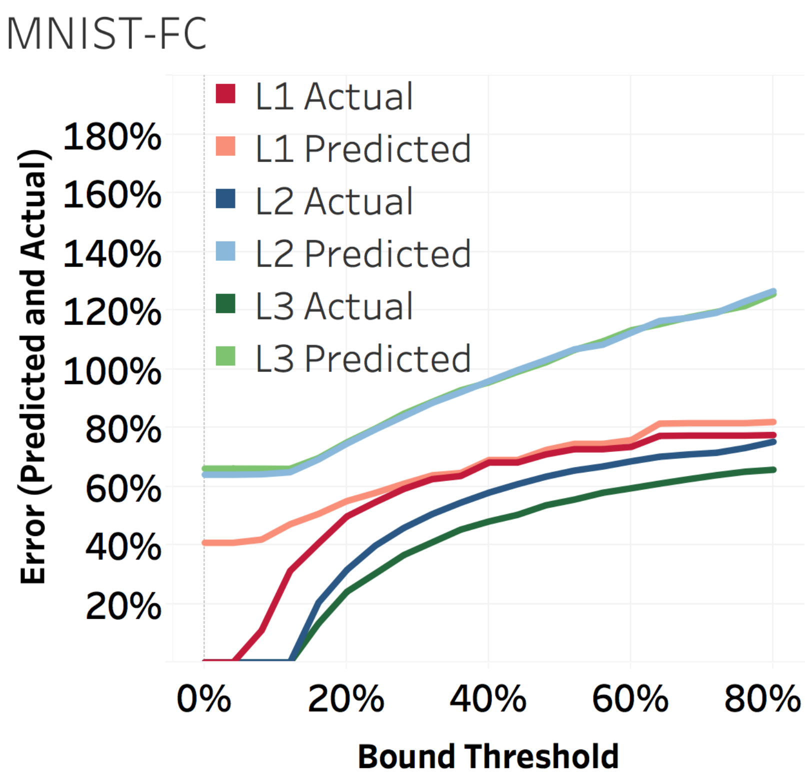
Error Bounds in Practice

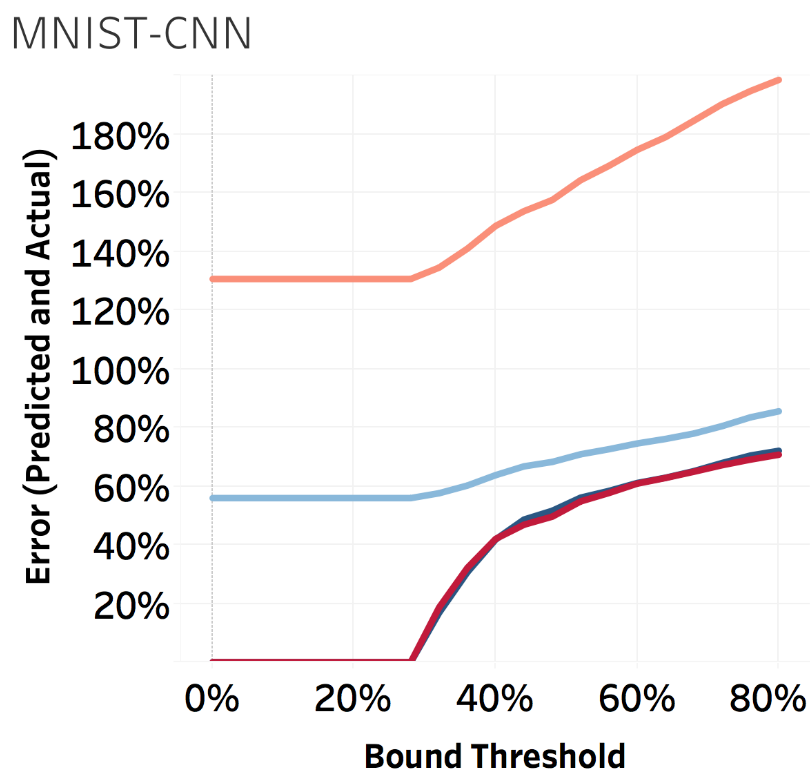
- Asymptotic < Asympotic Relative < Absolute
- Want to skip computation of product AB for bound

D-SLAM: Bounds

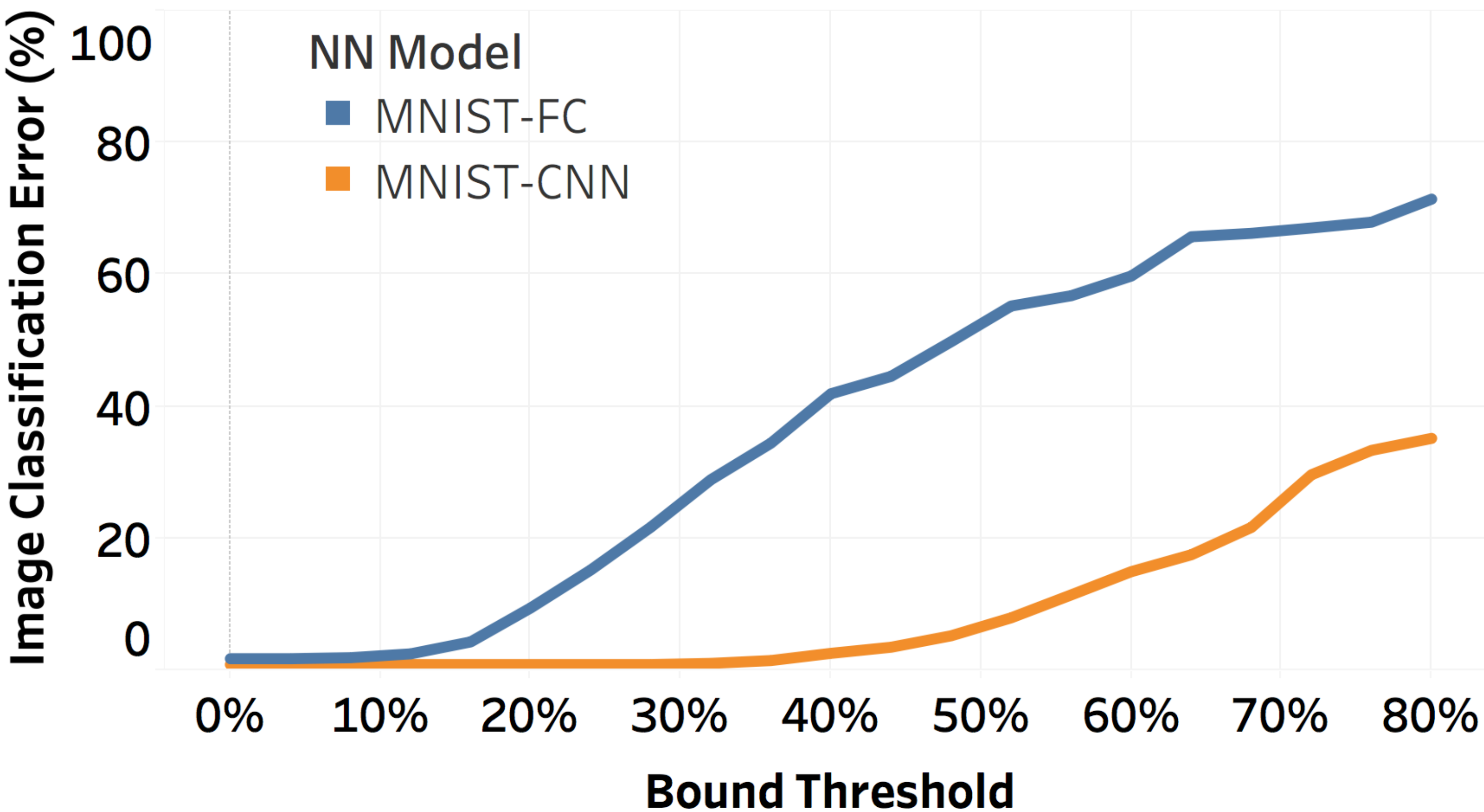
- Too conservative (predicted error ~200%)
- Future work

Neural Networks: Bounds





Neural Networks: Bounds



Outline

1. Overview of approximate linear algebra
2. Evaluating end-to-end sampling strategies
3. Predicting end-to-end error bounds

Future Work

- Other linear algebra approximations
- Fine-tuning adaptive control of approximation

Conclusion

- Practical limitations to applying approximations…
  - Errors cascade in larger systems
  - Global stability
- …but randomized approximation appears promising

Thanks to our sponsors!

This work was partially supported by the Applications Driving Architectures (ADA) Research Center, a JUMP Center co-sponsored by SRC and DARPA, the DARPA DSSoC program, and the National Science Foundation Graduate Research Fellowship (under grant DGE1745303).

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the funding organizations.

Q&A