Fast Computation of Uncertainty in Deep Learning

Mohammad Emtiyaz Khan

Approximate Bayesian Inference Team

RIKEN Center for AI Project, Tokyo

http://emtiyaz.github.io

The Goal of My Research

“To understand the fundamental principles of learning from data and use them to develop algorithms that can learn like living beings.”

Learning by exploring

at the age of 6 months

Converged
at the age
of
12 months

Transfer Learning at 14 months

The Goal of My Research

"To understand the fundamental principles of learning from data and use them to develop algorithms that can learn like living beings."

Human learning $\neq$ Deep learning

Can we fix this?

My current research is focused on reducing this gap.

Approximate Bayesian Inference

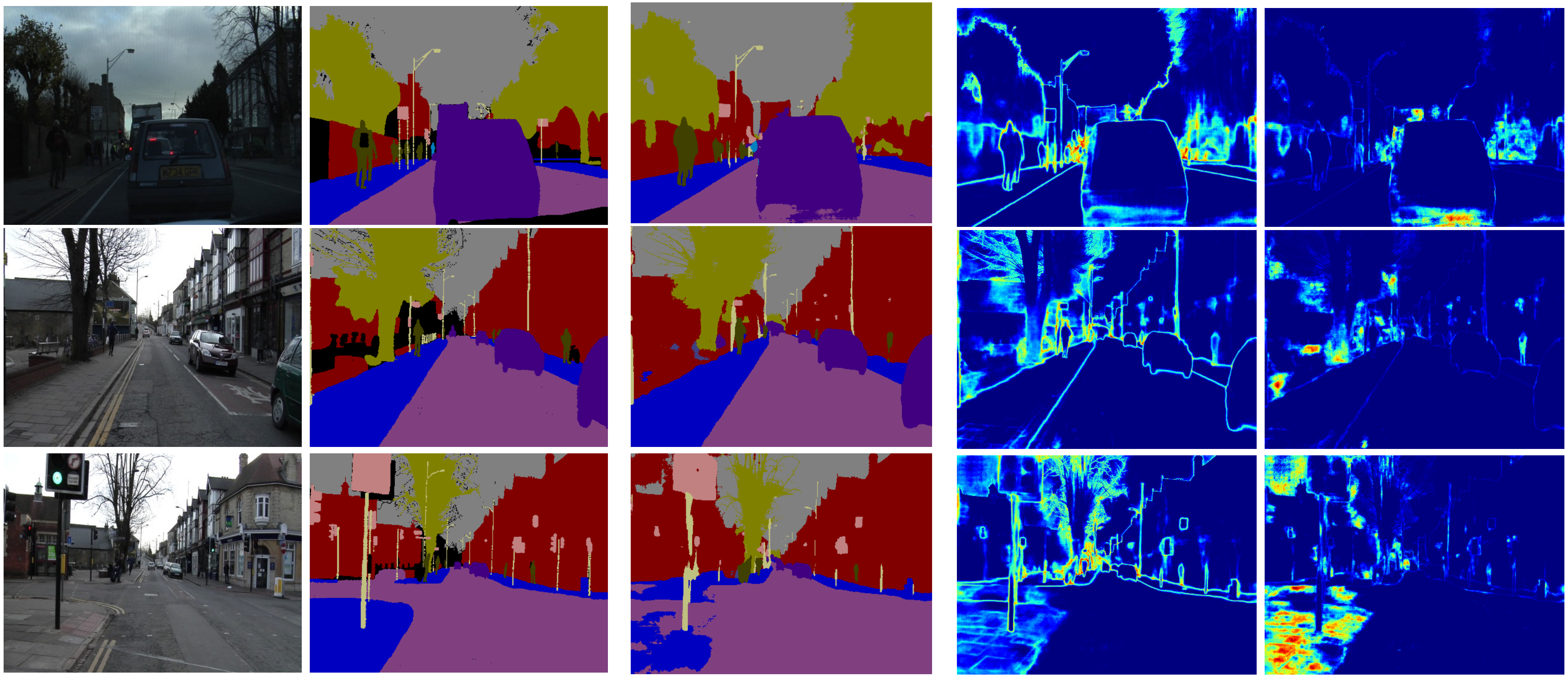
- Bayesian Learning $\approx$ human learning (Tannenbaum 1999)
  - Estimate posterior distribution over unknowns,
  - But computationally very difficult!
- Algorithms that generalize well-known algorithms.
- Natural-Gradient Variational Inference
  - A generalization of least-squares, Newton's method, Expectation Maximization, Kalman filters ….
  - Also deep learning algorithms (Adam).
  - Combines ideas from Bayesian Statistics, Continuous Optimization, Information geometry, Deep Learning.

Uncertainty in Deep Learning

To estimate the confidence in the predictions of a deep-learning system

Uncertainty for Image Segmentation

Image Truth Prediction Uncertainty



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

(taken from Kendall et al. 2017)

Challenges and Solution

The data and model are both extremely large.

Data DNN Parameters

$$\min_{\theta} \ \ell(\mathcal{D}, \theta)$$

Loss

Bayesian solution: Estimate a distribution over theta

$$\max_{\lambda} \ \underbrace{-\mathbb{E}_{q_\lambda(\theta)}[\ell(\mathcal{D}, \theta)] - \mathcal{H}(q)}_{\mathcal{L}(\lambda)}$$

Parameters (e.g., mean and variance)

Distribution (e.g. Gaussian)

Entropy

New Algorithms!

AIstats 2017

Stochastic Gradient Descent:

$$\theta \leftarrow \theta - \rho \nabla_\theta \ell(\theta)$$

Natural Gradient Descent for approximate Inference

$$\lambda \leftarrow \lambda + \rho \nabla_\mu \mathcal{L}$$

Moments of q
(e.g. mean & correlation)

A generalization of least-squares, Newton's method, Expectation Maximization, Kalman filters ….

Variational Adam ICML 2018

Deep learning optimizer (e.g. Adam)

Variational Adam (Vadam)

0. Sample ϵ from a standard normal distribution

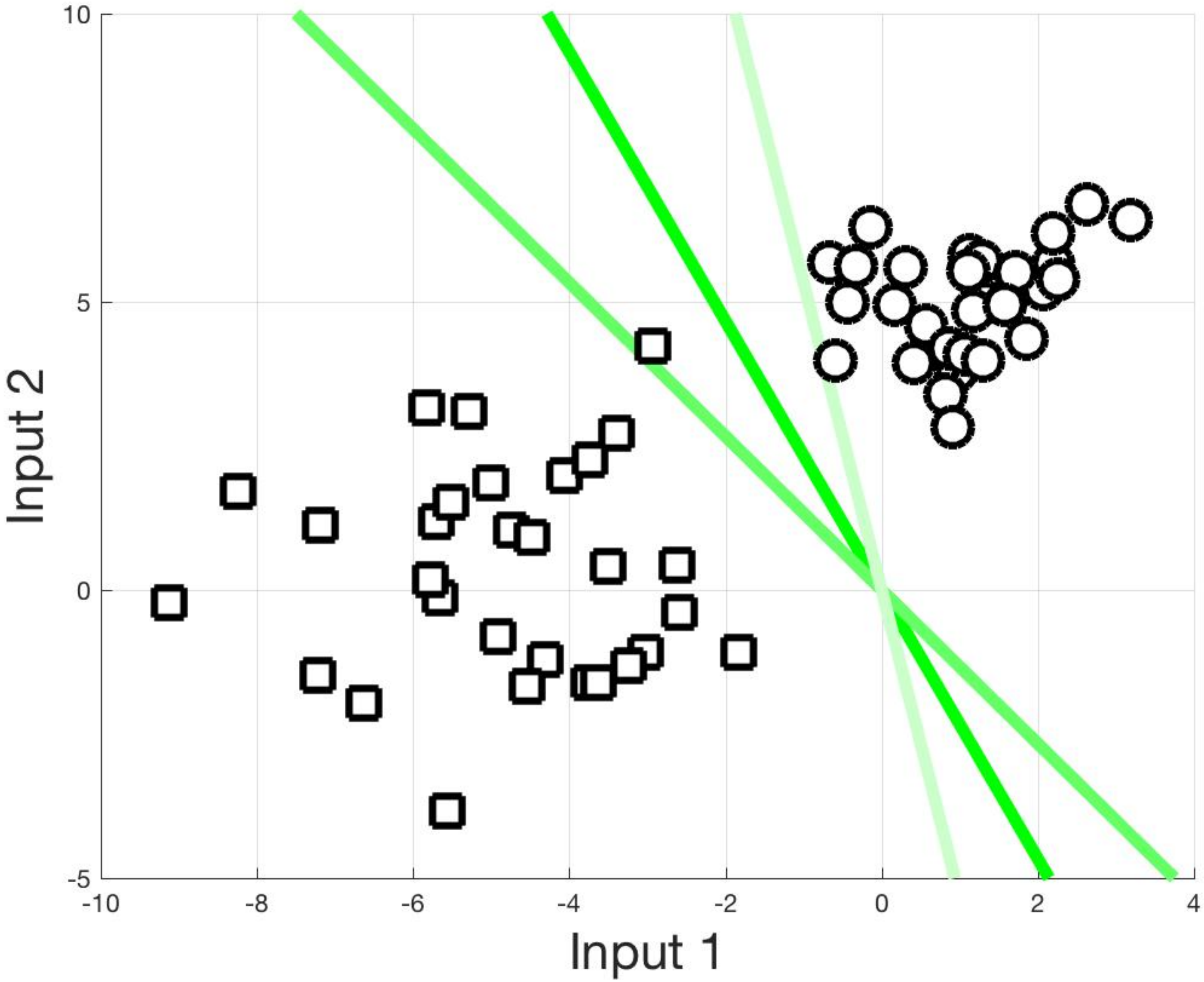
$$\theta_{\text{temp}} \leftarrow \theta + \epsilon * \sqrt{\underbrace{N * \text{scale} + 1}_{\text{Variance}}}$$

1. Select a minibatch
2. Compute gradient using backpropagation
3. Compute a scale vector to adapt the learning rate
4. Take a gradient step

Mean $$\theta \leftarrow \theta + \text{learning\_rate} * \frac{\text{gradient}}{\sqrt{\text{scale}} + 10^{-8}}$$

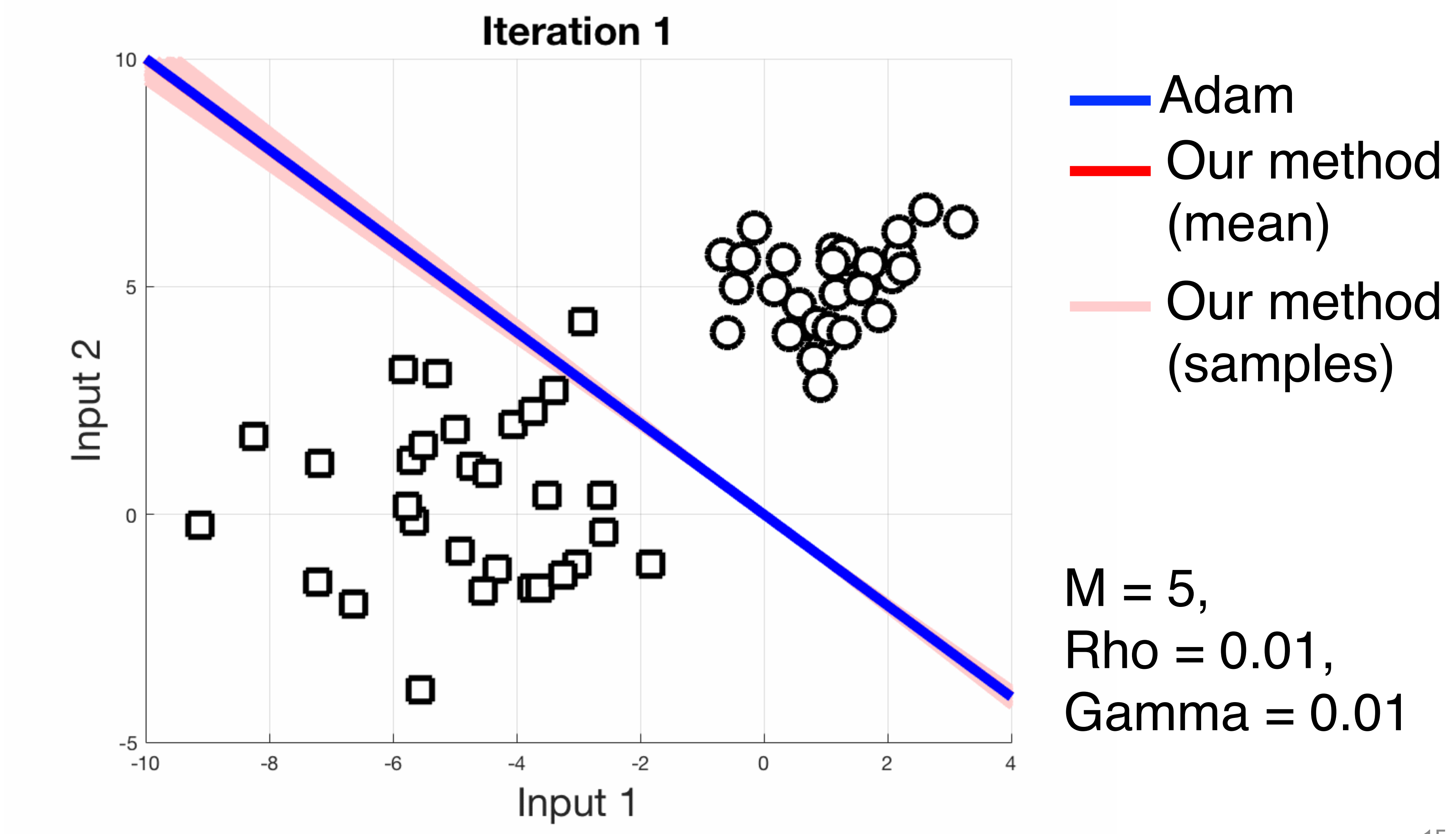
$$\theta \leftarrow \theta + \text{learning\_rate} * \frac{\text{gradient} + \theta/N}{\sqrt{\text{scale}} + 1/N}$$

Illustration: Classification



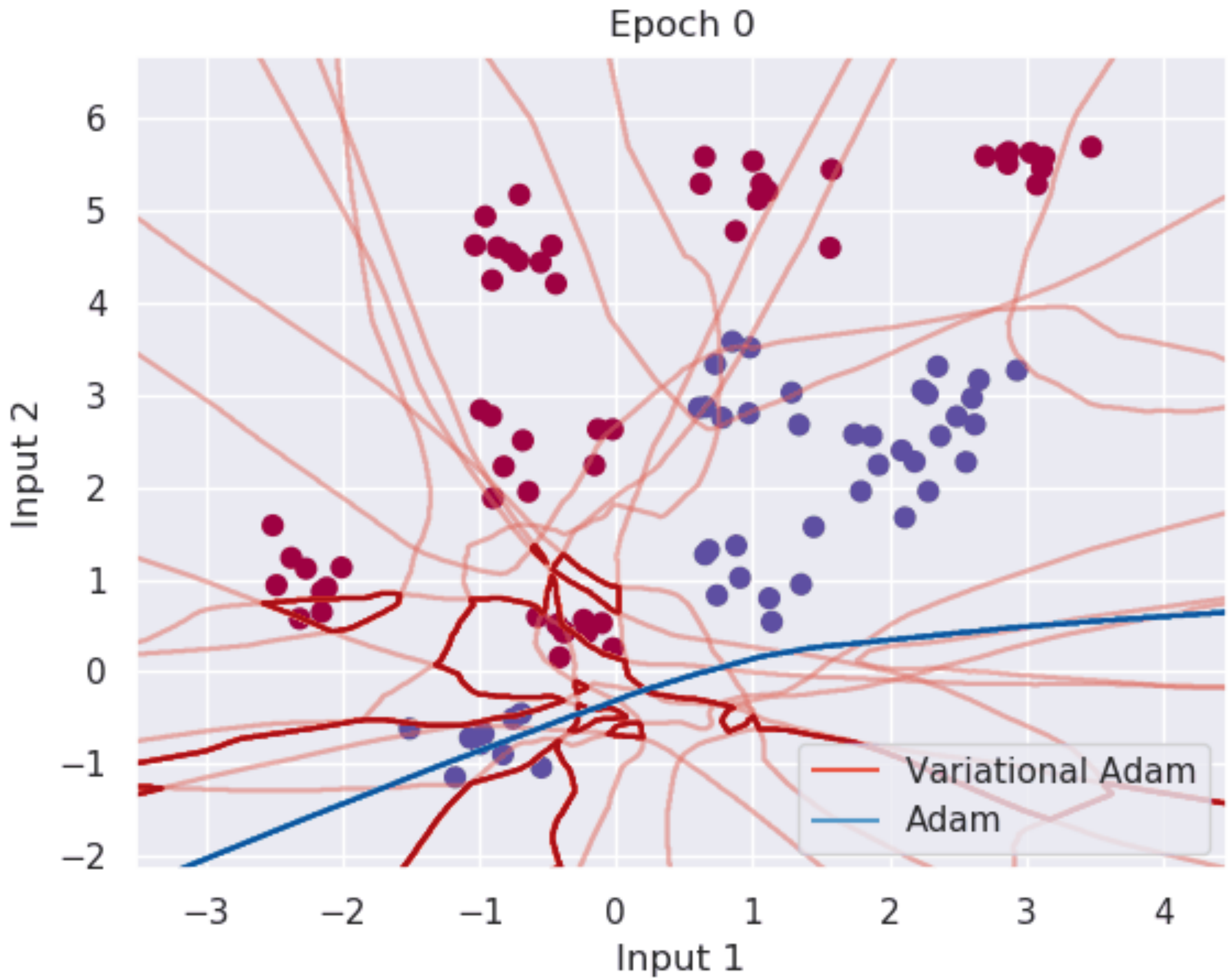
Logistic regression (30 data points, 2 dimensional input). Sampled from Gaussian mixture with 2 components

Adam vs Vadam (on Logistic-Reg)



M = 5,
Rho = 0.01,
Gamma = 0.01

Adam vs Vadam (on Neural Nets)



Adam

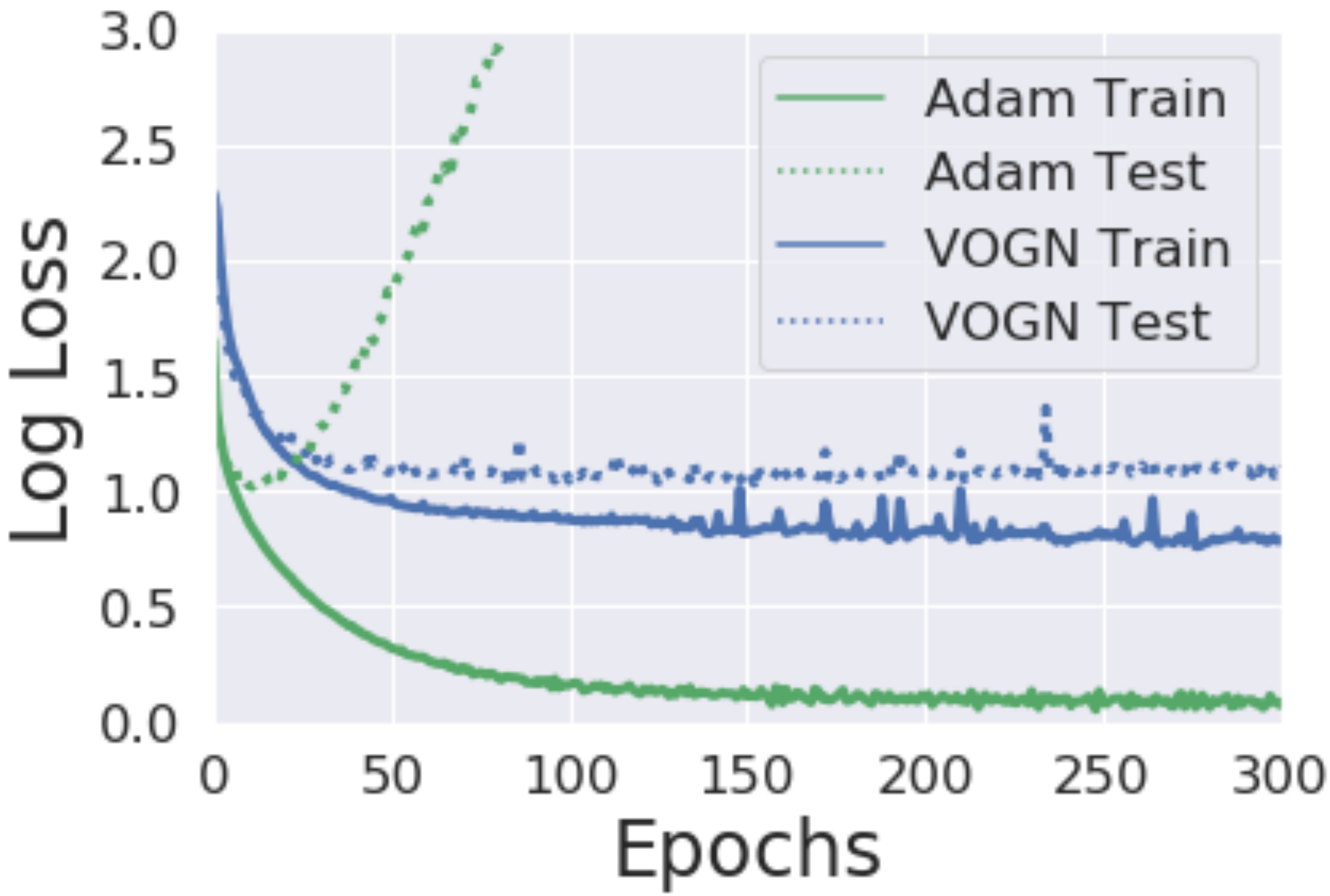
Ours (mean)

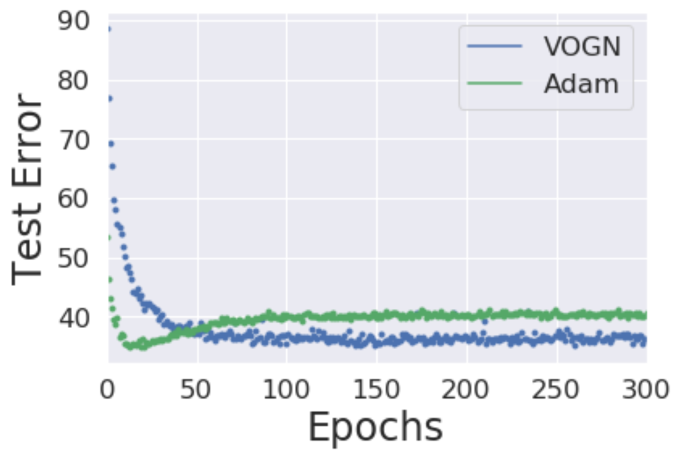
Ours (samples)

(By Runa E.)

LeNet-5 on CIFAR10

VOGN is our method

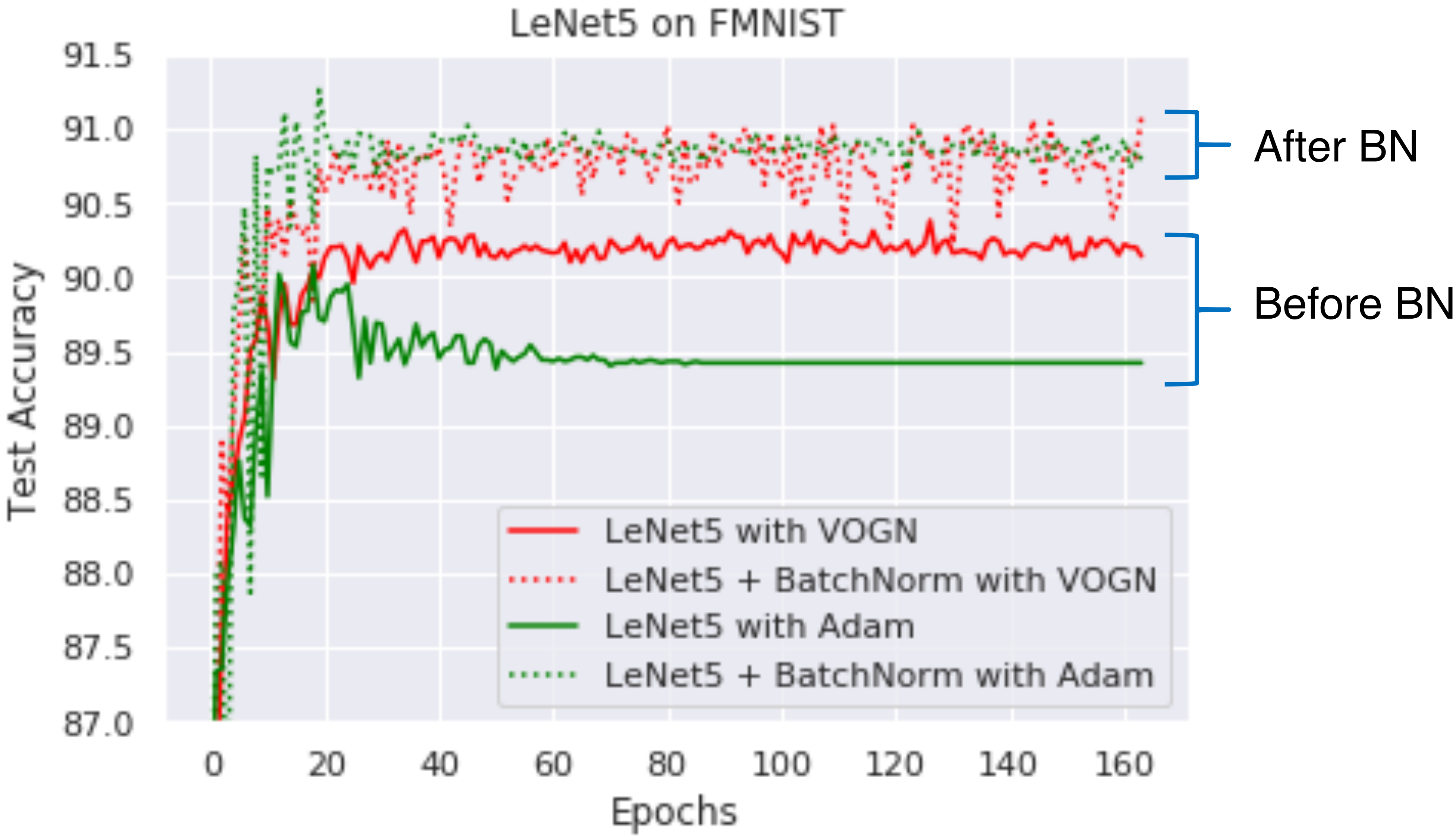




	VOGN	Adam
Log Loss	**1.130**	8.341
Error	**37.01**	40.47

(By Anirudh Jain)

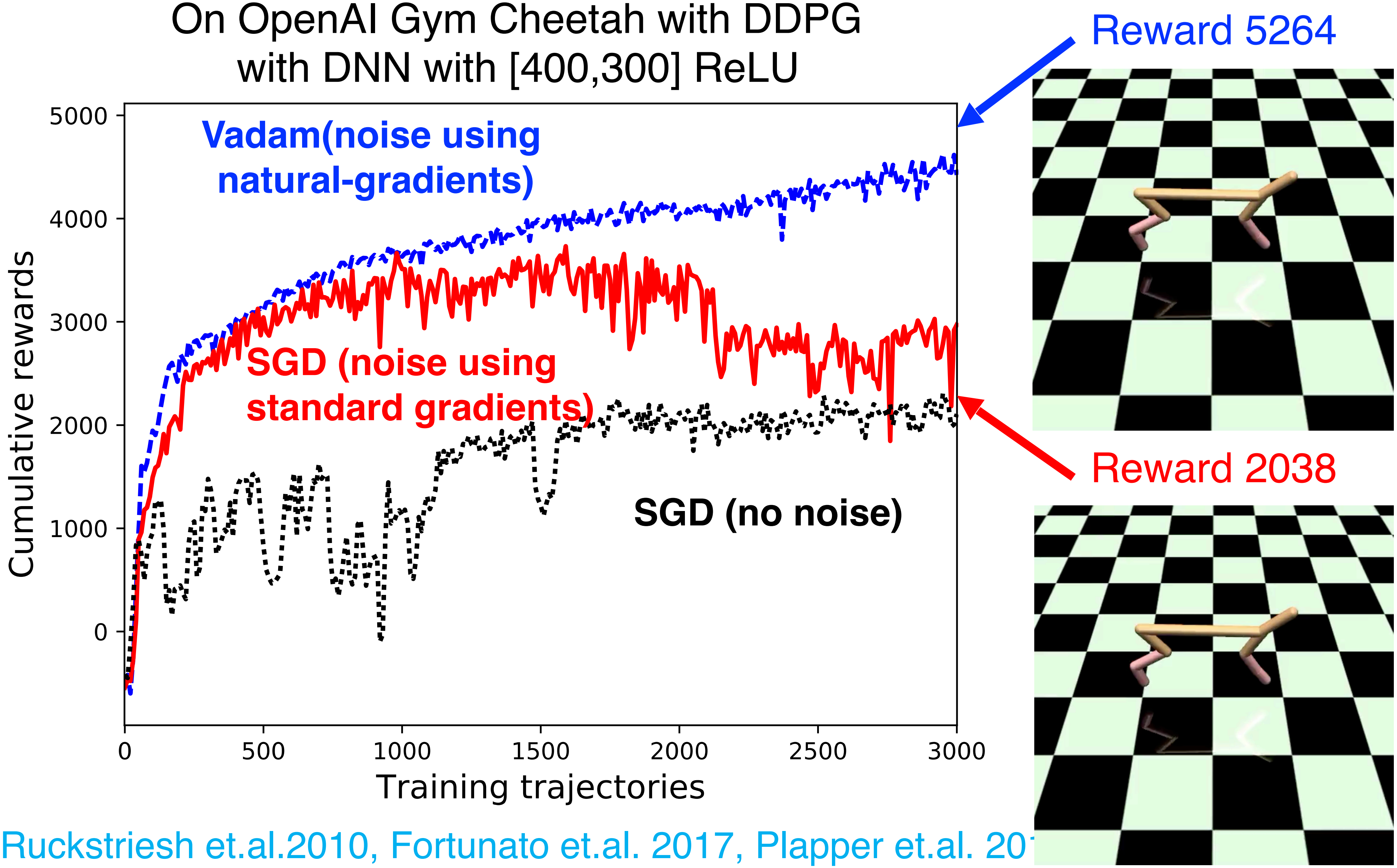
With BatchNorm



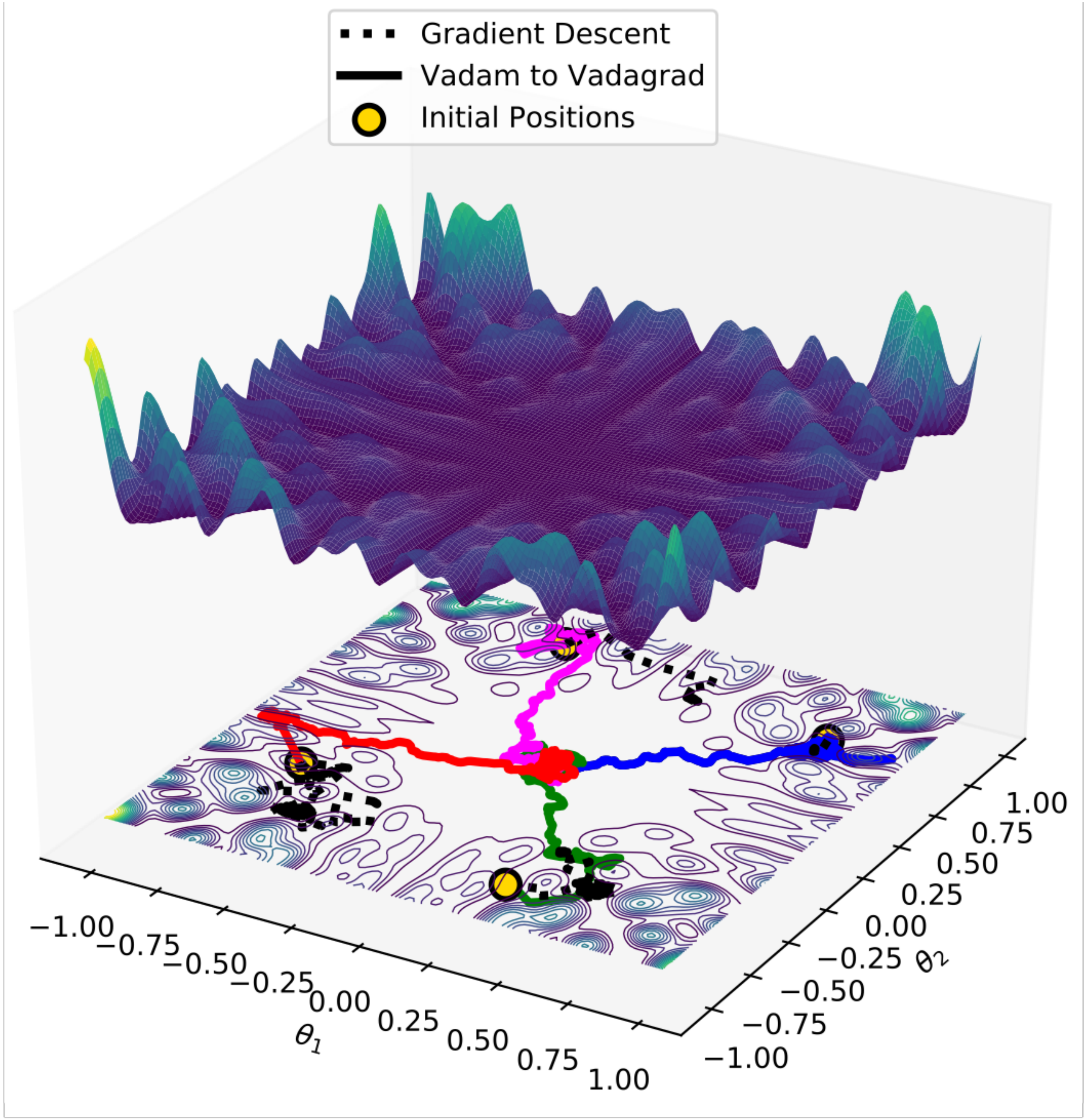
(By Anirudh Jain)

Parameter-Space Noise for Deep RL

On OpenAI Gym Cheetah with DDPG
with DNN with [400,300] ReLU



Ruckstriesh et.al.2010, Fortunato et.al. 2017, Plapper et.al. 20



Avoiding Local Minima

An example taken from Casella and Robert's book.

Vadam reaches the flat minima, but GD gets stuck at a local minima.

Optimization by smoothing, Gaussian homotopy/blurring etc., Entropy SGLD etc.

Summary of the Talk

- Approximate Bayesian inference
  - Uncertainty computation in deep learning
  - Generalization of many well-known algorithms
  - Works for deep nets.
- Generalizations and Extensions,
  - VAEs, Mixture of Exponential Family, Evolution strategy etc.
  - Convergence and regret bounds.

On-Going Work

- Very large problems (Imagenet)
- Built-in optimizer in PyTorch
- Modifications to enable online/continual learning
  - Theory of life-long learning
- Posterior approximations using DNNs
- Active learning
- Reinforcement Learning

Collaboration Areas

- Applications of deep learning
  - Computer vision, NLP, Audio, Multimodal data
- Interpretable/explainable/causal models
- Sequential learning
  - Continual learning, Active learning, reinforcement learning, online learning.
  - Generalization bounds
- Discrete optimization/ nonconvex optimization

Related Works

- Sato (1998), *Fast Learning of On-line EM Algorithm.*
- Sato (2001), *Online Model Selection Based on the Variational Bayes*.
- Jordan et al. (1999), *An Introduction to Variational Methods for Graphical Models*.
- Winn and Bishop (2005), *Variational Message Passing*.
- Honkela et al. (2007), *Natural Conjugate Gradient in Variational Inference*.
- Honkela et al. (2010), *Approximate Riemannian Conjugate Gradient Learning for Fixed-Form Variational Bayes*.
- Knowles and Minka (2011), *Non-conjugate Variational Message Passing for Multinomial and Binary Regression*.
- Hensman et al. (2012), *Fast Variational Inference in the Conjugate Exponential Family*.
- Hoffman et al. (2013), *Stochastic Variational Inference*.
- Salimans and Knowles (2013), *Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression*.
- Seth and Khardon (2016), *Monte Carlo Structured SVI for Two-Level Non-Conjugate Models*.
- Salimbani et al. (2018), *Natural Gradients in Practice: Non-Conjugate Variational Inference in Gaussian Process Models.*
- Zhang et al. (2018), *Noisy Natural Gradient as Variational Inference*

References

Available at https://emtiyaz.github.io/publications.html

Conjugate-Computation Variational Inference : Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models, (AIstats 2017) **M.E. Khan** and W. Lin [Paper] [Code

Faster Stochastic Variational Inference using Proximal-Gradient Methods with General Divergence Functions, (UAI 2016) **M.E. Khan**, R. Babanezhad, W. Lin, M. Schmidt, M. Sugiyama [Paper + Appendix] [Code]

References

Available at https://emtiyaz.github.io/publications.html

Variational Message Passing with Structured Inference Networks,
(ICLR 2018) W. Lin, N. Hubacher, and **M.E. Khan**, [Paper] [ArXiv Version]

Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam,
(ICML 2018) **M.E. Khan**, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, and A. Srivastava, [ArXiv Version] [Code] [Slides]

Fast yet Simple Natural-Gradient Descent for Variational Inference in Complex Models,
Invited paper at (ISITA 2018) **M.E. Khan** and D. Nielsen, [Pre-print]

SLANG: Fast Structured Covariance Approximations for Bayesian Deep Learning with Natural Gradient,
(NIPS 2018) A. Miskin, F. Kunstner, D. Nielsen, M. Schmidt, **M.E. Khan**.

Fast and Simple Natural-Gradient Variatioinal Inference with Mixture of Exponential Family,
(Under Submission) W. Lin, M. Schmidt, **M.E. Khan**.

A 5 page review

Fast yet Simple Natural-Gradient Descent for Variational Inference in Complex Models

Mohammad Emtiyaz Khan
RIKEN Center for Advanced Intelligence Project
Tokyo, Japan
emtiyaz.khan@riken.jp

Didrik Nielsen
RIKEN Center for Advanced Intelligence Project
Tokyo, Japan
didrik.nielsen@riken.jp

***Abstract*—Bayesian inference plays an important role in advancing machine learning, but faces computational challenges when applied to complex models such as deep neural networks. Variational inference circumvents these challenges by formulating Bayesian inference as an optimization problem and solving it using gradient-based optimization. In this paper, we argue in favor of *natural-gradient* approaches which, unlike their *gradient-based* counterparts, can improve convergence by exploiting the information geometry of the solutions. We show how to derive fast yet simple natural-gradient updates by using a duality associated with exponential-family distributions. An attractive feature of these methods is that, by using natural-gradients, they are able to extract accurate local approximations for individual model components. We summarize recent results for Bayesian deep learning showing the superiority of natural-gradient approaches over their gradient counterparts.**

***Index Terms*—Bayesian inference, variational inference, natural gradients, stochastic gradients, information geometry, exponential-family distributions, nonconjugate models.**

prove the rate of convergence [7]–[9]. Unfortunately, these approaches only apply to a restricted class of models known as *conditionally-conjugate* models, and do not work for non-conjugate models such as Bayesian neural networks.

This paper discusses some recent methods that generalize the use of natural gradients to such large and complex non-conjugate models. We show that, for exponential-family approximations, a duality between their natural and expectation parameter-spaces enables a simple natural-gradient update. The resulting updates are equivalent to a recently proposed method called Conjugate-computation Variational Inference (CVI) [10]. An attractive feature of the method is that it naturally obtains *local* exponential-family approximations for individual model components. We discuss the application of the CVI method to Bayesian neural networks and show some recent results from a recent work [11] demonstrating

Acknowledgement

- RIKEN AIP
  - Wu Lin (now at UBC), Didrik Nielsen (now at DTU), Voot Tangkaratt, Nicolas Hubacher, Masashi Sugiyama, Shunichi-Amari.
- Interns at RIKEN AIP
  - Zuozhu Liu (SUTD, Singapore), Aaron Mishkin (UBC), Frederik Kunstner (EPFL).
- Collaborators
  - Mark Schmidt (UBC), Yarin Gal (University of Oxford), Akash Srivastava (University of Edinburgh), Reza Babanezhad (UBC).

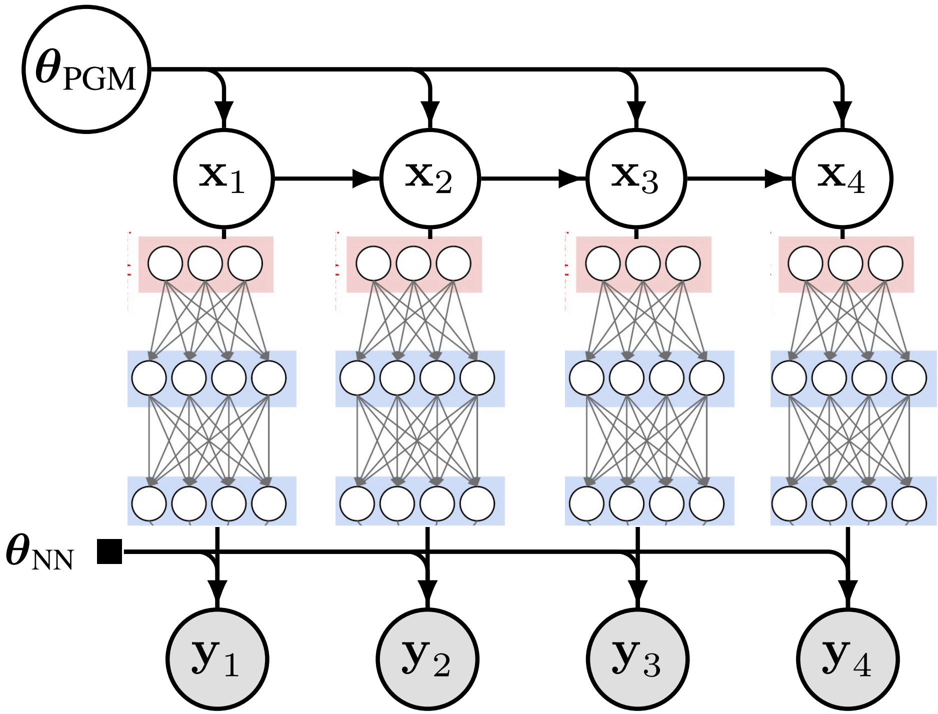
Thanks!

Slides, papers, and code available at
https://emtiyaz.github.io

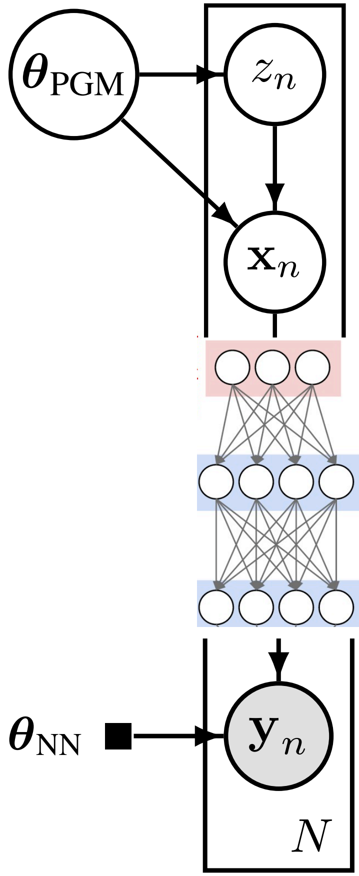
Generalization and Extensions

Deep Nets + Graphical Models

Neural Nets +
Linear Dynamical System



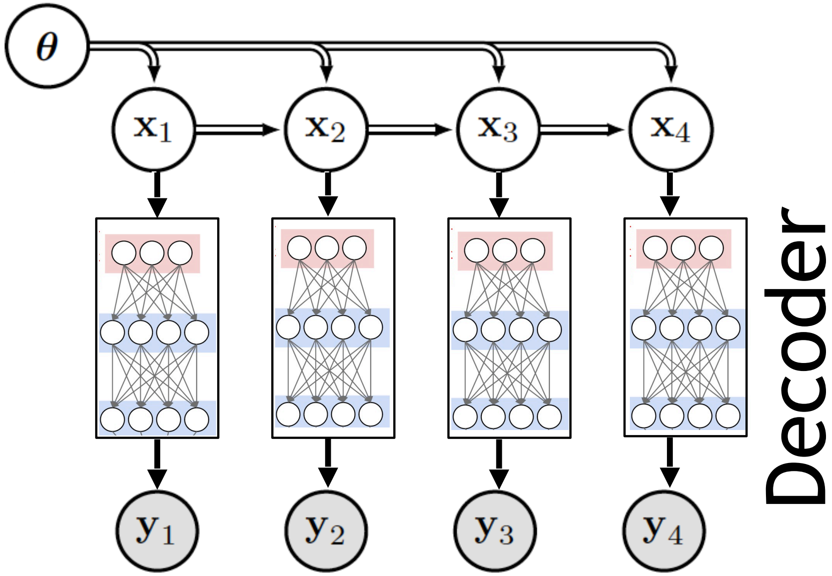
Neural Nets + GMM



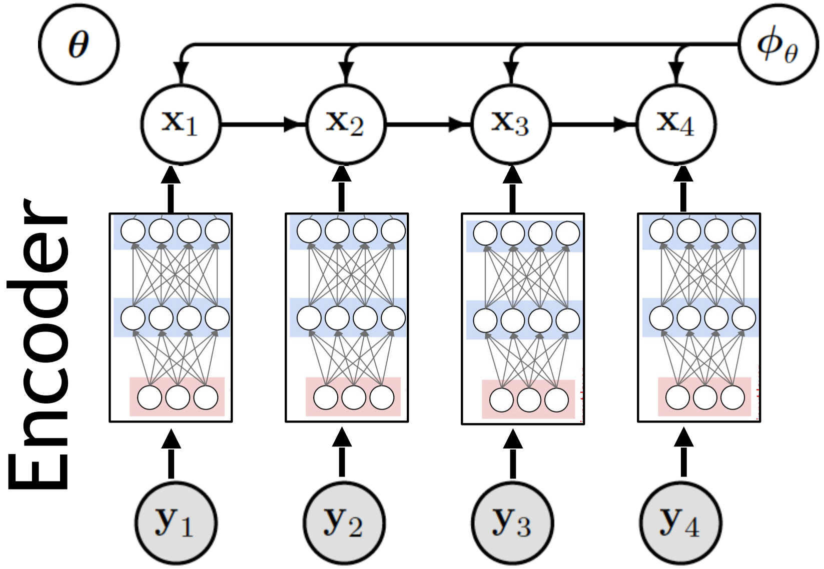
Amortized Inference on VAE + Probabilistic Graphical Models (PGM)

ICLR 2018

Graphical model + Deep Model



Structured Inference Network



Backprop on DNN, and forward-backward on PGM.

Going Beyond Exponential Family

- Fast and Simple NGD for approximations outside exponential family (under submission),
  - Scale mixture of Gaussians, e.g., T-distribution,
  - Finite mixture of Gaussian,
  - Matrix Variate Gaussian,
  - Skew-Gaussians.
- The updates can be implemented using message passing and back-propagation.

Convergence Rates

UAI 2016

Lipschitz constant of (nonconvex) ELBO

Gradient noise variance

$$\mathbb{E}\left[\|(\lambda_k - \lambda_{k+1})/\rho\|^2\right] \leq \left[\frac{2LC_0}{\alpha_*^2 t} + \frac{c\sigma^2}{M\alpha_*}\right]$$

Strong convexity of the Fisher Information Matrix

Mini-batch size

See Khan et al. UAI 2016. The proof is based on Ghadimi, Lan, and Zhang (2014)

Bound Generalization Error

ICLR 2018

$$\sum_{t=1}^{T} \ell_t(\hat{\theta}_t) \leq \inf_{\mu \in \mathcal{M}} \left\{ \mathbb{E}_{\theta \sim q_\mu} \left[\sum_{t=1}^{T} \ell_t(\theta) \right] + \frac{\eta L^2 T}{\alpha} + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \right\}.$$