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 Our Research

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

Human Learning:

At the age of 6 months, learning by actively and sequentially collecting limited and correlated data.

Converged

at the age

of

12 months

Transfer Knowledge at the age of 14 months

Human learning $\neq$ Deep learning

Humans can learn from limited, sequential, correlated data, with a clear understanding of the world.

Machines require large amount of IID data, and don’t really understand the world and cannot reason about it.

Our current research focuses on reducing this gap!

Approximate Bayesian Inference

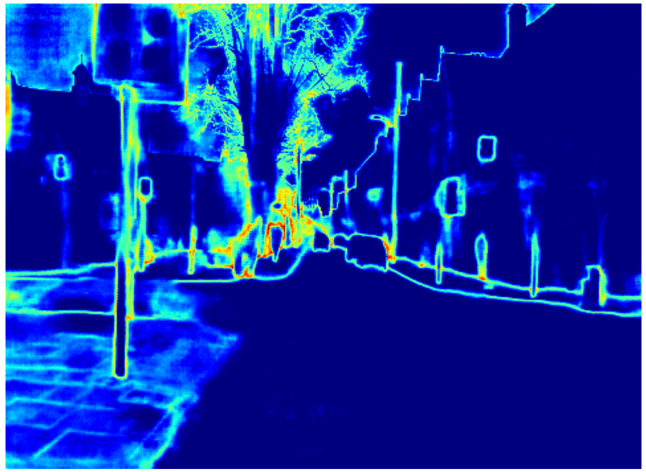
- Bayesian Learning $\approx$ human learning (Tannenbaum 1999)
  - But computationally very difficult!
- Scalable approximation algorithms
  - with principles of human learning
  - while generalizing existing algorithms.
- Today's talk
  - New deep-learning algorithms that "know how much they don't know" (uncertainty).

Uncertainty in Deep Learning

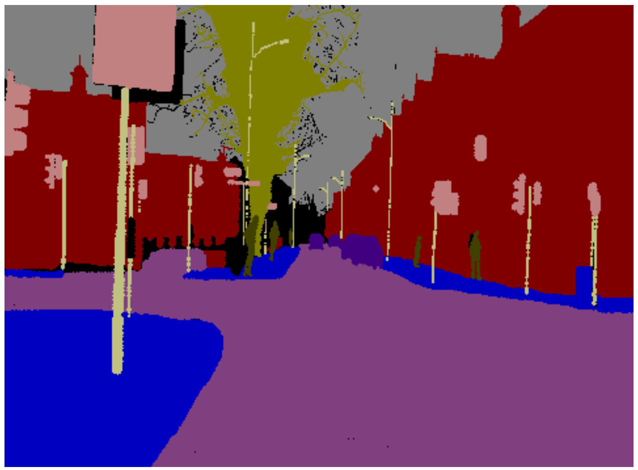
(by Kendall et al. 2017)

Image

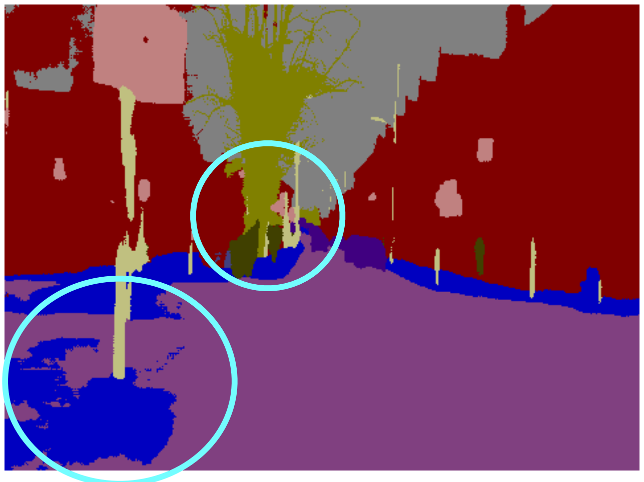
Uncertainty

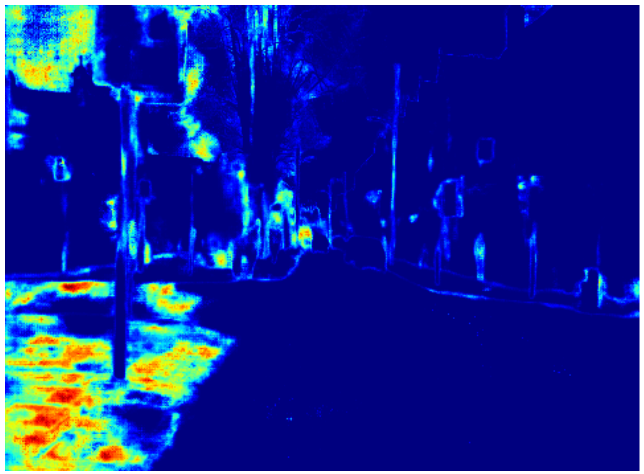


True Segments



Prediction





Challenges

The data and model are both extremely large.

Data DNN Parameters

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

A simple solution (ensemble method):
- Predict using multiple networks.
- Where they agree, we are more certain.
- Where they disagree, we are less certain.

This is very expensive!

A Bayesian Solution

- Estimate a distribution over model parameters.
- Draw multiple networks from the distribution.

Distribution (e.g. Gaussian)

Entropy

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

Parameters
(e.g., mean and variance)

Rest of the talk: Estimate mean and variance when training just "one" (or a few) deep network.

Contribution I : CVI

(Khan and Lin, Conjugate-Computation VI, AIstats 2017)

Deep Learning: SGD

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

Bayesian Deep Learning: CVI

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

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

CVI is a generalization of many existing algorithms: least-squares, Newton's method, EM, Kalman filters, HMM, Forward-backward,…. and SGD.

Contribution II : Vadam and VOGN

(Khan et al., Fast and scalable Bayesian deep learning, ICML 2018)

Deep learning optimizer (e.g. Adam)

Vadam/VOGN

0. Sample ϵ from a standard normal distribution

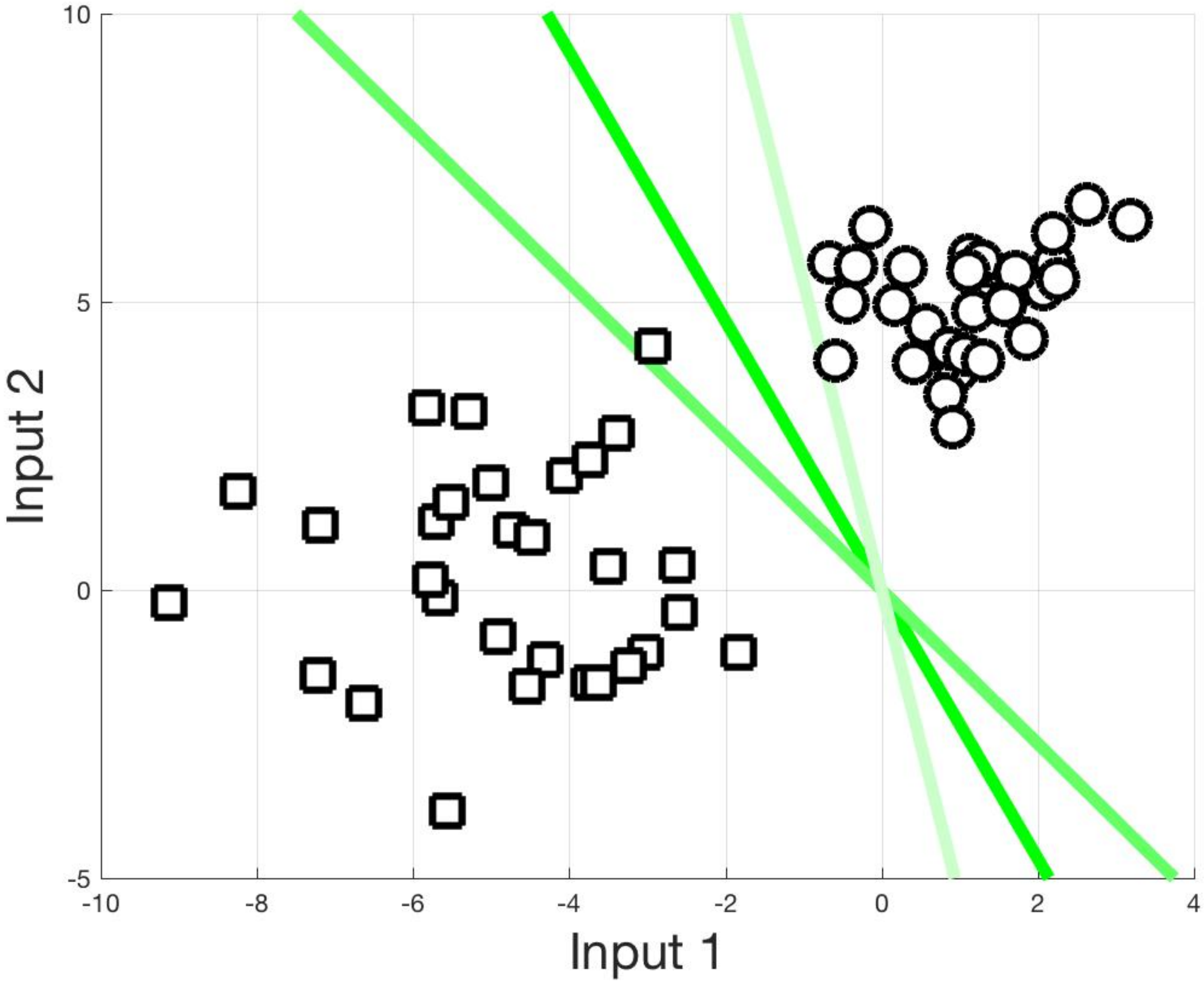
$$\theta_{\text{temp}} \leftarrow \theta + \epsilon * \underbrace{\sqrt{N * \text{scale} + 1}}_{\text{Variance}}$$

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

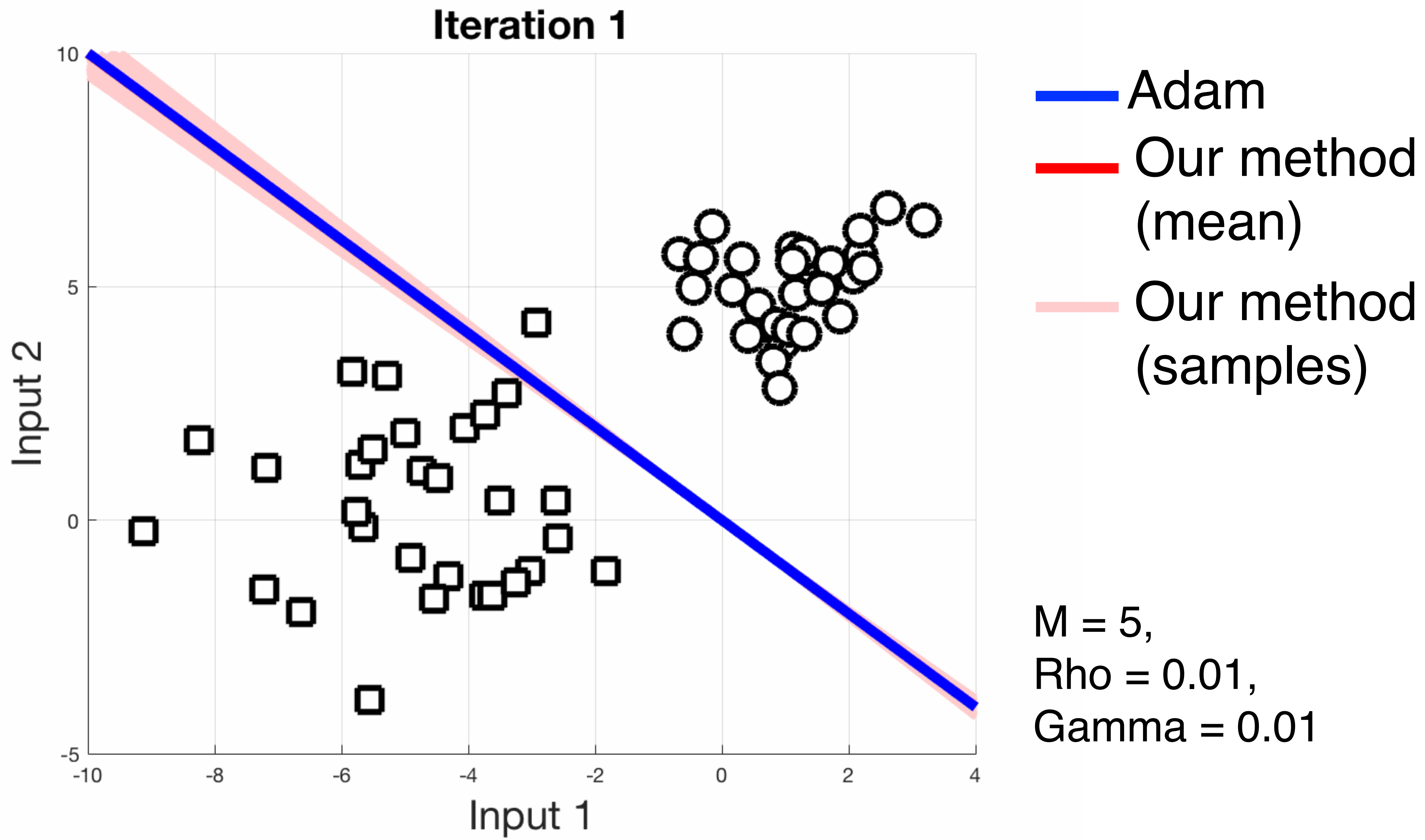
$$\theta \leftarrow \theta + \text{learning\_rate} * \frac{\text{gradient} + \theta/N}{\sqrt{\text{scale}} + 1/N}$$

Illustration: Classification



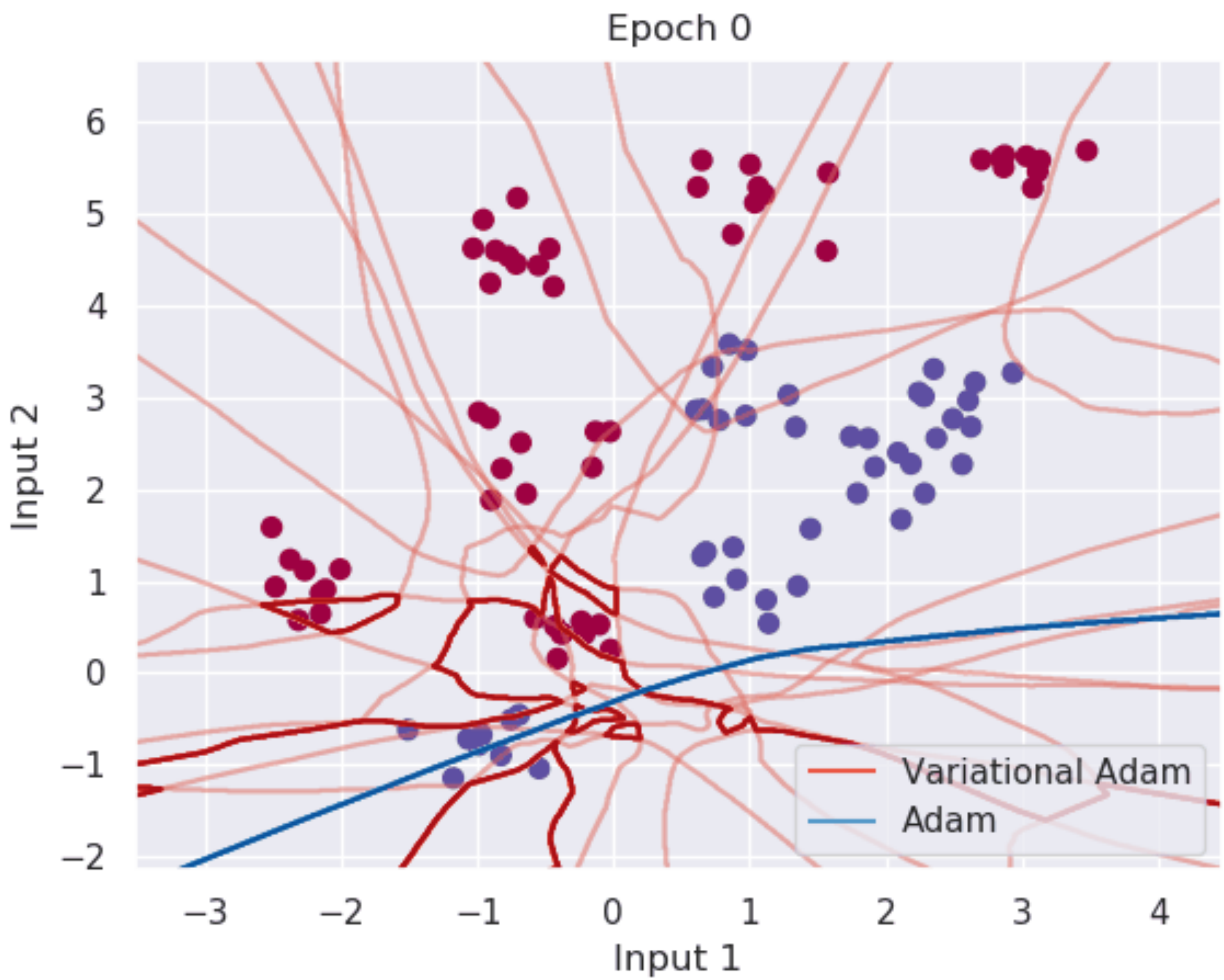
Logistic regression (30 data points, 2 dimensional input). Sampled from Gaussian mixture with 2 components

Adam vs Our Method (on Logistic-Reg)



M = 5,
Rho = 0.01,
Gamma = 0.01

Adam vs Our Method (on Neural Nets)



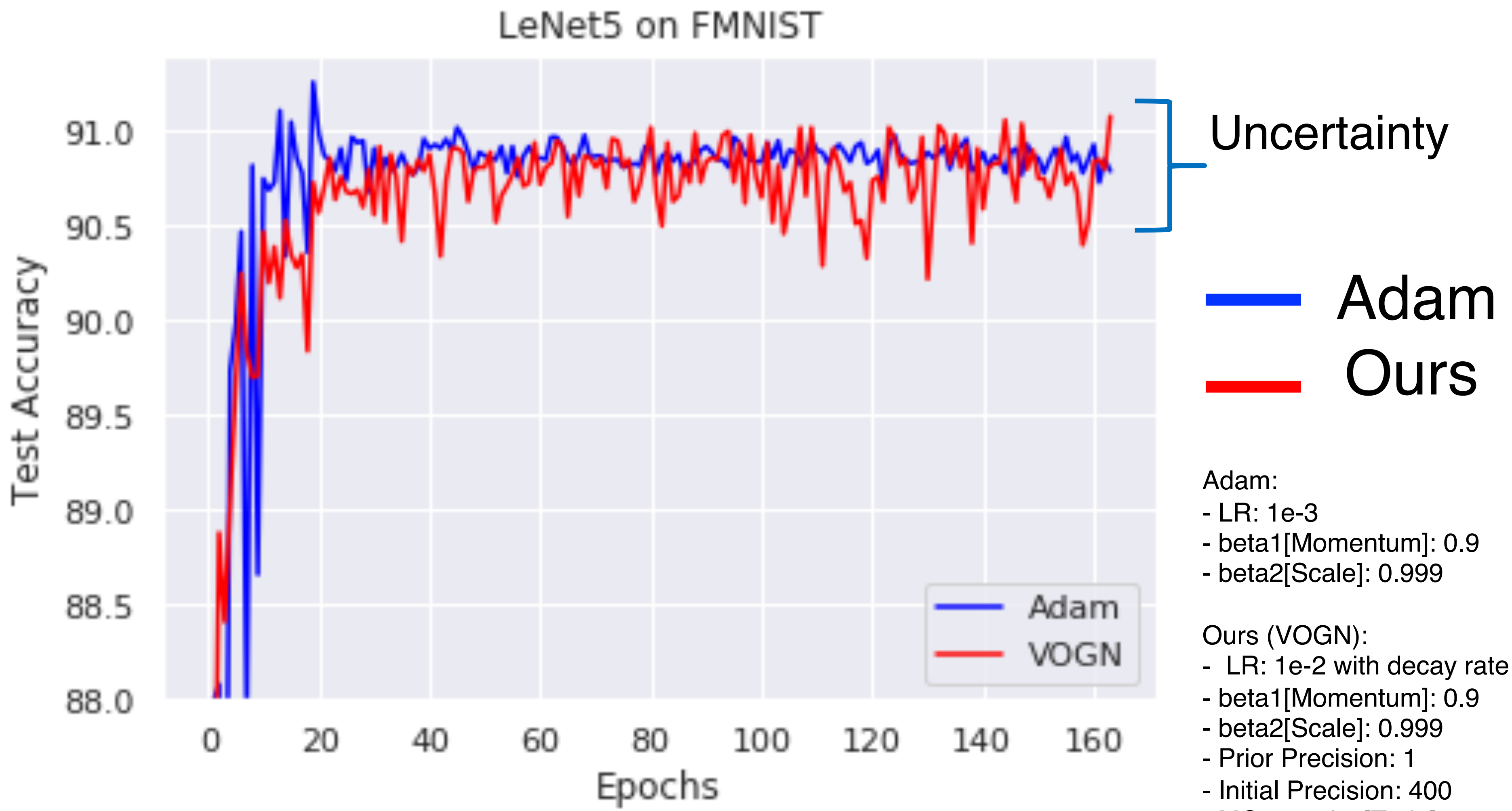
Adam

Ours (mean)

Ours (samples)

(By Runa E.)

Adam vs Our Method (Real Data)



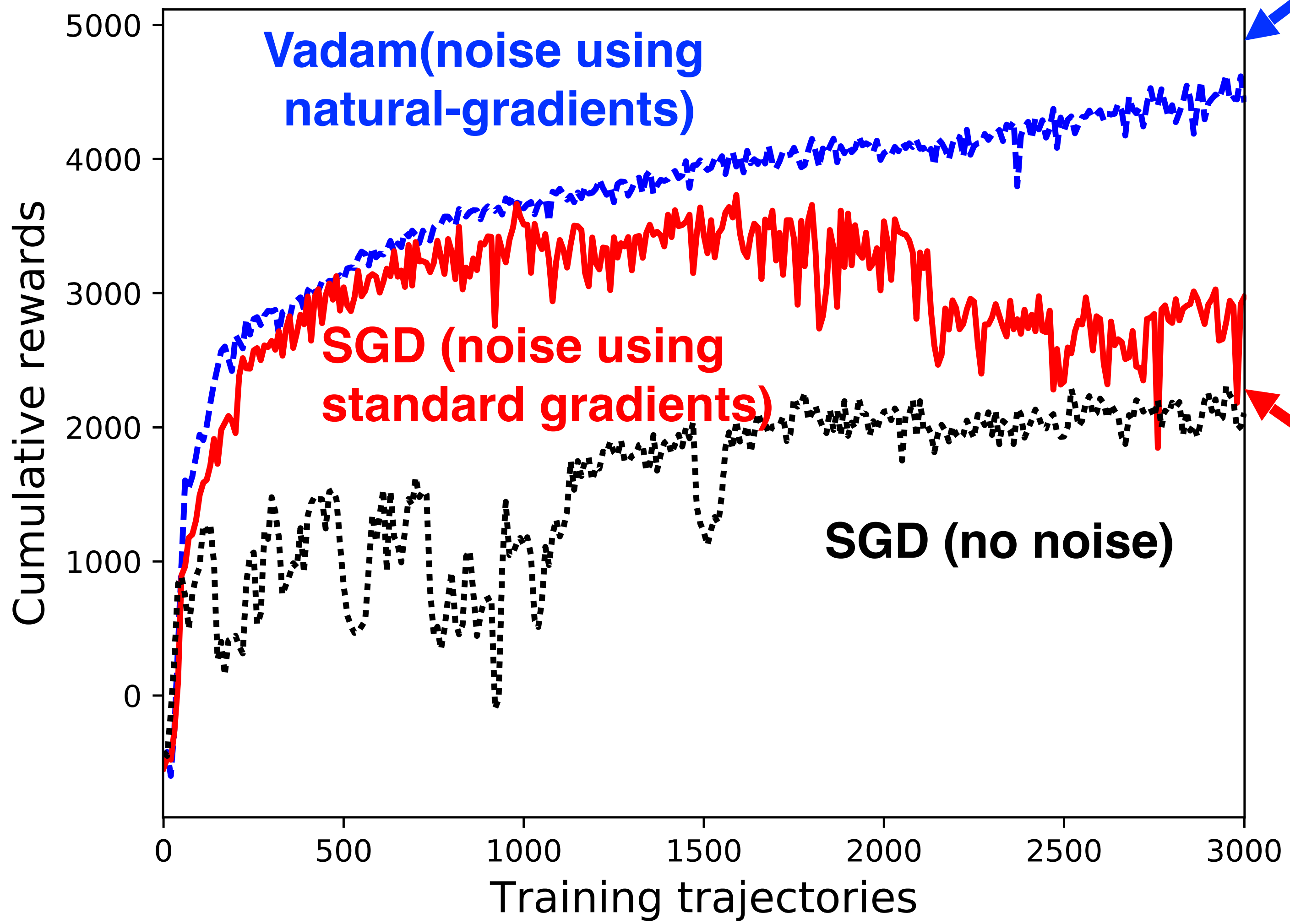
Adam:
- LR: 1e-3
- beta1[Momentum]: 0.9
- beta2[Scale]: 0.999

Ours (VOGN):
- LR: 1e-2 with decay rate: 0.9
- beta1[Momentum]: 0.9
- beta2[Scale]: 0.999
- Prior Precision: 1
- Initial Precision: 400
- MC samples[Train]: 10
- MC samples[Test]: 100

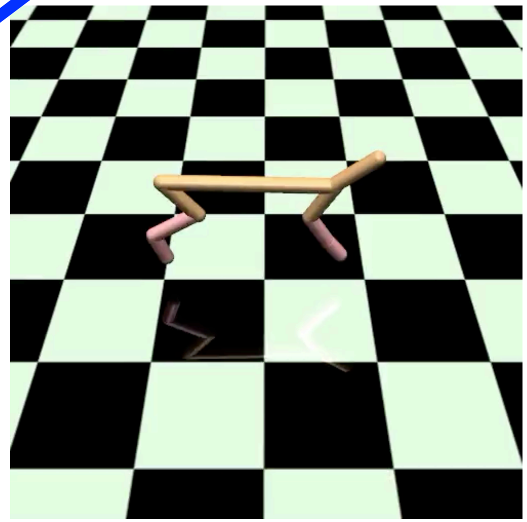
(By Anirudh Jain)

Deep Reinforcement Learning

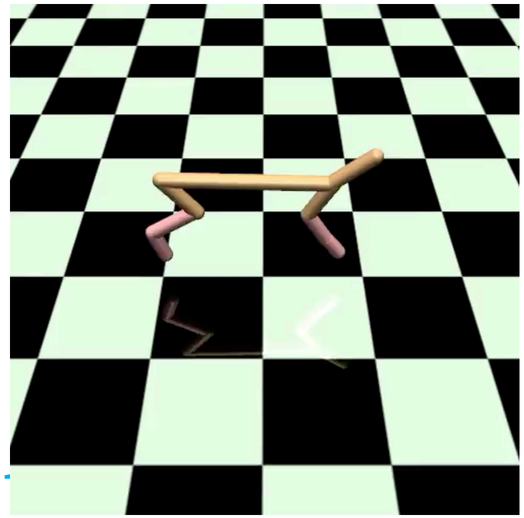
On OpenAI Gym Cheetah with DDPG
with DNN with [400,300] ReLU



Reward 5264



Reward 2038



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

Summary

- Approximate Bayesian inference
  - Fast uncertainty computation in deep learning
  - Generalization of many well-known algorithms
- Many generalizations and Extensions!

On-Going Work (for 2019)

- Scaling it up!
  - Bayesian inference on Imagenet in “x” minutes
  - Built-in VOGN optimizer in PyTorch.
- Enable sequential learning (online/ continual/ life-long/ Active/ Reinforcement learning)

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 (2018)

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

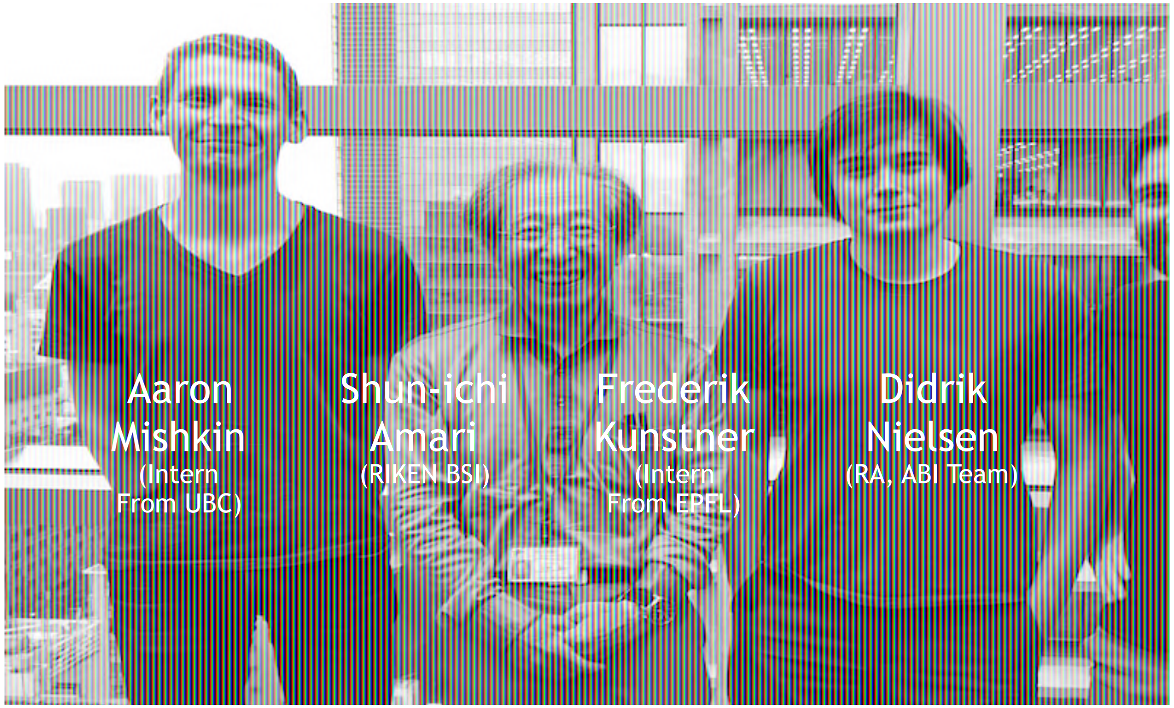
Slides, papers, & code are at emtiyaz.github.io

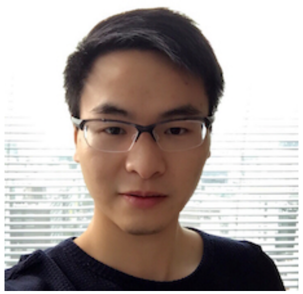
Wu Lin
(RA, ABI team)

Nicolas Hubacher
(RA, ABI team)

Masashi Sugiyama
(Director RIKEN-AIP)

Voot Tangkaratt
(Postdoc, Limited Information team at RIKEN-AIP)





Zuozhu Liu
(Intern from SUTD)

RAIDEN

External Collaborators

Mark Schmidt
(UBC)

Reza Babanezhad
(UBC)

Yarin Gal
(UOxford)

Akash Srivastava
(UEdinburgh)