Bayesian Deep Learning

Mohammad Emtiyaz Khan

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."

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**

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.

My current research focuses on reducing this gap!

Learning-Algorithms from Bayesian Principles

- Practical Bayesian principles
  - To design/improve/generalize learning-algorithms.
  - Distribution over unknowns.
- Generalization of many existing algorithms,
  - Classical (least-squares, HMM, Kalman.. etc).
  - Deep Learning (SGD, RMSprop, Adam).
- Helps us design new algorithms
  - Reinforcement, online, continual learning, reasoning..
- Impact: Everything with one common principle.

Learning Goals

- Keywords
  - Statistics (Gaussian distribution, Bayes' rule)
  - Optimization (Gradient descent, Least-squares)
  - Deep Learning (Stochastic gradient descent, RMSprop)
- What will you learn
  - Some issues with deep learning.
  - A Bayesian principle to fix it.
  - Least squares from Bayesian principles.
  - RMSprop optimizer from Bayesian principles.
  - Applications of Bayesian principles to deep learning.

Uncertainty in Deep Learning

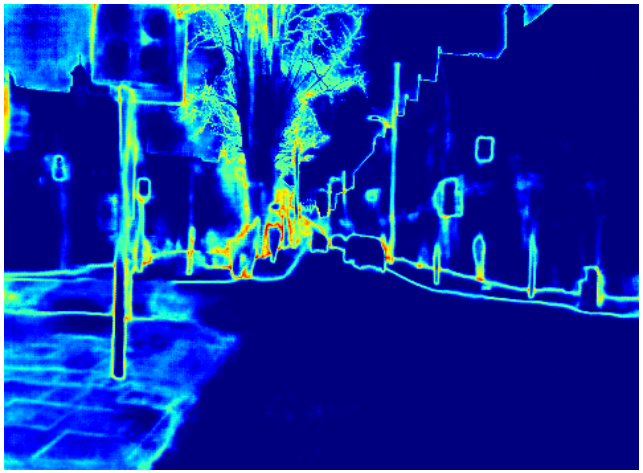
To estimate the confidence in the predictions of a deep-learning system

Uncertainty in Deep Learning

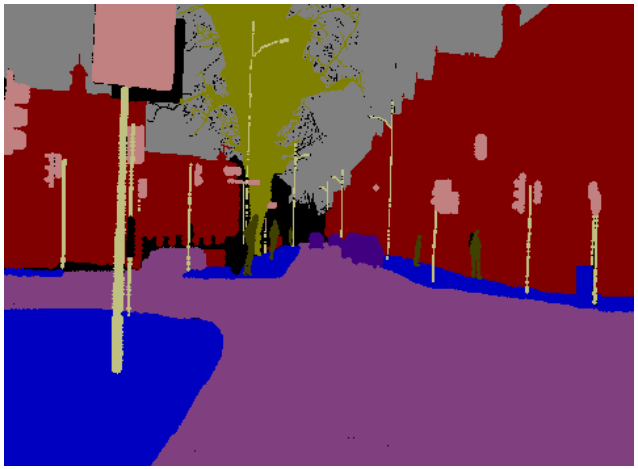
(by Kendall et al. 2017)

Image

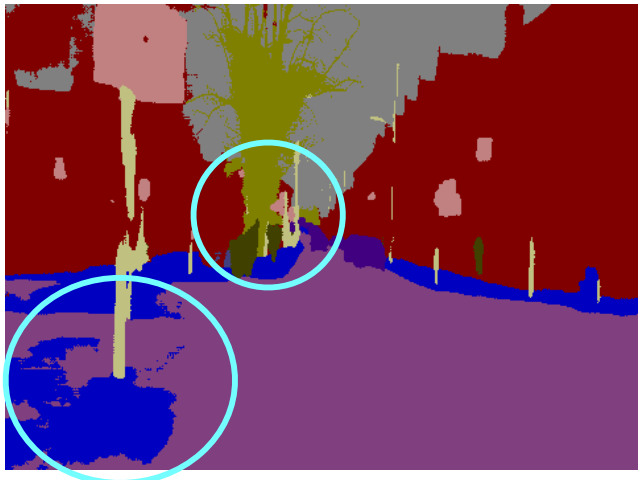
Uncertainty

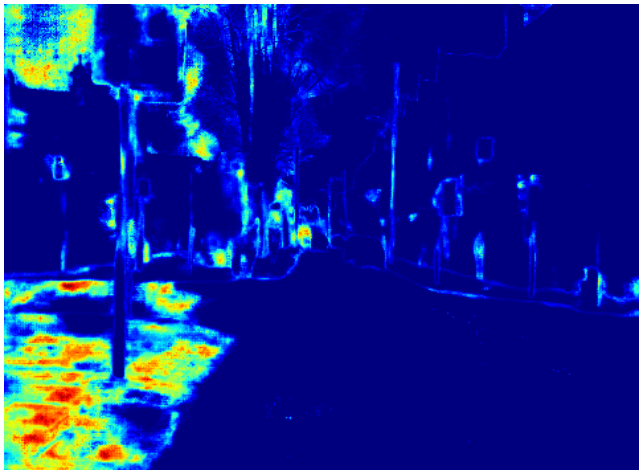


True Segments



Prediction





Learning by Optimization

Empirical Risk Minimization (ERM)

Data Params

$$\min_{\theta} \ \underbrace{\ell(\mathcal{D}, \theta)}_{\text{Loss}} := \sum_{i=1}^{N} \ell_i(\theta) + r(\theta)$$

$$\sum_{i=1}^{N} [y_i - f_\theta(x_i)]^2 + \gamma \theta^T \theta$$

Deep Learning: SGD/RMSprop/Adam/Newton etc.

$$\theta \leftarrow \theta - \rho H^{-1} \nabla_\theta \ell(\theta)$$

Learning by Bayes

Estimate a distribution over model parameters.

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

Optimization formulation (Zellner, 1988)

Distribution (e.g. Gaussian)

Entropy

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

Parameters
(e.g., mean and variance)

Exploitation

Exploration

Exponential Family Distribution

Sufficient Statistics

$$q_\lambda(\theta) \propto \exp\left[\lambda^\top \phi(\theta)\right]$$

Natural parameters

e.g. Gaussian distribution

$$\exp\left[m^\top V^{-1}\theta - \frac{1}{2}\theta^\top V^{-1}\theta\right]$$

$$\{V^{-1}m, V^{-1}\}$$

Expectation/moment/ mean parameters

$$\mu := \mathbb{E}_{q_\lambda}[\phi(\theta)]$$

$$\{\mathbb{E}(\theta), \mathbb{E}(\theta\theta^\top)\}$$

Learning-Algorithms by Bayesian Principles

Learning by optimization: $\theta \leftarrow \theta - \rho H^{-1} \nabla_\theta \ell(\theta)$

Learning by Bayes: $\lambda \leftarrow (1-\rho)\lambda - \rho \nabla_\mu \mathbb{E}_q [\ell(\theta)]$

Natural and Expectation parameters of q

e.g., Gaussian distribution

Natural parameters $\{V^{-1}m, V^{-1}\}$

$q(\theta) := \mathcal{N}(\theta|m, V)$

Expectation/moment/ mean parameters $\{\mathbb{E}(\theta), \mathbb{E}(\theta\theta^\top)\}$

$\exp\left[m^\top V^{-1}\theta - \frac{1}{2}\theta^\top V^{-1}\theta\right]$

Learning by Bayes

Learning by optimization: $\theta \leftarrow \theta - \rho H^{-1} \nabla_\theta \ell(\theta)$

Learning by Bayes: $\lambda \leftarrow (1-\rho)\lambda - \rho \nabla_\mu \mathbb{E}_q \left[\ell(\theta)\right]$

Natural and Expectation parameters of q

AIstats 2017
ICML 2017
- Classical algorithms: Least-squares, Newton's method, Kalman filters, Baum-Welch, Forward-backward, etc.
- Bayesian inference: EM, Laplace's method, SVI, VMP.

ICML 2018
NeurIPS 2018
ISITA 2018
ICLR 2018
- Deep learning: SGD, RMSprop, Adam.
- Reinforcement learning: parameter-space exploration, natural policy-search.
- Continual learning: Elastic-weight consolidation.
- Online learning: Exponential-weight average.

NIPS 2017
- Global optimization: Natural evolutionary strategies, Gaussian homotopy, continuation method & smoothed optimization.
- List incomplete…

Least Squares

$$\lambda \leftarrow (1-\rho)\lambda - \rho \nabla_{\mu} \mathbb{E}_q \left[\ell(\theta)\right] \qquad q_\lambda(\theta) := \mathcal{N}(m, V)$$

$$\mathbb{E}_q \Big[\underbrace{(y - X\theta)^\top (y - X\theta)}_{\text{likelihood}} + \underbrace{\gamma \theta^\top \theta}_{\text{prior}} \Big] := \ell(\theta)$$

$$\nabla_{\mathbb{E}_{q_\lambda}[\theta]} =$$

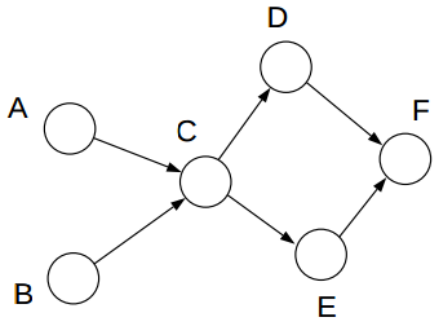
$$\nabla_{\mathbb{E}_{q_\lambda}[\theta\theta^\top]} =$$

Expectation params

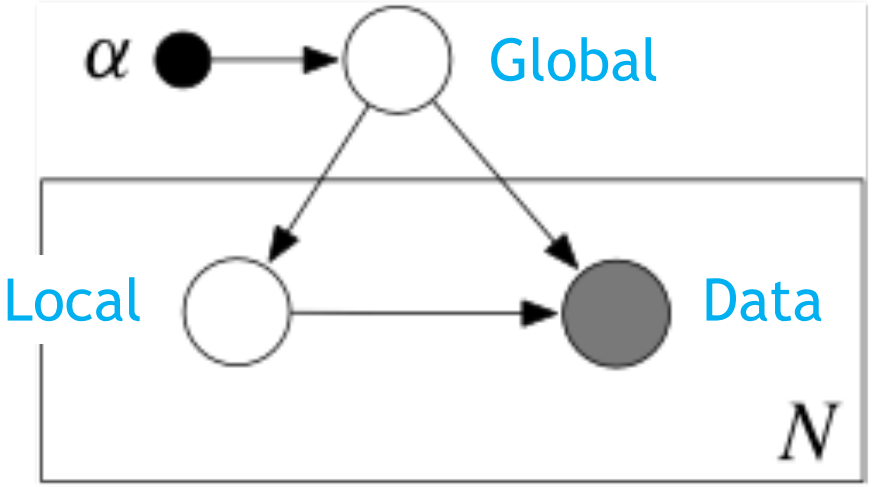
$$\left[X^\top X + \gamma I\right]^{-1} X^\top y$$

Bayesian inference on Conditionally-Conjugate Models

VMP: Sequential update with rho =1



SVI: Update local variable with rho=1 and global variable with rho in (0,1)



Learning by Bayes is a generalization of both of these algorithms.

Images taken from Hoffman et al. (2013) and https://www.zybuluo.com/nanmeng/note/369145

Neural Network

$$\mathbb{E}_q \left(\sum_{i=1}^{N} \ell(y_i, f_\theta(x_i)) + \gamma \theta \theta^\top \right)$$

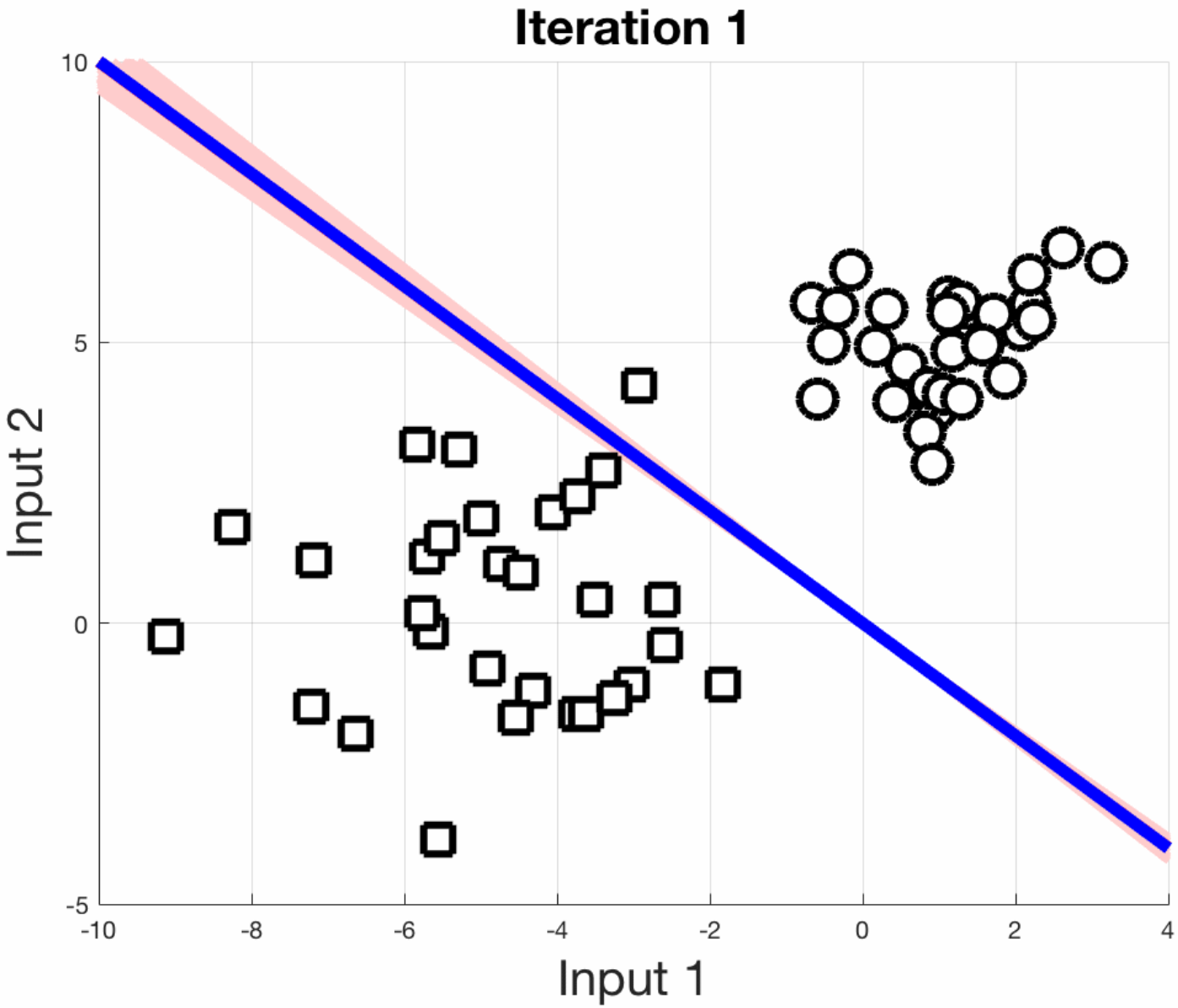
likelihood

prior

neural network

Adam vs Our Method (on Logistic-Reg)

ICML 2018



Adam

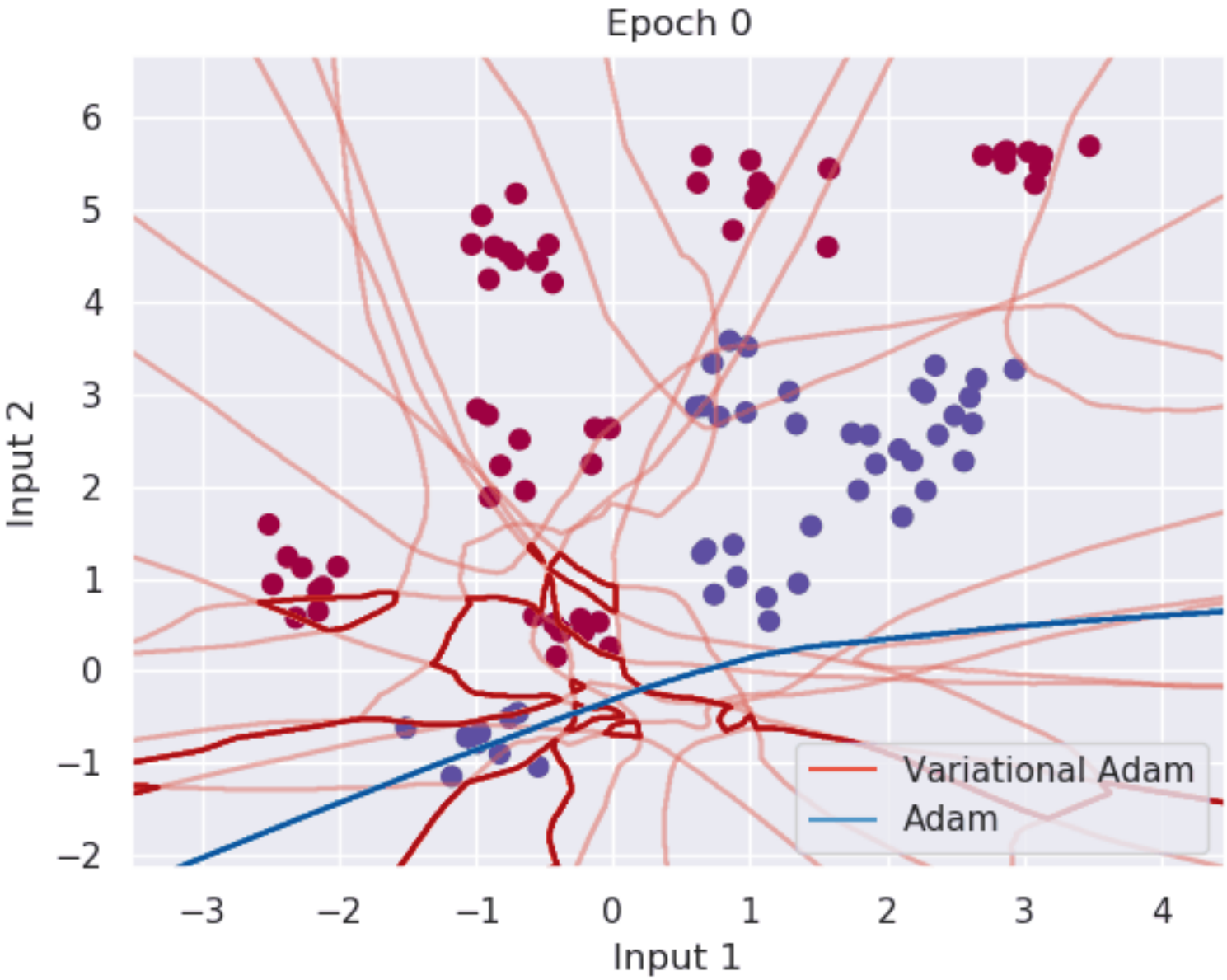
Our method (mean)

Our method (samples)

M = 5,
Rho = 0.01,
Gamma = 0.01

Adam vs Our Method (on Neural Nets)

ICML 2018



Adam

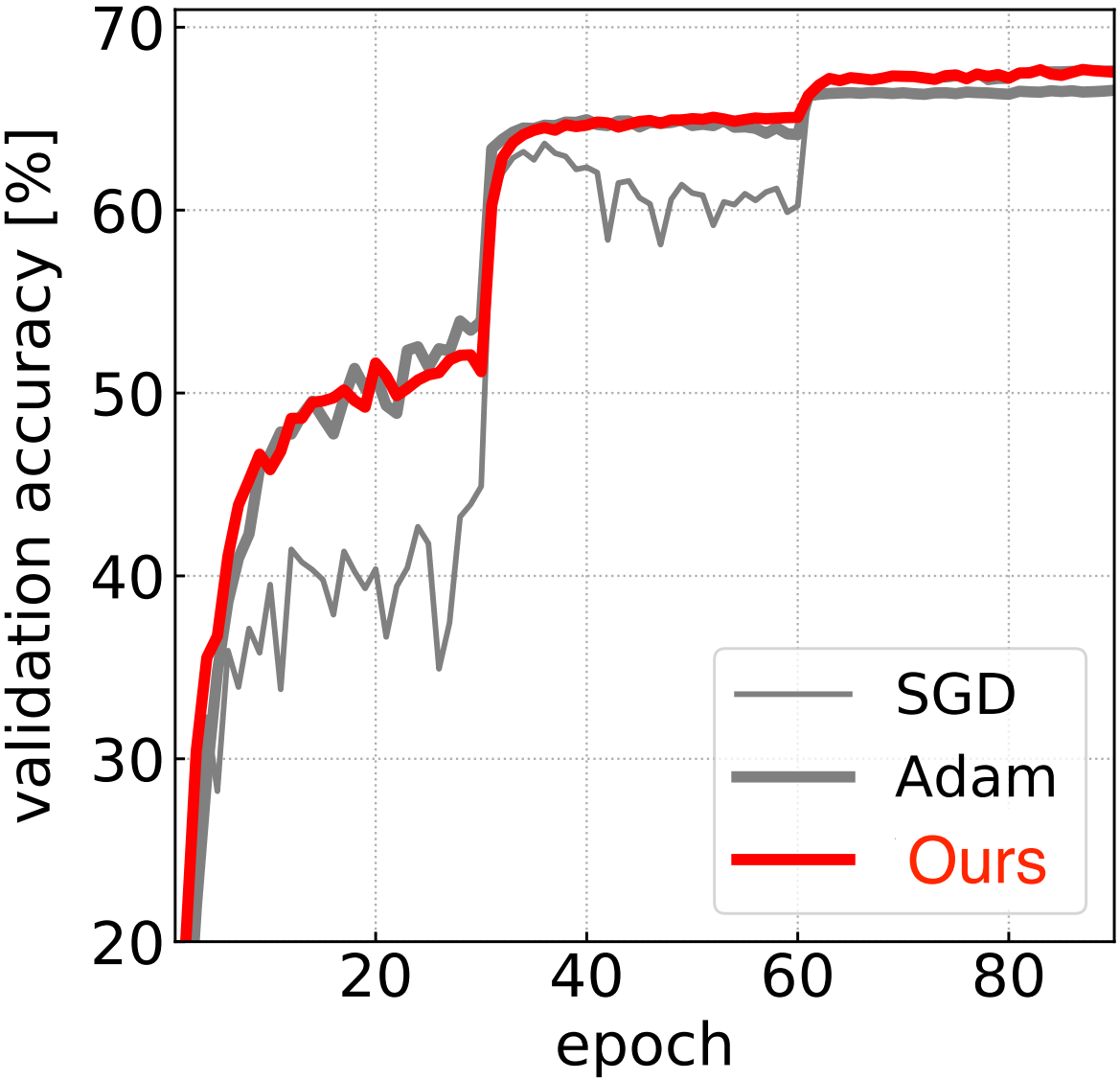
Ours (mean)

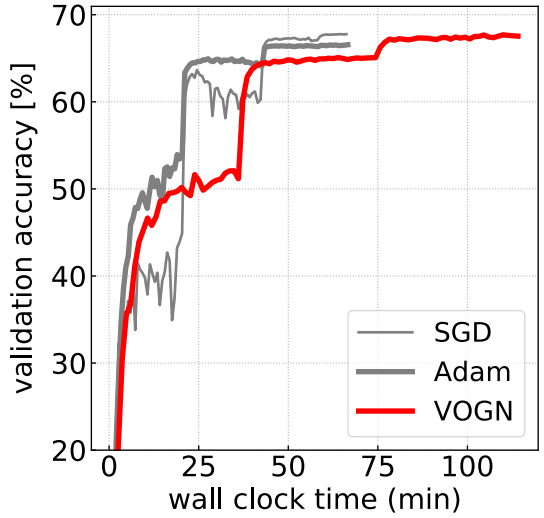
Ours (samples)

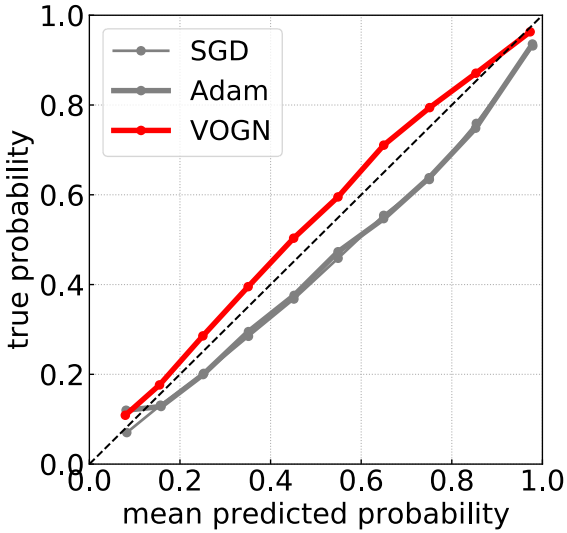
(By Runa E.)

Practical DL with Bayes (on ImageNet)

Under review

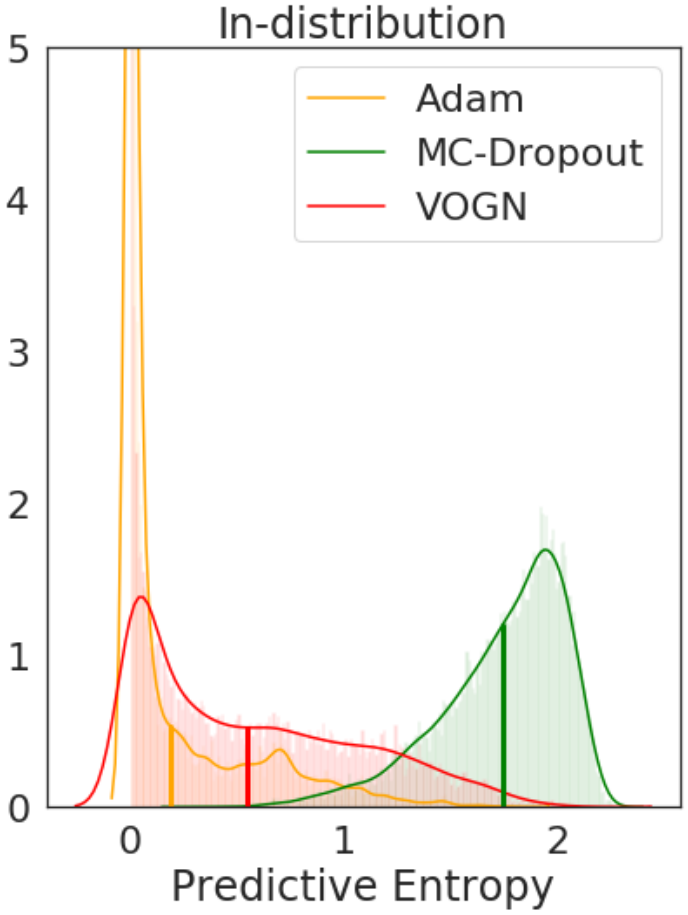


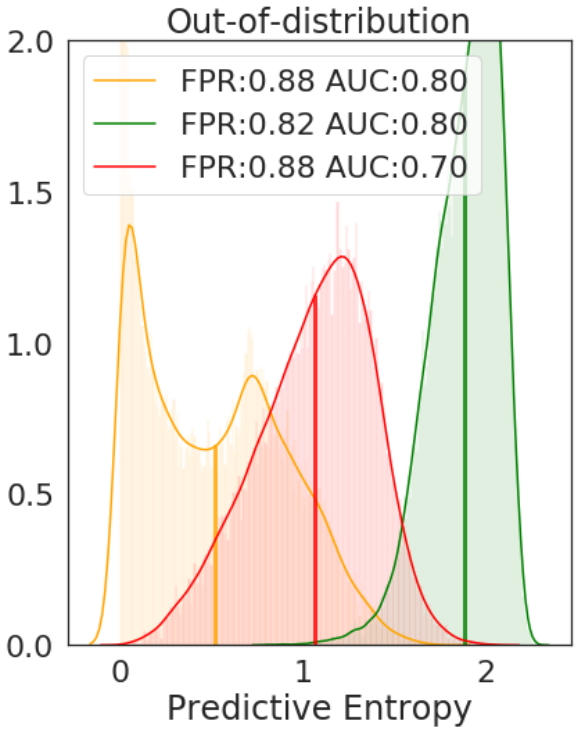


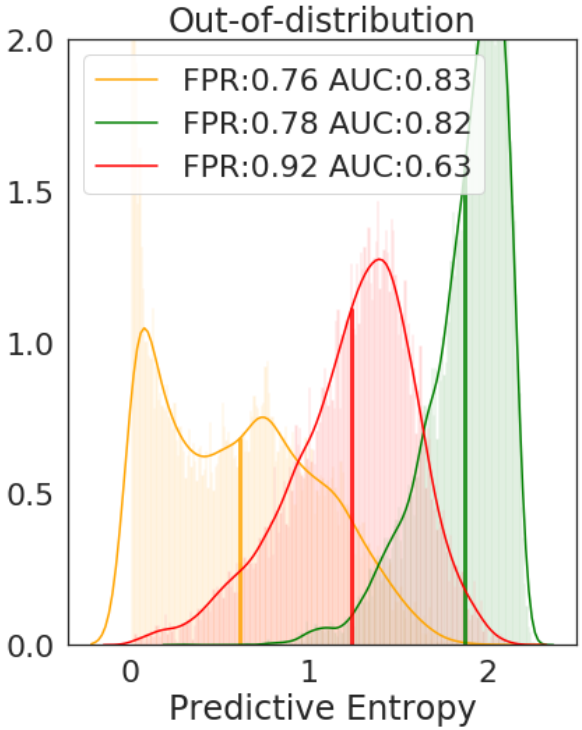


Out-of-Distributions Test

Under review

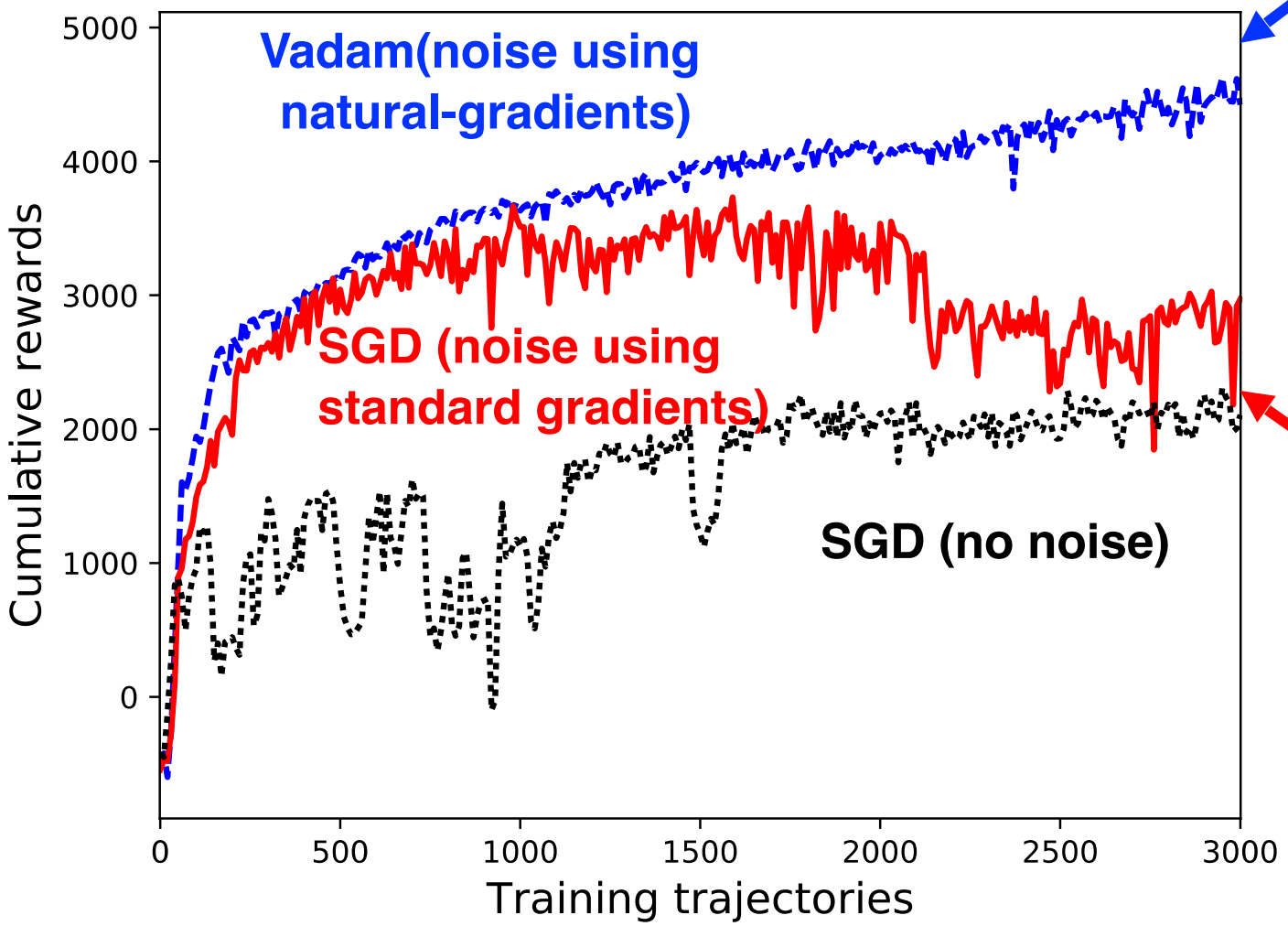




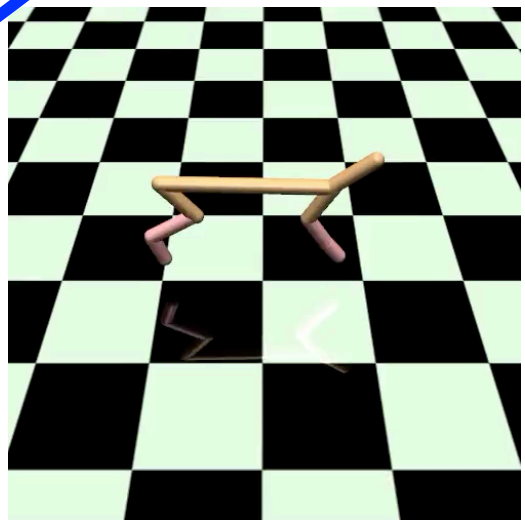


Deep Reinforcement Learning

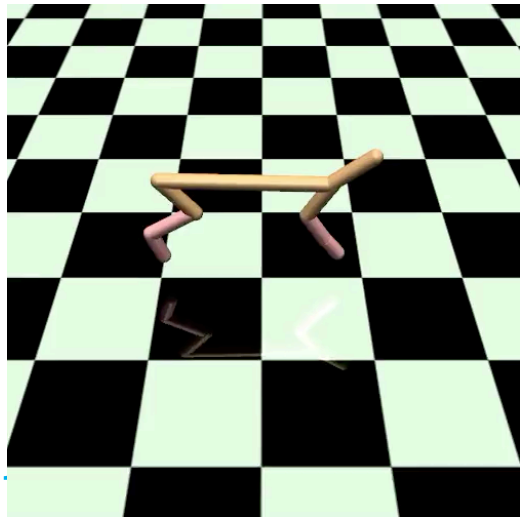
On OpenAI Gym Cheetah with DDPG
with DNN with [400,300] ReLU



Reward 5264



Reward 2038



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

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*

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

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 for Logistic Reg

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]

Practical Deep Learning with Bayesian Principles, (UNDER REVIEW) K. OSAWA, S. SWAROOP, A. JAIN, R. ESCHENHAGEN, R.E. TURNER, R. YOKOTA, **M.E. KHAN**. [arXiv]

Approximate Inference Turns Deep Networks into Gaussian Processes, (UNDER REVIEW) **M.E. KHAN**, A. IMMER, E. ABEDI, M. KORZEPA. [arXiv]

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

Acknowledgements

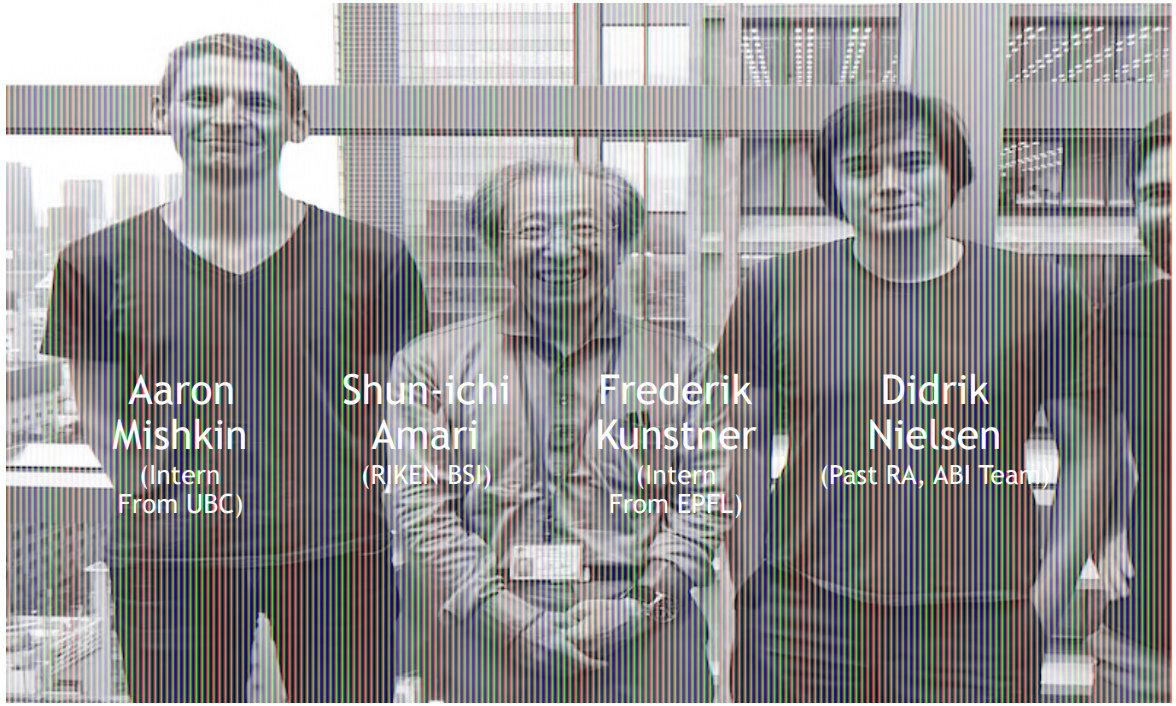
Slides, papers, & code are at emtiyaz.github.io

Wu Lin
(Past: RA, ABI team)

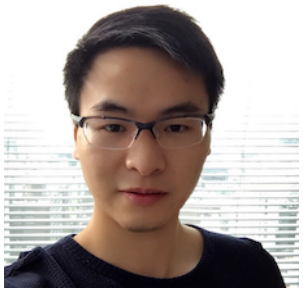
Nicolas Hubacher
(Past: RA, ABI team)

Masashi Sugiyama
(Director RIKEN-AIP)

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



External Collaborators



Zuozhu Liu
(Intern from SUTD)

RAIDEN

Mark Schmidt
(UBC)

Reza Babanezhad
(UBC)

Yarin Gal
(UOxford)

Akash Srivastava
(UEdinburgh)

The ABI Team

Homework

- Derive the Bayes update for least-squares.
- Derive the same for neural networks
- Can you think of ways to RMSprop and Bayes update even more similar?
  - RMSprop vs Bayes with diagonal Gaussian.
  - Justify why your way is reasonable, and also when will it work and when it won't.

Thanks!

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