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

Mohammad Emtiyaz Khan
RIKEN Center for AI Project, Tokyo, Japan
Slides available at http://emtiyaz.github.io

Joint work with

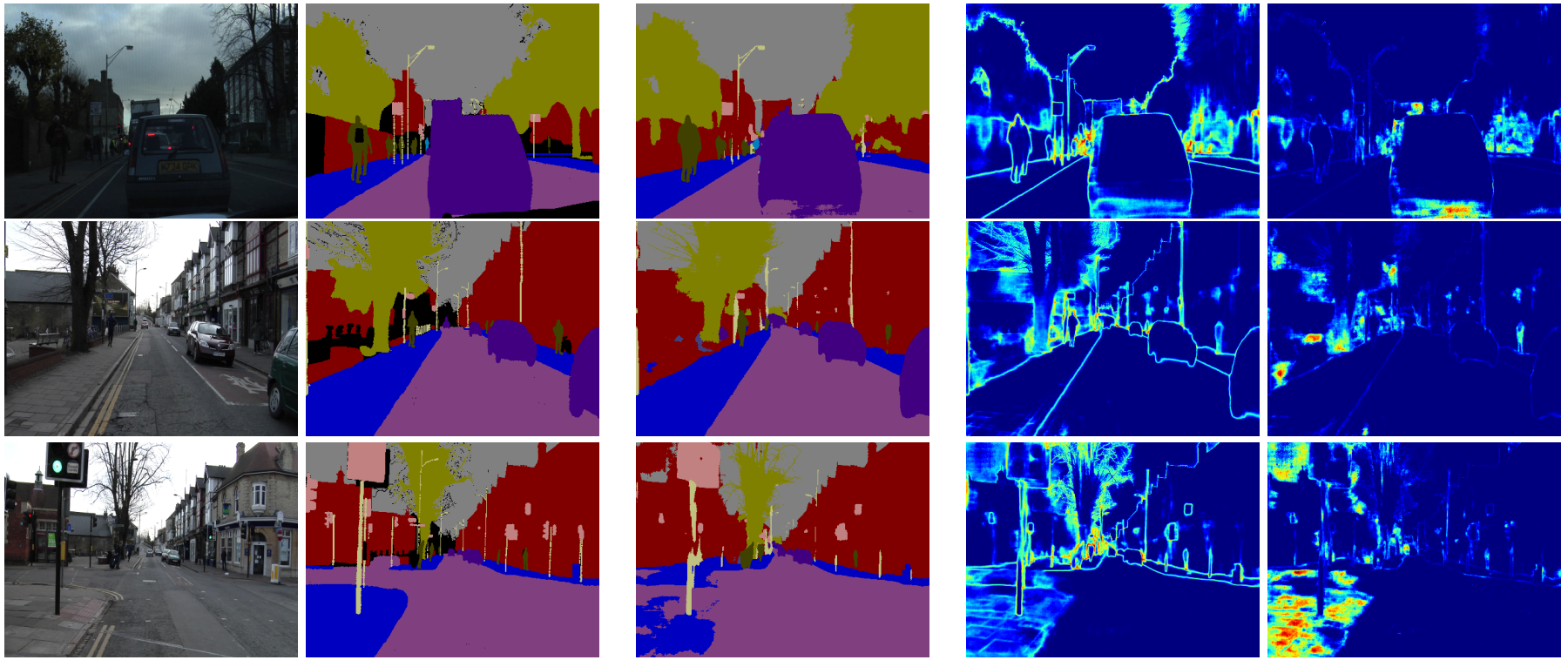
Didrik Nielsen
RIKEN Center for AI Project, Tokyo Japan

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)

Variational Inference

- Compute an approximate posterior distribution using an optimization algorithm
- Variational Inference (VI) using gradient methods (SGD/Adam)
  - Gaussian VI: Bayes by Backprop (Blundell et al. 2015), Practical VI (Graves et al. 2011), Black-box VI (Rangnathan et al. 2014) and many more….
- This talk: VI using natural-gradient methods (faster and simpler than gradients methods)
  - Khan & Lin (AIstats 2017), Khan et al. (ICML 2018), Khan & Nielsen (ISITA2018)

Outline

- Backgound
  - Bayesian model and Variational Inference (VI)
  - VI using gradient descent
  - VI using natural-gradient descent
- Simple natural-gradients
- Fast computation of natural-gradients
- Results on Bayesian deep learning and RL

Variational Inference

Gradients

Natural Gradients

BACKGROUND

Variational Inference

Bayes' rule:

Neural Network

weights

$$p(w|\mathcal{D}) = \frac{\prod_i p(\mathcal{D}_i|w)p(w)}{\int \prod_i p(\mathcal{D}_i|w)p(w)dw}$$

ExpFamily(η_0)

$$\approx \quad q_\lambda(w) = \text{ExpFamily}(\lambda)$$

$$\min_\lambda \mathbb{D}_{KL}\Big[q_\lambda(w)\|p(w|\mathcal{D})\Big]$$

$$\equiv \max_\lambda \mathcal{L}(\lambda)$$

Variational objective

VI using Natural-Gradient Descent

Gradient

Gradient Descent: $\lambda \leftarrow \lambda + \rho \nabla_\lambda \mathcal{L}(\lambda)$

Natural-Gradient Descent: $\lambda \leftarrow \lambda + \rho F(\lambda)^{-1} \nabla_\lambda \mathcal{L}(\lambda)$

Natural Gradients

$\tilde{\nabla}_\lambda \mathcal{L}(\lambda)$

Fisher Information Matrix (FIM)

$$F(\lambda) := \mathbb{E}_{q_\lambda}\left[\nabla \log q_\lambda(w) \nabla \log q_\lambda(w)^\top\right]$$

“Simple” Natural-Gradients

By using expectation parameterization of ExpFamily

Expectation/Moment Parameters

Wainwright and Jordan, 2006

$$\mu(\lambda) := \mathbb{E}_{q_\lambda}\left[\text{Sufficient Statistics}\right]$$

E.g., Gaussian has two moments

$$\mathbb{E}_{q_\lambda}[w] := \mu_1$$
$$\mathbb{E}_{q_\lambda}[ww^\top] := \mu_2$$

Gradient

$$\nabla_\mu \mathcal{L} = F(\lambda)^{-1}\nabla_\lambda \mathcal{L} := \tilde{\nabla}_\lambda \mathcal{L}$$

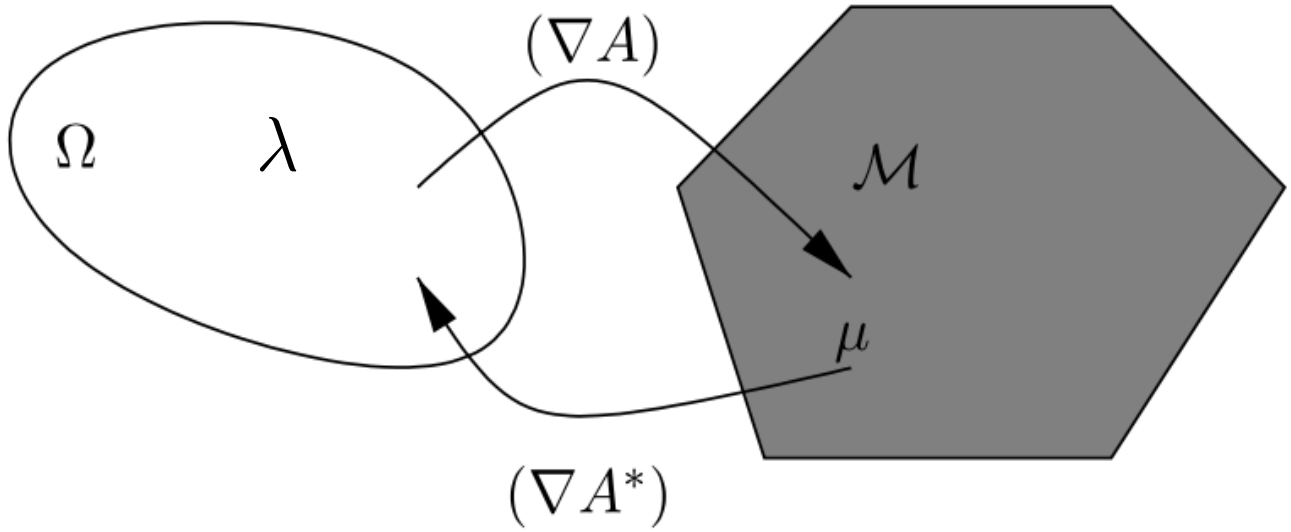
Natural-Gradient

$$\nabla_\lambda \mathcal{L} = F(\mu)^{-1}\nabla_\mu \mathcal{L} := \tilde{\nabla}_\mu \mathcal{L}$$

Dually-Flat Riemannian Structure

See Amari's book 2016

Figure from Wainwright and Jordan, 2006



Gradient

Natural-Gradient

$$\nabla_\mu \mathcal{L} = F(\lambda)^{-1} \nabla_\lambda \mathcal{L} := \tilde{\nabla}_\lambda \mathcal{L}$$ Easy

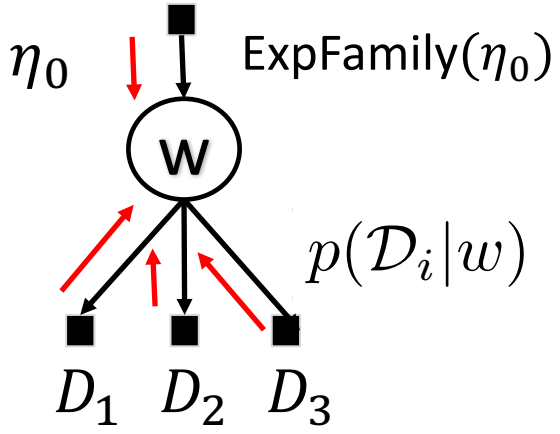
$$\nabla_\lambda \mathcal{L} = F(\mu)^{-1} \nabla_\mu \mathcal{L} := \tilde{\nabla}_\mu \mathcal{L}$$ Difficult

NatGrad Descent as Message Passing

Khan and Lin 2017, Khan and Nielsen 2018

$$\lambda \leftarrow \lambda + \rho F(\lambda)^{-1} \nabla_\lambda \mathcal{L}(\lambda)$$

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



$$\lambda \leftarrow (1-\rho)\lambda + \rho \left[\eta_0 + \sum_{i=1}^{N} \nabla_\mu \mathbb{E}_{q_\lambda} [\log p(\mathcal{D}_i|w)]|_{\mu=\mu(\lambda)} \right]$$

Locally, add all the natural gradients

A generalization of Kalman filtering, Sum-product, etc., Variational Message Passing (Winn and Bishop 2005), Stochastic variational inference (Hoffman et al. 2013).

Fast Gaussian Approximation for Deep Neural Networks

$$\prod_{i=1}^{N} p(\mathcal{D}_i|\theta) \qquad \theta \sim \mathcal{N}(\theta|0, I) \qquad (\mathrm{w} = \theta)$$

Adaptive learning-rate method (e.g., Adam)

Natural Gradient VI => Variational Adam (Vadam)

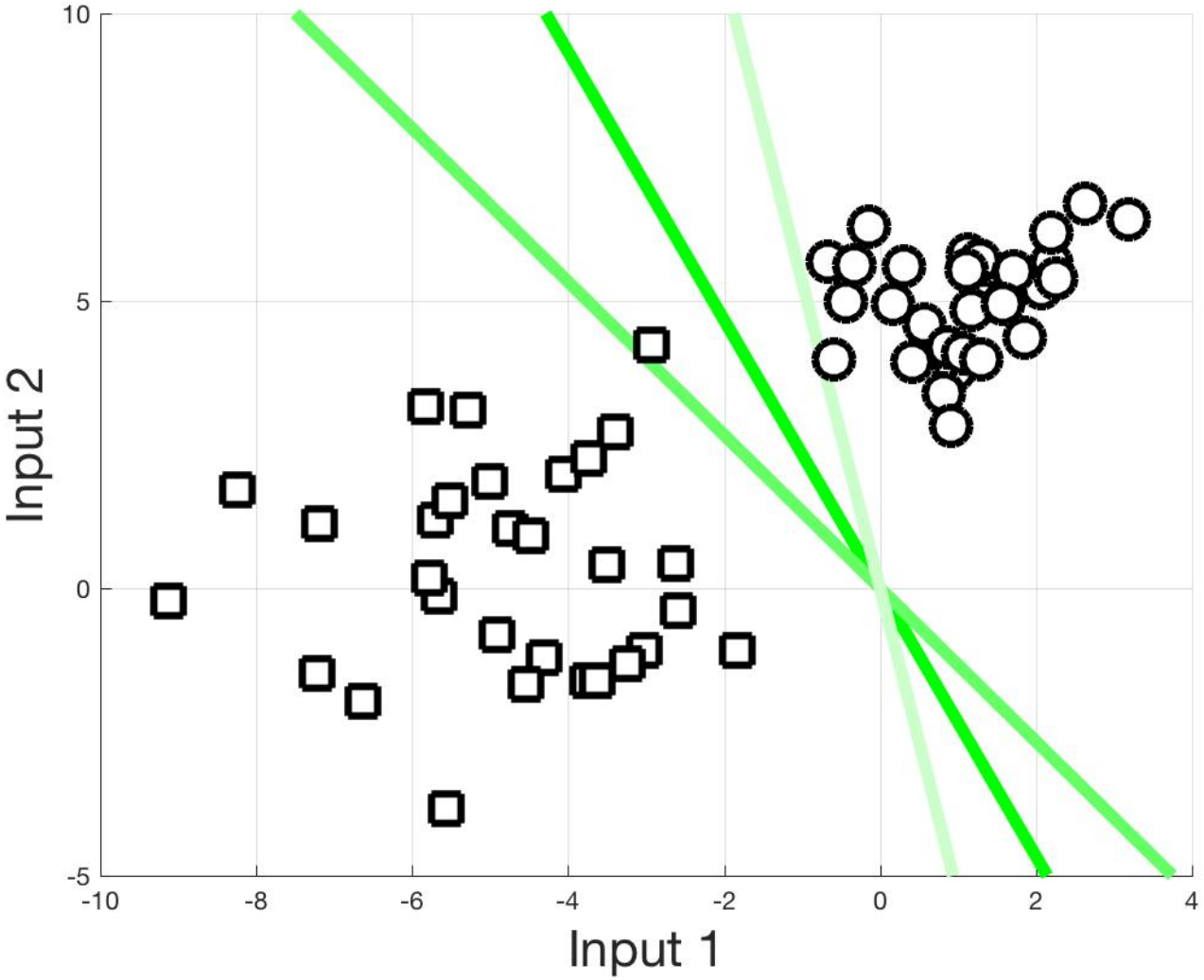
0. Sample ϵ from a standard normal distribution

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

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

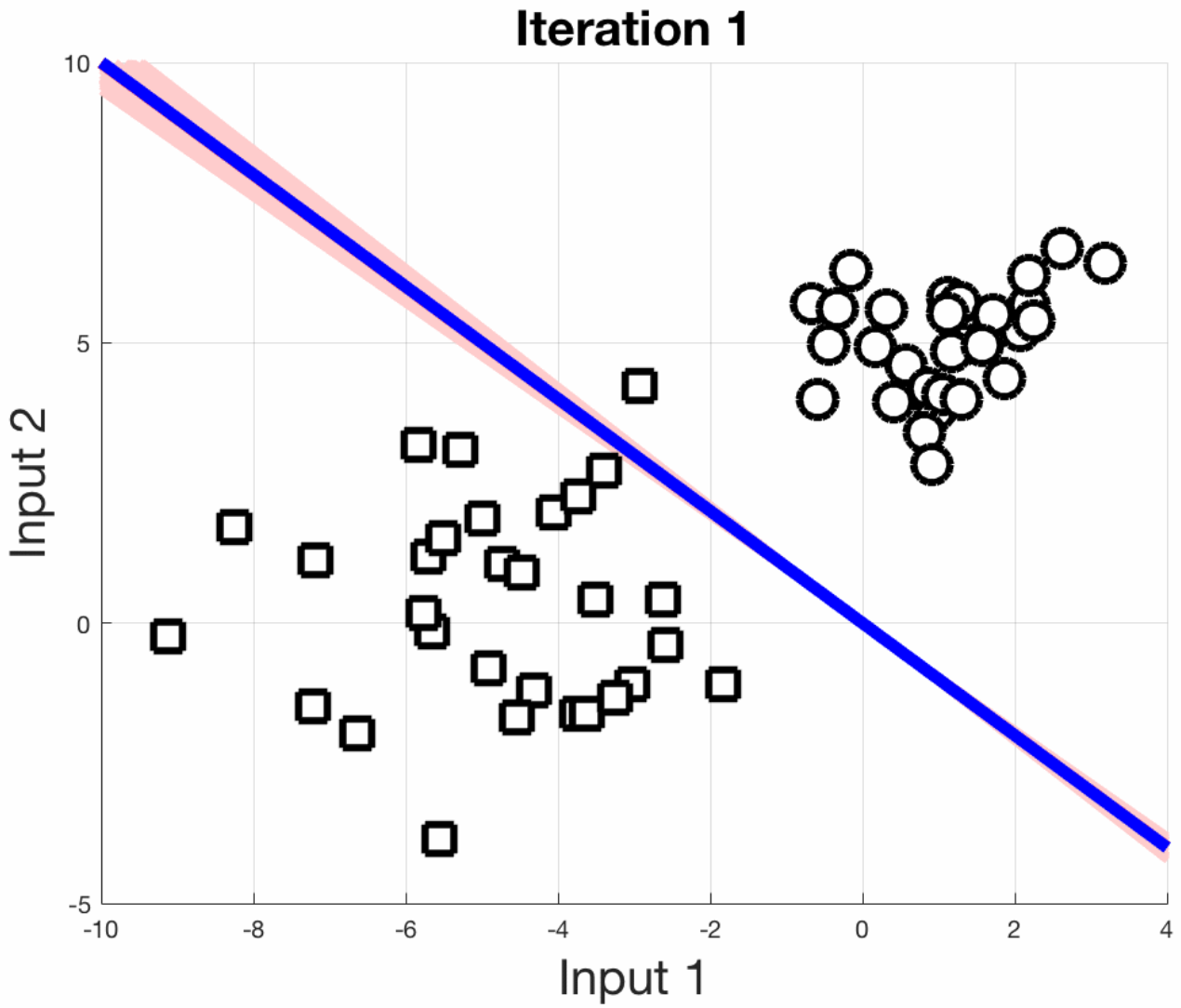
$$\theta \leftarrow \theta + \text{learning\_rate} * \frac{\text{gradient} + \theta/N}{\sqrt{\text{scale}} + 1/N^8}$$

Illustration: Classification



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

Adam vs Vadam



Adam

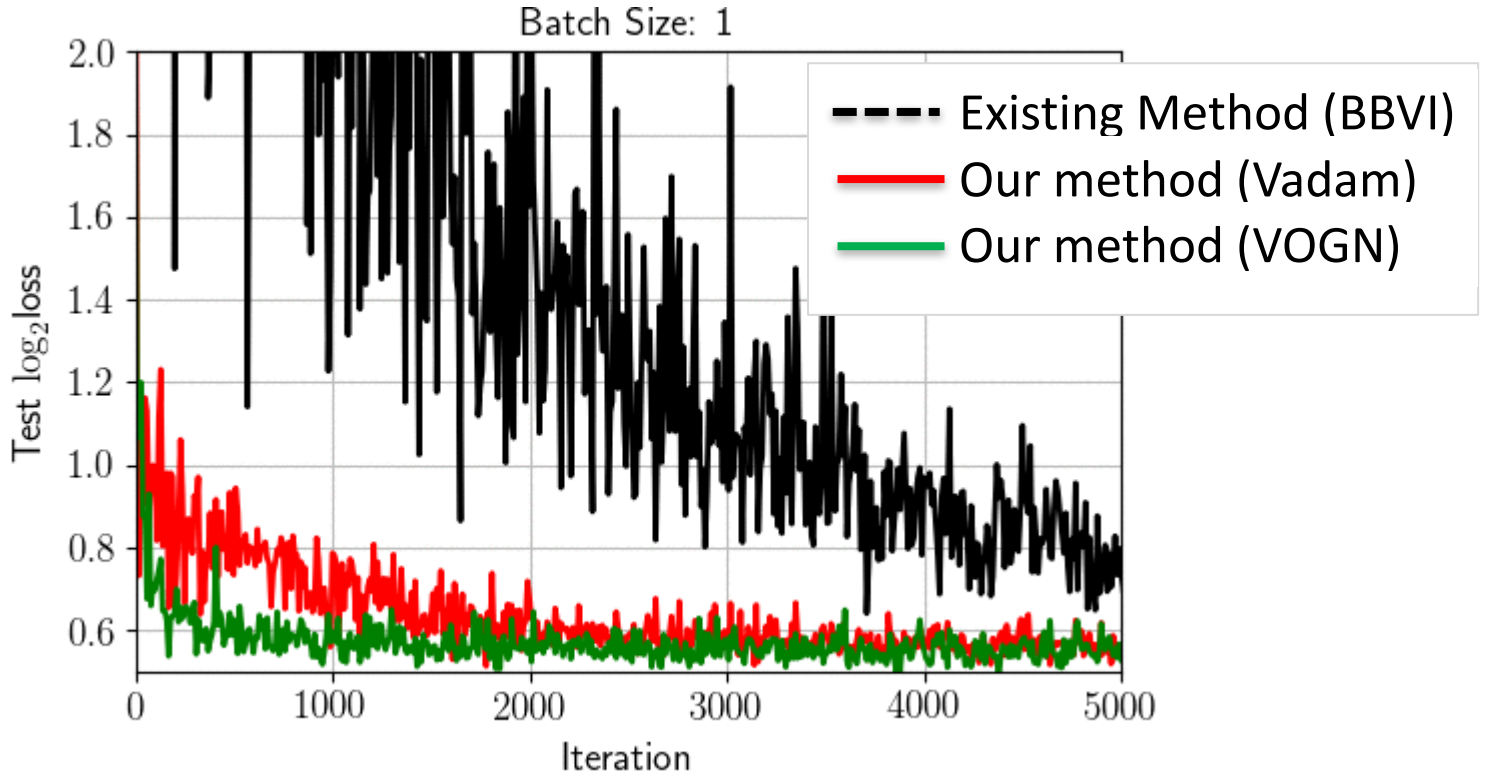
Vadam (mean)

Vadam (samples)

For both algorithms,
Minibatch of 5
Learning_rate = 0.01
Prior precision = 0.01

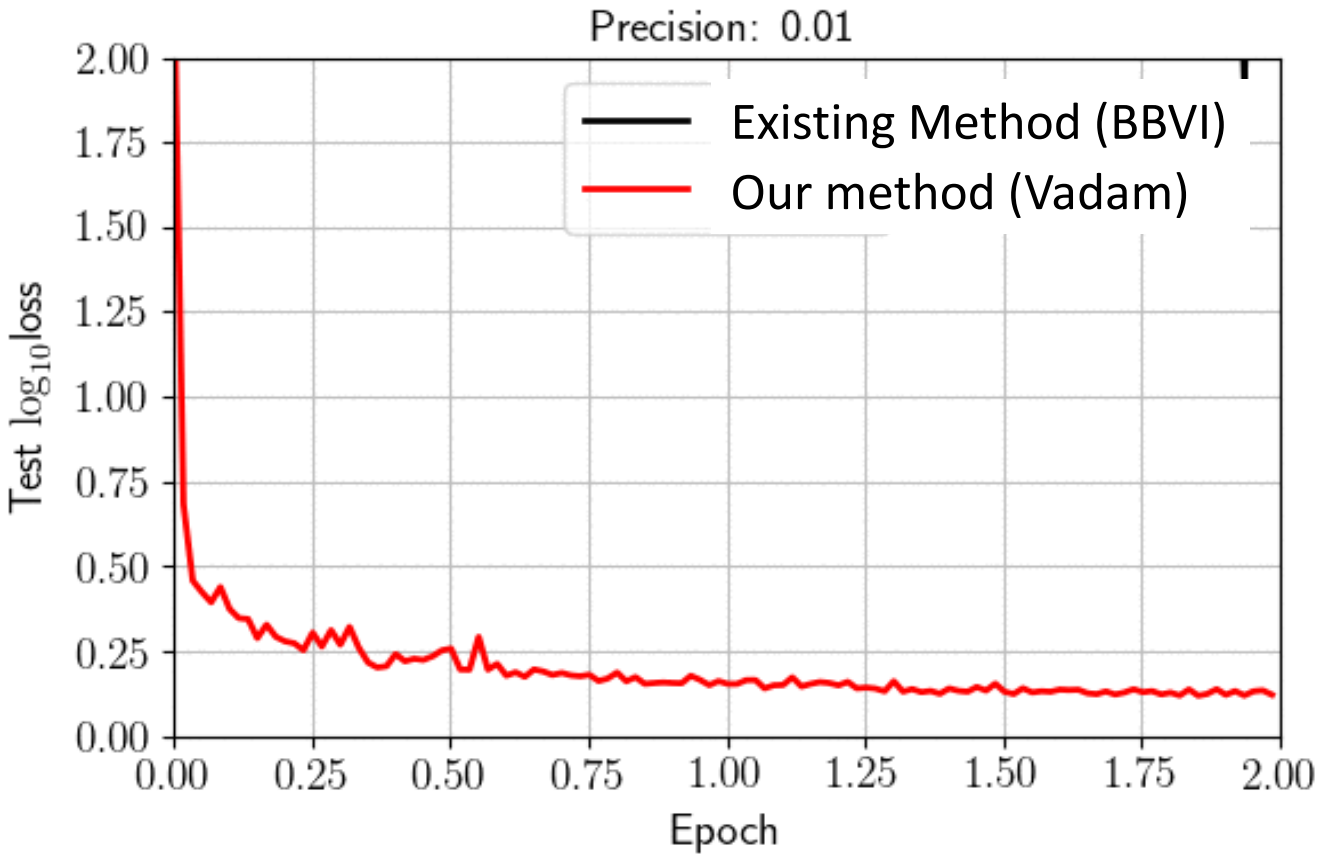
Faster, Simpler, and More Robust

Regression on Australian-Scale dataset using deep neural nets for various number of minibatch size.



Faster, Simpler, and More Robust

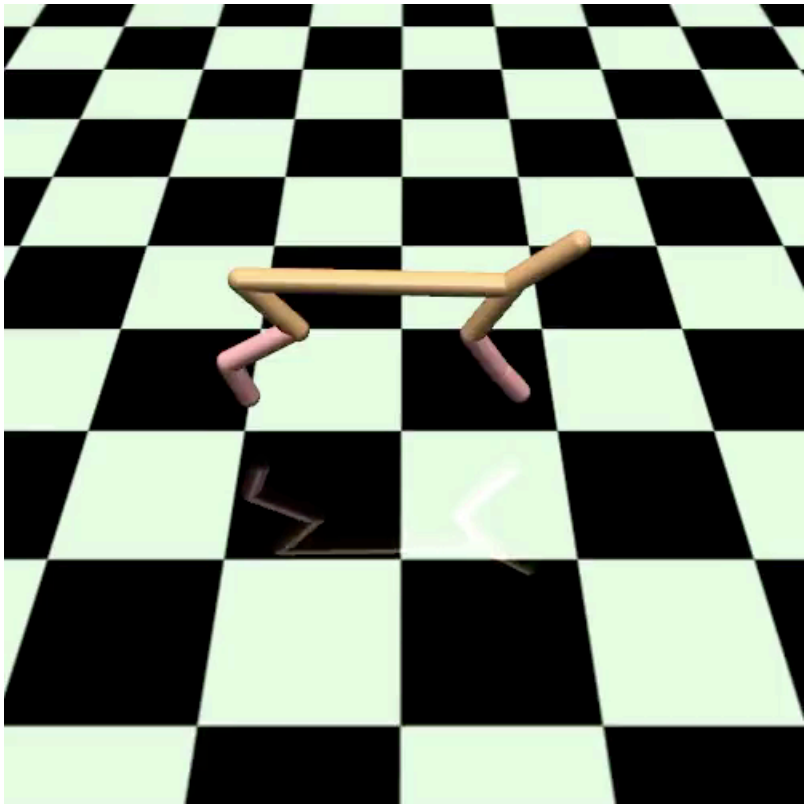
Results on MNIST digit classification (for various values of Gaussian prior precision parameter λ)



Deep Reinforcement Learning

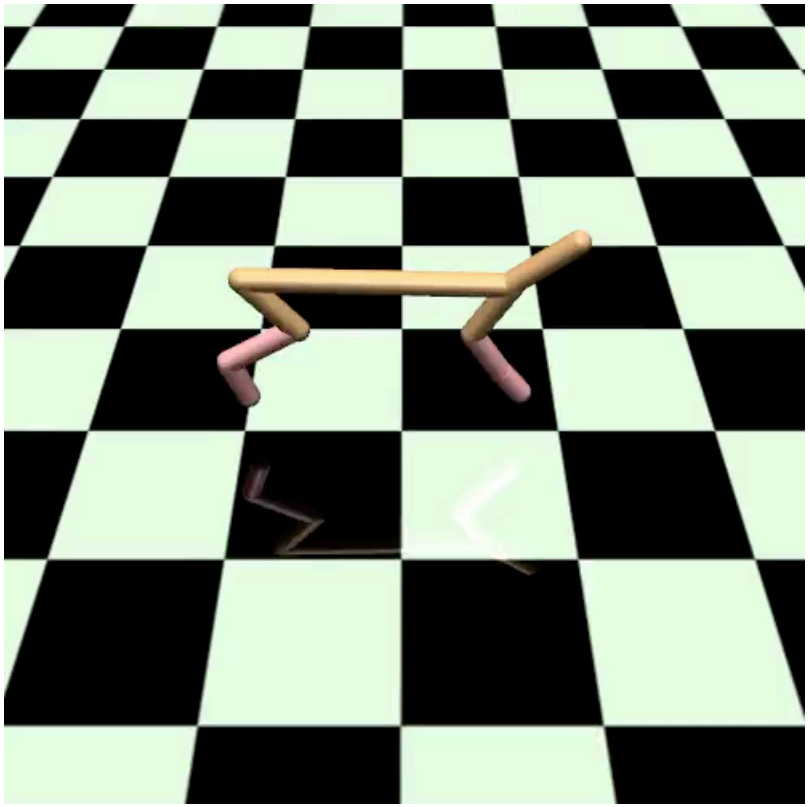
No Exploration (SGD)

Reward = 2860



Exploration using Vadam

Reward = 5264



Summary

- For exp-family approximation, natural-gradients can be computed using expectation parameters ("simple" updates in many cases)
- Messages in variational message-passing are natural-gradients in the expectation parameters
  - Extends to non-conjugate factors.
  - Gives local approximations such that variational objective is maximized
- Fast computation is possible in some cases (e.g., Gaussian approximation).

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]

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**.

Thanks!

Slides, paper, and code available at
http://emtiyaz.github.io