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 ≠ 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).
The supplementary video for examples). We show how modeling aleatoric uncertainty in regression
is important for different inputs, and uncertainty can further be categorized into
equally. uncertainty can be explained away with the large amounts of data often available in machine vision. We further show
that modeling aleatoric uncertainty alone comes at a cost. Out-of-data examples, which can be
explained away. This is in comparison to epistemic uncertainty which is mostly
identified with epistemic uncertainty, cannot be identified with aleatoric uncertainty alone.

Figure 1: Illustrating the difference between aleatoric and epistemic uncertainty

We study the trade-offs between modeling aleatoric or epistemic uncertainty by character-
ing the properties of each uncertainty and comparing model performance and inference
We improve model performance by effect of noisy data with the implied attenuation obtained from explicitly representing
with a novel approach for classification,

The main contributions of this work are:

1. Capturing an accurate understanding of aleatoric and epistemic uncertainties, in particular
2. Improving model performance by over non-Bayesian baselines by reducing the
3. exhibiting increased aleatoric uncertainty on object boundaries and for objects far from the camera.

Uncertainty Approximations. We derive our framework for both regression and classification applications
which captures our ignorance about which model generated our collected data. This uncertainty
exhibits increased epistemic uncertainty for semantically

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

(b) Ground Truth
(d) Aleatoric
(e) Epistemic

Uncertainty in Deep Learning

(by Kendall et al. 2017)
Challenges

The data and model are both extremely large.

\[ \min_{\theta} \ell(\mathcal{D}, \theta) \rightarrow \text{Loss} \]

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.

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, Alstats 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)

Vadam/VOGN

0. Sample $\epsilon$ from a standard normal distribution

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

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

Mean

$$\theta \leftarrow \theta + \text{learning rate} \times \frac{\text{gradient} \times \theta/N}{\sqrt{\text{scale} + 10/N^8}}$$
Logistic regression
(30 data points, 2 dimensional input).
Sampled from Gaussian mixture with 2 components.
Adam vs Our Method (on Logistic-Reg)

Iteration 1

- Adam
- Our method (mean)
- Our method (samples)

M = 5,
Rho = 0.01,
Gamma = 0.01
Adam vs Our Method (on Neural Nets)

(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

- Vadam (noise using natural-gradients)
- SGD (noise using standard gradients)
- SGD (no noise)

Reward 5264

Ruckstiesh et.al. 2010, Fortunato et.al. 2017, Plapper et.al. 2017
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 (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.*
- Hensman et al. (2012), *Fast Variational Inference in the Conjugate Exponential Family.*
- Hoffman et al. (2013), *Stochastic Variational Inference.*
- 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,

Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam,

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,

Fast and Simple Natural-Gradient Variatioinal Inference with Mixture of Exponential Family,
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

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

Mark Schmidt  
(UBC)

Reza Babanezhad  
(UBC)

Yarin Gal  
(UOxford)

Akash Srivastava  
(UEdinburgh)

External Collaborators

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