Bayesian Deep Learning

Mohammad Emtiyaz Khan

RIKEN Center for AI Project, Tokyo
http://emtiyaz.github.io
Keywords

• Bayesian Statistics
  – Gaussian distribution, Bayes’ rule
• Continues Optimization
  – Gradient descent, Least-squares.
• Deep Learning
  – Stochastic gradient descent, RMSprop
• Information Geometry
  – Riemannian Manifold, Natural Gradients
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.”
Bayesian Inference

• Compute the posterior distribution
  – Instead of just a point estimate (e.g. MLE).
• A natural representation of all the past information which can then be sequentially updated with new information
  – Useful for active learning, sequential experiment design, continual learning, RL.
  – But also for global optimization, causality, etc.
  – Eventually, for ML methods which can learn like humans (data efficient, robust, causal).
Uncertainty in Deep Learning

To estimate the confidence in the predictions of a deep-learning system
Example: Which is a Better Fit?

Real data from Tohoku (Japan). Example taken from Nate Silver’s book “The signal and noise”
Example: Which is a Better Fit?

When the data is scarce and noisy, e.g., in medicine, and robotics.
The main contributions of this work are;

- demonstrating the efficacy of our approach on difficult and large scale tasks.

We can be used to learn loss attenuation, and develop a complimentary approach for the classification task.

We show how modeling aleatoric uncertainty in regression tasks can be explained away with the large amounts of data often available in machine vision. We further show that modeling epistemic uncertainty alone comes at a cost. Out-of-data examples, which can be identified with epistemic uncertainty, cannot be identified with aleatoric uncertainty alone.

Uncertainty accounts for our ignorance about which model generated our collected data. This is a notably different measure of uncertainty and in (e) our model exhibits increased epistemic uncertainty for semantically visually challenging pixels. The bottom row shows a failure case of the segmentation model when the model fails to segment the footpath due to increased epistemic uncertainty, but not aleatoric uncertainty.

Uncertainty can further be categorized into two types:

- **Aleatoric uncertainty**: captures noise inherent in the observations. Heteroscedastic uncertainty depends on different properties of each uncertainty and comparing model performance and inference time.
- **Epistemic uncertainty**: captures our ignorance about which model generated our collected data.

We capture an accurate understanding of aleatoric and epistemic uncertainties, in particular heteroscedastic uncertainty, which cannot be explained away given enough data, and is often referred to as aleatoric uncertainty.

We improve model performance by effect of noisy data with the implied attenuation obtained from explicitly representing aleatoric uncertainty, and with a novel approach for classification.

We study the trade-offs between modeling aleatoric or epistemic uncertainty by characterizing the properties of each uncertainty and comparing model performance and inference time.

For example, for depth regression, highly textured input images with strong vanishing lines are expected to result in heteroscedastic uncertainty.

Uncertainty for Image Segmentation

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

(taken from Kendall et al. 2017)
Variational Inference (VI)

• Approximate the posterior using optimization
  – Popular in reinforcement learning, unsupervised learning, online learning, active learning etc.

• We need accurate VI algorithms that are
  – general (apply to many models),
  – scalable (for large data and models),
  – fast (converge quickly),
  – simple (easy to implement).

• This talk: New algorithms with such features.
Gradient vs Natural-Gradient

• Gradient Descent (GD)
  – Rely on stochastic and automatic gradients.
  – Simple, general, and scalable, but can have suboptimal convergence.
    – Practical VI (2011), Black-box VI (2014), Bayes by backprop (2015), ADVI (2015), and many more.

• Natural-Gradient Descent (NGD)
  – Fast convergence, but computationally difficult, therefore not simple, general, and scalable
    – (Sato (2001), Riemannian CG (2010), Stochastic VI (2013), etc.

• Fast and simple NGD for complex models, such as those containing deep networks.
Talk Outline

• Variational Inference with gradient descent and natural-gradient descent.

• NGD with Conjugate-Computation VI

• Generalizations and Extensions,
  – Structured VAEs (ICLR 2018), Mixture of Exponential Family approximations, Evolution strategy (ArXiv 2017), etc.
Variational Inference

Gradient Descent (GD)

Vs

Natural-Gradient Descent (NGD)
A Naïve Method

\[
p(D|\theta) = \prod_{i=1}^{N} p(y_i | f_{\theta}(x_i))
\]

Data  \rightleftharpoons  Parameters  \rightleftharpoons  Neural network

Input  \rightleftharpoons  Output

\[
\theta \sim p(\theta)
\]

Prior distribution

Generate
Bayesian Inference

Bayes’ rule:

\[ p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)\,d\theta} \]

- Posterior distribution
- Intractable integral

Narrow

Wide
Variational Inference

\[ p(\theta | D) = \frac{p(D | \theta)p(\theta)}{\int p(D | \theta)p(\theta) d\theta} \]

Variational Approximation

\[ \approx q_\lambda(\theta) = \text{ExpFamily}(\lambda) \]

Maximize the Evidence Lower Bound (ELBO):

\[ \max_{\lambda} \mathcal{L}(\lambda) := \mathbb{E}_{q_\lambda} \left[ \log p(D, \theta) - \log q_\lambda(\theta) \right] \]
Relationship to Other Fields

• Perturbation to avoid local minima.
  – Gaussian homotopy and Continuation method.
  – Smoothed/graduated optimization.

• Online learning.
  – Exponentiated weighted averaging.

• Reinforcement learning.
  – Structured distribution.
  – q is the policy x environment (Levin 2018).
Gradient Descent

Maximize the Evidence Lower Bound (ELBO):

$$\max_{\lambda} \mathcal{L}(\lambda) := \mathbb{E}_{q_\lambda} \left[ \log p(\mathcal{D}, \theta) - \log q_\lambda(\theta) \right]$$

Gradient descent (GD) :

$$\lambda \leftarrow \lambda + \rho \nabla_\lambda \mathcal{L}$$
VI with Natural-Gradient Descent


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

Natural Gradient

Fisher Information Matrix (FIM)

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

- Fast convergence due to optimization in Riemannian manifold (not Euclidean space).
- But requires additional computations.
- Can we simplify/reduce this computation?
Can we simplify NGD computation? Yes, by using algorithms such as message passing/ backprop.

Conjugate-Computation VI
Khan and Lin, AI-STATS 2017
The key idea: Expectation Parameters

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

Sufficient statistics

**For Gaussians, it’s mean and correlation matrix**

$$\mathbb{E}_{q_\lambda} [\theta] = m$$
$$\mathbb{E}_{q_\lambda} [\theta \theta^\top] = mm^\top + V$$

A key relationship: $F(\lambda)^{-1} \nabla_\lambda \mathcal{L} = \nabla_\mu \mathcal{L}$

Natural Gradient wrt natural parameter

Gradient wrt expectation parameter

NGD: $\lambda \leftarrow \lambda + \rho \nabla_\mu \mathcal{L}$
Conjugate-Computation VI (CVI)

\[ \lambda \leftarrow \lambda + \rho \nabla_{\mu} \mathcal{L} \]

- In a “conjugate” model, this is equivalent to simply adding the natural parameters of the factors of a model.
- This is a type of conjugate computation, and enables “simple” updates for complex models.
**CVI on Bayesian Linear Regression**

\[ q_\lambda(\theta) := \mathcal{N}(m, V) \]

\[
\mathbb{E}_q \left[ (y - X\theta)^\top (y - X\theta) + \gamma \theta^\top \theta - \log q_\lambda(\theta) \right]
\]

\[-\mathbb{E}_{q_\lambda}[\theta]^\top X^\top y + \text{trace}\left[ X^\top X\mathbb{E}_{q_\lambda}[\theta\theta^\top] \right] \]

\[
\nabla \mathbb{E}_{q_\lambda}[\theta] = \begin{pmatrix} -X^\top y + 0 - V^{-1}m \\ X^\top X + \gamma I - V^{-1} \end{pmatrix}
\]

Expectation params | Natural Gradient
NGD == Newton’s Method

\[ m \leftarrow (1 - \rho)m - \rho [X^\top X + \gamma I]^{-1} X^\top y \]

Least-square solution

For \( \rho = 1 \), converges in 1 step (Newton’s method).

Gradient descent is suboptimal:

\[ m \leftarrow m - \alpha \left[ (X^\top X + \gamma I)m - X^\top y \right] \]

This property generalizes to all “conjugate” models, where forward-backward algorithm returns the natural-gradients of ELBO.
Conditionally-Conjugate Models

VMP: Sequential update with rho = 1

For CVI, rho can follow any schedule, and updates can be sequential or parallel.

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

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

- Lipschitz constant of (nonconvex) ELBO
- Gradient noise variance
- 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)
NGD for Deep Learning

Using CVI on Bayesian deep learning with Gaussian approximation. Reduces to a Newton step.
CVI for Bayesian Neural Network

\[ \mathbb{E}_q \left( \sum_{i=1}^{N} \log p(y_i | f_{\theta}(x_i)) + \gamma \theta^\top \theta - \log q_{\lambda}(\theta) \right) \]

\[
\begin{align*}
m &\leftarrow m - \beta (S + \gamma I)^{-1} [g_i(\theta) + \gamma m] \\
S &\leftarrow (1 - \beta) S + \beta H_i(\theta)
\end{align*}
\]

\[
\begin{align*}
\theta &\sim q_{\lambda}(\theta), \\
g_i(\theta) &:= -\nabla_{\theta} \log p(y_i | f_{\theta}(x_i)), \\
V^{-1} &\leftarrow S + \gamma I, \\
H_i(\theta) &:= -\nabla_{\theta}^2 \log p(y_i | f_{\theta}(x_i))
\end{align*}
\]
CVI for Bayesian Neural Network

\[(X^\top X + \gamma I)^{-1} X^\top y\]

\[m \leftarrow m - \beta (S + \gamma I)^{-1} [g_i(\theta) + \gamma m]\]

\[S \leftarrow (1 - \beta) S + \beta H_i(\theta)\]

\[\theta \sim q_{\lambda}(\theta), \quad g_i(\theta) := -\nabla_\theta \log p(y_i | f_\theta(x_i)),\]

\[V^{-1} \leftarrow S + \gamma I, \quad H_i(\theta) := -\nabla^2_\theta \log p(y_i | f_\theta(x_i))\]
Variational Adam for Mean-Field

Approximate the Hessian by square of gradients.

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

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

$$\theta_{\text{temp}} \leftarrow \theta + \varepsilon \cdot \sqrt{N \cdot \text{scale} + 1}$$

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

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

Iteration 1

- Adam
- Vadam (mean)
- Vadam (samples)

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

(by Runa E.)
LeNet-5 on CIFAR10

Figure 2: Evaluation metrics on Train and Test sets for both optimizers. Adam overfits while VOGN does a good job of keeping test and train errors close. VOGN outperforms Adam on CIFAR10 but underperforms on MNIST for test accuracy. For test log loss, VOGN is better than Adam in both cases. Model architectures given in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>VOGN</th>
<th>Adam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Loss</td>
<td>1.130</td>
<td>8.341</td>
</tr>
<tr>
<td>Error</td>
<td>37.01</td>
<td>40.47</td>
</tr>
</tbody>
</table>

(By Anirudh Jain)
Faster, Simpler, and More Robust

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

![Graph showing comparison between existing method and our methods (Vadam and VOGN).](image)
Faster, Simpler, and More Robust

Results on MNIST digit classification (for various values of Gaussian prior precision parameter)
Parameter-Space Noise for Deep RL

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
Reward 2038

Ruckstriesh et.al. 2010, Fortunato et.al. 2017, Plapper et.al. 2017
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.
Stochastic, Low-Rank, Approximate, Natural-Gradient (SLANG)

NeurIPS 2018

• Low-rank + diagonal covariance matrix.
• SLANG is linear in D!

Low-Rank + diagonal

\[ m \leftarrow m - \rho \left[ UU^\top + D \right]^{-1} [g_i + \gamma m] \]

\[(1 - \beta)S + \beta H_i(\theta)\]
SLANG is Faster than GD

Classification on USPS with BNNs
Generalization and Extensions
Deep Nets + Graphical Models

Neural Nets +
Linear Dynamical System

Neural Nets + GMM

\( \theta_{\text{PGM}} \)

\( \theta_{\text{NN}} \)

\( \theta_{\text{PGM}} \)

\( \theta_{\text{NN}} \)

\( x_1 \)

\( x_2 \)

\( x_3 \)

\( x_4 \)

\( y_1 \)

\( y_2 \)

\( y_3 \)

\( y_4 \)

\( z_n \)

\( x_n \)

\( \mathbf{y}_n \)

\( N \)
Amortized Inference on VAE + Probabilistic Graphical Models (PGM)

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,
  – 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.
Summary of the Talk

• Fast yet simple NGD for VI using Conjugate-Computation VI (AI-STATS 2017),
  – Generalization of forward-backward algorithm, Stochastic VI, Variational Message Passing etc.
  – Beyond conjugacy: Extends fast and simple NGD to deep nets (ICML 2018, NeurIPS 2018).

• Generalizations and Extensions,
  – VAEs (ICLR 2018), Mixture of Exponential Family, Evolution strategy (ArXiv 2017), etc.
Related Works

Sorry, if I miss some important work! Please email me.
EM, Forward-Backward, and VI

• 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.*
NGD: Author Name Starting with an H

• 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*. 
NGD: Author Name Starting with an S

• Salimans and Knowles (2013), *Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression.*
  – Approximate Natural-Gradient steps.
• Seth and Khardon (2016), *Monte Carlo Structured SVI for Two-Level Non-Conjugate Models.*
  – Applies to models with two level of hierarchy.
• Salimbani et al. (2018), *Natural Gradients in Practice: Non-Conjugate Variational Inference in Gaussian Process Models.*
  – Fast convergence on GP models
NGD for Bayesian Deep Learning

• Zhang et al. (2018), *Noisy Natural Gradient as Variational Inference*
  – For Bayesian deep learning (similar to Variational Adam).
Issues and Open Problems

• Automatic natural-gradient computation.
• Good implementation of message passing.
  – Gradient with respect to covariance matrices.
• Structured approximation for covariance.
• Comparisons on really large problems.
• Applications.
• Flexible posterior approximations.
References

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


References

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,
(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,

Fast and Simple Natural-Gradient Variational 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

• RIKEN AIP
  – Wu Lin (now at UBC), Didrik Nielsen (now at DTU), Voot Tangkaratt, Nicolas Hubacher, Masashi Sugiyama, Sunichi-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