Fast yet Simple Natural-Gradient Variational Inference in Complex Models

Mohammad Emtiyaz Khan
RIKEN Center for AI Project
http://emtiyaz.github.io

Joint work with
Wu Lin (UBC)
Didrik Nielsen, Voot Tangkaratt (RIKEN)
Yarin Gal (University of Oxford)
Akash Srivastava (University of Edinburgh)
Zuozhu Liu (SUTD, Singapore)
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.”
Current Focus: Methods to Improve Deep Learning

Data-efficiency, robustness, active learning, continual/online learning, exploration
Bayesian Inference

Compute a **probability distribution** over the unknowns given the data

“to know how much we don’t know”
Uncertainty Estimation

Scene

Uncertainty of depth estimates

(taken from Kendall et al. 2017)
Bayesian Inference is Difficult!

Bayes’ rule: \( p(w|D) = \frac{p(D|w)p(w)}{\int p(D|w)p(w)dw} \)

- 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 methods than gradients-based methods)
  - Khan & Lin (Aistats 2017), Khan et al. (ICML 2018), Khan & Nielsen (ISITA2018)
Natural-Gradient Descent for Gaussian Mean-Field VI

With weight-perturbation in Adam

(add noise to the weights during backprop)
Outline

• Background
  – Bayesian model and Variational Inference (VI)
  – VI using gradient descent
  – VI using natural-gradient descent

• Fast and simple natural-gradient VI

• Results on Bayesian deep learning and RL
Bayesian model
VI using gradient descent
Euclidean distance is inappropriate
VI using natural-gradient descent

BACKGROUND
A Bayesian Model

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

\[ p(w) = \text{ExpFamily}(\eta_0) \quad \eta_0 = \left\{ -\frac{1}{2} \Sigma^{-1}, \Sigma^{-1} \mu \right\} \]

\[ p(w|D) = \frac{p(D|w)p(w)}{\int p(D|w)p(w)dw} \]

Neural network

Intractable integral
Variational Inference with Gradients

\[ p(w|\mathcal{D}) \approx q_\lambda(w) = \text{ExpFamily}(\lambda) \]

\[ \lambda = \left\{ -\frac{1}{2} V^{-1}, V^{-1} m \right\} \]

\[ \max_\lambda \mathcal{L}(\lambda) := \mathbb{E}_{q_\lambda} \left[ \log \frac{p(w)}{q_\lambda(w)} \right] + \sum_{i=1}^{N} \mathbb{E}_{q_\lambda} [\log p(\mathcal{D}_i|w)] \]

Regularizer \quad Data-fit term

\[ \lambda_{t+1} = \lambda_t + \rho_t \nabla_\lambda \mathcal{L}_t \]

\[ = \arg \max_\lambda \lambda^T \nabla_\lambda \mathcal{L}_t - \frac{1}{2\rho_t} \| \lambda - \lambda_t \|^2 \]
Euclidean Distance is inappropriate!

Two Gaussians with mean 1 and 10 respectively and variances equal to $\sigma_1$ have Euclidean distance = 10

Same as the top row but with the variance $\sigma_2 > \sigma_1$ but still Euclidean distance = 10

VI using Natural-Gradient Descent

Fisher Information Matrix (FIM)

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

\[
\max_{\lambda} \lambda^T \nabla_\lambda \mathcal{L}_t - \frac{1}{2\rho_t} (\lambda - \lambda_t)^T F(\lambda_t)(\lambda - \lambda_t)
\]

\[ \lambda_{t+1} = \lambda_t + \rho_t F(\lambda_t)^{-1} \nabla_\lambda \mathcal{L}_t \]

Natural Gradients: \[ \tilde{\nabla}_\lambda \mathcal{L}_t \]
Natural-Gradients require computation of the FIM

Can we avoid this?

Yes, by computing the gradient \textit{w.r.t.} the expectation parameter of exponential family
“Simple” Natural-Gradients

Part I
Expectation Parameters of Exp-Family

Wainwright and Jordan, 2006

Mean/expectation /moment parameters

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

Sufficient statistics

\[ \mathbb{E}_{q_\lambda} [w] = m \]
\[ \mathbb{E}_{q_\lambda} [ww^\top] = mm^\top + V \]
NatGrad Descent == Mirror Descent

Raskutti and Mukherjee, 2015, Khan and Lin 2017

$$
\lambda_{t+1} = \lambda_t + \rho_t F(\lambda_t)^{-1} \nabla_\lambda \mathcal{L}_t
$$

$$
\max_\mu \mu^T \nabla_\mu \mathcal{L}_t - \frac{1}{\rho_t} K L[q_\mu \| q_{\mu_t}] = 0
$$

$$
\nabla_\mu \mathcal{L}_t - \frac{1}{\rho_t} (\lambda - \lambda_t) = 0
$$

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

$$
\nabla_\lambda \mathcal{L}_t = F(\mu_t)^{-1} \nabla_\mu \mathcal{L}_t := \tilde{\nabla}_\mu \mathcal{L}_t
$$
Dually-Flat Riemannian Structure

See Amari’s book 2016

Figure from Wainwright and Jordan, 2006

For VI, natural gradient in natural-parameter is computationally simpler than in the expectation parameter space.
Natural-Gradient Descent in the Natural-Parameter Space

\[
\max_{\lambda} \mathcal{L}(\lambda) := \mathbb{E}_{q_\lambda} \left[ \log \frac{p(w)}{q_\lambda(w)} \right] + \sum_{i=1}^{N} \mathbb{E}_{q_\lambda} \left[ \log p(D_i|w) \right]
\]

\[
\tilde{\nabla}_{\lambda} \mathcal{L} = \nabla_\mu \mathcal{L}
\]

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

Conjugate

Nonconjugate

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

\[
p(w) = \text{ExpFamily}(\eta_0)
\]

\[
q_\lambda(w) = \text{ExpFamily}(\lambda)
\]
Natural-Gradient Descent for VI
Khan and Lin 2017, Khan and Nielsen 2018

\[ \lambda_{t+1} = (1 - \rho_t) \lambda_t + \]

\[ \rho_t \left[ \eta_0 + \sum_{i=1}^{N} \nabla_{\mu} \mathbb{E}_{q_\lambda} [\log p(D_i|w)] \big|_{\mu=\mu(\lambda_t)} \right] \]

This is a generalization of Variational Message Passing (Winn and Bishop 2005) and stochastic variational inference (Hoffman et al. 2013) to nonconjugate models, such as, Bayesian neural networks.

Convergence proof is in Khan et al. UAI 2018
Approximate Bayesian Filter

\[ q(w | \lambda_{t+1}) \propto [q(w | \lambda_t)]^{(1-\rho_t)} \left[ p(w) e^{\nabla_\mu \mathbb{E}[\log p(D_i | w)]} \right]^{\rho_t} \]

New Approximation  Previous Approximation  Exponential Prior  DNN Likelihood

Optimal natural-parameter == natural-gradient

\[ \lambda_* = \eta_0 + \sum_{i=1}^{N} \nabla \lambda \mathbb{E}_{q_*^\lambda} [\log p(D_i | w)] \]

Similar to EP, we get local approximations, but now they are natural-gradients of the local factors.

\[ q(w | \lambda_*) \propto p(w) \left[ \prod_{i=1}^{N} e^{\phi(w)^\top \nabla \lambda \mathbb{E}_{q_*^\lambda} [\log p(D_i | w)]} \right] \]
“Fast” Natural-Gradients

Part II: Application to Bayesian deep learning
VI as Weight-Perturbed Adam (Vadam)

Adam
1: while not converged do
2: \( \theta \leftarrow \mu \)
3: Randomly sample a data example \( D_i \)
4: \( g \leftarrow -\nabla \log p(D_i | \theta) \)
5: \( m \leftarrow \gamma_1 m + (1 - \gamma_1) g \)
6: \( s \leftarrow \gamma_2 s + (1 - \gamma_2) (g \circ g) \)
7: \( \hat{m} \leftarrow m / (1 - \gamma_1^t), \hat{s} \leftarrow s / (1 - \gamma_2^t) \)
8: \( \mu \leftarrow \mu - \alpha \hat{m} / (\sqrt{\hat{s}} + \delta) \)
9: \( t \leftarrow t + 1 \)
10: end while

Vadam
1: while not converged do
2: \( \theta \leftarrow \mu + \sigma \circ \epsilon \), where \( \epsilon \sim \mathcal{N}(0, I) \), \( \sigma \leftarrow 1 / \sqrt{Ns + \lambda} \)
3: Randomly sample a data example \( D_i \)
4: \( g \leftarrow -\nabla \log p(D_i | \theta) \)
5: \( m \leftarrow \gamma_1 m + (1 - \gamma_1) (g + \lambda \mu / N) \)
6: \( s \leftarrow \gamma_2 s + (1 - \gamma_2) (g \circ g) \)
7: \( \hat{m} \leftarrow m / (1 - \gamma_1^t), \hat{s} \leftarrow s / (1 - \gamma_2^t) \)
8: \( \mu \leftarrow \mu - \alpha \hat{m} / (\sqrt{\hat{s}} + \lambda / N) \)
9: \( t \leftarrow t + 1 \)
10: end while

Figure 1. Comparison of Adam (left) and one of our proposed method Vadam (right). Adam performs maximum-likelihood estimation while Vadam performs variational inference, yet the two pseudocodes differ only slightly (differences highlighted in red). A major difference is in line 2 where, in Vadam, weights are perturbed during the gradient evaluations.
Natural-gradient vs gradients

**Natural-Gradient VI**

\[
\mu \leftarrow \mu - \beta \sigma^2 \nabla_\mu \mathcal{L} \\
\frac{1}{\sigma^2} \leftarrow \frac{1}{\sigma^2} + 2\beta \nabla_{\sigma^2} \mathcal{L}
\]

**Existing Methods**

\[
\mu \leftarrow \mu + \alpha \frac{\nabla_\mu \mathcal{L}}{\sqrt{s_\mu + \delta}} \\
\sigma \leftarrow \sigma + \alpha \frac{\nabla_\sigma \mathcal{L}}{\sqrt{s_\sigma + \delta}}
\]

(Graves et al. 2011, Blundell et al. 2015)
Approximate Natural-Gradient

Natural-Gradient VI

$$\mu \leftarrow \mu - \beta \sigma^2 \nabla_{\mu} \mathcal{L}$$

$$\frac{1}{\sigma^2} \leftarrow \frac{1}{\sigma^2} + 2\beta \nabla_{\sigma^2} \mathcal{L}$$

Gauss-Newton Gradient-Magnitude

$$\nabla_{\sigma^2} \mathbb{E}[f(\theta)] \approx \sum_i \nabla^2_{\theta\theta} f_i(\theta) \approx \sum_i [\nabla_{\theta} f_i(\theta)]^2 \approx \left[ \sum_i \nabla_{\theta} f_i(\theta) \right]^2$$

MC

Gauss-Newton

Gradient-Magnitude

Hard to implement

Easy to implement and accurate

Easy to implement

Inaccurate

Gauss-Newton is proposed by Graves et al. 2011 for SG-VI methods
Vprop: Natural-Gradient VI via Weight-Perturbation in RMSprop

RMSprop for Max-likelihood

\[
\begin{align*}
\theta & \leftarrow \mu \\
g & \leftarrow \hat{\nabla}_\theta f(\theta) \\
\varepsilon & \leftarrow (1 - \beta)s + \beta g^2 \\
\mu & \leftarrow \mu + \alpha \frac{g}{\sqrt{s + \delta}}
\end{align*}
\]

Solves max f

Vprop for variational inference

\[
\begin{align*}
\theta & \leftarrow \mu + \frac{\epsilon}{\sqrt{s + \lambda}} \\
g & \leftarrow \hat{\nabla}_\theta f(\theta) \\
\varepsilon & \leftarrow (1 - \beta)s + \beta g^2 \\
\mu & \leftarrow \mu + \alpha \frac{g + \lambda \mu}{\sqrt{s + \lambda}}
\end{align*}
\]

Approximately solves max \( \mathcal{L}(\mu, \sigma^2) \)
Weight-Perturbed Adam (Vadam)

We can derive Adam like update by adding a “natural-momentum”

\[ m \leftarrow \langle m, \nabla m \mathcal{L}_t \rangle + \frac{1}{\beta} KL(q\|q_t) - \frac{\gamma}{\beta} KL(q\|q_{t-1}) \]

Expectation parameter

We can perform VI using weight-perturbation in Adam \(^{Khan et al. ICML 2018}\)
Error in the Uncertainty Estimates

**Theorem 1.** Denote the full-batch gradient with respect to $\theta_j$ by $g_j(\theta)$ and the corresponding full-batch GGN approximation by $h_j(\theta)$. Suppose minibatches $\mathcal{M}$ are sampled from the uniform distribution $p(\mathcal{M})$ over all $\binom{N}{M}$ minibatches, and denote a minibatch gradient by $\hat{g}_j(\theta; \mathcal{M})$, then the expected value of the GM approximation is the following,

$$
\mathbb{E}_{p(\mathcal{M})} \left[ \hat{g}_j(\theta; \mathcal{M})^2 \right] = wh_j(\theta) + (1 - w)[g_j(\theta)]^2, \quad (15)
$$

where $w = \frac{1}{M} \left( N - M \right)/(N - 1)$. 

---

Batch Gauss-Newton

Gradient-Magnitude

Batch gradient
Related Work

• Natural-Gradient Methods for VI

• Gradient methods for VI

• Zhang et al. ICML 2018
  – Very similar to our ICML paper and our previous work on Variational Adaptive Newton method.

• Mandt et al. 2017, SGD as VI.

• Global optimization methods
  – Optimization by smoothing, graduated optimization, Gaussian homotopy, etc.
  – Entropy-SGD, noisy networks for exploration etc.
Results
Illustration

VOGN uses Gauss-Newton
Vadam uses Gradient-Magnitude
Quality of Posterior Approximation

VOGN-1 uses Gauss-Newton with minibatch of size 1

Vadam uses Gradient-Magnitude with minibatch > 1
Effect of Minibatch on the Accuracy

As we decrease minibatch size, the accuracy improves, but the stochastic noise increases which might slow-down the algorithm.
Vadam performs comparable to BBVI

1 layer 50 hidden units with ReLU on “energy” (N=768, D=8) and “kin8nm” (N=8192, D=8), 5 MC samples for Vadam, 10 for BBVI, minibatch of 32
Gauss-Newton Converges fast

1 layer 50 hidden units with tanh on Breast Cancer [N=683, D=10], minibatch of size 1 with 16 MC samples and step-sizes = 0.01

![Graph showing log2 loss vs Epoch for BBVI, Vadam, and VOGN]
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.
Parameter-Space Noise for Deep RL

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

VAdaGrad (noise updated with natural-gradients)

SGD (noise updated with standard gradients)

SGD (no noise injection)

Reward 5264

Reward 3674

Ruckstriesh et.al.2010, Fortunato et.al. 2017, Plapper et.al. 2017
Improving the “Marginal-value” of Adam/AdaGrad

SGD and Vadam reach a better minimum than Adam and AdaGrad

2-layer LSTM on War & Peace dataset.

The example taken from Wilson et.al. 2017 “Marginal-value of adaptive-gradient method”
Summary
References

https://emtiyaz.github.io

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 ]

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

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 + GPs ] [ Code for Correlated Topic Model ]
Thanks!

I am looking for post-docs, research assistants, and interns

See details at https://emtiyaz.github.io