Fast yet Simple Natural-Gradient Variational Inference in Complex Models

Mohammad Emtiyaz Khan

RIKEN Center for AI Project, Tokyo, Japan

http://emtiyaz.github.io

Joint work with

Wu Lin (UBC)

Didrik Nielsen, Voot Tangkaratt (RIKEN)

Yarin Gal (University of Oxford)

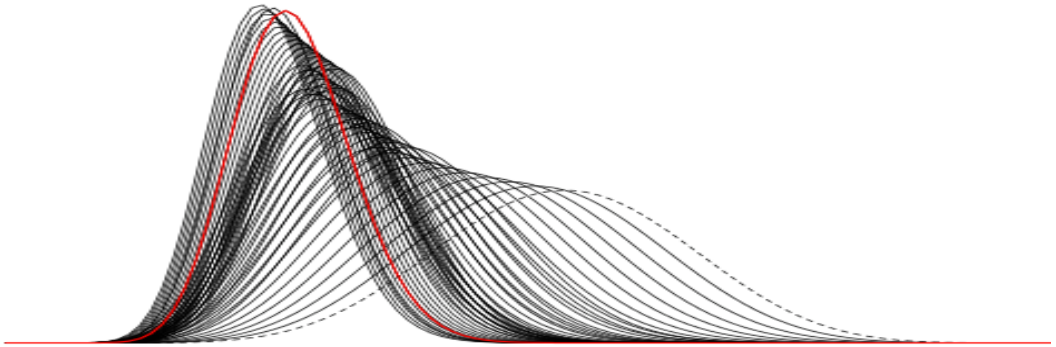
Akash Srivastava (University of Edinburgh)

Zuozhu Liu (SUTD, Singapore)

Bayesian Inference

Compute a probability distribution over the unknowns given the knowns

"to know how much we don't know".



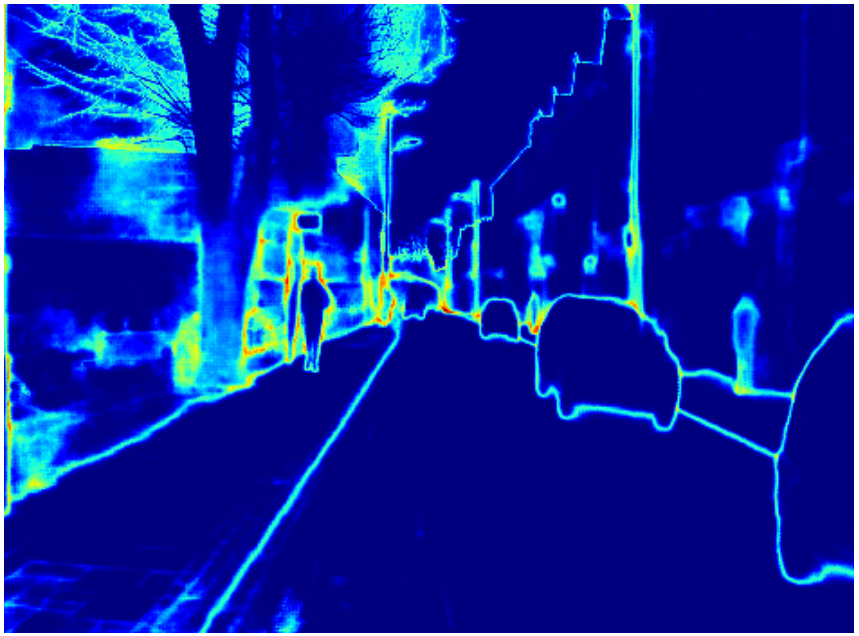
Bayesian Deep Learning

To improve many aspects of deep learning:
data-efficiency, robustness, active learning,
continual/online learning, exploration

Example

Scene

Uncertainty of depth estimates



(taken from Kendall et al. 2017)

Bayesian Inference is Difficult!

Bayes' rule: $$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{\underbrace{\int p(\mathcal{D}|w)p(w)dw}_{\text{Intractable integral}}}$$

Data

Parameters

- 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)

Distribution Function

$$\min_{\lambda} \mathbb{E}_{q_\lambda(w)} \Big[f(w, \lambda) \Big]$$

Distribution parameters

Model parameters

Arises in approximate Bayesian inference, reinforcement learning, stochastic search, discrete optimization

Bayesian model

VI using gradient descent

Euclidean distance is inappropriate

VI using natural-gradient descent

BACKGROUND

Approximate Bayesian Inference

Neural network

$$p(\mathcal{D}|w) = \prod_{i=1}^{N} p(y_i|f_w(x_i)) \qquad p(w) = \text{ExpFamily}(\eta_0)$$

$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{\int p(\mathcal{D}|w)p(w)dw} \qquad \eta_0 = \left\{-\frac{1}{2}\Sigma^{-1}, \Sigma^{-1}\mu\right\}$$

$$\approx \; q_\lambda(w) = \text{ExpFamily}(\lambda) \qquad \lambda = \left\{-\frac{1}{2}V^{-1}, V^{-1}m\right\}$$

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

Regularizer

Data-fit term

Distribution Function

$$\min_{\lambda} \mathbb{E}_{q_\lambda(w)} \Big[f(w, \lambda) \Big]$$

Distribution parameters

Model parameters

Arises in approximate Bayesian inference, reinforcement learning, stochastic search, discrete optimization

Optimization using Gradient Descent

Distribution Function

$$\min_{\lambda} \mathbb{E}_{q_\lambda(w)} \Big[f(w, \lambda) \Big]$$

Distribution parameters

Model parameters

$$\lambda_{t+1} = \lambda_t + \rho_t \nabla_\lambda \mathcal{L}_t \quad \Leftarrow \quad \nabla_\lambda \mathcal{L}_t - \frac{1}{\rho_t}(\lambda - \lambda_t) = 0$$

$$= \arg\max_{\lambda} \lambda^T \nabla_\lambda \mathcal{L}_t - \frac{1}{2\rho_t} \|\lambda - \lambda_t\|^2$$

Using Natural-Gradient Descent

Fisher Information Matrix (FIM)

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

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

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

Natural Gradients: $\tilde{\nabla}_\lambda \mathcal{L}_t$

“Simple” Natural-Gradients

Part I

Natural-Gradients require computation of the FIM. Can we avoid this?

Khan & Lin (AI-Stats 2017), Khan & Nielsen (ISITA2018)

Expectation Parameters of Exp-Family

Wainwright and Jordan, 2006

Mean/expectation /moment parameters

Sufficient statistics

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

$$\mathbb{E}_{q_\lambda}[w] = m$$

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

NatGrad Descent == Mirror Descent

Raskutti and Mukherjee, 2015, Khan and Lin 2017, Hensman et al. 2012

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

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

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

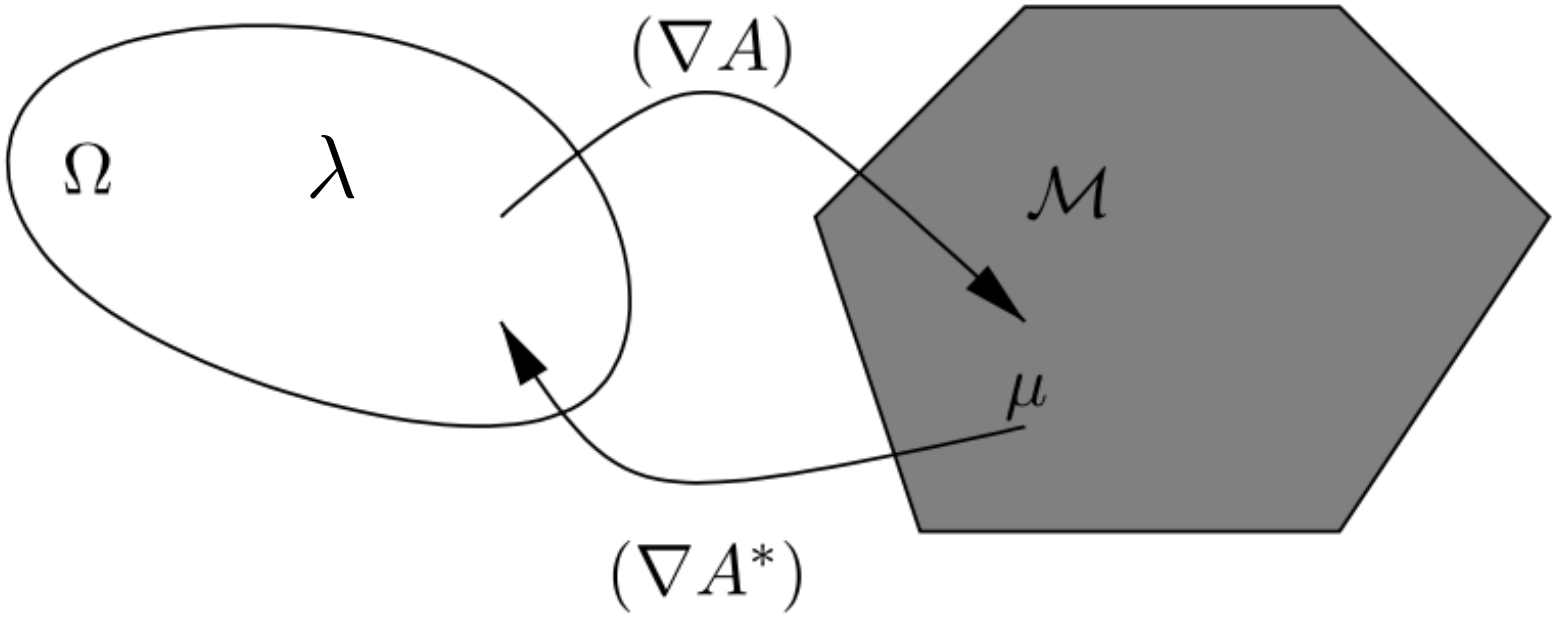
$$\nabla_\mu \mathcal{L}_t = F(\lambda_t)^{-1} \nabla_\lambda \mathcal{L}_t := \tilde{\nabla}_\lambda \mathcal{L}_t$$

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

Dually-Flat Riemannian Structure

See Amari's book 2016

Figure from Wainwright and Jordan, 2006



For variational inference, natural gradient wrt natural-parameters is computationally simpler than in the expectation parameter space (Hoffman et al. 2013)

Example: Penalized Linear Regression

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

$$\mathbb{E}_{q_\lambda}\left[-\log q_\lambda(w) + \frac{\tau}{2} w^T w + \frac{1}{2}\sum_{i=1}^{N}(y_i - x_i^T w)^2\right]$$

For terms in exp-family, this is linear in the expectation parameters!

Mean of q $\underbrace{(X^T X + \tau I)^{-1}}_{\text{Covariance of q}} X^T y$

Gradient wrt Expectation Parameters

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

Conjugate

Nonconjugate

$$p(\mathcal{D}|w) = \prod_{i=1}^{N} p(y_i|f_w(x_i))$$
$$p(w) = \text{ExpFamily}(\eta_0)$$
$$q_\lambda(w) = \text{ExpFamily}(\lambda)$$

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

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

Conjugate

Nonconjugate

(similar to SVI, Hoffmann et al. 2013)

Natural-Gradients as Message Passing

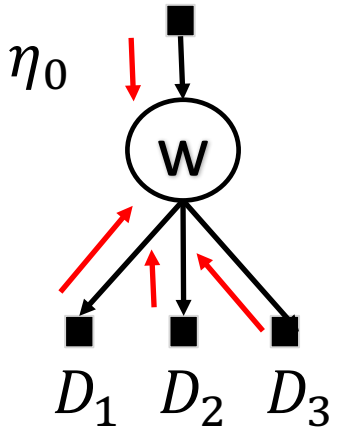
Khan and Lin 2017, Khan and Nielsen 2018

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

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

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

Natural gradients
== Local natural parameters



A generalization of Variational Message Passing (Winn and Bishop 2005) and stochastic variational inference (Hoffman et al. 2013) to nonconjugate models.

“Fast” Natural-Gradients

Deep Learning with Gaussian mean-field approximation

Natural-Gradient VI ≈ Adam for MLE

Notation change: $\theta = w$, and $\lambda \neq$ natural parameter, but precision of the Gaussian prior, μ is not the expectation parameter but the mean of the Gaussian approximation

Gaussian Approximation

Gaussian prior

Non-Gaussian (DNN)

$$\max_{\mu,\sigma^2} \mathcal{L}(\mu,\sigma^2) := \mathbb{E}_q\Big[\log \frac{\mathcal{N}(\theta|0,\lambda I)}{\mathcal{N}(\theta|\mu,\sigma^2)}\Big] + \sum_{i=1}^{N} \mathbb{E}_q[\log p(\mathcal{D}_i|\theta)]$$

Gaussian variational approximation

Natural-Gradient VI

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

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

Doubly Stochastic Approximation

$$\max_{\mu,\sigma^2} \mathcal{L}(\mu,\sigma^2) := \mathbb{E}_q\Big[\log \frac{\mathcal{N}(\theta|0,\lambda I)}{\mathcal{N}(\theta|\mu,\sigma^2)}\Big] + \sum_{i=1}^{N} \mathbb{E}_q[\log p(\mathcal{D}_i|\theta)]$$

Gaussian prior

Gaussian variational approximation

Non-Gaussian (DNN)

$f_i(\theta)$

Natural-Gradient VI

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

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

Sample a θ from q and a minibatch

$$\approx \lambda\mu + \frac{N}{M}\sum_{i\in\mathcal{M}} \nabla_\theta f_i(\theta)$$

$$\approx \lambda - \frac{1}{\sigma^2} + \frac{N}{M}\sum_{i\in\mathcal{M}} \nabla^2_{\theta\theta} f_i(\theta)$$

MLE vs Natural-Gradient VI

RMSprop for Max-likelihood

$$\theta \leftarrow \mu$$
$$g \leftarrow \frac{1}{M} \sum_i \nabla_\theta \log p(\mathcal{D}_i|\theta)$$
$$s \leftarrow (1-\beta)s + \beta g^2$$
$$\mu \leftarrow \mu + \alpha \frac{g}{\sqrt{s}+\delta}$$

Natural-Gradient VI (Khan, Lin 2017, Khan, Nielsen 2018)

$$\theta \leftarrow \mu + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, Ns + \lambda)$$
$$g \leftarrow \frac{1}{M} \sum_i \nabla_\theta \log p(\mathcal{D}_i|\theta)$$
$$s \leftarrow (1-\beta)s + \beta \frac{1}{M} \sum_i \left[\nabla^2_{\theta\theta} \log p(\mathcal{D}_i|\theta)\right]^2$$
$$\mu \leftarrow \mu + \alpha \frac{g + \lambda\mu/N}{\sqrt{s} + \lambda/N}$$

Variational Online-Newton (VON)
Khan et al. 2017

Variational Online Gauss-Newton (VOGN)

Variational RMSprop (Vprop)

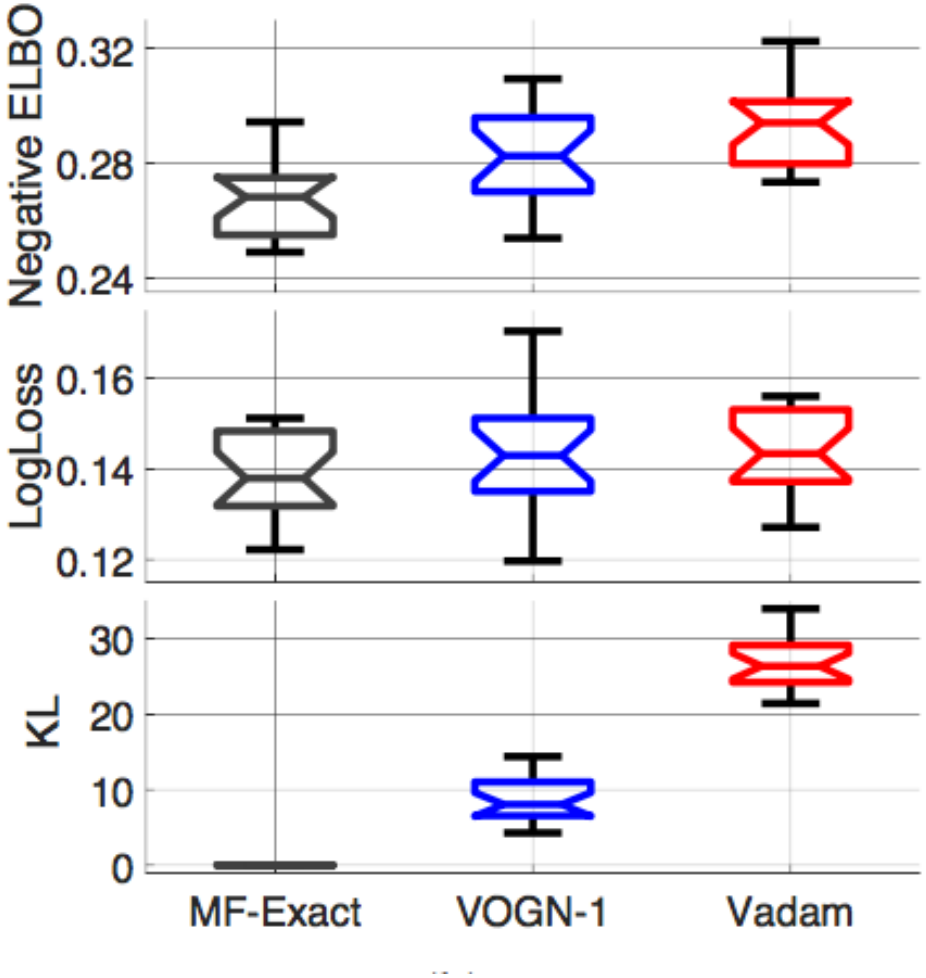
Variational Adam (Vadam)

Estimate Distributions using Weight-Perturbation in Adam

Choose a small minibatch size (see Theorem 1 in the ICML paper)

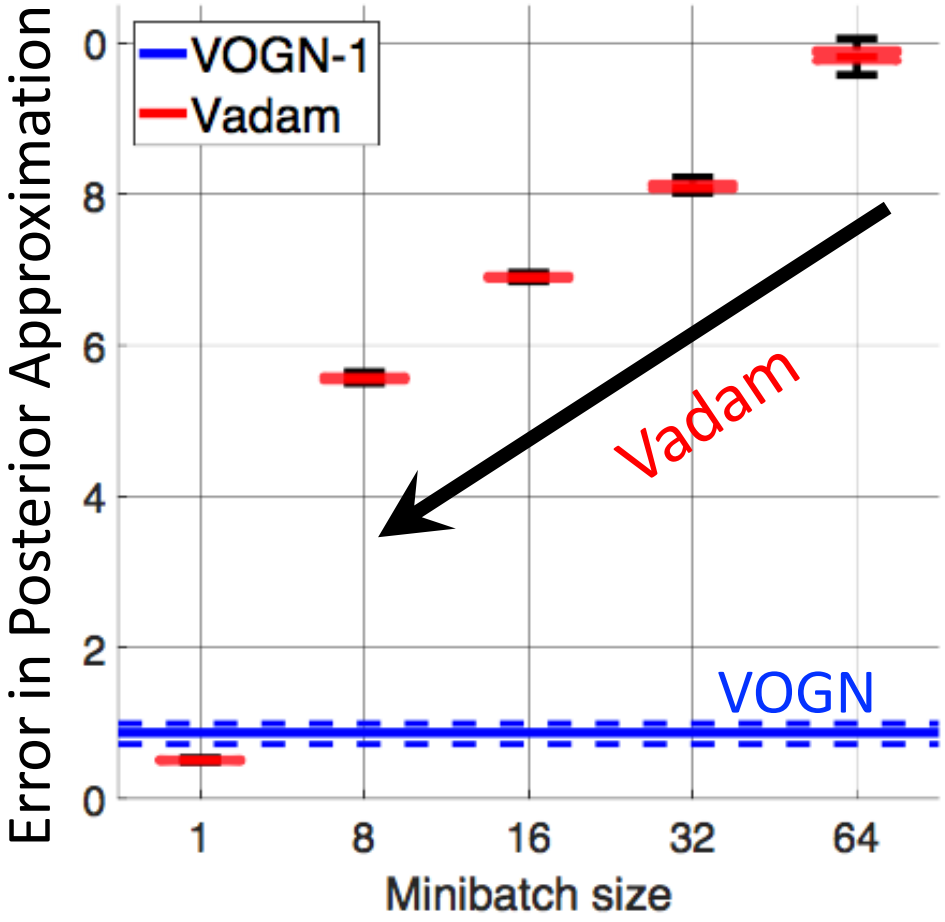
Results for Variational Inference

Quality of Posterior Approximation



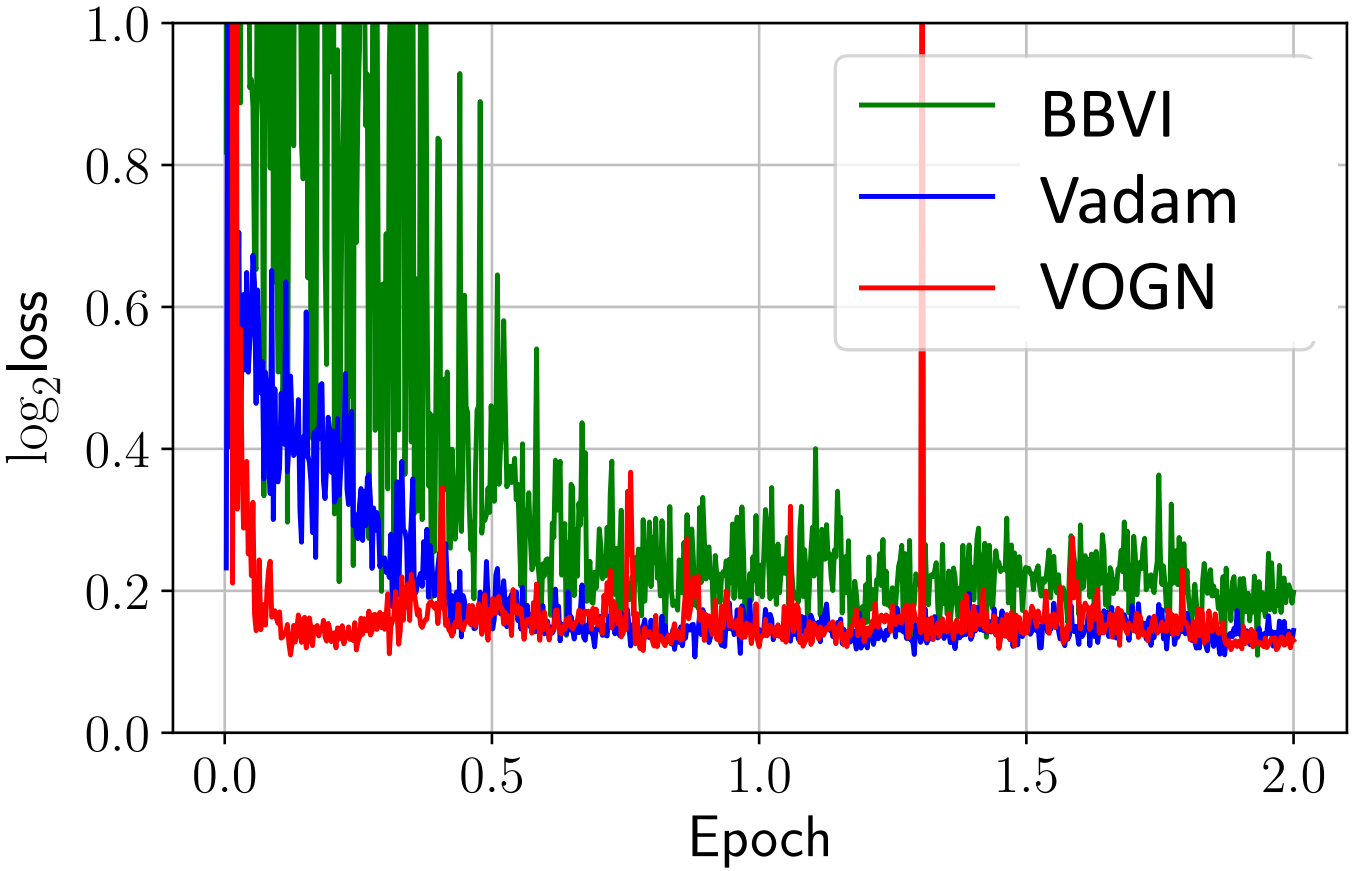
VOGN uses Gauss-Newton with minibatch of size 1

Vadam uses Gradient-Magnitude with minibatch > 1



Bayesian logistic regression on "Breast-Cancer" (N=683, D=8)

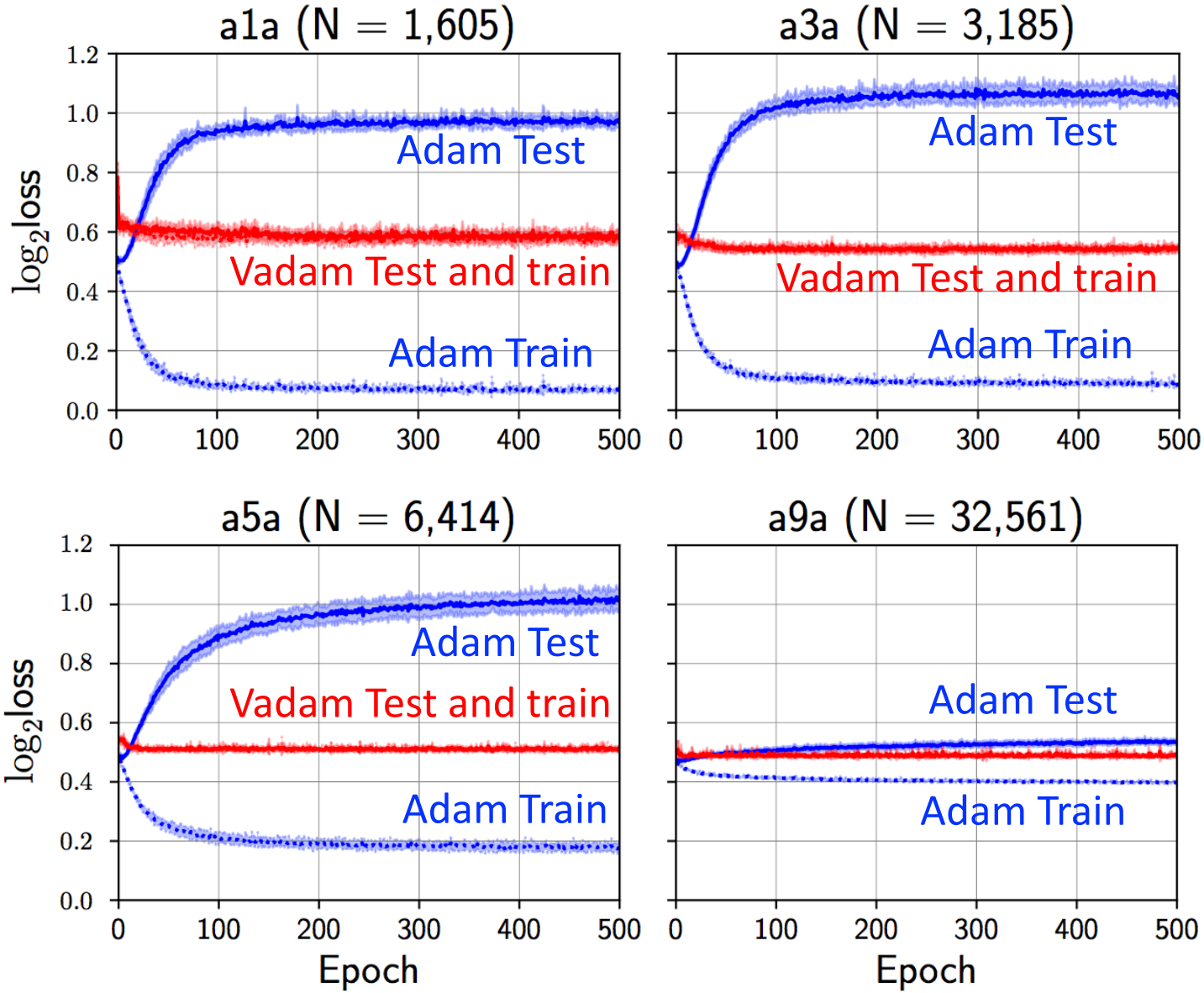
As we reduce the minibatch size, Vadam gives similar performance as VOGN.



1 layer 64 hidden Units with ReLu on Breast Cancer [N=683, D=10]

VOGN shows fast convergence

Reduce Overfitting with VI

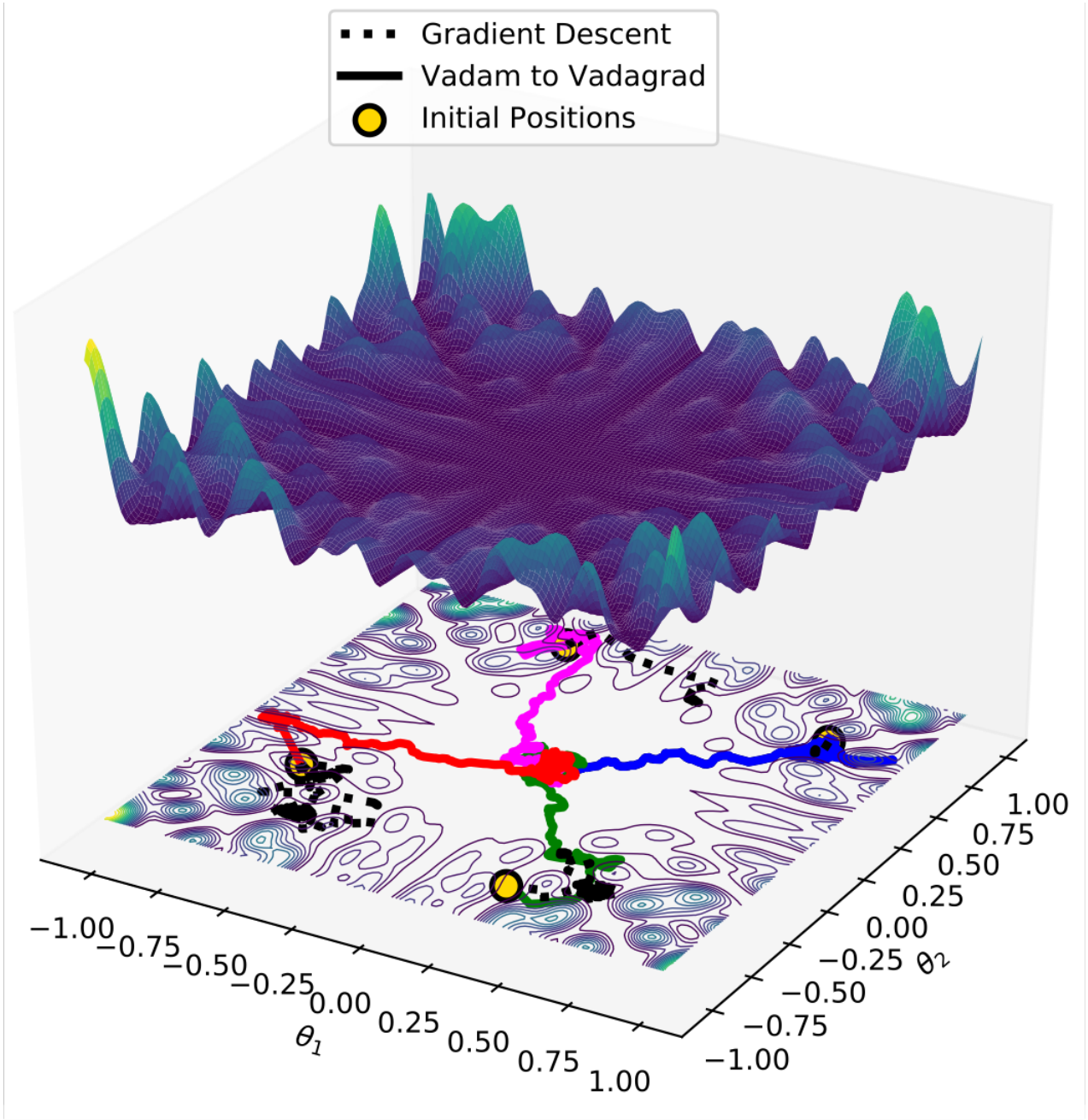


VI shows consistent train-test performance, while MLE overfits when N is small

BNN classification on a1a - a9a datasets

Results for Optimization

$$\max_{\lambda} \mathbb{E}_{q_\lambda(w)}[f(w)]$$



Avoiding Local Minima

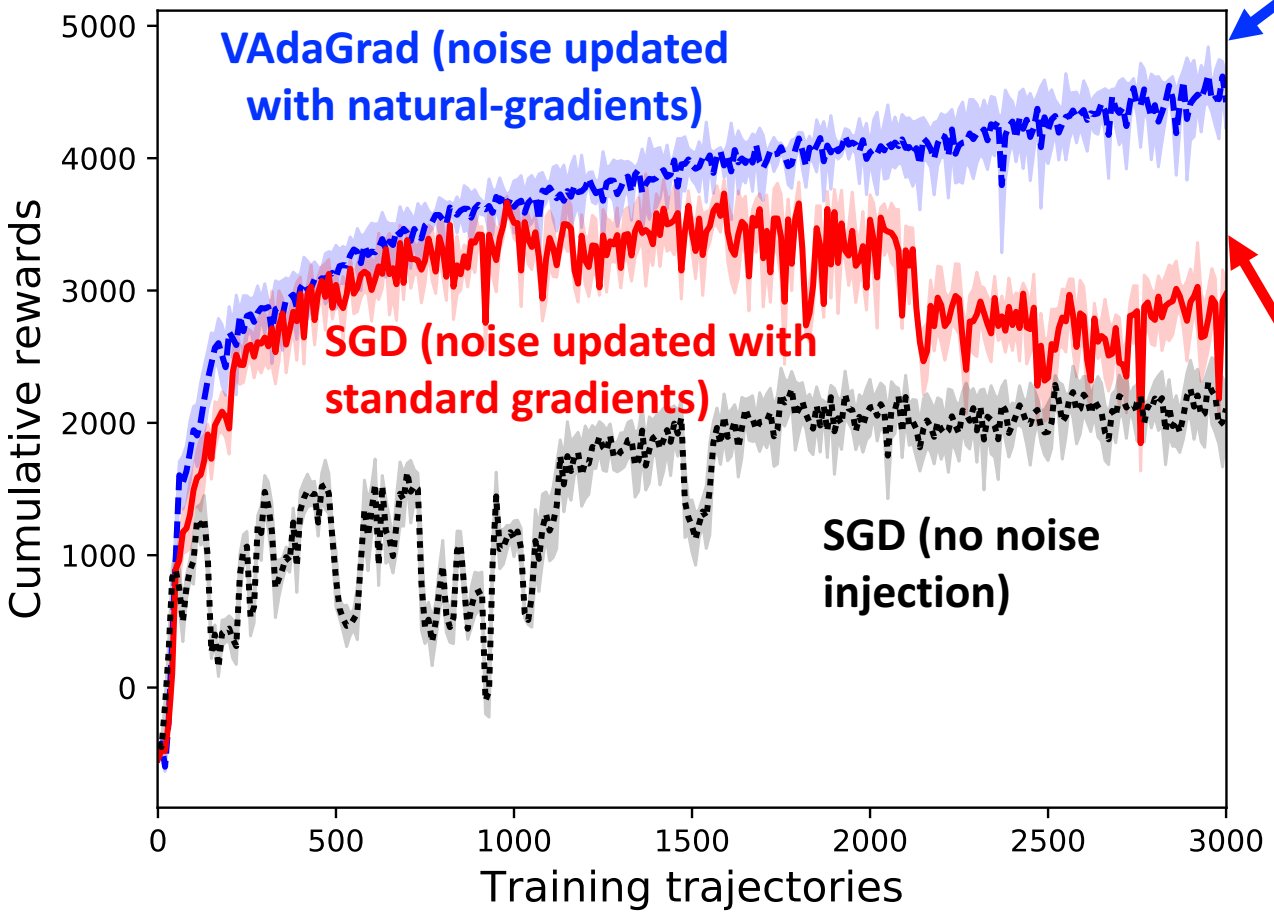
An example taken from Casella and Robert's book.

Vadam reaches the flat minima, but GD gets stuck at a local minima.

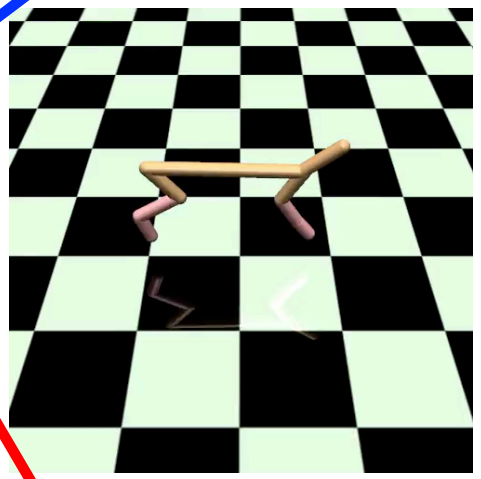
Optimization by smoothing, Gaussian homotopy/blurring etc., Entropy SGLD etc.

Parameter-Space Noise for Deep RL

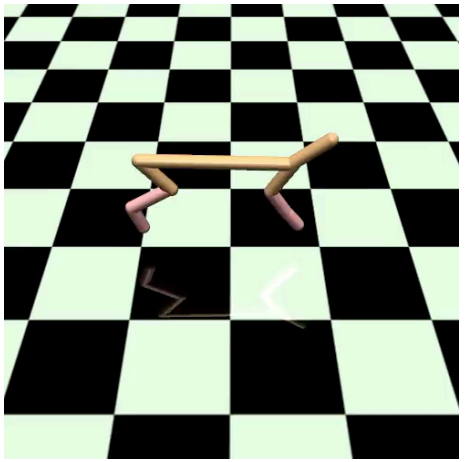
On OpenAI Gym Cheetah with DDPG
with DNN with [400,300] ReLU



Reward 5264



Reward 3674



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

Summary

- Natural gradients exploit information geometry, but could be difficult to compute
- Simple updates can be obtained by using expectation parameterization
  - Messages passing
- Fast updates for Bayesian neural networks
  - Implement using Adam

Related Work

- Natural-Gradient Methods for VI
  - Sato 2001, Honkela et al. 2010, Hoffman et al. 2013
- Gradient methods for VI
  - Rangnathan et al. 2014, Graves et al. 2011, Blundell et al. 2015, Salimans and Knowles 2013
- 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.

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,
(ICML 2018) **M.E. KHAN**, D. NIELSEN, V. TANGKARATT, W. LIN, Y. GAL, AND A. SRIVASTAVA, [ArXiv Version] [Code]

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!

https://emtiyaz.github.io