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)

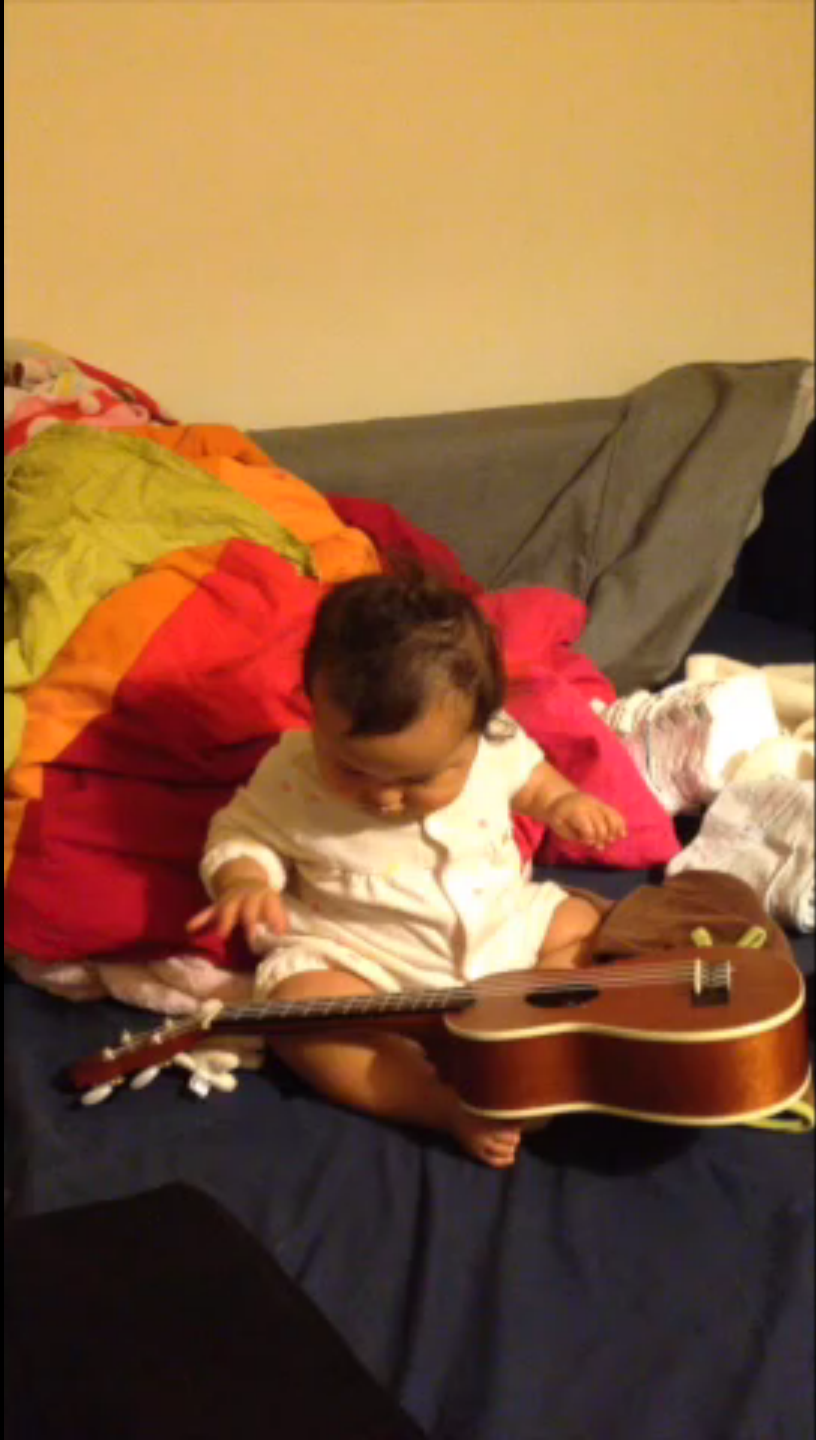
Zuozhu Liu (SUTD, Singapore)

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."

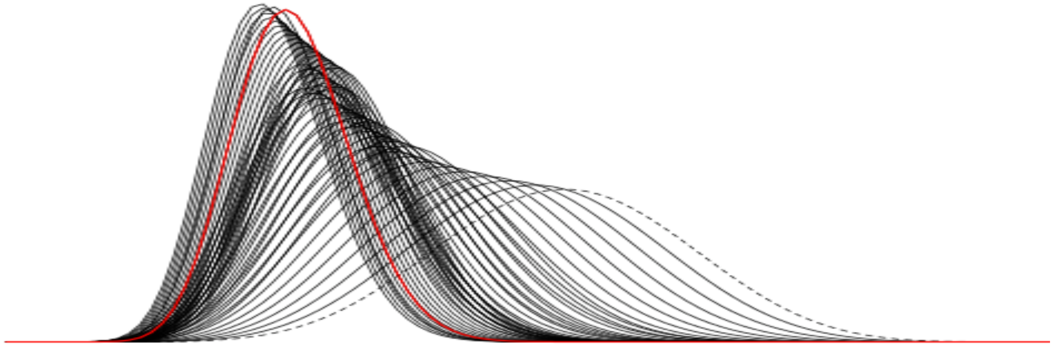
Current Focus: Methods to Improve Deep Learning

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

Bayesian Inference

Compute a probability distribution over the unknowns given the data

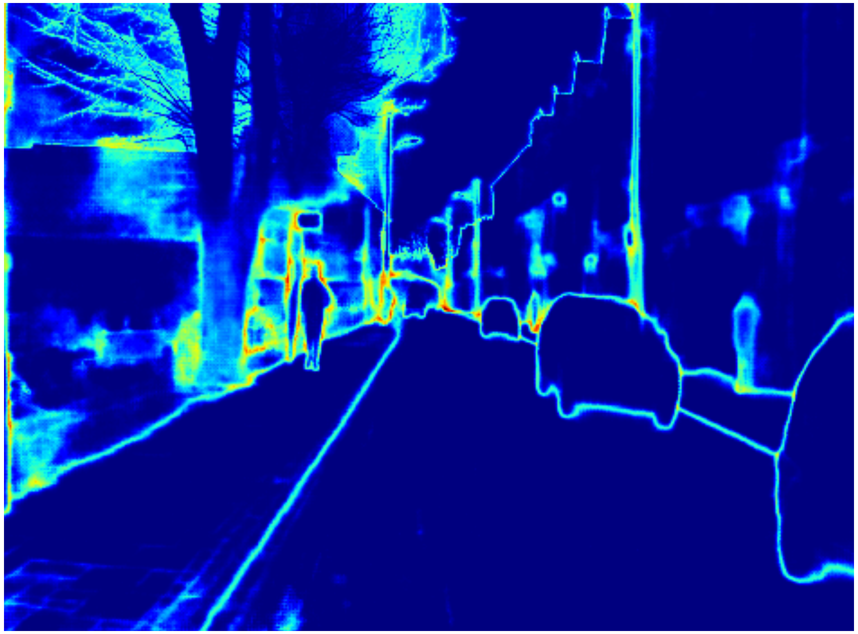
"to know how much we don't know"



Uncertainty Estimation

Scene

Uncertainty of depth estimates



(taken from Kendall et al. 2017)

Bayesian Inference is Difficult!

Data — Parameters

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

Intractable integral

- 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)

Outline

- Backgound
  - 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

Neural network

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

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

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

Variational Inference with Gradients

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

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

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

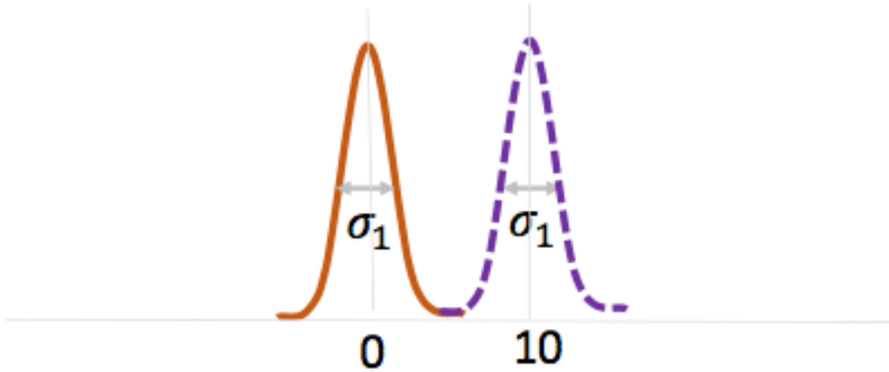
Regularizer   Data-fit term

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

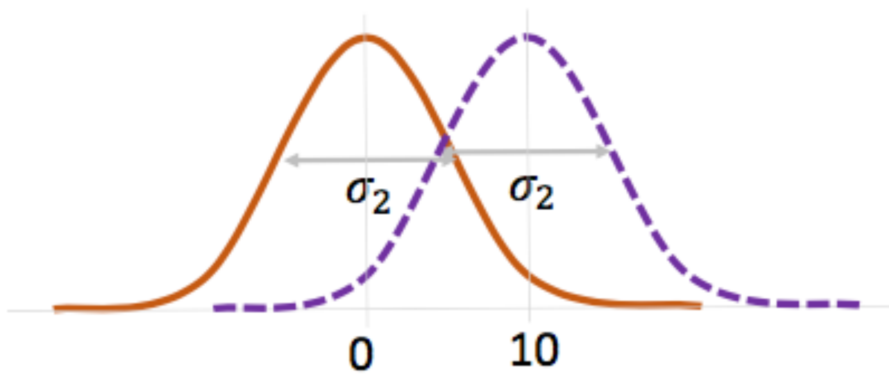
$$= \arg\max_\lambda \lambda^T \nabla_\lambda \mathcal{L}_t - \frac{1}{2\rho_t} \|\lambda - \lambda_t\|^2$$

Euclidean Distance is inappropriate!

Two Gaussians with mean 1 and 10 respectively
and variances equal to σ_1 have Euclidean distance = 10



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



(Amari 1999, Sato 2001, Honkela et.al. 2010, Hoffman et.al. 2013, Khan and Lin 2017)

VI using Natural-Gradient Descent

Fisher Information Matrix (FIM)

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

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

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

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

Natural-Gradients require computation of the FIM

Can we avoid this?

Yes, by computing the gradient 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 — 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

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

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

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

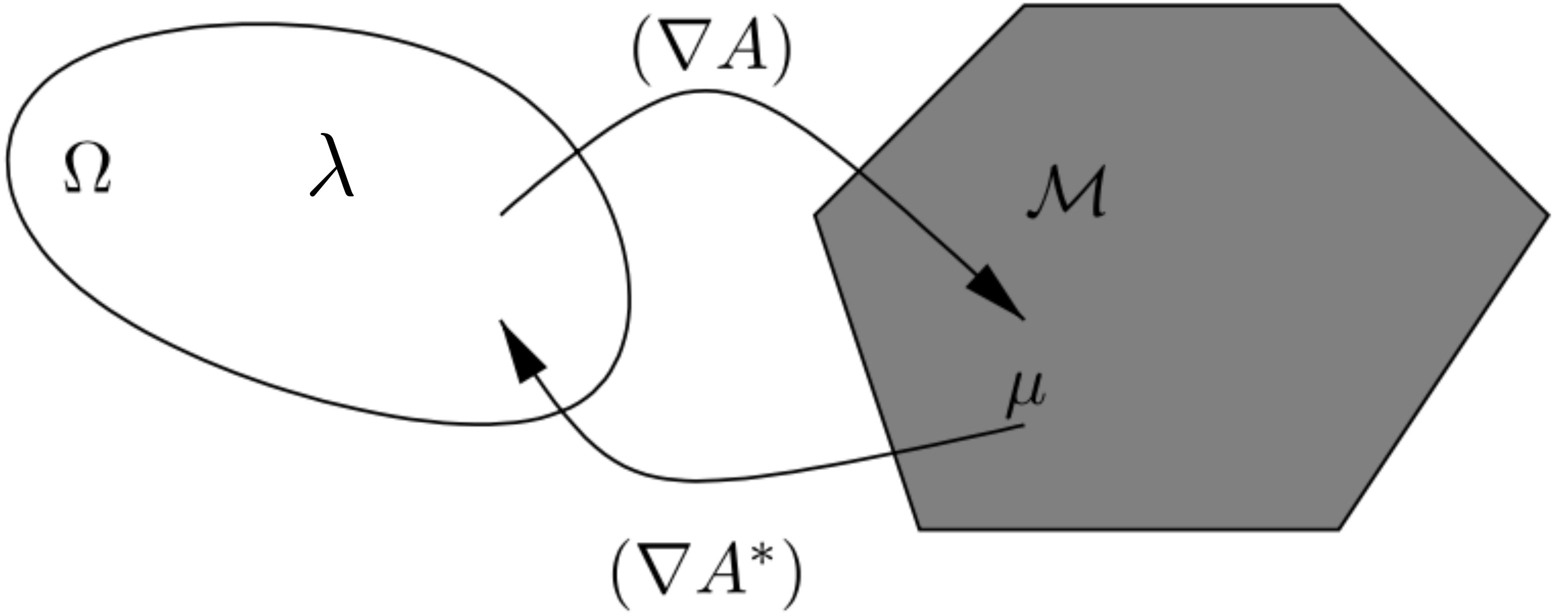
$$\nabla_\mu \mathcal{L}_t = F(\lambda_t)^{-1} \nabla_\lambda \mathcal{L}_t := \tilde{\nabla}_\lambda \mathcal{L}_t$$

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

Dually-Flat Riemannian Structure

See Amari's book 2016

Figure from Wainwright and Jordan, 2006



For VI, natural gradient in natural-parameter is computationally simpler than in the expectation parameter space

Natural-Gradient Descent in the Natural-Parameter Space

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

Conjugate   Nonconjugate

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

$$p(w) = \text{ExpFamily}(\eta_0)$$

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

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

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

Conjugate (similar to SVI, Hoffmann et al. 2013)

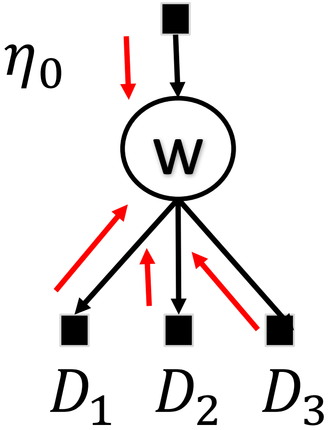
Nonconjugate

NGVI as Message Passing

Khan and Lin 2017, Khan and Nielsen 2018

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

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



A generalization of Variational Message Passing (Winn and Bishop 2005) and stochastic variational inference (Hoffman et al. 2013) to nonconjugate models. Convergence proof is in Khan et al. UAI 2016

Approximate Bayesian Filter

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

New Approximation | Previous Approximation | ExpFamily Prior | DNN Likelihood

Optimal natural-parameter == natural-gradient

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

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

$$q(w|\lambda_*) \propto p(w) \left[\prod_{i=1}^{N} e^{\phi(w)^\top \tilde{\nabla}_\lambda \mathbb{E}_{q_\lambda^*}[\log p(\mathcal{D}_i|w)]}\right]$$

“Fast” Natural-Gradients

Part II: Application to Bayesian deep learning

Notation change $\theta = w$

Natural-Gradient VI by using Weight-Perturbed Adam (Vadam)

Vadam

while not converged **do**
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, $\boldsymbol{\sigma} \leftarrow 1/\sqrt{N\mathbf{s} + \lambda}$
    Randomly sample a data example $\mathcal{D}_i$
    $\mathbf{g} \leftarrow -\nabla \log p(\mathcal{D}_i|\boldsymbol{\theta})$
    $\mathbf{m} \leftarrow \gamma_1 \mathbf{m} + (1 - \gamma_1) \, (\mathbf{g} + \lambda\boldsymbol{\mu}/N)$
    $\mathbf{s} \leftarrow \gamma_2 \mathbf{s} + (1 - \gamma_2) \, (\mathbf{g} \circ \mathbf{g})$
    $\hat{\mathbf{m}} \leftarrow \mathbf{m}/(1 - \gamma_1^t), \quad \hat{\mathbf{s}} \leftarrow \mathbf{s}/(1 - \gamma_2^t)$
    $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \alpha \, \hat{\mathbf{m}}/(\sqrt{\hat{\mathbf{s}}} + \lambda/N)$
    $t \leftarrow t + 1$
end while

Natural-gradient vs gradients

Gaussian prior

Non-Gaussian (DNN)

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

Gaussian variational approximation

Natural-Gradient VI

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

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

Existing Methods

$$\mu \leftarrow \mu + \alpha \; \frac{\hat{\nabla}_\mu \mathcal{L}}{\sqrt{s_\mu} + \delta}$$

$$\sigma \leftarrow \sigma + \alpha \; \frac{\hat{\nabla}_\sigma \mathcal{L}}{\sqrt{s_\sigma} + \delta}$$

(Graves et al. 2011, Blundell et al. 2015)

Doubly Stochastic Approximation

Gaussian prior

Non-Gaussian (DNN)

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

Gaussian variational approximation

Natural-Gradient VI

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

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

Sample a minibatch and a θ from q

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

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

Hessian Approximation

Gauss-Newton

$$\frac{N}{M}\sum_{i\in\mathcal{M}}\nabla^2_{\theta\theta}f_i(\theta) \approx \frac{N}{M}\sum_{i\in\mathcal{M}}[\nabla_\theta f_i(\theta)]^2$$

$$\approx N\left[\frac{1}{M}\sum_{i\in\mathcal{M}}\nabla_\theta f_i(\theta)\right]^2$$

Gradient-Magnitude

“Natural” Momentum Method

Polyak’s heavy-ball method

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \bar{\alpha}_t \nabla_\theta f_1(\boldsymbol{\theta}_t) + \bar{\gamma}_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}),$$

Our natural-momentum method

$$\lambda_{t+1} = \lambda_t + \bar{\alpha}_t \nabla_\mu \mathcal{L}_t + \bar{\gamma}_t(\lambda_t - \lambda_{t-1})$$

Natural-Gradient VI by using Weight-Perturbed Adam (Vadam)

Vadam

while not converged **do**

$\boldsymbol{\theta} \leftarrow \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, $\boldsymbol{\sigma} \leftarrow 1/\sqrt{N\mathbf{s} + \lambda}$

Randomly sample a data example $\mathcal{D}_i$

$\mathbf{g} \leftarrow -\nabla \log p(\mathcal{D}_i|\boldsymbol{\theta})$

$\mathbf{m} \leftarrow \gamma_1\, \mathbf{m} + (1 - \gamma_1)\, (\mathbf{g} + \lambda\boldsymbol{\mu}/N)$

$\mathbf{s} \leftarrow \gamma_2\, \mathbf{s} + (1 - \gamma_2)\, (\mathbf{g} \circ \mathbf{g})$

$\hat{\mathbf{m}} \leftarrow \mathbf{m}/(1 - \gamma_1^t), \quad \hat{\mathbf{s}} \leftarrow \mathbf{s}/(1 - \gamma_2^t)$

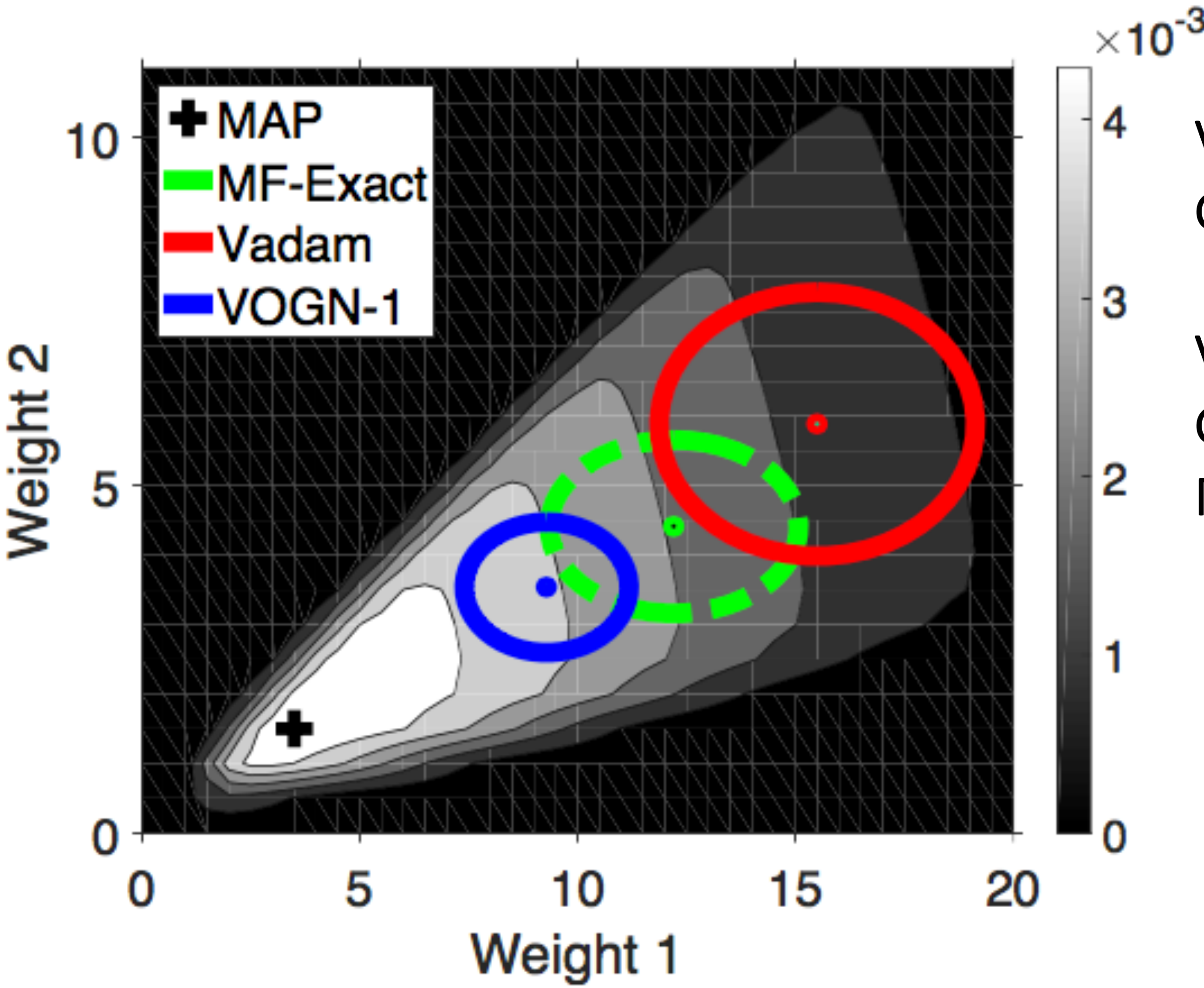
$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \alpha\, \hat{\mathbf{m}}/(\sqrt{\hat{\mathbf{s}}} + \lambda/N)$

$t \leftarrow t + 1$

end while

Results

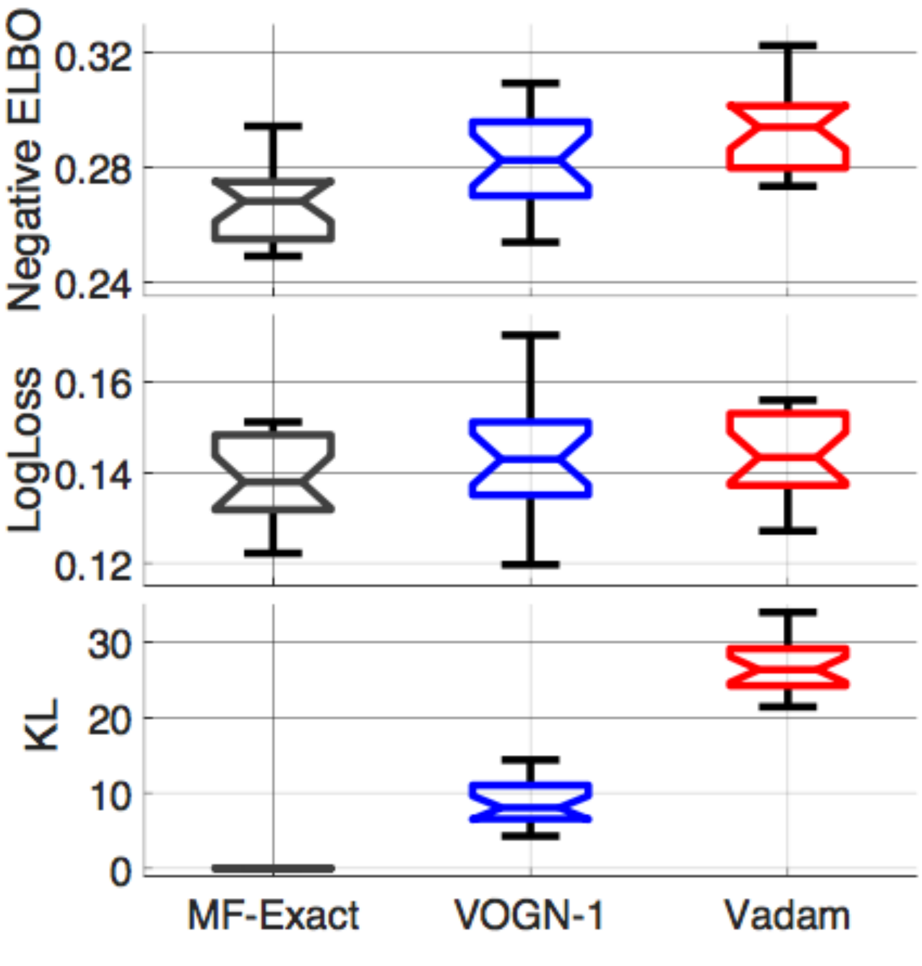
Illustration



VOGN uses Gauss-Newton

Vadam uses Gradient-Magnitude

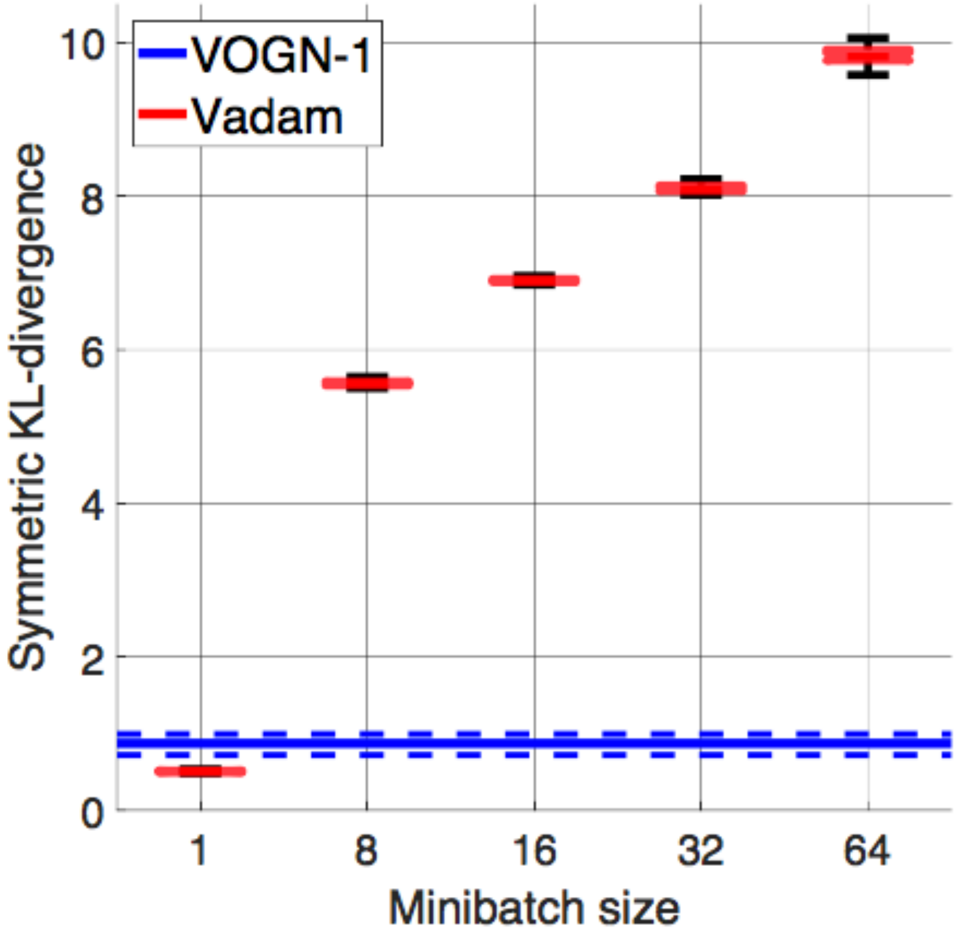
Quality of Posterior Approximation



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

Vadam uses Gradient-Magnitude with minibatch > 1

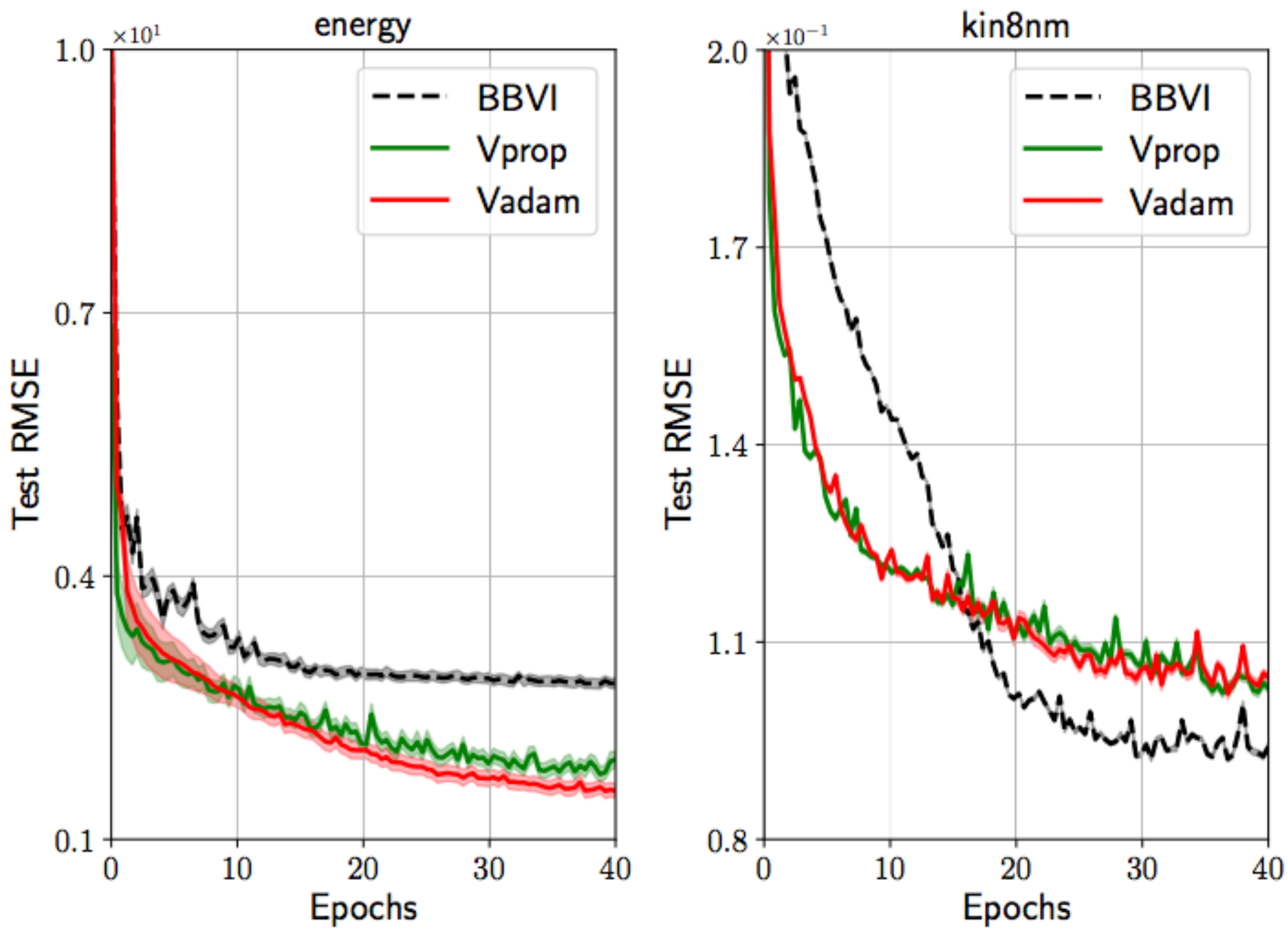
Effect of Minibatch on the Accuracy



As we decrease mini-batch size, the accuracy improves, but the stochastic noise increases which might slow-down the algorithm.

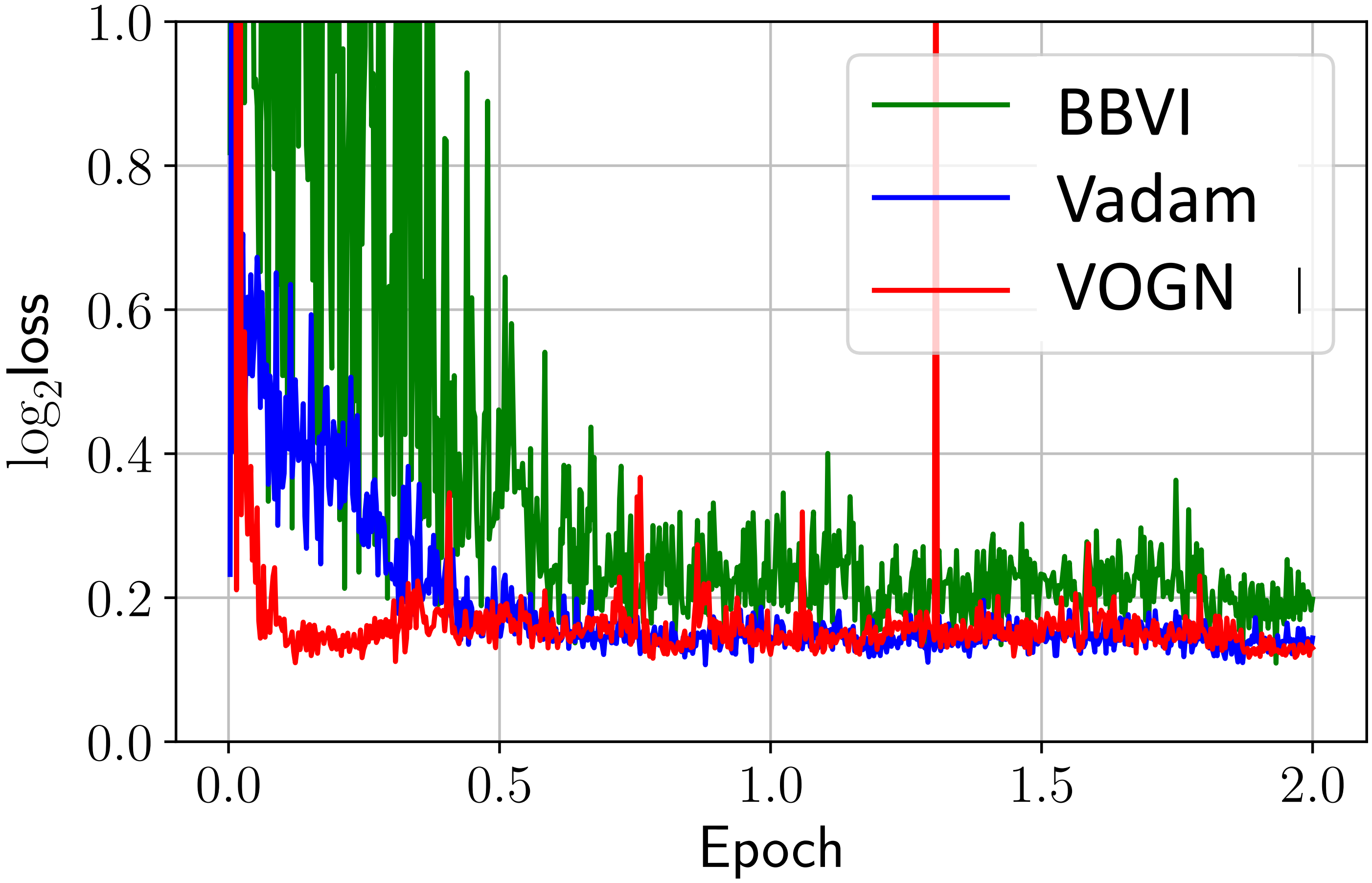
Vadam performs comparable to BBVI

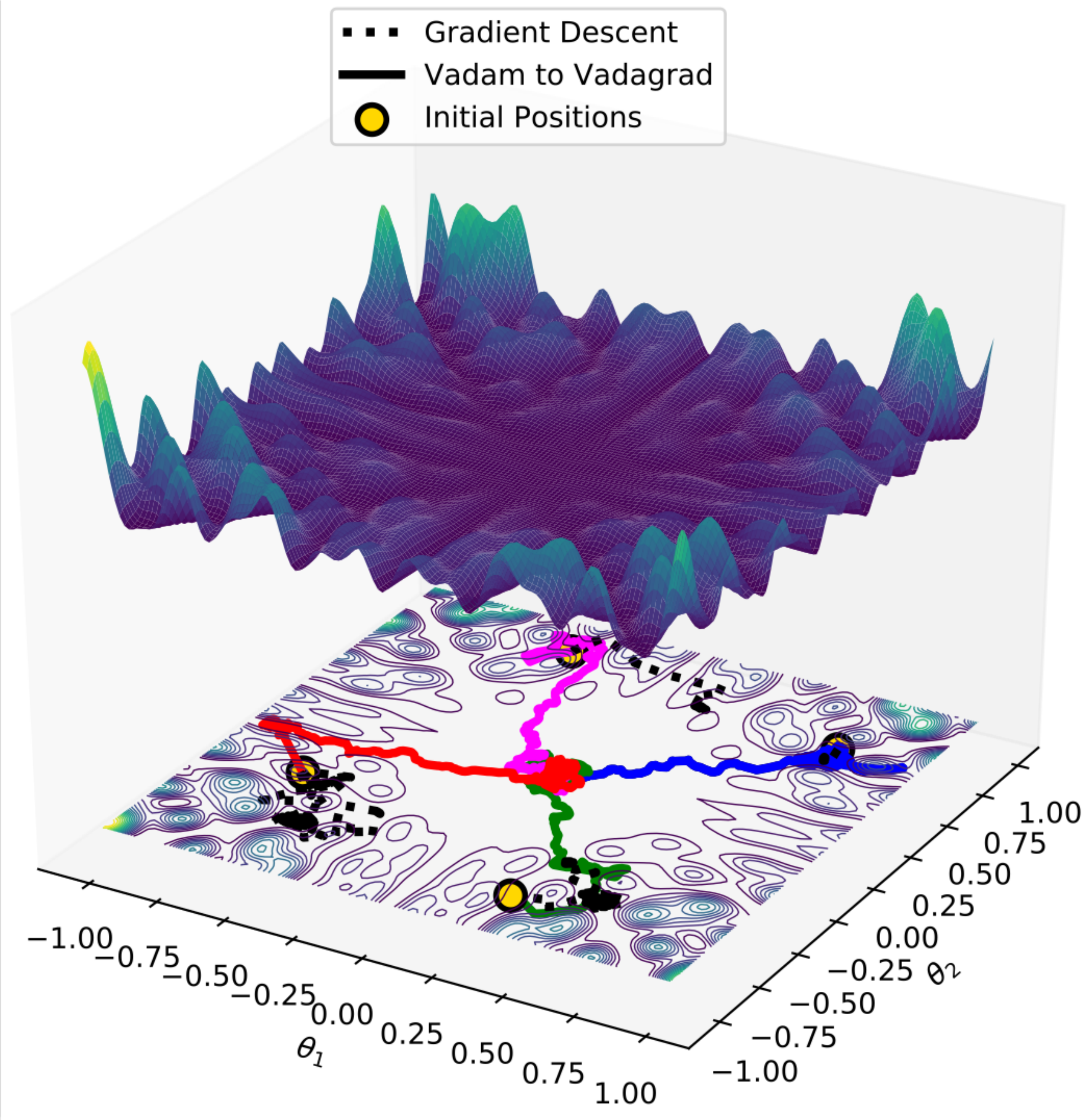
1 layer 50 hidden units with ReLU on “energy” (N=768, D= 8) and “kin8nm” (N=8192, D=8), 5 MC samples for Vadam, 10 for BBVI, minibatch of 32



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





Avoiding Local Minima

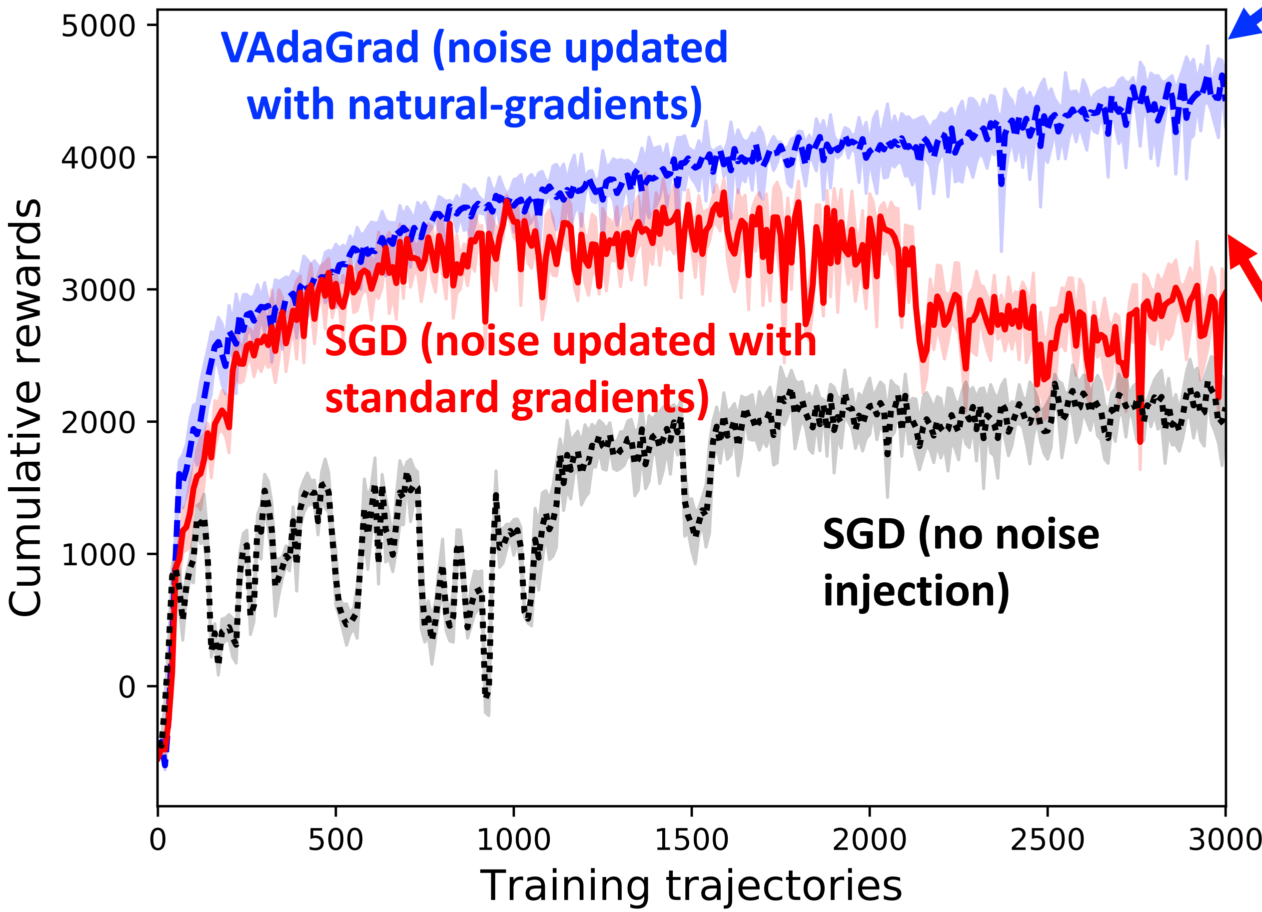
An example taken from Casella and Robert's book.

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

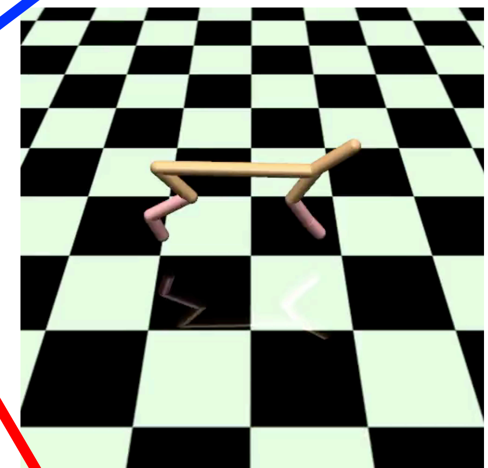
Optimization by smoothing, Gaussian homotopy/blurring etc., Entropy SGLD etc.

Parameter-Space Noise for Deep RL

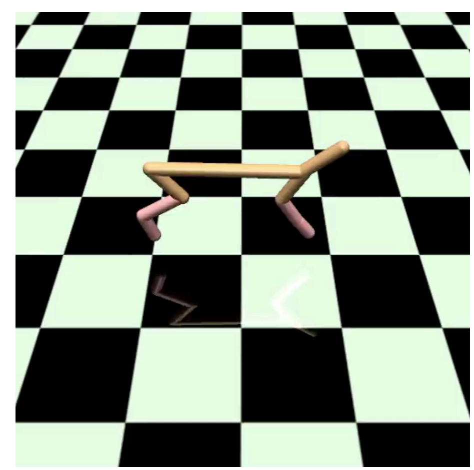
On OpenAI Gym Cheetah with DDPG
with DNN with [400,300] ReLU



Reward 5264



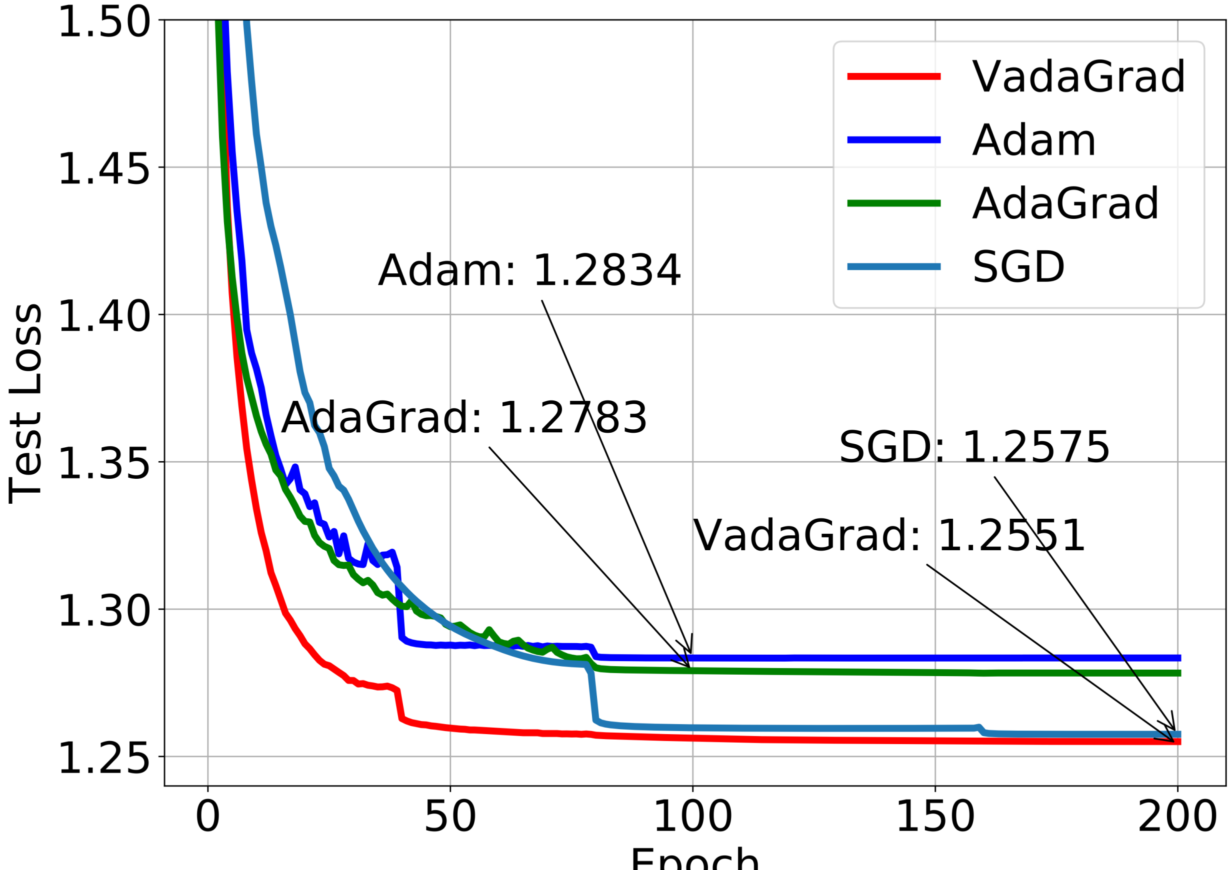
Reward 3674



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

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

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!

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

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