The Variational Adaptive-Newton Method

Mohammad Emtiyaz Khan[†], Wu Lin[†], Voot Tangkaratt[†], Zuozhu Liu[*], Didrik Nielsen[†]

[†]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan.
[*]Singapore University of Technology and Design, Singapore.

1 Introduction

Exploration based on Bayesian inference is extremely popular, but it is also computationally demanding. Here we present Variational Adaptive Newton (VAN), which

- is a black-box optimization method.
- is particularly useful for explorative-learning tasks such as active learning and reinforcement learning.
- is a second-order method and is related to adaptive-gradient methods.
- requires computations that are similar to continuous optimization methods.

2 Variational Optimization using Gaussians

Variational optimization (VO) (Staines and Barber, 2012) optimizes an objective function f by optimizing its expectation w.r.t. a distribution q. We will here consider Gaussian $q(\boldsymbol{\theta}) := \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\text{Standard optimization}: \quad \boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \tag{1}$$

$$\text{Variational optimization}: \quad \{\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*\} = \arg\min_{\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}} \mathbb{E}_{\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\Sigma})}[f(\boldsymbol{\theta})] := \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2}$$

One straightforward approach to optimize $\mathcal{L}$ is to use SGD:

$$\text{V-SGD}: \quad \boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t - \rho_t \left[\widehat{\nabla}_{\mu}\mathcal{L}_t\right] \tag{3}$$

$$\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t - \rho_t \left[\widehat{\nabla}_{\Sigma}\mathcal{L}_t\right]. \tag{4}$$

Since $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ are parameters of a distribution, natural-gradient updates are preferred. Wiestra et al. (2008) proposed such a method and shows that it improves stability. However, their update requires computation of the Fisher information matrix, which has memory complexity $O(D^4)$, where D is the length of $\boldsymbol{\theta}$. Our method is also a natural-gradient method, but with much simpler updates that requires $O(D^2)$ memory.

3 Variational Adaptive-Newton

For a Gaussian with parameters $\boldsymbol{\eta} := \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ we have:

- Mean parameters: $\mathbf{m} := \{\boldsymbol{\mu}, \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}\}$
- Natural parameters: $\boldsymbol{\lambda} := \{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\}$

Mirror-descent in mean parameters $\iff$ natural gradients in natural parameters:

$$\mathbf{m}_{t+1} = \underbrace{\arg\min_{m} \left\{\mathbf{m}^T \nabla_m \mathcal{L}_t + \frac{1}{\beta}\mathbb{D}_{KL}[q \,\|\, q_t]\right\}} \iff \boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \beta \underbrace{I(\boldsymbol{\lambda}_t)^{-1}\nabla_\lambda \mathcal{L}_t}_{\text{Natural gradient}}$$

$$\implies \underbrace{\nabla_m \mathcal{L}_t}_{\text{Natural gradient}} + \tfrac{1}{\beta}(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) = 0.$$

Rewriting this and using Bonnet's theorem, we obtain the VAN updates:

$$\text{VAN}: \quad \begin{aligned} \boldsymbol{\mu}_{t+1} &= \boldsymbol{\mu}_t - \beta_t \, \mathbf{P}_{t+1}^{-1}\mathbb{E}_{q_t}[\nabla_\theta f(\boldsymbol{\theta})] \\ \mathbf{P}_{t+1} &= \mathbf{P}_t + \beta_t \, \mathbb{E}_{q_t}\left[\nabla^2_{\theta\theta} f(\boldsymbol{\theta})\right], \end{aligned}$$

where $\mathbf{P}_t = \boldsymbol{\Sigma}_t^{-1}$ is the precision matrix and $q_t = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$.

Connections to Newton's Method:

VAN is related to Newton's Method:

$$\text{Newton's Method}: \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \rho_t \left[\nabla^2_{\theta\theta} f(\boldsymbol{\theta}_t)\right]^{-1} \left[\nabla_\theta f(\boldsymbol{\theta}_t)\right]. \tag{5}$$

Instead of scaling the gradients by Hessian, VAN scales the averaged gradients by the precision matrix P_t which contains a weighted sum of the past averaged Hessians.

A Large-Scale Variant:

By using a mean-field approximation for q, we obtain a diagonal version of VAN:

$$\text{VAN-D}: \quad \begin{aligned} \boldsymbol{\mu}_{t+1} &= \boldsymbol{\mu}_t - \beta_t \, \text{diag}(\mathbf{s}_{t+1})^{-1}\mathbb{E}_{q_t}[\nabla_\theta f(\boldsymbol{\theta})] \\ \mathbf{s}_{t+1} &= \mathbf{s}_t + \beta_t \mathbb{E}_{q_t}[\mathbf{h}(\boldsymbol{\theta})] \end{aligned}$$

Connections to AdaGrad:

VAN-D is very similar to AdaGrad (Duchi et al., 2011) shown below:

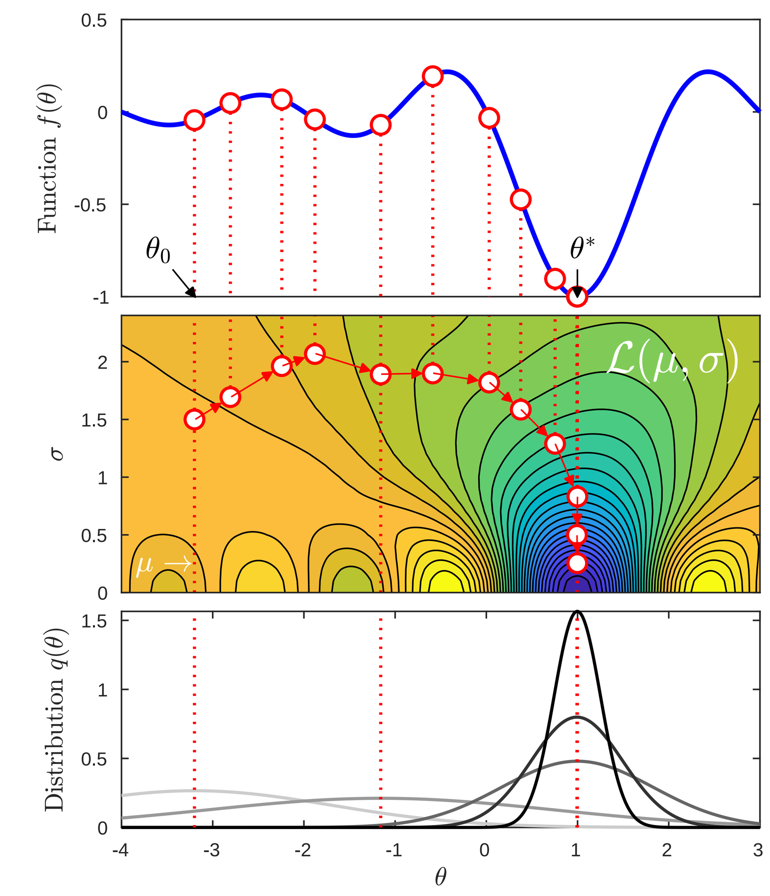
$$\text{AdaGrad}: \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \rho_t \text{diag}(\mathbf{s}_{t+1})^{-1/2}\mathbf{g}(\boldsymbol{\theta}_t) \tag{6}$$

$$\mathbf{s}_{t+1} = \mathbf{s}_t + [\mathbf{g}(\boldsymbol{\theta}_t) \odot \mathbf{g}(\boldsymbol{\theta}_t)] \tag{7}$$

Connections to Variational Inference:

Using VAN for VI is equivalent to Conjugate-Computation Variational Inference (CVI) (Khan and Lin, 2017). A direct consequence of this is that CVI also is a second-order method when q is a Gaussian distribution.

5 Results



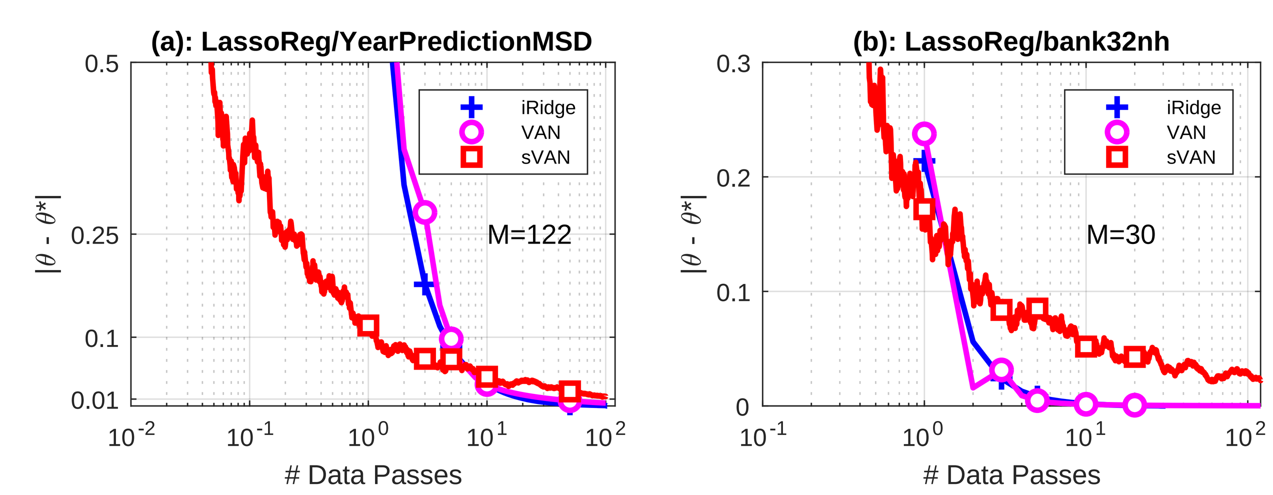
Exploration to Avoid Local Minima:

We show that VAN can avoid local minima. The top figure shows the function $f(\theta) = sinc(\theta)$ with a blue curve with a global minimum at $\theta^* = 1$. The second plot shows the VO objective $\mathcal{L}(\mu, \sigma) = E_q[f(\theta)]$. The red points and arrows show the iterations of VAN. The progression of the distribution q is shown in the bottom figure, where darker curves indicate higher iterations. As desired,the distribution peaks around θ^* as iterations increase.

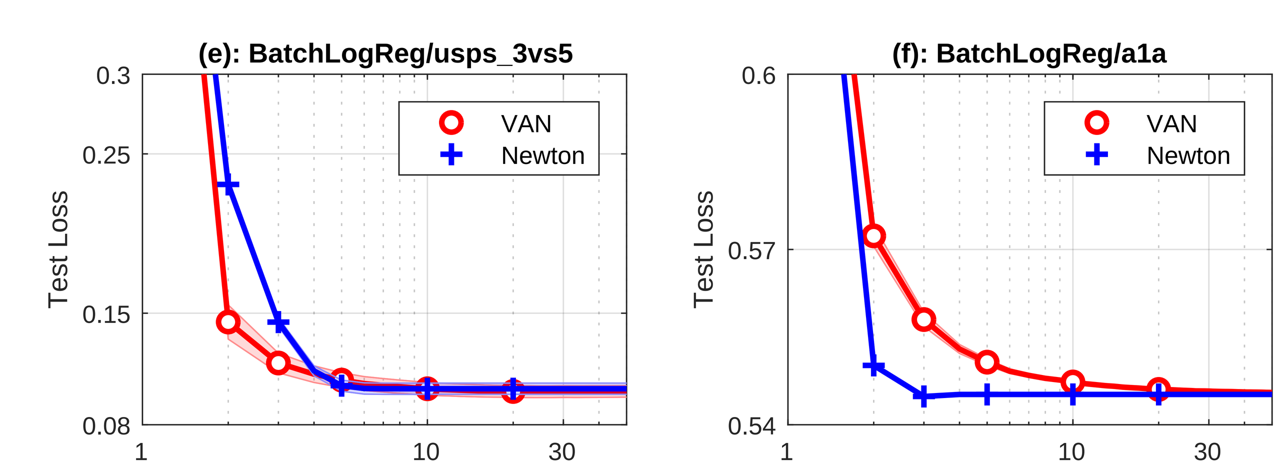
Supervised & Unsupervised Learning:

We show that VAN is a general-purpose algorithm and gives comparable results to existing methods. Figures show experimental results on different learning tasks. Datasets are specified in the title.

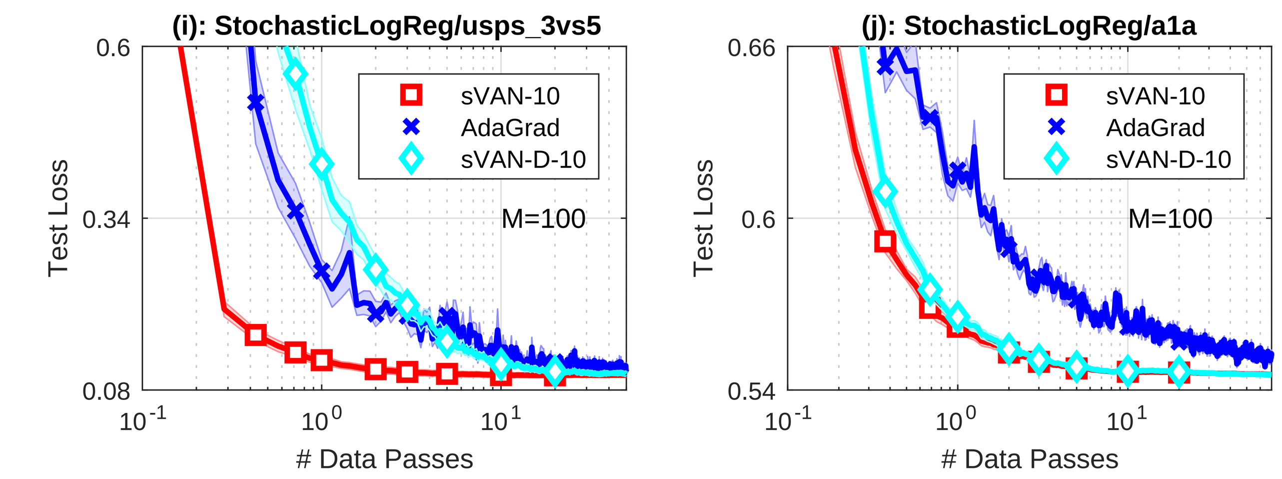
Lasso regression with VAN, sVAN and iRidge.



Logistic regression with VAN and Newton's method.

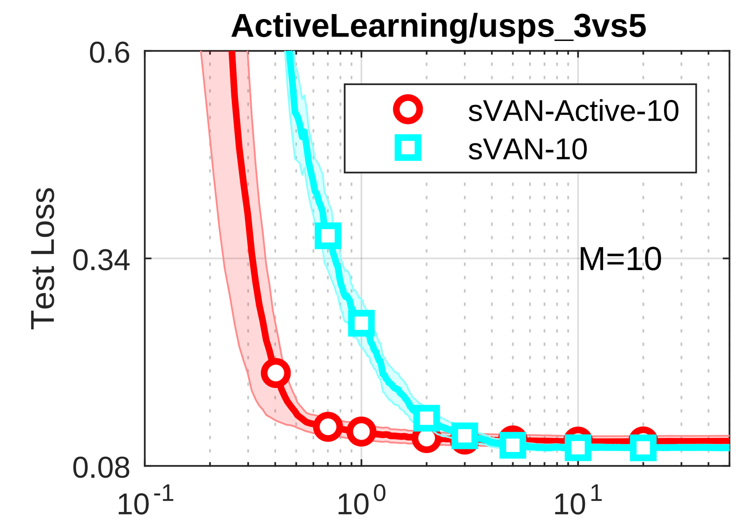


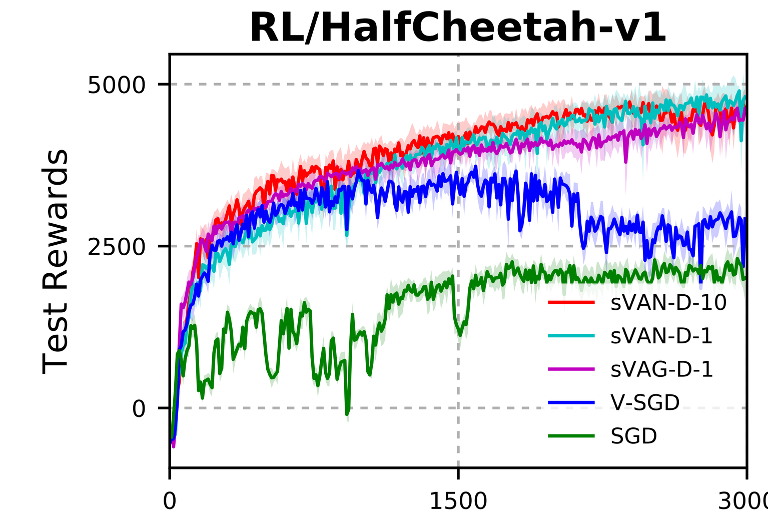
Logistic regression with sVAN, sVAN-D and AdaGrad.
M refers to the mini-batch size for stochastic methods.



Example of Explorative Learning:

We show that VAN gives better results than methods without exploration for active learning and reinforcement learning. The left figures shows data-space exploration using active learning. The right figure shows parameter-based exploration (Ruckstieß et al., 2010) in reinforcement learning.





References

- Duchi, Hazan and Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 2011.
- Khan and Lin. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. *AIStats*, 2017.
- Ruckstieß, Sehnke, Schaul, Wierstra, Sun and Schmidhuber. Exploring parameter space in reinforcement learning. *Paladyn*, 2010.
- Staines and Barber. Variational Optimization. *ArXiv e-prints*, 2012.
- Wierstra, Schaul, Glasmachers, Sun and Schmidhuber. Natural evolution strategies. *Evolutionary Computation*, 2008.