Variational Hamiltonian Monte Carlo via Score Matching

Cheng Zhang
Joint work with Babak Shahbaba and Hongkai Zhao

July 25, 2018

Computational Biology Program
Fred Hutchinson Cancer Research Center

Introduction

Bayesian Inference

- **Model setup**
  - Data: $\mathcal{D} = \{y^1, \ldots, y^N\}$
  - Model: $p(\mathcal{D}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the model parameter
  - Prior: $p(\boldsymbol{\theta})$
- **Goal**: learn the posterior distribution

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathcal{D})} \propto p(\mathcal{D}, \boldsymbol{\theta})$$

- **Difficulty**: for most useful models, e.g. Bayesian logistic regression, Bayesian neural networks, and topic models, $p(\mathcal{D})$ is unknown.
- **Current approaches**
  - *Markov chain Monte Carlo* (MCMC) [Metropolis et al., 1953].
  - *Variational inference* (VI) [Jordan et al., 1999].

Markov chain Monte Carlo

- Main idea: construct a *Markov chain* that converges to the target posterior $p(\boldsymbol{\theta}|\mathcal{D})$
- Metroplis-Hastings:
  1. sample $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta})$
  2. accept $\boldsymbol{\theta}'$ with probability

$$\alpha(\boldsymbol{\theta} \to \boldsymbol{\theta}') = \min\left(1, \frac{p(\mathcal{D}, \boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{p(\mathcal{D}, \boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right)$$

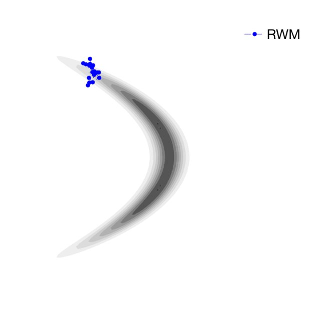
Simple examples

- Random walk Metropolis (RWM)

$$\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$$

- Gibbs sampling

$$\theta_i' \sim p(\theta_i|\boldsymbol{\theta}_{-i}, \mathcal{D}), \; i = 1, \ldots$$



Fixed-form Variational Bayes

Variational inference (VI) seeks the best candidate from a family of tractable distributions that minimizes a statistical distance measure to the target posterior, usually the Kullback-Leibler (KL) divergence

$$\hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} D_{KL}(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}|\mathcal{D}))$$

equivalent to maximizing the evidence lower bound (ELBO)

$$L(\boldsymbol{\eta}, \mathcal{D}) = \mathbb{E}_{q_{\boldsymbol{\eta}}(\boldsymbol{\theta})} \log \left(\frac{p(\boldsymbol{\theta}, \mathcal{D})}{q_{\boldsymbol{\eta}}(\boldsymbol{\theta})} \right) \leq \log p(\mathcal{D})$$

VI tends to be faster than MCMC. Fixed-form VI further assumes

$$q_{\boldsymbol{\eta}}(\boldsymbol{\theta}) = \exp(T(\boldsymbol{\theta})\boldsymbol{\eta} - A(\boldsymbol{\eta}))$$

- Potentially more accurate than using mean-field assumptions
- Still requires tractable approximating distributions which usually have limited expressive power.

Hamiltonian Monte Carlo

Hamiltonian Dynamics

- Main idea: suppress the random walk behavior using a Hamiltonian dynamical system.
- The **Hamiltonian** energy function

$$H(\boldsymbol{\theta}, \boldsymbol{r}) = U(\boldsymbol{\theta}) + K(\boldsymbol{r})$$

  - Potential: $U(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}, \mathcal{D})$
  - Kinetic: $K(\boldsymbol{r}) = \frac{1}{2}\boldsymbol{r}^\intercal \boldsymbol{M}^{-1}\boldsymbol{r}$

  The joint density

$$p(\boldsymbol{\theta}, \boldsymbol{r}) \propto \exp(-U(\boldsymbol{\theta}) - K(\boldsymbol{r})) \propto p(\boldsymbol{\theta}|\mathcal{D}) \cdot \mathcal{N}(\boldsymbol{r}|\mathbf{0}, \boldsymbol{M})$$

- *Hamilton's equations*

$$\frac{d\boldsymbol{\theta}}{dt} = \nabla_{\boldsymbol{r}} H = \nabla_{\boldsymbol{r}} K(\boldsymbol{r}), \quad \frac{d\boldsymbol{r}}{dt} = -\nabla_{\boldsymbol{\theta}} H = -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$$

Hamiltonian Flow

Let $\boldsymbol{z} = (\boldsymbol{\theta}, \boldsymbol{r}) \in \mathbb{R}^{2d}$, a Hamiltonian flow $\phi(\boldsymbol{z}, t)$ is a solution to the Hamilton's equations such that $\phi(\boldsymbol{z}, 0) = \boldsymbol{z}$.

- Reversibility

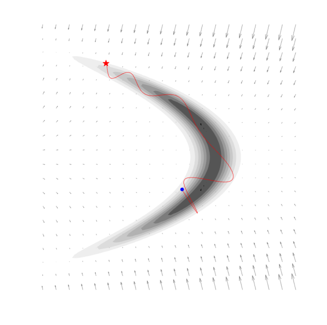
$$\phi((\boldsymbol{\theta}_0, \boldsymbol{r}_0), T) = (\boldsymbol{\theta}_T, \boldsymbol{r}_T)$$
$$\Leftrightarrow$$
$$\phi((\boldsymbol{\theta}_T, -\boldsymbol{r}_T), T) = (\boldsymbol{\theta}_0, -\boldsymbol{r}_0)$$

- Volume preservation

$$\left| \det \frac{\partial \phi(\boldsymbol{z}, t)}{\partial \boldsymbol{z}} \right| = 1$$

- Energy preservation

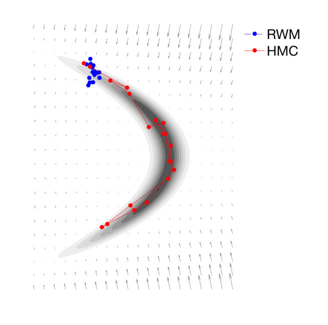
$$H(\phi(\boldsymbol{z}, t)) = H(\boldsymbol{z})$$



Hamiltonian Monte Carlo

- Numerical integrator (leap-frog)

$$\boldsymbol{r}(t+\epsilon/2) = \boldsymbol{r}(t) - \epsilon/2 \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}(t))$$
$$\boldsymbol{\theta}(t+\epsilon) = \boldsymbol{\theta}(t) + \epsilon \boldsymbol{M}^{-1} \boldsymbol{r}(t+\epsilon/2)$$
$$\boldsymbol{r}(t+\epsilon) = \boldsymbol{r}(t+\epsilon/2) - \epsilon/2 \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}(t+\epsilon))$$



Leap-frog scheme is time reversible and volume preserving, but **does not** preserve the Hamiltonian.

- Hamiltonian Monte Carlo
  - $\boldsymbol{z}' = \hat{\phi}(\boldsymbol{z}, T)$
  - accept $\boldsymbol{z}'$ with probability

$$\alpha_{hmc}(\boldsymbol{z} \to \boldsymbol{z}') = \min\left(1, \exp(H(\boldsymbol{z}) - H(\boldsymbol{z}'))\right)$$

Scalable Markov chain Monte Carlo

Stochastic Gradient MCMC

- Using stochastic gradient

$$\nabla_{\boldsymbol{\theta}} \tilde{U}(\boldsymbol{\theta}) = -\frac{N}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log p(y^{t_i} | \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) \sim \mathcal{O}(n)$$

- Examples:
    - SGLD [Welling and Teh, 2011]
    - SGHMC [Chen et al., 2014]
    - SGNHT [Ding et al., 2014]
    - etc.
- Convergence is based on SDE theory (Fokker-Planck equation)
    - Require small stepsize to reduce the noise introduced by stochastic gradients
    - Sacrifice exploration efficiency for scalability [Betancourt, 2015]

Surrogate Method

- Function Approximation [Neal, 1995, Liu, 2001]

$$U^S(\boldsymbol{\theta}) \approx U(\boldsymbol{\theta}) \quad \Rightarrow \quad \nabla_{\boldsymbol{\theta}} U^S(\boldsymbol{\theta}) \approx \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$$

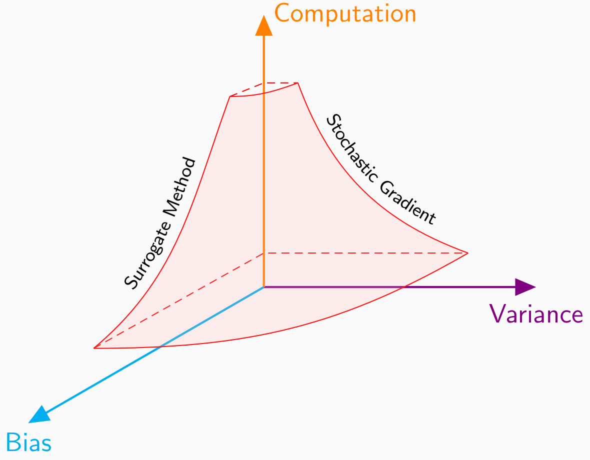
  $U^S(\boldsymbol{\theta})$ should be

  - cheap to compute
  - flexible enough for good approximation

  Stochastic gradients can be viewed as unbiased function approximations.

- Current Approaches
  - Gaussian process [Rasmussen, 2003, Lan et al., 2015]
  - Reproducing kernel Hilbert space [Strathmann et al., 2015]
  - Random network [Zhang et al., 2015]

Bias, Variance, and Computation Trade-off



Variational Hamiltonian Monte Carlo

Surrogate Method: A Variational Perspective

- Surrogate induced distribution

$$q_{\boldsymbol{\psi}}(\boldsymbol{\theta}) \propto \exp\left(-U^S_{\boldsymbol{\psi}}(\boldsymbol{\theta})\right), \quad \boldsymbol{\psi} \in \Omega$$

where

$$\Omega := \{\boldsymbol{\psi} \text{ s.t. } \int \exp\left(-U^S_{\boldsymbol{\psi}}(\boldsymbol{\theta})\right) d\boldsymbol{\theta} < \infty\}$$

- Free-form variational inference to improve approximation

$$\hat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}\in\Omega}{\arg\min}\, D\left(q_{\boldsymbol{\psi}}(\boldsymbol{\theta}), p(\boldsymbol{\theta}, \mathcal{D})\right)$$

D is some statistical distance measure between unnormalized densities.

Remark: Unlike fixed-form VI, $q_{\boldsymbol{\psi}}(\boldsymbol{\theta})$ does not have to be tractable and $U^S_{\boldsymbol{\psi}}(\boldsymbol{\theta})$ enjoys free style construction.

Random Bases Surrogate

Random Bases Surrogate: $U_{\boldsymbol{\psi}}^{S}(\boldsymbol{\theta}) = \sum_{i=1}^{s} \psi_i a(\boldsymbol{\theta}; \gamma_i)$

Theorem (Rahimi and Recht 2008)

Let μ be any probability measure, $\|f\|_\mu^2 = \int f^2(\boldsymbol{\theta})\mu(d\boldsymbol{\theta})$. Suppose $\sup_{\boldsymbol{\theta},\gamma} |a(\boldsymbol{\theta};\gamma)| \leq 1$. Fix $f \in \mathcal{F}_p$. $\forall \delta > 0$, with probability at least $1 - \delta$ over $\gamma_i \overset{\text{iid}}{\sim} p(\gamma)$, there exist $\psi_1, \ldots, \psi_s$ such that

$$U_{\boldsymbol{\psi}}^{S}(\boldsymbol{\theta}) = \sum_{i=1}^{s} \psi_i a(\boldsymbol{\theta}; \gamma_i)$$

satisfies

$$\|U_{\boldsymbol{\psi}}^{S} - f\|_\mu < \frac{\|f\|_p}{\sqrt{s}} \left(1 + \sqrt{2 \log \frac{1}{\delta}}\right)$$

where $\|f\|_p = \sup_\gamma \left|\frac{\psi_f(\gamma)}{p(\gamma)}\right|$

Score Matching

- "Score matching" [Hyvärinen, 2005]

$$\tilde{D}_{SM}(q_{\psi}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}, \mathcal{D})) = \frac{1}{2} \int q_{\psi}(\boldsymbol{\theta}) \| \nabla_{\boldsymbol{\theta}} U^{S}_{\boldsymbol{\psi}}(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \|^2 d\boldsymbol{\theta}$$

- Consistency

$$\tilde{D}_{SM}(q_{\psi}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}, \mathcal{D})) = 0$$

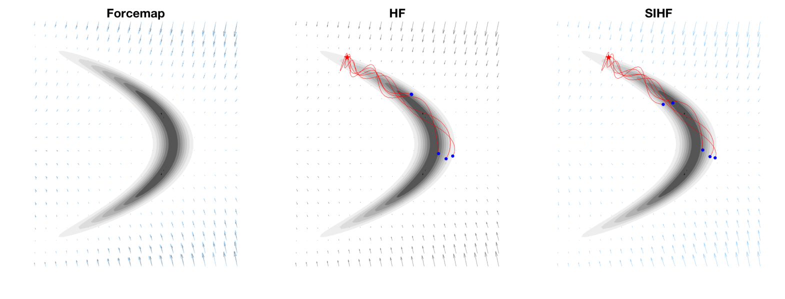
$$\Rightarrow U^{S}_{\psi}(\boldsymbol{\theta}) = U(\boldsymbol{\theta}) + \text{Constant} \Rightarrow q_{\psi}(\boldsymbol{\theta}) = p(\boldsymbol{\theta} | \mathcal{D})$$

Surrogate Induced Hamiltonian Flow

Define $H^S_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{r}) = U^S_{\boldsymbol{\psi}}(\boldsymbol{\theta}) + K(\boldsymbol{r})$, a surrogate induced Hamiltonian flow (SIHF) is a solution $\phi^S_{\boldsymbol{\psi}}(\boldsymbol{z}, t)$ to the modified Hamilton's equations

$$\frac{d\boldsymbol{\theta}}{dt} = \boldsymbol{M}^{-1}\boldsymbol{r}, \quad \frac{d\boldsymbol{r}}{dt} = -\nabla_{\boldsymbol{\theta}} U^S_{\boldsymbol{\psi}}(\boldsymbol{\theta})$$

such that $\phi^S_{\boldsymbol{\psi}}(\boldsymbol{z}, 0) = \boldsymbol{z}$



Variational Hamiltonian Monte Carlo

- Sample by HMC from the current surrogate induced distribution
  - $\boldsymbol{z}' = \hat{\phi}^S_{\boldsymbol{\psi}}(\boldsymbol{z}, T)$
  - accept $\boldsymbol{z}'$ with probability

$$\alpha_{vhmc}(\boldsymbol{z} \to \boldsymbol{z}') = \min\left(1, \exp(H^S_{\boldsymbol{\psi}}(\boldsymbol{z}) - H^S_{\boldsymbol{\psi}}(\boldsymbol{z}'))\right)$$

- Empirical score matching distance minimization (with regularization)

$$\hat{\boldsymbol{\psi}}^{(t)} = \arg\min_{\boldsymbol{\psi}} \frac{1}{2}\sum_{n=1}^{t} \| \sum_{i=1}^{s} \nabla_{\boldsymbol{\theta}} a(\boldsymbol{\theta}^{(n)};\gamma_i)\psi_i - \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(n)})\|^2 + \frac{\lambda}{2}\|\boldsymbol{\psi}\|^2$$

- Regularized surrogate can be helpful at the start, e.g.

$$V^S_{\boldsymbol{\psi}^{(t)}}(\boldsymbol{\theta}) = \mu_t U^S_{\boldsymbol{\psi}^{(t)}}(\boldsymbol{\theta}) + (1-\mu_t)\cdot \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^L)^\intercal \nabla^2_{\boldsymbol{\theta}} U(\boldsymbol{\theta})^L(\boldsymbol{\theta} - \boldsymbol{\theta}^L)$$

where μ_t is a transition schedule that goes from 0 to 1 as t increases.

Online Variational Hamiltonian Monte Carlo

ψ can be updated online. Denote $A(\boldsymbol{\theta}) = (\nabla_{\boldsymbol{\theta}} a(\boldsymbol{\theta}; \gamma_1), \ldots, \nabla_{\boldsymbol{\theta}} a(\boldsymbol{\theta}; \gamma_s))$

Online Variational HMC

Set λ, μ_t, s and HMC parameters ε, L. Initialize $\boldsymbol{\theta}^{(0)}$ to a first guess, $\psi^{(0)} = \mathbf{0}$, $\boldsymbol{C}^{(0)} = \frac{1}{\lambda}\boldsymbol{I}_s$. Find $\boldsymbol{\theta}^L$ and compute $\nabla^2_{\boldsymbol{\theta}} U(\boldsymbol{\theta})^L$.
for $t = 1$ to T **do**
    Perform one HMC iteration for the regularized surrogate induced distribution $q_{\psi^{(t)}}(\boldsymbol{\theta}) \propto \exp(-V^S_{\psi^{(t)}}(\boldsymbol{\theta}))$ to draw $(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{r}^{(t+1)})$
    Acquire $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(t+1)})$ and $A_{t+1} = A(\boldsymbol{\theta}^{(t+1)})$
    Compute $W^{(t+1)} = \boldsymbol{C}^{(t)} A^{\intercal}_{t+1} [\boldsymbol{I}_d + A_{t+1} \boldsymbol{C}^{(t)} A^{\intercal}_{t+1}]^{-1}$
    Update $\psi^{(t+1)}, \boldsymbol{C}^{(t+1)}$ as follows
$$\psi^{(t+1)} = \psi^{(t)} + W^{(t+1)}(\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(t+1)}) - A_{t+1}\psi^{(t)})$$
$$\boldsymbol{C}^{(t+1)} = \boldsymbol{C}^{(t)} - W^{(t+1)} A_{t+1} \boldsymbol{C}^{(t)}$$
end for

Connections to Related Work

- Compared to Stochastic linear regression [Salimans and Knowles, 2013], VHMC allows free-form intractable approximate distributions.
- Compared to RNSHMC [Zhang et al., 2015] and Kamiltonian Monte Carlo [Strathmann et al., 2015], VHMC enables variational approximation that further reduces the computation in the correction step.

Experiments

A Beta-binomial Model for Overdispersion

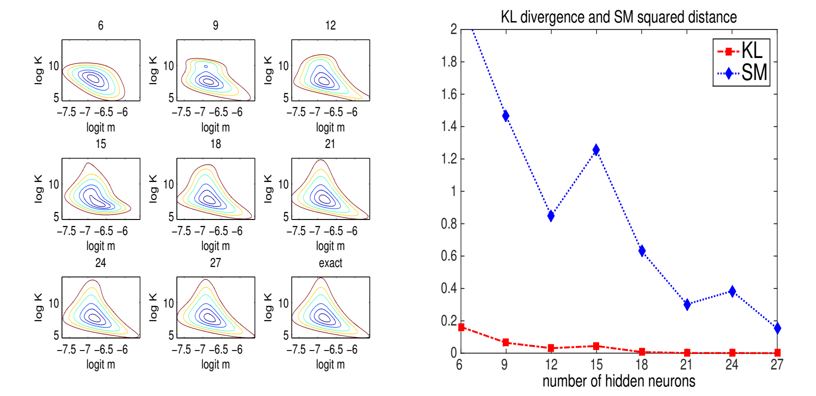


Figure 1: Left: Approximate posteriors for a varying number of hidden neurons. Exact posterior at bottom right. **Right**: KL-divergence and score matching squared distance between the surrogate approximation and the exact posterior density using an increasing number of hidden neurons.

Bayesian Probit Regression

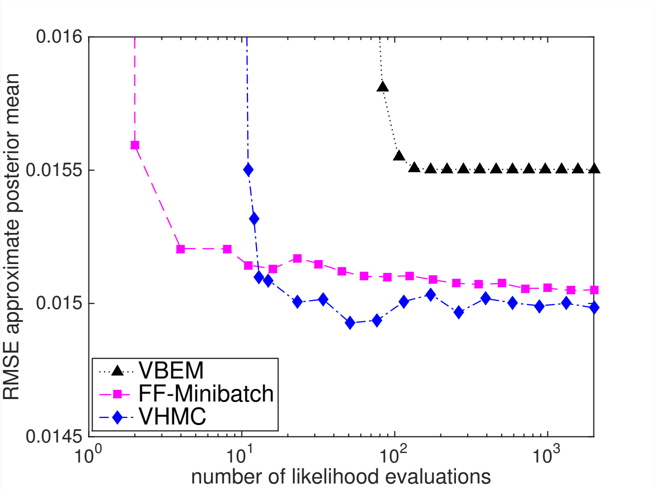


Figure 2: RMSE of the approximate posterior mean as a function of the number of likelihood evaluations for different variational Bayesian approaches and VHMC algorithm.

Bayesian Logistic Regression

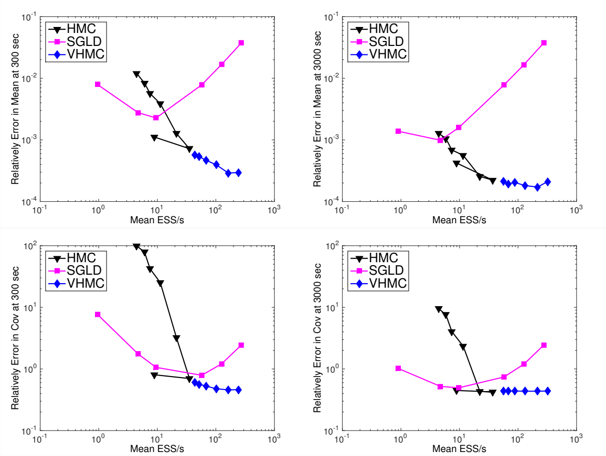


Figure 3: Final error of logistic regression at time T versus mixing rate for the mean (top) and covariance (bottom) estimates after 300 (left) and 3000 (right) seconds of computation. Each algorithm is run using different setting of parameters.

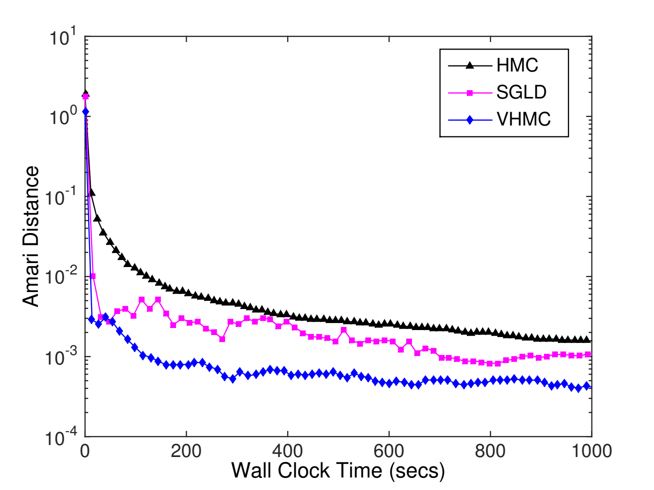


Figure 4: Convergence of Amari distance on the MEG data for HMC, SGLD and our Variational HMC algorithm.

Conclusion

Summary and Future Work

VHMC provides a general framework that combines variational inference and MCMC for scalable Bayesian inference.

- Accelerate HMC via efficient surrogate

$$\frac{d\boldsymbol{\theta}}{dt} = \boldsymbol{M}^{-1}\boldsymbol{r}, \quad \frac{d\boldsymbol{r}}{dt} = -\nabla_{\boldsymbol{\theta}} U_{\boldsymbol{\psi}}^{S}(\boldsymbol{\theta})$$

- Improve surrogate via free-form variational inference

$$\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi}\in\Omega} \tilde{D}_{SM}\left(q_{\boldsymbol{\psi}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}, \mathcal{D})\right)$$

Future work:

- Extension to high dimensional problems using more sophisticate structures (e.g., deep neural networks).
- Other distance measure for unnormalized densities (e.g. stein's discrepancy).

Questions?

Stochastic Linear Regression

- Linear relaxation

$$\tilde{q}_{\tilde{\boldsymbol{\eta}}}(\boldsymbol{\theta}) = \exp(\tilde{T}(\boldsymbol{\theta})\tilde{\boldsymbol{\eta}}), \quad \tilde{T}(\boldsymbol{\theta}) = (1, T(\boldsymbol{\theta})), \quad \tilde{\boldsymbol{\eta}} = (\eta_0, \boldsymbol{\eta}^\intercal)^\intercal$$

- Minimizing unnormalized KL divergence

$$\begin{aligned}\hat{\tilde{\boldsymbol{\eta}}} &= \arg\min_{\tilde{\boldsymbol{\eta}}} \tilde{D}_{KL}(\tilde{q}_{\tilde{\boldsymbol{\eta}}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}, \mathcal{D})) \\ &= \left(\mathbb{E}_q(\tilde{T}(\boldsymbol{\theta})^\intercal \tilde{T}(\boldsymbol{\theta}))\right)^{-1} \mathbb{E}_q(\tilde{T}(\boldsymbol{\theta})^\intercal \log p(\boldsymbol{\theta}, \mathcal{D}))\end{aligned}$$

- Fixed-point update

$$\hat{\tilde{\boldsymbol{\eta}}}^{(n+1)} = \left(\mathbb{E}_{q_{\hat{\tilde{\boldsymbol{\eta}}}^{(n)}}}(\tilde{T}(\boldsymbol{\theta})^\intercal \tilde{T}(\boldsymbol{\theta}))\right)^{-1} \mathbb{E}_{q_{\hat{\tilde{\boldsymbol{\eta}}}^{(n)}}}(\tilde{T}(\boldsymbol{\theta})^\intercal \log p(\boldsymbol{\theta}, \mathcal{D}))$$

See Salimans and Knowles [2013] for more details and variations.

A Dense Subset in RKHS

Reproducing kernel

$$k(\boldsymbol{\theta}, \boldsymbol{\theta}') = \int p(\gamma) a(\boldsymbol{\theta}; \gamma) a(\boldsymbol{\theta}'; \gamma) d\gamma$$

related reproducing kernel Hilbert space (RKHS)

$$\mathcal{H} := \left\{ f(\boldsymbol{\theta}) = \int \psi_f(\gamma) a(\boldsymbol{\theta}; \gamma) d\gamma \text{ s.t. } \int \frac{\psi_f^2(\gamma)}{p(\gamma)} d\gamma < \infty \right\}$$

with an inner product $\langle f, g \rangle_{\mathcal{H}} = \int \frac{\psi_f(\gamma)\psi_g(\gamma)}{p(\gamma)} d\gamma$ between $f(\boldsymbol{\theta})$ and $g(\boldsymbol{\theta}) = \int \psi_g(\gamma) a(\boldsymbol{\theta}; \gamma) d\gamma$.

A Dense Subset

Define

$$\mathcal{F}_p := \left\{ f(\boldsymbol{\theta}) \in \mathcal{H} \text{ s.t. } \sup_{\gamma} \left| \frac{\psi_f(\gamma)}{p(\gamma)} \right| < \infty \right\}$$

$\mathcal{F}_p$ is dense in $\mathcal{H}$. See [Rahimi and Recht, 2008] for more details.

References

M. Betancourt. The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, 2015.

T. Chen, E. B. Fox, and C. Guestrin. Stochastic gradient hamiltonian monte carlo. In *Proceedings of 31st International Conference on Machine Learning (ICML 2014)*, 2014.

N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skell, and H. Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical methods. In *Machine Learning*, pages 183–233. MIT Press, 1999.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *The 2nd International Conference on Learning Representations (ICLR)*, 2013.

S. Lan, T. Bui, M. Christie, and M. Girolami. Emulation of higher-order tensors in manifold Monte Carlo methods for Bayesian inverse problems. arxiv.org/abs/1507.06244, 2015.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

Y. A. Ma, T. Chen, and E. Fox. A complete recipe for stochastic gradient mcmc. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2015.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

R. M. Neal. *Bayesian learning for neural networks*. PhD thesis, Department of Computer Science, University of Toronto, 1995.

A. Rahimi and B. Recht. Uniform approximation of functions with random bases. In *Proc. 46th Ann. Allerton Conf. Commun., Contr. Comput.*, 2008.

C. E. Rasmussen. Gaussian processes to speed up hybrid monte carlo for expensive bayesian integrals. *Bayesian Statistics*, 7:651–659, 2003.

T. Salimans and D. A. Knowles. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szabo, and A. Gretton. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. In *Advances in Neual Information Processing Systems*, Cambridge, MA, 2015. MIT Press.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, 2011.

C. Zhang, B. Shahbaba, and H. K. Zhao. Hamiltonian Monte Carlo Acceleration Using Surrogate Functions with Random Bases. arxiv.org/abs/1506.05555, 2015.