Variational Hamiltonian Monte Carlo via Score Matching

Cheng Zhang (Joint work with Prof. Shahbaba and Prof. Zhao)

Department of Mathematics
University of California, Irvine

Jan 6, 2017

Outline

UCIrvine

Bayesian Inference

- Bayesian inference model
  - $\mathcal{D} = \{y^1, \ldots, y^N\}$: observed data
  - $\boldsymbol{\theta} \in \mathbb{R}^d$: model parameter
  - $p(\mathcal{D}|\boldsymbol{\theta})$: model density
  - $p(\boldsymbol{\theta})$: prior

Bayesian Inference

- Bayesian inference model
  - $\mathcal{D} = \{y^1, \ldots, y^N\}$: observed data
  - $\boldsymbol{\theta} \in \mathbb{R}^d$: model parameter
  - $p(\mathcal{D}|\boldsymbol{\theta})$: model density
  - $p(\boldsymbol{\theta})$: prior
- Goal : learning parameter $\boldsymbol{\theta}$ from data

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

UCIrvine

Bayesian Inference

- Bayesian inference model
  - $\mathcal{D} = \{y^1, \ldots, y^N\}$: observed data
  - $\boldsymbol{\theta} \in \mathbb{R}^d$: model parameter
  - $p(\mathcal{D}|\boldsymbol{\theta})$: model density
  - $p(\boldsymbol{\theta})$: prior
- Goal : learning parameter $\boldsymbol{\theta}$ from data

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

- Difficulty : $p(\mathcal{D})$ unknown $\Rightarrow$ intractable posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$
  e.g., probabilistic graphical models, Bayesian hierarchical models

UCIrvine

Bayesian Inference

- Bayesian inference model
  - $\mathcal{D} = \{y^1, \ldots, y^N\}$: observed data
  - $\boldsymbol{\theta} \in \mathbb{R}^d$: model parameter
  - $p(\mathcal{D}|\boldsymbol{\theta})$: model density
  - $p(\boldsymbol{\theta})$: prior
- Goal : learning parameter $\boldsymbol{\theta}$ from data

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

- Difficulty : $p(\mathcal{D})$ unknown $\Rightarrow$ intractable posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$
  e.g., probabilistic graphical models, Bayesian hierarchical models
- Two popular approximations
  - **Markov chain Monte Carlo**. Sample by running a Markov chain : asymptotically unbiased but computationally slow
  - *Variational Bayes*. Approximate via tractable distributions : computationally fast but may result in poor approximation.

UCIrvine

Markov chain Monte Carlo

- Intuitive idea : evolve a Markov chain to sample from a target distribution $\pi(\boldsymbol{\theta})$ (METROPOLIS et al. 1953).

Markov chain Monte Carlo

- Intuitive idea : evolve a Markov chain to sample from a target distribution $\pi(\boldsymbol{\theta})$ (METROPOLIS et al. 1953).
- Conditions for transition kernel $T(\cdot|\cdot)$
  - *Irreducibility* : any state has positive probability of visiting any other state.
  - *Aperiodicity* : The chain should not get trapped in cycles.
  - *Detailed Balance condition (sufficient)* :

$$\pi(\boldsymbol{\theta})T(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}')T(\boldsymbol{\theta}|\boldsymbol{\theta}')$$

UCIrvine

Markov chain Monte Carlo

- Intuitive idea : evolve a Markov chain to sample from a target distribution $\pi(\boldsymbol{\theta})$ (METROPOLIS et al. 1953).
- Conditions for transition kernel $T(\cdot|\cdot)$
  - *Irreducibility* : any state has positive probability of visiting any other state.
  - *Aperiodicity* : The chain should not get trapped in cycles.
  - *Detailed Balance condition (sufficient)* :

$$\pi(\boldsymbol{\theta})T(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}')T(\boldsymbol{\theta}|\boldsymbol{\theta}')$$

- Metropolis-Hastings algorithms (one iteration)
  1. sample $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta})$
  2. update the current state to $\boldsymbol{\theta}'$ with probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min[1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}]$

UCIrvine

Markov chain Monte Carlo

- Intuitive idea : evolve a Markov chain to sample from a target distribution $\pi(\boldsymbol{\theta})$ (METROPOLIS et al. 1953).
- Conditions for transition kernel $T(\cdot|\cdot)$
  - *Irreducibility* : any state has positive probability of visiting any other state.
  - *Aperiodicity* : The chain should not get trapped in cycles.
  - *Detailed Balance condition (sufficient)* :

$$\pi(\boldsymbol{\theta})T(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}')T(\boldsymbol{\theta}|\boldsymbol{\theta}')$$

- Metropolis-Hastings algorithms (one iteration)
  1. sample $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta})$
  2. update the current state to $\boldsymbol{\theta}'$ with probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min[1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}]$
- Pros & Cons for simple MCMCs (e.g., RWM and Gibbs sampling)
  - Pro : easy to implement and computationally cheap
  - Con : slow mixing due to random walk behaviors, especially in complicate, high-dimensional models.

UCIrvine

Hamiltonian Monte Carlo

- Intuition : Leveraging a Hamiltonian dynamical system to generate trial moves in MCMC samplers. (DUANE et al. 1987, NEAL 2011)

Hamiltonian Monte Carlo

- Intuition : Leveraging a Hamiltonian dynamical system to generate trial moves in MCMC samplers. (DUANE et al. 1987, NEAL 2011)
- Model based energy function : the Hamiltonian

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

  - Potential : $U(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}, \mathcal{D}) = -[\log p(\boldsymbol{\theta}) + \log p(\mathcal{D}|\boldsymbol{\theta})]$
  - Kinetic : $K(\boldsymbol{r}) = \frac{1}{2}\boldsymbol{r}^\intercal M^{-1}\boldsymbol{r} \Rightarrow \pi(\boldsymbol{r}) \sim \mathcal{N}(\mathbf{0}, M)$

  The joint density of $\boldsymbol{z} = (\boldsymbol{\theta}, \boldsymbol{r})$ is

$$\pi(\boldsymbol{z}) \propto \exp(-U(\boldsymbol{\theta}) - K(\boldsymbol{r})) \propto p(\boldsymbol{\theta}|\mathcal{D}) \cdot \pi(\boldsymbol{r})$$

UCIrvine

Hamiltonian Monte Carlo

- Intuition : Leveraging a Hamiltonian dynamical system to generate trial moves in MCMC samplers. (DUANE et al. 1987, NEAL 2011)
- Model based energy function : the Hamiltonian

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

  - Potential : $U(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}, \mathcal{D}) = -[\log p(\boldsymbol{\theta}) + \log p(\mathcal{D}|\boldsymbol{\theta})]$
  - Kinetic : $K(\boldsymbol{r}) = \frac{1}{2}\boldsymbol{r}^\intercal M^{-1}\boldsymbol{r} \Rightarrow \pi(\boldsymbol{r}) \sim \mathcal{N}(\mathbf{0}, M)$

  The joint density of $\boldsymbol{z} = (\boldsymbol{\theta}, \boldsymbol{r})$ is

$$\pi(\boldsymbol{z}) \propto \exp(-U(\boldsymbol{\theta}) - K(\boldsymbol{r})) \propto p(\boldsymbol{\theta}|\mathcal{D}) \cdot \pi(\boldsymbol{r})$$

- 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 $\phi_s^H : \mathbb{R}^{2d} \to \mathbb{R}^{2d}, \quad \boldsymbol{z}(0) = \boldsymbol{z} \mapsto \boldsymbol{z}^* = \boldsymbol{z}(s)$

UCIrvine

Hamiltonian Monte Carlo

- Intuition : Leveraging a Hamiltonian dynamical system to generate trial moves in MCMC samplers. (DUANE et al. 1987, NEAL 2011)
- Model based energy function : the Hamiltonian

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

  - Potential : $U(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}, \mathcal{D}) = -[\log p(\boldsymbol{\theta}) + \log p(\mathcal{D}|\boldsymbol{\theta})]$
  - Kinetic : $K(\boldsymbol{r}) = \frac{1}{2}\boldsymbol{r}^\intercal M^{-1}\boldsymbol{r} \Rightarrow \pi(\boldsymbol{r}) \sim \mathcal{N}(\mathbf{0}, M)$

  The joint density of $\boldsymbol{z} = (\boldsymbol{\theta}, \boldsymbol{r})$ is

$$\pi(\boldsymbol{z}) \propto \exp(-U(\boldsymbol{\theta}) - K(\boldsymbol{r})) \propto p(\boldsymbol{\theta}|\mathcal{D}) \cdot \pi(\boldsymbol{r})$$

- 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 $\phi_s^H : \mathbb{R}^{2d} \to \mathbb{R}^{2d}, \quad \boldsymbol{z}(0) = \boldsymbol{z} \mapsto \boldsymbol{z}^* = \boldsymbol{z}(s)$
- Properties : **reversibility**, **volume preservation** and **constant Hamiltonian** over time t

Hamiltonian Monte Carlo

- Numerical integrators : the leap-frog scheme

$$\begin{aligned} \boldsymbol{r}^{(t+1/2)} &= \boldsymbol{r}^{(t)} - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(t)}) \\ \boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} + \varepsilon M^{-1}\boldsymbol{r}^{(t+1/2)} \\ \boldsymbol{r}^{(t+1)} &= \boldsymbol{r}^{(t+1/2)} - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(t+1)}) \end{aligned} \tag{1}$$

The leap-frog scheme (1) remains **time reversible** and **volume preserving.**☺
Constant Hamiltonian does not hold.☹

UCIrvine

Hamiltonian Monte Carlo

- Numerical integrators : the leap-frog scheme

$$\begin{aligned} \boldsymbol{r}^{(t+1/2)} &= \boldsymbol{r}^{(t)} - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(t)}) \\ \boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} + \varepsilon M^{-1}\boldsymbol{r}^{(t+1/2)} \\ \boldsymbol{r}^{(t+1)} &= \boldsymbol{r}^{(t+1/2)} - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(t+1)}) \end{aligned} \tag{1}$$

  The leap-frog scheme (1) remains **time reversible** and **volume preserving.**☺ **Constant Hamiltonian does not hold.**☹

- Metropolis correction

$$\alpha_{hmc}(\boldsymbol{z}, \boldsymbol{z}^*) = \min\{1, \exp(-H(\boldsymbol{z}^*) + H(\boldsymbol{z}))\} \tag{2}$$

Hamiltonian Monte Carlo

- Numerical integrators : the leap-frog scheme

$$\begin{aligned} \boldsymbol{r}^{(t+1/2)} &= \boldsymbol{r}^{(t)} - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(t)}) \\ \boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} + \varepsilon M^{-1}\boldsymbol{r}^{(t+1/2)} \\ \boldsymbol{r}^{(t+1)} &= \boldsymbol{r}^{(t+1/2)} - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(t+1)}) \end{aligned} \tag{1}$$

  The leap-frog scheme (1) remains **time reversible** and **volume preserving.**☺ **Constant Hamiltonian does not hold.**☹

- Metropolis correction

$$\alpha_{hmc}(\boldsymbol{z}, \boldsymbol{z}^*) = \min\{1, \exp(-H(\boldsymbol{z}^*) + H(\boldsymbol{z}))\} \tag{2}$$

- Some variants.
  - Automatically tuning of the hyper-parameters (e.g., step size ε, number of leap-frog steps L) (Hoffman et Gelman 2011 ; Wang, Mohamed et Nando 2013
  - Riemannian Manifold HMC (Girolami et Calderhead 2011)

UCIrvine

Stochastic Gradient MCMC

- Challenge from massive data.

$$\nabla_{\theta} U(\theta) = -\sum_{y \in \mathcal{D}} \nabla_{\theta} \log p(y|\theta) - \nabla_{\theta} \log p(\theta) \quad \sim \mathcal{O}(N)$$

Stochastic Gradient MCMC

- Challenge from massive data.

$$\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) = -\sum_{y \in \mathcal{D}} \nabla_{\boldsymbol{\theta}} \log p(y|\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) \quad \sim \mathcal{O}(N)$$

- Stochastic gradient MCMC : use stochastic gradients

$$\nabla_{\boldsymbol{\theta}} \tilde{U}(\boldsymbol{\theta}) = -\frac{|\mathcal{D}|}{|\tilde{\mathcal{D}}|} \sum_{y \in \tilde{\mathcal{D}}} \nabla_{\boldsymbol{\theta}} \log p(y|\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}), \quad \tilde{\mathcal{D}} \subset \mathcal{D}$$

e.g., SGLD (Welling et Teh 2011), SGHMC (Chen, E. B. Fox et Guestrin 2014), SGNHT (Ding et al. 2014)

UCIrvine

Stochastic Gradient MCMC

- Challenge from massive data.

$$\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) = -\sum_{y \in \mathcal{D}} \nabla_{\boldsymbol{\theta}} \log p(y|\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) \quad \sim \mathcal{O}(N)$$

- Stochastic gradient MCMC : use stochastic gradients

$$\nabla_{\boldsymbol{\theta}} \tilde{U}(\boldsymbol{\theta}) = -\frac{|\mathcal{D}|}{|\tilde{\mathcal{D}}|} \sum_{y \in \tilde{\mathcal{D}}} \nabla_{\boldsymbol{\theta}} \log p(y|\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}), \quad \tilde{\mathcal{D}} \subset \mathcal{D}$$

  e.g., SGLD (Welling et Teh 2011), SGHMC (Chen, E. B. Fox et Guestrin 2014), SGNHT (Ding et al. 2014)

- Properties of stochastic gradient MCMC
  - Convergence relies on stochastic differential equation of the form

$$d\boldsymbol{z} = f(\boldsymbol{z})dt + \sqrt{2D(\boldsymbol{z})}dW(t)$$

    with appropriate $f(\boldsymbol{z})$ and $D(\boldsymbol{z})$. See a complete recipe (Ma, Chen et E. Fox 2015)
  - Need to anneal (or use small) step size, sacrificing the exploration efficiency to compromise scalability (Betancourt 2015).

UCIrvine

Surrogate Method

Find cheap surrogate function $U^*(\boldsymbol{\theta})$ and transition kernel $T_s(\cdot|\cdot)$ that leaves $q^*(\boldsymbol{\theta}) \propto \exp(-U^*(\boldsymbol{\theta}))$ invariant.

$$q^*(\boldsymbol{\theta})T_s(\boldsymbol{\theta}'|\boldsymbol{\theta}) = q^*(\boldsymbol{\theta}')T_s(\boldsymbol{\theta}|\boldsymbol{\theta}')$$

Use T_s to generate proposals

Proposition (Liu 2001)

The target distribution $p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp(-U(\boldsymbol{\theta}))$ is the stationary distribution for a Markov chain simulated according to the following procedure : given the current state $\boldsymbol{\theta}$, let $\boldsymbol{\vartheta}_0 = \boldsymbol{\theta}$ and recursively sample $\boldsymbol{\vartheta}_i \sim T_s(\cdot|\boldsymbol{\vartheta}_{i-1})$ for $i = 1, \ldots, k$. Then, accept the proposal $\boldsymbol{\theta}^* = \boldsymbol{\vartheta}_k$ with the following probability

$$\alpha_s(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\mathcal{D})q^*(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})q^*(\boldsymbol{\theta}^*)}\right\}$$

Some of the existing surrogate methods : Gaussian process surrogate (RASMUSSEN 2003 ; LAN et al. 2015), reproducing kernel Hilbert space surrogate (STRATHMANN et al. 2015) and random network surrogate (ZHANG, SHAHBABA et ZHAO 2015).

UCIrvine

Lower Bounds and Free Energy

Fixed-Form Variational Bayes (HONKELA et al. 2010 ; SAUL et JORDAN 1996) uses a parametrized distribution

$$q_\eta(\boldsymbol{\theta}) = \exp[T(\boldsymbol{\theta})\boldsymbol{\eta} - A(\boldsymbol{\eta})] \tag{3}$$

to approximate the target posterior $p(\boldsymbol{\theta}|\mathcal{D})$

Lower Bounds and Free Energy

Fixed-Form Variational Bayes (HONKELA et al. 2010 ; SAUL et JORDAN 1996) uses a parametrized distribution

$$q_\eta(\boldsymbol{\theta}) = \exp[T(\boldsymbol{\theta})\boldsymbol{\eta} - A(\boldsymbol{\eta})] \tag{3}$$

to approximate the target posterior $p(\boldsymbol{\theta}|\mathcal{D})$

- Distance between distributions

$$\begin{aligned} D_{KL}(q_{\boldsymbol{\eta}}(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\mathcal{D})) &= \int q_{\boldsymbol{\eta}}(\boldsymbol{\theta}) \log\left(\frac{q_{\boldsymbol{\eta}}(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})}\right) d\boldsymbol{\theta} \\ &= \log p(\mathcal{D}) - \underbrace{\int q_{\boldsymbol{\eta}}(\boldsymbol{\theta}) \log\left(\frac{p(\boldsymbol{\theta},\mathcal{D})}{q_{\boldsymbol{\eta}}(\boldsymbol{\theta})}\right) d\boldsymbol{\theta}}_{\text{Free Energy (Lower bound)}} \end{aligned}$$

UCIrvine

Lower Bounds and Free Energy

Fixed-Form Variational Bayes (HONKELA et al. 2010 ; SAUL et JORDAN 1996) uses a parametrized distribution

$$q_\eta(\boldsymbol{\theta}) = \exp[T(\boldsymbol{\theta})\boldsymbol{\eta} - A(\boldsymbol{\eta})] \tag{3}$$

to approximate the target posterior $p(\boldsymbol{\theta}|\mathcal{D})$

- Distance between distributions

$$\begin{aligned} D_{KL}(q_{\boldsymbol{\eta}}(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\mathcal{D})) &= \int q_{\boldsymbol{\eta}}(\boldsymbol{\theta}) \log\left(\frac{q_{\boldsymbol{\eta}}(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})}\right) d\boldsymbol{\theta} \\ &= \log p(\mathcal{D}) - \underbrace{\int q_{\boldsymbol{\eta}}(\boldsymbol{\theta}) \log\left(\frac{p(\boldsymbol{\theta},\mathcal{D})}{q_{\boldsymbol{\eta}}(\boldsymbol{\theta})}\right) d\boldsymbol{\theta}}_{\text{Free Energy (Lower bound)}} \end{aligned}$$

- An optimization problem

$$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} \mathbb{E}_{q_{\boldsymbol{\eta}}}[\log p(\boldsymbol{\theta},\mathcal{D}) - \log q_{\boldsymbol{\eta}}(\boldsymbol{\theta})] \tag{4}$$

more accurate than using mean-field assumptions ; requires analytically evaluation of $\mathbb{E}_{q_{\boldsymbol{\eta}}} \log q_{\boldsymbol{\eta}}(\boldsymbol{\theta})$, $\mathbb{E}_{q_{\boldsymbol{\eta}}} \log p(\boldsymbol{\theta},\mathcal{D})$ and the derivatives.

Stochastic Linear Regression

Optimization in (4) can be solved using stochastic linear regression (SALIMANS et KNOWLES 2013). Rewrite (3) in the unnormalized form

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

Unnormalized KL divergence :

$$\begin{aligned} \tilde{D}_{KL}(\tilde{q}_{\tilde{\eta}} \| p(\theta, \mathcal{D})) &= \int \tilde{q}_{\tilde{\eta}}(\theta) \log \frac{\tilde{q}_{\tilde{\eta}}(\theta)}{p(\theta, \mathcal{D})} d\theta - \int \tilde{q}_{\tilde{\eta}}(\theta) d\theta \\ &= \int \exp[\tilde{T}(\theta)\tilde{\eta}][\tilde{T}(\theta)\tilde{\eta} - \log p(\theta, \mathcal{D})] d\theta - \int \exp[\tilde{T}(\theta)\tilde{\eta}] d\theta \end{aligned} \tag{5}$$

Find the minimum by differentiation

$$\nabla_{\tilde{\eta}} \tilde{D}_{KL}(\tilde{q}_{\tilde{\eta}} \| p(\theta, \mathcal{D})) = \int \tilde{q}_{\tilde{\eta}}(\theta)[\tilde{T}(\theta)^\intercal \tilde{T}(\theta)\tilde{\eta} - \tilde{T}(\theta)^\intercal \log p(\theta, \mathcal{D})] d\theta = \mathbf{0}$$

The minimum :

$$\tilde{\eta} = \left(\mathbb{E}_q[\tilde{T}(\theta)^\intercal \tilde{T}(\theta)]\right)^{-1} \mathbb{E}_q[\tilde{T}(\theta)^\intercal \log p(\theta, \mathcal{D})] \tag{6}$$

UCIrvine

Monte Carlo Estimation

(6) is not a solution yet, but can be used to derive a fix point iteration. Let

$$C = \mathbb{E}_q[\tilde{T}(\boldsymbol{\theta})^\intercal \tilde{T}(\boldsymbol{\theta})], \quad g = \mathbb{E}_q[\tilde{T}(\boldsymbol{\theta})^\intercal \log p(\boldsymbol{\theta}, \mathcal{D})]$$

then $\tilde{\boldsymbol{\eta}} = C^{-1} g$.

Stochastic Optimization for Fixed-Form VB

Initialize $\tilde{\boldsymbol{\eta}}_1, C_1$, $g_1 = C_1 \tilde{\boldsymbol{\eta}}_1$, set step-size $w \in [0, 1]$
for $t = 1$ to T **do**
  Drawing a single sample $\boldsymbol{\theta}_t^*$ from the current approximation $q_{\boldsymbol{\eta}_t}(\boldsymbol{\theta})$
  Update C_{t+1}, g_{t+1} as follows :

$$g_{t+1} = (1 - w) g_t + w \hat{g}_t, \quad \hat{g}_t = \tilde{T}(\boldsymbol{\theta}_t^*)^\intercal \log p(\boldsymbol{\theta}_t^*, \mathcal{D})$$

$$C_{t+1} = (1 - w) C_t + w \hat{C}_t, \quad \hat{C}_t = \tilde{T}(\boldsymbol{\theta}_t^*)^\intercal \tilde{T}(\boldsymbol{\theta}_t^*)$$

  Update the parameters :

$$\tilde{\boldsymbol{\eta}}_{t+1} = C_{t+1}^{-1} g_{t+1}$$

end for

See SALIMANS et KNOWLES 2013 for more variants.

UCIrvine

Combine VB and HMC via Surrogate

- Surrogate method revisit

$$\text{Target} : p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp(-U(\boldsymbol{\theta})), \quad \text{Surrogate} : q^*(\boldsymbol{\theta}) \propto \exp(-U^*(\boldsymbol{\theta}))$$

Acceptance probability

$$\alpha_s(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\mathcal{D})q^*(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})q^*(\boldsymbol{\theta}^*)}\right\}$$

High approximation quality $\Rightarrow$ high acceptance rate.

UCIrvine

Combine VB and HMC via Surrogate

- Surrogate method revisit

$$\text{Target} : p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp(-U(\boldsymbol{\theta})), \quad \text{Surrogate} : q^*(\boldsymbol{\theta}) \propto \exp(-U^*(\boldsymbol{\theta}))$$

  Acceptance probability

$$\alpha_s(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\mathcal{D})q^*(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})q^*(\boldsymbol{\theta}^*)}\right\}$$

  High approximation quality $\Rightarrow$ high acceptance rate.

- Flexible and efficient surrogate based on random network

$$U^*(\boldsymbol{\theta}) = z(\boldsymbol{\theta}) = \sum_{i=1}^{s} v_i a(\boldsymbol{\theta}; \boldsymbol{\gamma_i}) \quad \sim \quad \mathcal{O}(s) \tag{7}$$

  where $\{\boldsymbol{\gamma}_i\}_{i=1}^n$ are random samples from some distribution.

UCIrvine

Combine VB and HMC via Surrogate

- Surrogate method revisit

$$\text{Target} : p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp(-U(\boldsymbol{\theta})), \quad \text{Surrogate} : q^*(\boldsymbol{\theta}) \propto \exp(-U^*(\boldsymbol{\theta}))$$

  Acceptance probability

$$\alpha_s(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\mathcal{D})q^*(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})q^*(\boldsymbol{\theta}^*)}\right\}$$

  High approximation quality $\Rightarrow$ high acceptance rate.

- Flexible and efficient surrogate based on random network

$$U^*(\boldsymbol{\theta}) = z(\boldsymbol{\theta}) = \sum_{i=1}^{s} v_i a(\boldsymbol{\theta}; \boldsymbol{\gamma_i}) \quad \sim \quad \mathcal{O}(s) \tag{7}$$

  where $\{\boldsymbol{\gamma}_i\}_{i=1}^n$ are random samples from some distribution.

- Use **variational Bayes** to improve approximation.

$$\hat{q}(\boldsymbol{\theta}) = \arg\min_{q^*} D(q^*(\boldsymbol{\theta}), p(\boldsymbol{\theta}, \mathcal{D}))$$

UCIrvine

A Rich Class of Functions : $\mathcal{F}_p$

Bases : $\{a(\boldsymbol{\theta};\boldsymbol{\gamma}) : \boldsymbol{\gamma} \in \boldsymbol{\Gamma}\}$, $\boldsymbol{\Theta} \subset \mathbb{R}^d$, a distribution on $\boldsymbol{\Gamma}$: $p(\boldsymbol{\gamma})$, consider functions

$$f(\boldsymbol{\theta}) = \int_{\boldsymbol{\Gamma}} \alpha(\boldsymbol{\gamma}) a(\boldsymbol{\theta};\boldsymbol{\gamma})\, d\boldsymbol{\gamma} \tag{8}$$

where $|\alpha(\boldsymbol{\gamma})| \leq C|p(\boldsymbol{\gamma})|$. Define a norm $\|f\|_p = \sup_{\boldsymbol{\gamma}} \left|\frac{\alpha(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma})}\right|$ and the set

$$\mathcal{F}_p \equiv \left\{ f(\boldsymbol{\theta}) = \int_{\boldsymbol{\Gamma}} \alpha(\boldsymbol{\gamma}) a(\boldsymbol{\theta};\boldsymbol{\gamma})\, d\boldsymbol{\gamma} \;\middle|\; \|f\|_p < \infty \right\}$$

Theorem (Rahimi 2008)

Let μ be any probability measure on $\boldsymbol{\Theta}$, $\|f\|_\mu^2 = \int_{\boldsymbol{\Theta}} f^2(\boldsymbol{\theta})\mu(d\boldsymbol{\theta})$. Suppose $\sup_{\boldsymbol{\theta},\boldsymbol{\gamma}} |a(\boldsymbol{\theta};\boldsymbol{\gamma})| \leq 1$. Fix $f \in \mathcal{F}_p$. $\forall \delta > 0$, with probability at least $1-\delta$ over $\boldsymbol{\gamma}_i \overset{iid}{\sim} p(\boldsymbol{\gamma})$, there exist $v_1, \ldots, v_s$ such that $z(\boldsymbol{\theta}) = \sum_{i=1}^s v_i a(\boldsymbol{\theta};\boldsymbol{\gamma}_i)$ satisfies

$$\|z - f\|_\mu < \frac{\|f\|_p}{\sqrt{s}} \left(1 + \sqrt{2 \log \frac{1}{\delta}}\right)$$

UCIrvine

$\mathcal{F}_p$ is dense in $\mathcal{H}$

A reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with kernel k on $\boldsymbol{\Theta} \times \boldsymbol{\Theta}$

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

$\mathcal{H}$ can be constructed based on functions of the form (8).

Proposition (Rahimi 2008)

Let $\hat{\mathcal{H}}$ be the completion of the set of functions of the form (8) *such that*

$$\int_{\boldsymbol{\Gamma}} \frac{\alpha(\boldsymbol{\gamma})^2}{p(\boldsymbol{\gamma})} \, d\boldsymbol{\gamma} < \infty \tag{9}$$

with the inner product

$$\langle f, g \rangle = \int_{\boldsymbol{\Gamma}} \frac{\alpha(\boldsymbol{\gamma}) \beta(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma})} \, d\boldsymbol{\gamma}$$

where $g(\boldsymbol{\theta}) = \int_{\boldsymbol{\gamma}} \beta(\boldsymbol{\gamma}) a(\boldsymbol{\theta}; \boldsymbol{\gamma}) \, d\boldsymbol{\gamma}$. Then $\hat{\mathcal{H}} = \mathcal{H}$

Note that $\forall f \in \mathcal{F}_p$, (9) is satisfied. Therefore, $\mathcal{F}_p$ is a subset of $\mathcal{H}$. In fact, $\mathcal{F}_p$ is dense in $\mathcal{H}$. See RAHIMI et RECHT 2008 for detailed proofs.

Free-Form Variational Bayes

- Random bases surrogate induced distribution

$$q_{\boldsymbol{v}}(\boldsymbol{\theta}) \propto \exp(-z(\boldsymbol{\theta})) = \exp[-\sum_{i=1}^{s} v_i a(\boldsymbol{\theta}; \boldsymbol{\gamma}_i) - \Phi(\boldsymbol{v})]$$

  where $\boldsymbol{\gamma}_i$ are drawn iid from some distribution. The natural parameters $\boldsymbol{v}$ of interest belong to the set

$$\Omega := \{\boldsymbol{v} \in \mathbb{R}^s \mid \Phi(\boldsymbol{v}) < +\infty\}$$

- A distance measure $D: (q, p) \to [0, +\infty)$, q, p are unnormalized densities.
- **Free-Form** variational inference

$$\hat{\boldsymbol{v}} = \arg\min_{\boldsymbol{v} \in \Omega} D(q_{\boldsymbol{v}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}, \mathcal{D})) \tag{10}$$

  - $q_{\boldsymbol{v}}(\boldsymbol{\theta})$ does not have to be tractable
  - Free style construction of the random network surrogate.

UCIrvine

Distance Between Unnormalized Densities

- Potential matching distance

$$
\begin{aligned}
D_{PM}(q_{\boldsymbol{v}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}, \mathcal{D})) &= \frac{1}{2} \min_{b} \int q_{\boldsymbol{v}}(\boldsymbol{\theta}) \| z(\boldsymbol{\theta}) - U(\boldsymbol{\theta}) - b \|^2 \, d\boldsymbol{\theta} \\
&= \frac{1}{2} \mathbb{V}\mathrm{ar}_{q_{\boldsymbol{v}}} \left(z(\boldsymbol{\theta}) - U(\boldsymbol{\theta}) \right)
\end{aligned} \tag{11}
$$

- **"Score matching"** distance

$$
\tilde{D}_{SM}(q_{\boldsymbol{v}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}, \mathcal{D})) = \frac{1}{2} \int q_{\boldsymbol{v}}(\boldsymbol{\theta}) \| \nabla_{\boldsymbol{\theta}} z(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \|^2 \, d\boldsymbol{\theta} \tag{12}
$$

The above distances are well-defined

$$
\begin{gathered}
D_{PM}(q_{\boldsymbol{v}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}, \mathcal{D})) = 0 \text{ or } \tilde{D}_{SM}(q_{\boldsymbol{v}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}, \mathcal{D})) = 0 \\
\Rightarrow z(\boldsymbol{\theta}) = U(\boldsymbol{\theta}) + \text{Constant} \Rightarrow q_{\boldsymbol{v}}(\boldsymbol{\theta}) = p(\boldsymbol{\theta} | \mathcal{D})
\end{gathered}
$$

In practice, (12) is usually intractable, use **empirical version** instead.

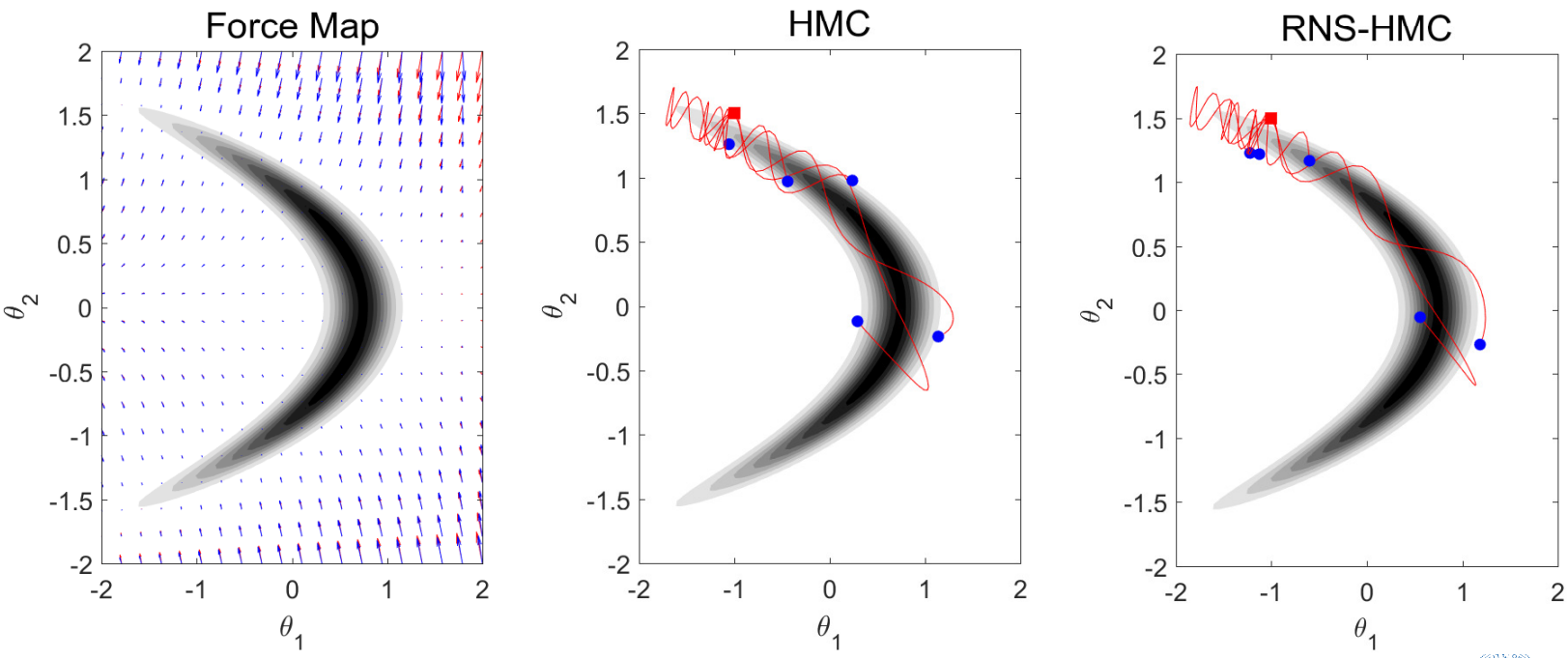
UCIrvine

Surrogate Induced Hamiltonian Flow

Define $\tilde{H}(\boldsymbol{\theta}, \boldsymbol{r}) = z(\boldsymbol{\theta}) + K(\boldsymbol{r})$, simulating the surrogate induced Hamiltonian Dynamics

$$\frac{d\boldsymbol{\theta}}{dt} = M^{-1}\boldsymbol{r}, \quad \frac{d\boldsymbol{r}}{dt} = -\nabla_{\boldsymbol{\theta}} z(\boldsymbol{\theta}) \tag{13}$$

defines a mapping $\phi_s^{\tilde{H}} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}, \quad (\boldsymbol{\theta}, \boldsymbol{r}) \mapsto (\boldsymbol{\theta}^*, \boldsymbol{r}^*)$



Variational Hamiltonian Monte Carlo

1. Simulate surrogate induced Hamiltonian flow to generate $(\boldsymbol{\theta}^*, \boldsymbol{r}^*)$ and accept with probability
$$\alpha_{vhmc} = \min\{1, \exp(\tilde{H}(\boldsymbol{\theta}, \boldsymbol{r}) - \tilde{H}(\boldsymbol{\theta}^*, \boldsymbol{r}^*))\}$$
2. Add to the training data set.
$$\mathcal{T}_s^{(t)} := \mathcal{T}_s^{(t-1)} \cup \{(\boldsymbol{\theta}_t, \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}_t))\}$$
3. Update surrogate by minimizing the empirical squared distance plus regularization.
$$\hat{\boldsymbol{v}}_t = \arg\min_{\boldsymbol{v}} \frac{1}{2} \sum_{n=1}^{t} \|\nabla_{\boldsymbol{\theta}} z(\boldsymbol{\theta}_n) - \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}_{\boldsymbol{n}})\|^2 + \frac{\lambda}{2} \|\boldsymbol{v}\|^2 \tag{14}$$

Regularized surrogate approximation to simulate Hamiltonian flow

$$V_t(\boldsymbol{\theta}) = \mu_t z_t(\boldsymbol{\theta}) + \frac{1}{2}(1 - \mu_t)(\boldsymbol{\theta} - \boldsymbol{\theta}^L)^\intercal \nabla^2_{\boldsymbol{\theta}} U(\boldsymbol{\theta})^L (\boldsymbol{\theta} - \boldsymbol{\theta}^L), \quad \mu_t : 0 \uparrow 1$$

Online Updating

Let $A(\boldsymbol{\theta}) = (A_1(\boldsymbol{\theta}), A_2(\boldsymbol{\theta}, \ldots, A_s(\boldsymbol{\theta})))$, where

$$A_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} a(\boldsymbol{\theta}; \gamma_i), \quad i = 1, \ldots, s$$

Variational Hamiltonian Monte Carlo

1: Set λ, μ_t, s and HMC parameters ε, L. Initialize $\boldsymbol{\theta}^{(0)}$ to a first guess, $\boldsymbol{v}^{(0)} = \mathbf{0}$, $C^{(0)} = \frac{1}{\lambda} I_s$. Find $\boldsymbol{\theta}^L$ and compute $\nabla^2_{\boldsymbol{\theta}} U(\boldsymbol{\theta})^L$.
2: **for** $t = 1$ to T **do**
3: Perform one HMC iteration for the regularized surrogate induced distribution $q_{\boldsymbol{v}}^{(t)}(\boldsymbol{\theta}) \propto \exp(-V_t(\boldsymbol{\theta}))$ to draw $(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{r}^{(t+1)})$
4: Acquire $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(t+1)})$ and $A_{t+1} = A(\boldsymbol{\theta}^{(t+1)})$
5: Compute $W^{(t+1)} = C^{(t)} A_{t+1}^{\intercal} [I_d + A_{t+1} C^{(t)} A_{t+1}^{\intercal}]^{-1}$
6: Update $\boldsymbol{v}^{(t+1)}, C^{(t+1)}$ as follows
7: $\boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)} + W^{(t+1)} (\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(t+1)}) - A_{t+1} \boldsymbol{v}^{(t)})$
8: $C^{(t+1)} = C^{(t)} - W^{(t+1)} A_{t+1} C^{(t)}$
9: **end for**

UCIrvine

Stopping time and Connections to Existing Work

- Inefficient updating when surrogate is well trained $\Rightarrow$ set up a stopping time t_0
  - $t < t_0$. Perform **Free-Form** variational Bayes to improve approximate quality of the surrogate
  - $t > t_0$. Perform standard **HMC** that samples from the surrogate induced distribution.
- Connections to some existing work
  - Stochastic linear regression for **Fix-Form** variational Bayes (SALIMANS et KNOWLES 2013) : allows **Free-Form** intractable approximate distributions, use HMC to draw samples.
  - Random bases surrogate HMC (ZHANG, SHAHBABA et ZHAO 2015) : further reduce the computation by allowing to use the fast surrogate in the Metropolis correction step, trade-off between approximation accuracy and computational efficiency.

UCIrvine

A Beta-binomial Model for Overdispersion

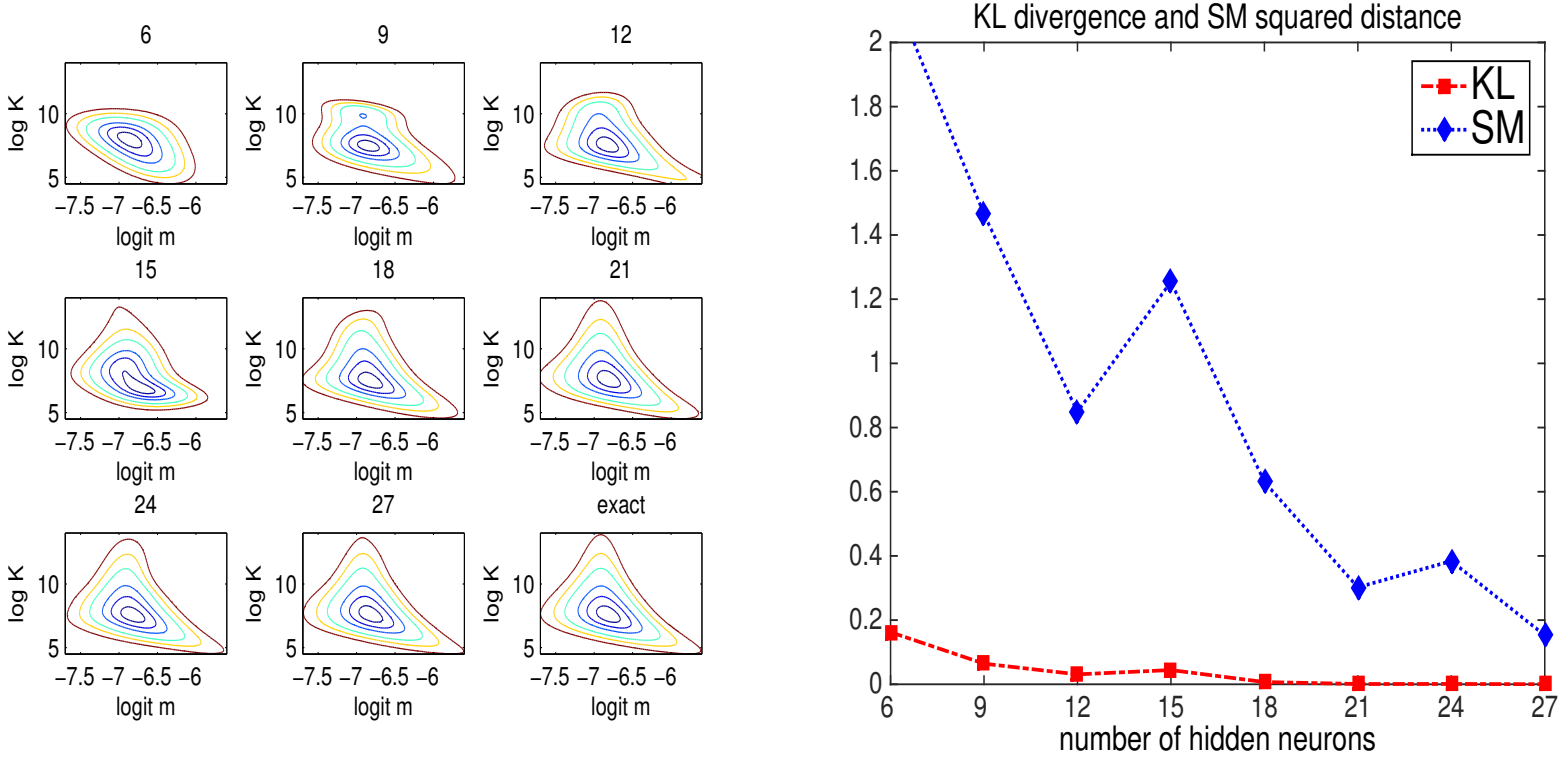


FIGURE – **Left** : Approximate posteriors for a varying number of hidden neurons. Exact posterior at bottom right. **Right** : KL-divergence and score matching squared distance between the surrogate approximation and the exact posterior density using an increasing number of hidden neurons.

UCIrvine

Bayesian Probit Regression

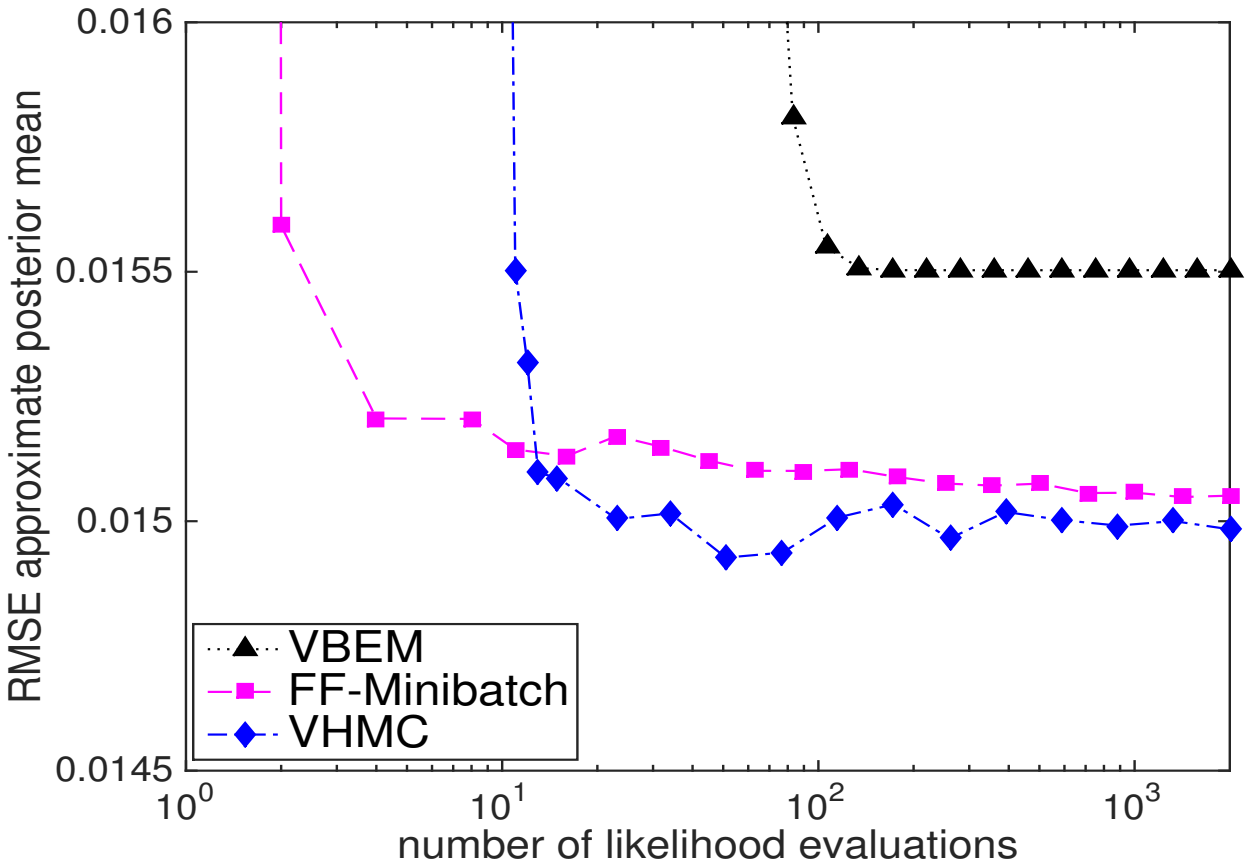


FIGURE – RMSE of the approximate posterior mean as a function of the number of likelihood evaluations for different variational Bayesian approaches and VHMC algorithm.

UCIrvine

Independent Component Analysis

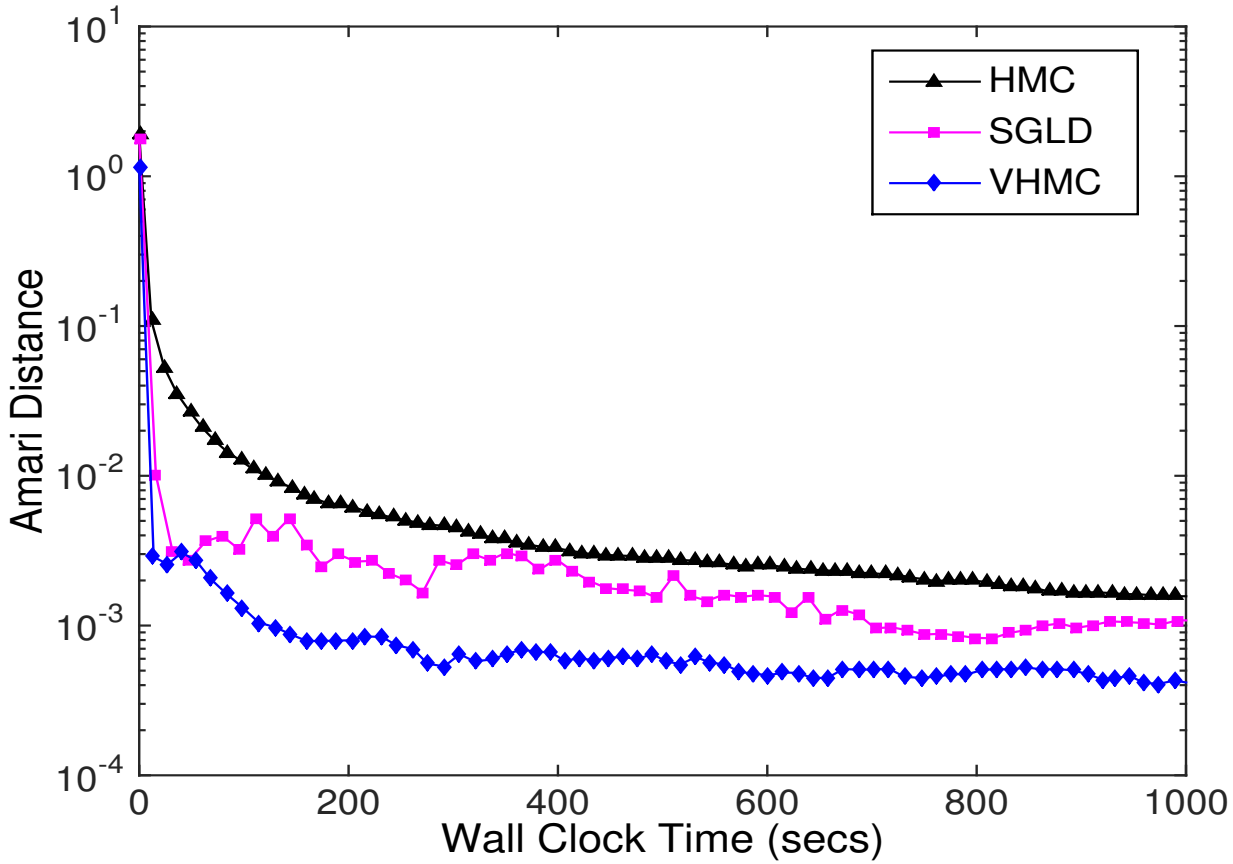


FIGURE – Convergence of Amari distance on the MEG data for HMC, SGLD and our variational HMC algorithm.

UCIrvine

Summary and Discussion

Combine variational Bayes and MCMC via random bases surrogate :

- Construct efficient surrogate to accelerate HMC

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

- Find good surrogate via **Free-Form** variational Bayes

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

Open future directions :

- The random bases surrogate is more effective in problems with costly likelihood and a moderate number of parameters. How about really high dimensional problems ? Would more sophisticated structures (e.g., deep networks) work ?
- Evaluating full-data gradient $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$ to collect training data is still expensive, how about stochastic gradient ? Be careful about overfitting !

UCIrvine

Reference I


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


CHEN, T., E. B. FOX et C. GUESTRIN (2014). "Stochastic Gradient Hamiltonian Monte Carlo". In : **Proceedings of 31st International Conference on Machine Learning (ICML 2014)**.


DING, N. et al. (2014). "Bayesian Sampling Using Stochastic Gradient Thermostats". In : **Advances in Neural Information Processing Systems 27 (NIPS 2014)**.


DUANE, S. et al. (1987). "Hybrid Monte Carlo". In : **Physics Letters B** 195.2, p. 216–222.


GIROLAMI, M. et B. CALDERHEAD (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In : **Journal of the Royal Statistical Society** (with discussion) 73.2, p. 123–214.


HOFFMAN, D. et A. GELMAN (2011). **The No-U-Turn Sampler : Adaptively Setting Path Lengths in Hamiltonian Monte Carlo**. arxiv.org/abs/1111.4246.


HONKELA, A. et al. (2010). "Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes". In : **Journal of Machine Learning Research** 11, p. 3235–3268.

Reference II


LAN, S. et al. (2015). **Emulation of higher-order tensors in manifold Monte Carlo methods for Bayesian inverse problems**. arxiv.org/abs/1507.06244.


MA, Y. A., T. CHEN et E. FOX (2015). "A complete recipe for stochastic gradient MCMC". In : **Advances in Neural Information Processing Systems 28 (NIPS 2015)**.


METROPOLIS, N. et al. (1953). "Equation of State Calculations by Fast Computing Machines". In : **The Journal of Chemical Physics** 21.6, p. 1087–1092.


NEAL, R. M. (2011). "MCMC using Hamiltonian dynamics". In : **Handbook of Markov Chain Monte Carlo**. Sous la dir. de S. BROOKS et al. Chapman et Hall/CRC, p. 113–162.


RAHIMI, A. et B. RECHT (2008). "Uniform approximation of functions with random bases". In : **Proc. 46th Ann. Allerton Conf. Commun., Contr. Comput.**


RASMUSSEN, C. E. (2003). "Gaussian Processes to Speed up Hybrid Monte Carlo for Expensive Bayesian Integrals". In : **Bayesian Statistics** 7, p. 651–659.


SALIMANS, T. et D. A. KNOWLES (2013). "Fixed-form variational posterior approximation through stochastic linear regression". In : **Bayesian Analysis** 8.4, p. 837–882.

UCIrvine

Reference III

- SAUL, L. et M. I. JORDAN (1996). "Exploiting tractable substructures in intractable networks". In : **Advance in neural information processing systems 7 (NIPS 1996)**. Sous la dir. de G. TESAURO, D. S. TOURETZKY et T. K. LEEN. Cambridge, MA : MIT Press, p. 486–492.
- STRATHMANN, H. et al. (2015). "Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families". In : **Advances in Neual Information Processing Systems**. Cambridge, MA : MIT Press.
- WANG, Z., S. MOHAMED et D. NANDO (2013). "Adaptive Hamiltonian and Riemann manifold Monte Carlo". In : **Proceedings of the 30th International Conference on Machine Learning (ICML 2013)**, p. 1462–1470.
- WELLING, M. et Y. W. TEH (2011). "Bayesian Learning via Stochastic Gradient Langevin Dynamics". In : **Proceedings of the International Conference on Machine Learning**.
- ZHANG, C., B. SHAHBABA et H. K. ZHAO (2015). **Hamiltonian Monte Carlo Acceleration Using Surrogate Functions with Random Bases**. arxiv.org/abs/1506.05555.

UCIrvine