

Stochastic Gradient Geodesic MCMC Methods

Chang Liu[†], Jun Zhu[†], Yang Song^{‡1}

[†] Dept. of Comp. Sci. & Tech., TNList Lab; Center for Bio-Inspired Computing Research

[†] State Key Lab for Intell. Tech. & Systems, Tsinghua University, Beijing, China

[‡] Dept. of Physics, Tsinghua University, Beijing, China

{chang-li14@mails., dcszj@}tsinghua.edu.cn;

songyang@stanford.edu

NIPS'16 @ Barcelona

¹ JZ is the corresponding author; YS is with Department of Computer Science, Stanford University, CA.

- 1 Introduction
- 2 Preliminaries
- 3 Stochastic Gradient Geodesic MCMC Methods
 - Technical Description of the Settings
 - The Dynamics
 - Simulation with 2nd-order Geodesic Integrators
- 4 Application to Spherical Admixture Model
 - Description of SAM
 - Posterior Sampling for SAM
- 5 Experiments
 - Toy Experiment
 - Synthetic Experiment
 - Real-world Experiment: SAM

Motivation

Motivation: the need of efficient sampling on manifold for Bayesian inference

- Bayesian inference: get access to the posterior $\pi(q|\mathcal{D})$ (e.g. by sampling from it).
+
- To describe data on manifold (e.g. normalized tf-idf feature is on hyperspheres), Bayesian models need to use q that is also on manifold (e.g. SAM [14] uses q also on hyperspheres)
=
- Inference for these models (sampling from $\pi(q|\mathcal{D})$) is a problem of sampling on manifold.

Challenges

Challenges of efficient sampling on manifold for Bayesian inference

- Constraint of manifold
 - To sample from hyperspheres, samples have to satisfy: $\|q\| = 1$.
- Non-scalability
 - To sample from a posterior $\pi(q|\mathcal{D}) \propto \pi_0(q) \prod_{d=1}^D \pi(x_d|q)$: $\mathcal{O}(D)!$

Current Stage

Current Stage (Related works)

- Constraint of manifold
 - MALA & RMHMC [9]: sample in coordinate space, iteratively simulate
 - CHMC [3]: iterative projection on manifold.
 - GMC [4]: sample in embedded space and use geodesic for simulation
- Non-scalability: use mini-batch to estimate stochastic gradient (SG-MCMC)
 - SGLD [15], SGHMC [6], SGNHT [7]
 - Ma et al. (2015) [12] give a framework for the dynamics of SG-MCMC, Chen et al. (2015) [5] give an analysis on simulation methods (integrators) of the dynamics.
- Combining both: SGRLD [13], SGRHMC [12]. But:
 - Both by sampling in coordinate space, so fail when manifold has no global coordinate systems (e.g. hyperspheres).
 - Hard to apply better integrators.

Our Work

To combine both in a better way:

Stochastic Gradient Geodesic Monte Carlo (SGGMC),

Geodesic Stochastic Gradient Nosé-Hoover Thermostats (gSGNHT).

- Release the requirement of global coordinate systems: sample in embedded space
- Use stochastic gradient: adopt the framework of [12].
- Avoid inner iteration: use geodesic for simulation.
- Better integrator: following the idea of [5].

Our Work

Table: A summary of related methods. –: sampling on manifold not supported; †: The integrators are not in the SSI scheme (It is unclear whether the claimed “2nd-order” is equivalent to ours); ‡: 2nd-order integrators for SGHMC and mSGNHT are developed by [5] and [11], respectively.

methods	stochastic gradient	no inner iteration	no global coordinates	order of integrator
GMC [4]	×	✓	✓	2nd
RMLD [9]	×	✓	×	1st
RMHMC [9]	×	×	×	2nd [†]
CHMC [3]	×	×	✓	2nd [†]
SGLD [15]	✓	✓	–	1st
SGHMC [6] / SGNHT [7]	✓	✓	–	1st [‡]
SGRLD [13] / SGRHMC [12]	✓	✓	×	1st
SGGMC / gSGNHT (proposed)	✓	✓	✓	2nd

- 1 Introduction
- 2 Preliminaries**
- 3 Stochastic Gradient Geodesic MCMC Methods
 - Technical Description of the Settings
 - The Dynamics
 - Simulation with 2nd-order Geodesic Integrators
- 4 Application to Spherical Admixture Model
 - Description of SAM
 - Posterior Sampling for SAM
- 5 Experiments
 - Toy Experiment
 - Synthetic Experiment
 - Real-world Experiment: SAM

Stochastic Gradient MCMC

- Prevalent MCMC: require the potential energy $U(q) \triangleq -\log \pi(q|\mathcal{D})$ and its gradient

$$\nabla_q U(q) = -\nabla_q \log \pi(q) - \sum_{d=1}^D \nabla_q \log \pi(x_d|q) \quad (1)$$

- Stochastic Gradient MCMC (SG-MCMC): use a randomly chosen mini-batch \mathcal{S} to estimate the gradient

$$\nabla_q \tilde{U}(q) = -\nabla_q \log \pi_0(q) - \frac{D}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \nabla_q \log \pi(x|q) \quad (2)$$

For i.i.d. data, we can approximate

$$\nabla_q \tilde{U}(q) = \nabla_q U(q) + \mathcal{N}(0, V(q)) \quad (3)$$

Stochastic Gradient MCMC

Theorem (The complete recipe for SG-MCMC dynamics ([12]))

For r.v. z , given a function (Hamiltonian) $H(z)$, a skew-symmetric matrix $Q(z)$ and a positive definite matrix $D(z)$, then the dynamics

$$dz = f(z)dt + \mathcal{N}(0, 2D(z)dt) \quad (4)$$

uniquely keeps $\pi(z) \propto \exp\{-H(z)\}$ invariant, where

$$\begin{aligned} f(z) &= -[D(z) + Q(z)] \nabla_z H(z) + \Gamma(z), \\ \Gamma_i(z) &= \sum_j \frac{\partial}{\partial z_j} (D_{ij}(z) + Q_{ij}(z)) \end{aligned} \quad (5)$$

Take $z = (q, p)$ and $H(z) = T(z) + U(q)$, then

$\int \exp\{-H(z)\} dp \propto \pi(q|\mathcal{D})$ provided that $\int \exp\{T(z)\} dp$ is independent of q .

Stochastic Gradient MCMC

The dynamics is compatible with stochastic gradient.

- $\nabla_z H(z) = \nabla_z T(z) + \nabla_q U(q),$
 $\nabla_z \tilde{H}(z) = \nabla_z T(z) + \nabla_q \tilde{U}(q) = \nabla_z H(z) + \mathcal{N}(0, V(q)),$
 $\tilde{f}(z) = f(z) + \mathcal{N}(0, B(z)).$

Then the dynamics can be expressed in another form:

$$\begin{aligned} dz &= f(z)dt + \mathcal{N}(0, 2D(z)dt) \\ &= \tilde{f}(z)dt + \mathcal{N}(0, 2D(z)dt - B(z)dt^2). \end{aligned} \quad (6)$$

- Estimation of B : empirical Fisher information ([2]), or just zero.

- 1 Introduction
- 2 Preliminaries
- 3 Stochastic Gradient Geodesic MCMC Methods**
 - Technical Description of the Settings
 - The Dynamics
 - Simulation with 2nd-order Geodesic Integrators
- 4 Application to Spherical Admixture Model
 - Description of SAM
 - Posterior Sampling for SAM
- 5 Experiments
 - Toy Experiment
 - Synthetic Experiment
 - Real-world Experiment: SAM

Technical Description of the Settings

Two ways to describe an m -dim Riemann manifold \mathcal{M} :

- Coordinate space ($\subset \mathbb{R}^m$)
 - (\mathcal{N}, Φ) : one of the local coordinate systems. Φ is a homomorphism.
 - $G(q)$: the Riemann metric tensor under (\mathcal{M}_1, Φ) .
 - Global one may not exist; easy constraint.
- Embedded space ($\subset \mathbb{R}^n$)
 - $\Xi : \mathcal{M} \rightarrow \mathbb{R}^n$ ($n \geq m$) injective, the embedding from \mathcal{M} in \mathbb{R}^n .
 - $\xi \triangleq \Xi \circ \Phi^{-1} : \mathbb{R}^m \rightarrow \mathbb{R}^n$ links the coordinate space and the embedded space.
 - Global description of \mathcal{M} and always exists; hard constraint.

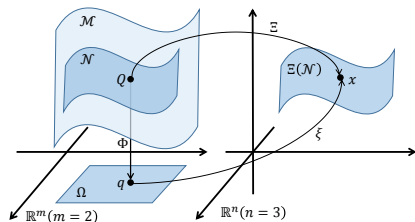


Figure: An illustration of a Riemann manifold \mathcal{M} and related concepts

Technical Description of the Settings

Define a distribution on \mathcal{M} :

- Coordinate space ($\subset \mathbb{R}^m$)
 $\pi(q)$ w.r.t. the Lebesgue measure $\lambda^m(dq)$ in \mathbb{R}^m .
- Embedded space ($\subset \mathbb{R}^n$)
 $\pi_{\mathcal{H}}(x)$ w.r.t. the Hausdorff measure $\mathcal{H}^m(dx)$ in $\Xi(\mathcal{M})$.
- $\pi_{\mathcal{H}}(\xi(q)) = \pi(q) / \sqrt{|G(q)|}$

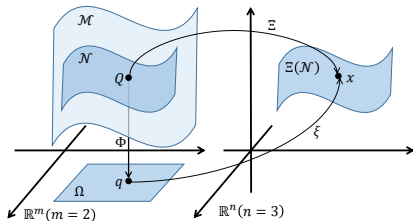


Figure: An illustration of a Riemann manifold \mathcal{M} and related concepts

We will derive the dynamics in the coordinate space for using the complete recipe, and simulate in the embedded space for releasing global coordinates requirement.

The Dynamics

Conceive the dynamics using the complete recipe.

- SGGMC

- Augment by $z = (q, p) \in \mathbb{R}^{2m}$.
- $H(z) = U(q) + \frac{1}{2} \log |G(q)| + \frac{1}{2} p^\top G(q)^{-1} p$, s.t. $\int \pi(z) dp \propto \pi(q)$, the target distribution.

$$D(z) = \begin{pmatrix} 0 & 0 \\ 0 & M(q)^\top C M(q) \end{pmatrix}, \quad Q(z) = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix},$$

where we define $M(q)_{n \times m} : M(q)_{ij} = \partial_{q_j} \xi_i(q)$, and C is some symmetric positive definite $n \times n$ matrix.

The Dynamics

- SGGMC

Then according to Eqn. (4, 5), the desired dynamics is

$$\left\{ \begin{array}{l} dq = G^{-1}p dt \\ dp = -\nabla_q U(q) dt - \frac{1}{2} \nabla_q \log |G(q)| dt \\ \quad - M^\top C M G^{-1} p dt - \frac{1}{2} \nabla_q [p^\top G^{-1} p] dt \\ \quad + \mathcal{N}(0, 2M^\top C M dt) \end{array} \right. \quad (7)$$

The Dynamics

Conceive the dynamics using the complete recipe.

- gSGNHT

- Augment by $z = (q, p, \xi) \in \mathbb{R}^{2m+1}$, where $\xi \in \mathbb{R}$ is the thermostats to adaptively balance the gradient noise [7].

- For $C > 0$, define

$H(z) = U(q) + \frac{1}{2} \log |G(q)| + \frac{1}{2} p^\top G(q)^{-1} p + \frac{m}{2} (\xi - C)^2$ s.t.
 $\int \pi(z) dp d\xi \propto \pi(q)$, the target distribution.

$$D(z) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & CG(q) & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad Q(z) = \begin{pmatrix} 0 & -I & 0 \\ I & 0 & p/m \\ 0 & -p^\top/m & 0 \end{pmatrix},$$

The Dynamics

- gSGNHT

Then according to Eqn. (4, 5), the desired dynamics is

$$\left\{ \begin{array}{l} dq = G^{-1}p dt \\ dp = -\nabla_q U dt - \frac{1}{2}\nabla_q \log |G| dt - \xi p dt \\ \quad - \frac{1}{2}\nabla_q [p^\top G^{-1}p] dt + \mathcal{N}(0, 2CG dt) \\ d\xi = \left(\frac{1}{m}p^\top G^{-1}p - 1\right) dt \end{array} \right. \quad (8)$$

- Comments

- The two dynamics are novel. They extend SGHMC [6] and SGNHT [7], respectively.
- They are suitable for better integrators to simulate.

Simulation with 2nd-order Geodesic Integrators

- Order of integrator: for a K th-order integrator, the bias of expected sample average at iteration L is $\mathcal{O}(L^{-K/(K+1)})$ and mean square error is $\mathcal{O}(L^{-2K/(2K+1)})$ [5].
- The typical Euler integrator is of 1st-order (used by SGRLD, SGRHMC).
- Symmetric Splitting Integrator (SSI) [5] is of 2nd-order. SSI: 1) split the dynamics into parts with each analytically solvable; 2) alternately simulate each exactly with the analytic solutions.

Simulation with 2nd-order Geodesic Integrators

Apply SSI to SGGMC.

(1) Split the dynamics (7):

$$A : \begin{cases} dq = G^{-1}p dt \\ dp = -\frac{1}{2}\nabla_q [p^\top G^{-1}p] dt \end{cases} \quad (9)$$

$$B : \begin{cases} dq = 0 \\ dp = -M^\top C M G^{-1}p dt \end{cases} \quad (10)$$

$$O : \begin{cases} dq = 0 \\ dp = -\nabla_q U(q) dt - \frac{1}{2}\nabla_q \log |G(q)| dt \\ \quad + \mathcal{N}(0, 2M^\top C M dt) \end{cases} \quad (11)$$

Simulation with 2nd-order Geodesic Integrators

(2) Solve each part analytically.

- Dynamics A :

Analytical solution: the geodesic flow ([4, 1]).

Example (Geodesic flow of hypersphere in the embedded space)

Geodesic flow (force-free motion) on the $(d - 1)$ -dim hypersphere $\mathbb{S}^{d-1} \triangleq \{x \in \mathbb{R}^d \mid \|x\| = 1\}$ is the rotation around the origin along the great circle:

$$\begin{cases} x(t) = x(0) \cos(\alpha t) + \frac{v(0)}{\alpha} \sin(\alpha t) \\ v(t) = -\alpha x(0) \sin(\alpha t) + v(0) \cos(\alpha t) \end{cases}, \quad (12)$$

where $x \in \mathbb{S}^{d-1}$, $v = \dot{x} \in T_x(\mathbb{S}^{d-1})$ a tangent vector, and $\alpha = \|v(0)\|$.

Simulation with 2nd-order Geodesic Integrators

(2) Solve each part analytically.

- Dynamics B and O : analytically solve in the embedded space by $x = \xi(q)$ and $v \triangleq \dot{x}$:

$$B : \begin{cases} x(t) = x(0) \\ v(t) = \Lambda(x(0)) \expm\{-Ct\}v(0) \end{cases}$$

$$O : \begin{cases} x(t) = x(0) \\ v(t) = v(0) + \Lambda(x(0))[-\nabla_x U_{\mathcal{H}}(x(0))t + \mathcal{N}(0, 2Ct)] \end{cases}$$

where $U_{\mathcal{H}}(x) \triangleq -\log \pi_{\mathcal{H}}(x)$, and $\Lambda(x)$ is the projection onto $T_x(\Xi(\mathcal{M}))$, the tangent space at x .

Example (The projection $\Lambda(x)$ for hypersphere in the embedded space)

$T_x(\Xi(\mathbb{S}^{d-1}))$ is a $(d-1)$ -dim plane, whose orthogonal complement is the line in the direction of x . So the projection onto the plane is

$$\Lambda(x) = I_d - xx^{\top}.$$

Simulation with 2nd-order Geodesic Integrators

(2) Solve each part analytically.

- Dynamics B and O : analytically solve in the embedded space by $x = \xi(q)$ and $v \triangleq \dot{x}$:

$$B : \begin{cases} x(t) = x(0) \\ v(t) = \Lambda(x(0)) \expm\{-Ct\} v(0) \end{cases}$$

$$O : \begin{cases} x(t) = x(0) \\ v(t) = v(0) + \Lambda(x(0)) [-\nabla_x U_{\mathcal{H}}(x(0))t + \mathcal{N}(0, 2Ct)] \end{cases}$$

To use stochastic gradient, only dynamics O is affected, which can be reformulated as

$$O : \begin{cases} x(t) = x(0) \\ v(t) = v(0) + \Lambda(x(0)) \cdot \left[-\nabla_x \tilde{U}_{\mathcal{H}}(x(0))t + \mathcal{N}\left(0, (2C - V(x(0)))t\right) \right] \end{cases},$$

where $V(x)$ can be estimated by empirical Fisher information, or just take as zero as discussed.

Simulation with 2nd-order Geodesic Integrators

- (3) Simulate the whole dynamics by alternatively simulate each sub-dynamics in closed form, in an “ABOBO” pattern.

Algorithm 1 Sampling procedure of SGGMC (for scalar C)

Randomly initialize $x^{(0)} \in \Xi(\mathcal{M})$.

Sample $v^* \sim \mathcal{N}(0, I)$ and project $v^{(0)} \leftarrow \Lambda(x^{(0)})v^*$.

for $n = 1, 2, \dots$, **do**

Sample a subset \mathcal{S} for computing $\nabla_x \tilde{U}_{\mathcal{H}}(x)$. $(x_0, v_0) \leftarrow (x^{(n-1)}, v^{(n-1)})$.

for $l = 1, 2, \dots, L$ **do**

A: Update $(x^*, v^*) \leftarrow (x_{l-1}, v_{l-1})$ by the geodesic flow for time step $\frac{\varepsilon_n}{2}$.

B: $v^* \leftarrow \exp\{-C\frac{\varepsilon_n}{2}\}v^*$.

O: $v^* \leftarrow v^* + \Lambda(x^*) \cdot \left[-\nabla_x \tilde{U}_{\mathcal{H}}(x^*)\varepsilon_n + \mathcal{N}(0, (2C - \varepsilon_n V(x^*))\varepsilon_n)\right]$.

B: $v^* \leftarrow \exp\{-C\frac{\varepsilon_n}{2}\}v^*$.

A: Update $(x_l, v_l) \leftarrow (x^*, v^*)$ by the geodesic flow for time step $\frac{\varepsilon_n}{2}$.

end for

$(x^{(n+1)}, v^{(n+1)}) \leftarrow (x_L, v_L)$. No M-H test.

end for

Simulation with 2nd-order Geodesic Integrators

Algorithm 2 Sampling procedure of gSGNHT (for scalar C)

Randomly initialize $x^{(0)} \in \Xi(\mathcal{M})$.

Sample $v^* \sim \mathcal{N}(0, I)$ and project $v^{(0)} \leftarrow \Lambda(x^{(0)})v^*$. $\xi^{(0)} \leftarrow C$.

for $n = 1, 2, \dots$, **do**

Sample a subset \mathcal{S} for $\nabla_x \tilde{U}_{\mathcal{H}}(x)$. $(x_0, v_0, \xi_0) \leftarrow (x^{(n-1)}, v^{(n-1)}, \xi^{(n-1)})$.

for $l = 1, 2, \dots, L$ **do**

A: Update $(x^*, v^*) \leftarrow (x_{l-1}, v_{l-1})$ by the geodesic flow for time step $\frac{\varepsilon_n}{2}$,

$$\xi^* \leftarrow \xi_{l-1} + \left(\frac{1}{m} v_{l-1}^\top v_{l-1} - 1\right) \frac{\varepsilon_n}{2}.$$

B: $v^* \leftarrow \exp\{-\xi^* \frac{\varepsilon_n}{2}\} v^*$.

O: $v^* \leftarrow v^* + \Lambda(x^*) \cdot \left[-\nabla_x \tilde{U}_{\mathcal{H}}(x^*) \varepsilon_n + \mathcal{N}(0, (2C - \varepsilon_n V(x^*)) \varepsilon_n)\right]$.

B: $v^* \leftarrow \exp\{-\xi^* \frac{\varepsilon_n}{2}\} v^*$.

A: Update $(x_l, v_l) \leftarrow (x^*, v^*)$ by the geodesic flow for time step $\frac{\varepsilon_n}{2}$,

$$\xi_l \leftarrow \xi^* + \left(\frac{1}{m} v^{*\top} v^* - 1\right) \frac{\varepsilon_n}{2}.$$

end for

$(x^{(n)}, v^{(n)}, \xi^{(n)}) \leftarrow (x_L, v_L, \xi_L)$. No M-H test.

end for

- 1 Introduction
- 2 Preliminaries
- 3 Stochastic Gradient Geodesic MCMC Methods
 - Technical Description of the Settings
 - The Dynamics
 - Simulation with 2nd-order Geodesic Integrators
- 4 Application to Spherical Admixture Model**
 - Description of SAM**
 - Posterior Sampling for SAM**
- 5 Experiments
 - Toy Experiment
 - Synthetic Experiment
 - Real-world Experiment: SAM

Description of SAM

- Spherical Admixture Model (SAM) ([14]) is a topic model for modeling spherical/directional data, such as tf-idf feature for text data.

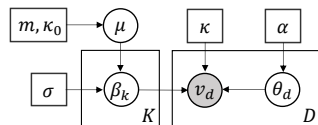


Figure: An illustration of SAM model structure.

- The generating process:
 - Draw $\mu \sim \text{vMF}(\mu|m, \kappa_0)$;
 - For $k = 1, \dots, K$, draw topic $\beta_k \sim \text{vMF}(\beta_k|\mu, \sigma)$;
 - For $d = 1, \dots, D$, draw $\theta_d \sim \text{Dir}(\theta_d|\alpha)$ and $v_d \sim \text{vMF}(v_d|\bar{v}(\beta, \theta_d), \kappa)$, where $\bar{v}(\beta, \theta_d) \triangleq \frac{\beta\theta_d}{\|\beta\theta_d\|}$ and $\text{vMF}(x|\mu, \kappa) = c_d(\kappa) \exp\{\kappa\mu^\top x\}$ is the p.d.f. (w.r.t. the Hausdorff measure) of the von Mises-Fisher distribution, a unimodal distribution on hyperspheres.
- $m, \mu, \beta_k \in \mathbb{S}^{V-1}$ where V is the vocabulary size, and θ_d lies on the simplex.
- μ can be collapsed so we know the exact form of $\pi(v, \beta, \theta)$, hence the joint posterior $\pi(\beta, \theta|v)$ up to a constant multiplier.

Posterior Sampling for SAM

- The task: sample from $\pi(\beta|v)$, which is entirely unknown.
- For SGGMC/gSGNHT, only $\nabla_{\beta}U(\beta) \triangleq -\nabla_{\beta} \log \pi(\beta|v)$ is needed, which can be reformed in the way provided by ([8]):

$$\begin{aligned} \nabla_{\beta} \log \pi(\beta|v) &= \frac{1}{\pi(\beta|v)} \nabla_{\beta} \int \pi(\beta, \theta|v) d\theta \\ &= \int \frac{\pi(\beta, \theta|v)}{\pi(\beta|v)} \frac{\nabla_{\beta} \pi(\beta, \theta|v)}{\pi(\beta, \theta|v)} d\theta = \mathbb{E}_{\pi(\theta|\beta, v)} [\nabla_{\beta} \log \pi(\beta, \theta|v)]. \end{aligned} \quad (13)$$

- The expectation can be estimated by samples $\{\theta^{(n)}\}_{n=1}^N$ from $\pi(\theta|\beta, v)$, which can be drawn using GMC with known $\pi(\theta|\beta, v)$ up to a constant multiplier.
- For the computationally cheaper stochastic gradient, we randomly select S documents from the total D and express the selected documents by indices $\{d(s)\}_{s=1}^S$. Finally,

$$\nabla_{\beta} \tilde{U}(\beta) \approx \nabla_{\beta} \log c_V(\|\bar{m}(\beta)\|) - \kappa \frac{D}{NS} \sum_{n=1}^N \sum_{s=1}^S v_{d(s)}^{\top} \bar{v}(\beta, \theta_{d(s)}^{(n)}). \quad (14)$$

Posterior Sampling for SAM

Algorithm 3 sampling inference for SAM using SGGMC/gSGNHT

for $m = 1, 2, \dots$ **do**

Sample a subset $\{d(s)\}_{s=1}^S$ from whole training data.

for $s = 1, 2, \dots, S$ **do**

Sample N times from $\pi(\theta_{d(s)} | \beta^{(m-1)}, v_{d(s)})$ using GMC to get $\{\theta_{d(s)}^{(n)}\}_{n=1}^N$.

end for

Sample once from $\pi(\beta | v)$ using SGGMC/gSGNHT to get $\beta^{(m)}$, with stochastic gradient computed by Eqn. (14).

end for

- 1 Introduction
- 2 Preliminaries
- 3 Stochastic Gradient Geodesic MCMC Methods
 - Technical Description of the Settings
 - The Dynamics
 - Simulation with 2nd-order Geodesic Integrators
- 4 Application to Spherical Admixture Model
 - Description of SAM
 - Posterior Sampling for SAM
- 5 Experiments**
 - Toy Experiment**
 - Synthetic Experiment**
 - Real-world Experiment: SAM**

Toy Experiment

- Check the correctness of SGGMC by a greenhouse experiment, where the stochastic gradient noise is a known Gaussian.

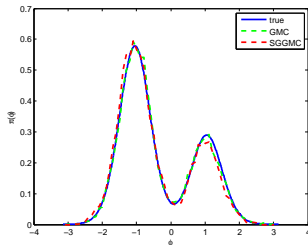
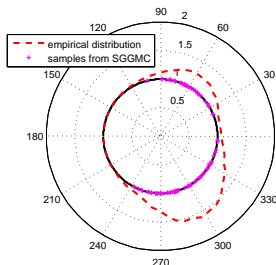


Figure: Toy experiment results: (left) samples and the empirical distribution of SGGMC in the embedded space; (right) comparison of true distribution and empirical distributions in the angle space.

- Settings:
sample on a circle embedded in \mathbb{R}^2 , with target distribution such that the potential energy is $U(x) = -\log(\exp\{\kappa\mu_1^\top x\} + 2\exp\{\kappa\mu_2^\top x\}) + \text{const.}$, where $\mu_1 = (\cos(\frac{\pi}{3}), \sin(\frac{\pi}{3}))$, $\mu_2 = (\cos(\frac{\pi}{3}), -\sin(\frac{\pi}{3}))$, and $\kappa = 5$.
A Gaussian noise $\mathcal{N}(0, 1000I)$ is injected into the exact gradient to produce a stochastic one.

Synthetic Experiment

- Settings: Bayesian posterior estimation for a mixture model of two vMF distributions with equal weights, with $e_1 = (1, 0)$ and $\mu \triangleq (v_1 + v_2) / \|v_1 + v_2\|$:

$$\pi(v_1) = \text{vMF}(v_1 | e_1, \kappa_1), \quad \pi(v_2) = \text{vMF}(v_2 | e_1, \kappa_2)$$

$$\pi(x_i | v_1, v_2) \propto \text{vMF}(x_i | v_1, \kappa_x) + \text{vMF}(x_i | \mu, \kappa_x)$$

- Data: generated by fixing $v_1 = v_1^{(g)}$ and $v_2 = v_2^{(g)}$, as shown in Figure.
- Estimate of the modes of the posterior: under weak prior, it is approximated by MLE for v_1 and v_2 . Estimate the MLE by matching the modes of x :

$$\begin{cases} v_1 = v_1^{(g)} \\ \mu = \mu^{(g)} \end{cases} \Rightarrow \begin{cases} v_1^{(1)} = v_1^{(g)} \\ v_2^{(1)} = v_2^{(g)} \end{cases}, \quad \begin{cases} v_1 = \mu^{(g)} \\ \mu = v_1^{(g)} \end{cases} \Rightarrow \begin{cases} v_1^{(2)} = \mu^{(g)} \\ v_2^{(2)} = v_2^{(g)} \end{cases}.$$

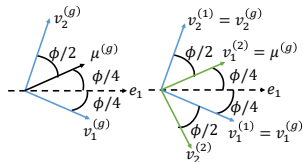


Figure: Left plot shows v_1, v_2, μ for generating data. Modes of the synthetic data are $v_1^{(g)}$ and $\mu^{(g)}$. Right plot shows the two modes of the posterior: $(v_1^{(1)}, v_2^{(1)})$ in blue and $(v_1^{(2)}, v_2^{(2)})$ in green.

Synthetic Experiment

- Results

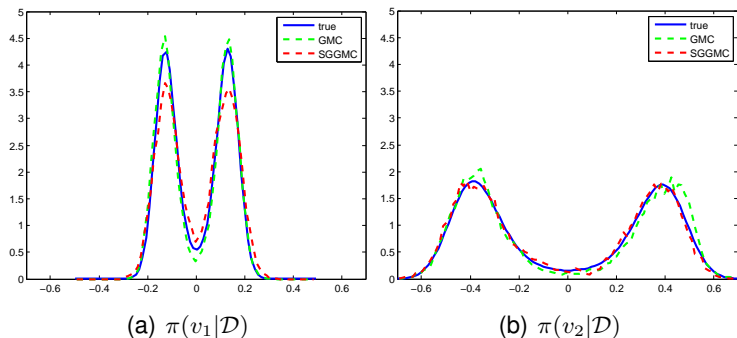


Figure: True marginal posteriors of v_1 (a) and v_2 (b) and their empirical distributions by samples from GMC and SGGMC.

Synthetic Experiment

- Results

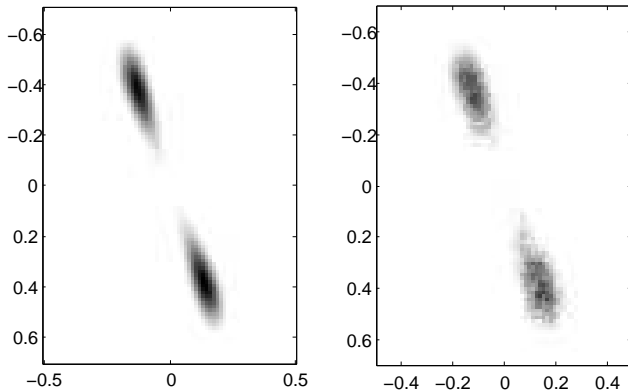


Figure: Joint posterior of v_1 and v_2 in gray scale. Left: true distribution; Right: empirical distribution by samples of SGGMC.

Real-world Experiment: SAM

- Our Methods
SGGMC and gSGNHT, with mini-batch and full-batch of data.
- Baselines
 - VI: variational inference by ([14]).
 - StoVI: stochastic variational inference, following ([10]).
 - GMC-bGibbs: blockwise Gibbs sampling that alternatively samples from $\pi(\beta|\theta, v)$ and $\pi(\theta|\beta, v)$ using GMC.
 - GMC-apprMH: samples β from $\pi(\beta|v)$ using GMC with potential energy estimated by $\{\theta^{(n)}\}$.

Real-world Experiment: SAM

- Evaluation Approach

The methods are compared for log-perplexity on a held-out test data set $\mathcal{D}_{\text{test}}$.

- For variational methods, evaluate the point estimate $\hat{\beta}$ by
$$\text{logperp} = -\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{d \in \mathcal{D}_{\text{test}}} \log \pi(v_d | \hat{\beta}).$$
- For sampling methods, evaluate the set of samples $\{\beta^{(m)}\}_{m=1}^M$ by
$$\text{logperp} = -\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{d \in \mathcal{D}_{\text{test}}} \log\left(\frac{1}{M} \sum_{m=1}^M \pi(v_d | \beta^{(m)})\right).$$
- To estimate $\pi(v_d | \beta)$,
$$\pi(v_d | \beta) = \int \pi(v_d, \theta_d | \beta) d\theta_d = \mathbb{E}_{\pi(\theta_d | \beta)}[\pi(v_d | \beta, \theta_d)],$$
 where $\pi(v_d | \beta, \theta_d)$ is known and samples from $\pi(\theta_d | \beta) = \pi(\theta_d) = \text{Dir}(\alpha)$ are easy to draw.

Real-world Experiment: SAM

- Results on the 20news-different dataset (1,666 training and 1,107 test, 20 topics)

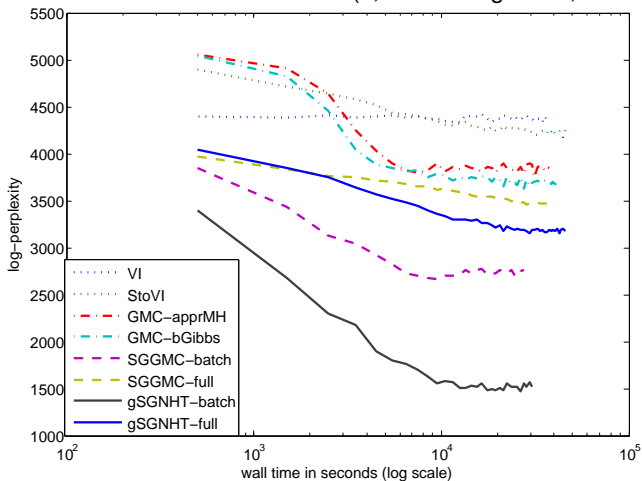


Figure: The evolution of log-perplexity of all the inference methods along wall time on the 20news-different dataset.

Real-world Experiment: SAM

- Results on the 150K Wikipedia subset (150K training and 1K test, 50 topics)

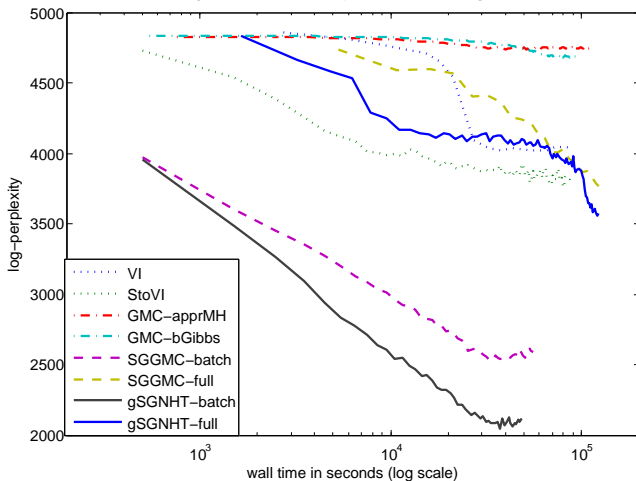


Figure: The evolution of log-perplexity of all the inference methods along wall time on the 150K Wikipedia subset.

Thanks!

- [1] Ralph Abraham, Jerrold E Marsden, and Jerrold E Marsden. *Foundations of mechanics*. Benjamin/Cummings Publishing Company Reading, Massachusetts, 1978.
- [2] Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. *arXiv preprint arXiv:1206.6380*, 2012.
- [3] Marcus A. Brubaker, Mathieu Salzmann, and Raquel Urtasun. A family of mcmc methods on implicitly defined manifolds. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 161–172, 2012.
- [4] Simon Byrne and Mark Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.
- [5] Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, pages 2269–2277, 2015.
- [6] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1683–1691, 2014.
- [7] Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D. Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems*, pages 3203–3211, 2014.

- [8] Chao Du, Jun Zhu, and Bo Zhang. Learning deep generative models with doubly stochastic mcmc. *arXiv preprint arXiv:1506.04557*, 2015.
- [9] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.
- [10] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.
- [11] Chunyuan Li, Changyou Chen, Kai Fan, and Lawrence Carin. High-order stochastic gradient thermostats for bayesian learning of deep models. *arXiv preprint arXiv:1512.07662*, 2015.
- [12] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. In *Advances in Neural Information Processing Systems*, pages 2899–2907, 2015.
- [13] Sam Patterson and Yee Whye Teh. Stochastic gradient riemannian langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, pages 3102–3110, 2013.
- [14] Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J. Mooney. Spherical topic models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 903–910, 2010.

- [15] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.