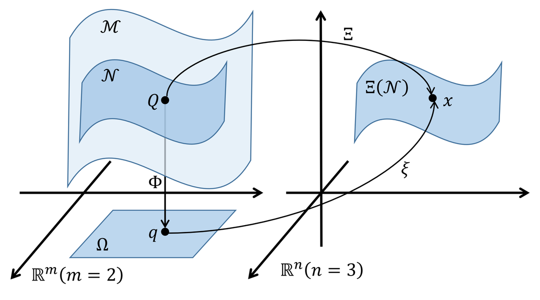


Stochastic Gradient Geodesic MCMC Methods

Chang Liu, Jun Zhu, Yang Song
{chang-li14@mails., dcszj@}tsinghua.edu.cn; songyang@stanford.edu

Introduction

Task

Scalable Bayesian inference for latent variables on *Riemann manifolds* by Monte Carlo.

Drawbacks of Existing Methods

- Unscalable: drawing one sample needs traversing the whole dataset.
- Inner iteration: iteration within one dynamics simulation step.
- Global coordinates requirement: limited applicability (fail for e.g. hypersphere).
- Lower order integrator.

Our Solution

Stochastic Gradient Geodesic Monte Carlo; geodesic SG Nosé-Hoover Thermostats.

Table: A summary of related methods (–: not for manifold r.v.; †: not SSI; ‡: 2nd-order versions appear afterwards)

methods	stochastic gradient	no inner iteration	no global coordinates	order of integrator
GMC	×	✓	✓	2nd
RMLD	×	✓	×	1st
RMHMC	×	×	×	2nd†
CHMC	×	×	✓	2nd†
SGLD	✓	✓	–	1st
SGHMC	✓	✓	–	1st‡
SGNHT	✓	✓	–	1st‡
SGRLD	✓	✓	×	1st
SGRHMC	✓	✓	×	1st
SGGMC	✓	✓	✓	2nd
gSGNHT	✓	✓	✓	2nd

Preliminaries

SG-MCMC

To sample from posterior $\pi(q|\mathcal{D})$, estimate the required *gradient* $\nabla U(q) \triangleq -\nabla \log \pi(q|\mathcal{D})$

$$= -\nabla \log \pi_0(q) - \sum_{d=1}^{D} \nabla \log \pi(x_d|q)$$

by *stochastic gradient* with random subset $\mathcal{S}$: $\nabla \tilde{U}(q) \triangleq -\nabla \log \pi_0(q) - (D/|\mathcal{S}|) \sum_{x \in \mathcal{S}} \nabla \log \pi(x|q)$.

- A complete recipe (Ma et al., 2015) for the dynamics of SG-MCMC:

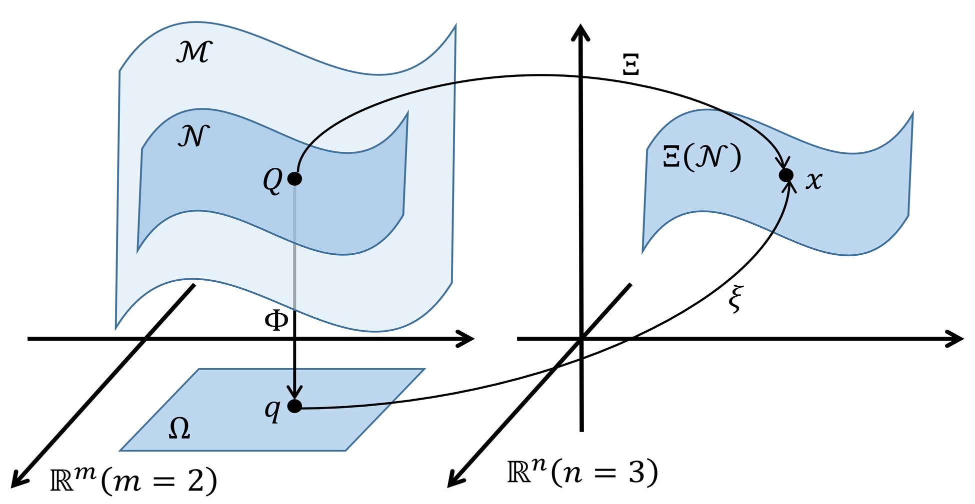
$$\mathrm{d}z = f(z)\mathrm{d}t + \mathcal{N}(0, 2D(z)\mathrm{d}t),$$

with f, D satisfying certain conditions.

Ability to handle SG noise $\tilde{f}(z) = f(z) + \mathcal{N}(0, B(z))$:

$$\mathrm{d}z = \tilde{f}(z)\mathrm{d}t + \mathcal{N}\big(0, 2D(z)\mathrm{d}t - B(z)\mathrm{d}t^2\big).$$

Riemann Manifold $\mathcal{M}$



$(\mathcal{N}, \Phi)$: local coordinate system.
Ξ: (isometric) embedding. $\xi \triangleq \Xi \circ \Phi^{-1}$.
$G(q)$: Riemann metric tensor.
Distribution on $\mathcal{M}$: $\pi_{\mathcal{H}}(x)|_{x=\xi(q)} = \pi(q)/\sqrt{|G(q)|}$,
$\pi(q)$: in the coordinate space;
$\pi_{\mathcal{H}}(x)$: in the embedded space.

Dynamics Construction

Design novel dynamics in coordinate space by the complete recipe, so that:

- the stationary distribution is desired;
- suitable for 2nd-order integrators.

SGGMC

Augment with momentum $p \in \mathbb{R}^m$: $z = (q, p)$.

$$\begin{cases} \mathrm{d}q = G^{-1}p\mathrm{d}t \\ \mathrm{d}p = -\nabla U \mathrm{d}t - (1/2)\nabla \log|G|\mathrm{d}t - M^\top C M G^{-1} p\,\mathrm{d}t \\ \quad - (1/2)\nabla\big[p^\top G^{-1} p\big]\mathrm{d}t + \mathcal{N}(0, 2M^\top C M \mathrm{d}t) \end{cases}$$

$M(q)_{ij} \triangleq \partial \xi_i(q)/\partial q_j$; choose $C_{n \times n}$ pos. def.

gSGNHT

$z = (q, p, \xi)$. Thermostats $\xi \in \mathbb{R}$: adaptive C.

2nd-order Integrators

Simulate in the Embedded Space

- to release global coordinates requirement;
- $(q, p) \to (x, v)$. (v also momentum)

Symmetric Splitting Integrator

- Guaranteed to be 2nd-order (Chen et al., 2015).
- Split the dynamics into parts and solve each in closed form:

A: $\mathrm{d}q = G^{-1}p\mathrm{d}t, \mathrm{d}p = -(1/2)\nabla[p^\top G^{-1}p]\mathrm{d}t$

B: $\mathrm{d}p = -M^\top C M G^{-1} p \mathrm{d}t$,

O: $\mathrm{d}p = -\nabla U(q)\mathrm{d}t - (1/2)\nabla \log|G|\mathrm{d}t + \mathcal{N}(0, 2M^\top C M \mathrm{d}t)$.

- Simulate the whole dynamics: "ABOBA".

A,B: $\varepsilon/2$; O: ε; use the closed-form solutions.

- A: geodesic flow (motion with no force), e.g. *great circles* on hypersphere $\mathbb{S}^{d-1} \triangleq \{x \in \mathbb{R}^d | \|x\| = 1\}$ ($\alpha \triangleq \|v(0)\|$):

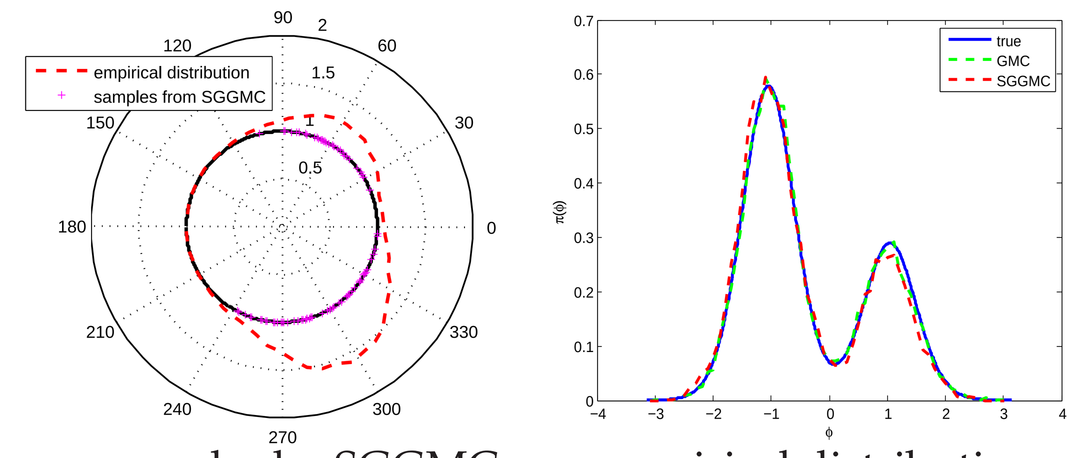
$$\begin{cases} x(t) = x(0)\cos(\alpha t) + \big(v(0)/\alpha\big)\sin(\alpha t) \\ v(t) = -\alpha x(0)\sin(\alpha t) + v(0)\cos(\alpha t) \end{cases}.$$

- B: $v(t) = \mathrm{expm}\{-\Lambda(x(0))Ct\}v(0)$. For $C \in \mathbb{R}^+$, $v(t) = e^{-Ct}v(0) \approx (1 - Ct)v(0)$: SGHMC friction. (Not for p)
- O with stochastic gradient: $v(t) = v(0) + \Lambda(x)\Big[-\nabla \tilde{U}_{\mathcal{H}}(x)t + \mathcal{N}\Big(0, 2Ct - V(x)t^2\Big)\Big]_{x=x(0)}$,

$\Lambda(x)$: projection onto $T_x\Xi(\mathcal{M})$.
For $\mathbb{S}^{d-1}$, $\Lambda(x) = I_d - xx^\top$.

Synthetic Experiments

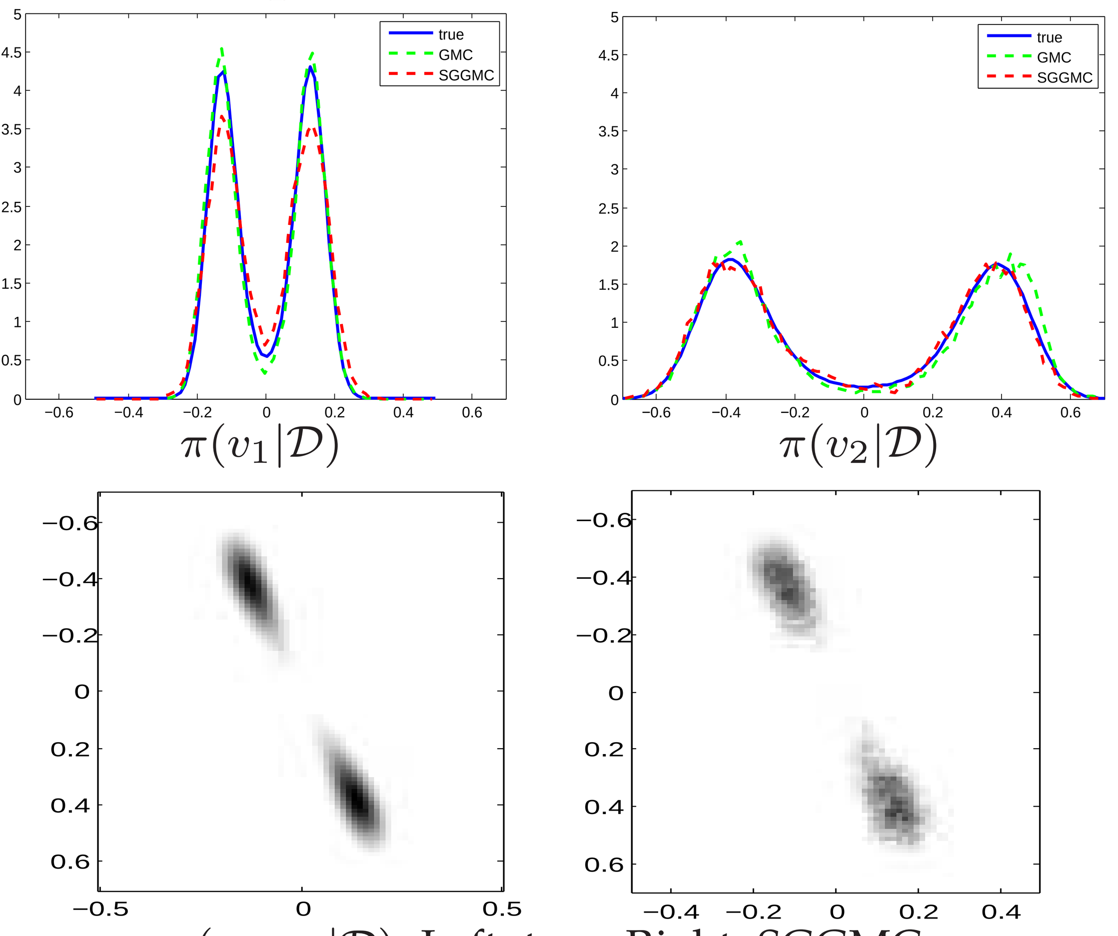
Toy Experiment



samples by SGGMC — empirical distribution

Sample $x \in \mathbb{S}^1$, $U(x) = -\log\big(e^{5\mu_1^\top x} + 2e^{5\mu_2^\top x}\big)$.
Known gradient noise: artificially added.

Synthetic Experiment



$\pi(v_1|\mathcal{D})$ — $\pi(v_2|\mathcal{D})$

$\pi(v_1, v_2|\mathcal{D})$. Left: true; Right: SGGMC

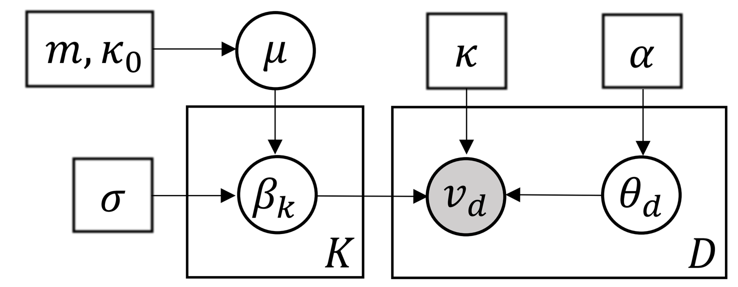
Inference for mixture of vMF:

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

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

with $\mu \triangleq \frac{(v_1+v_2)}{\|v_1+v_2\|}$. Synthetic data: drawn by GMC.

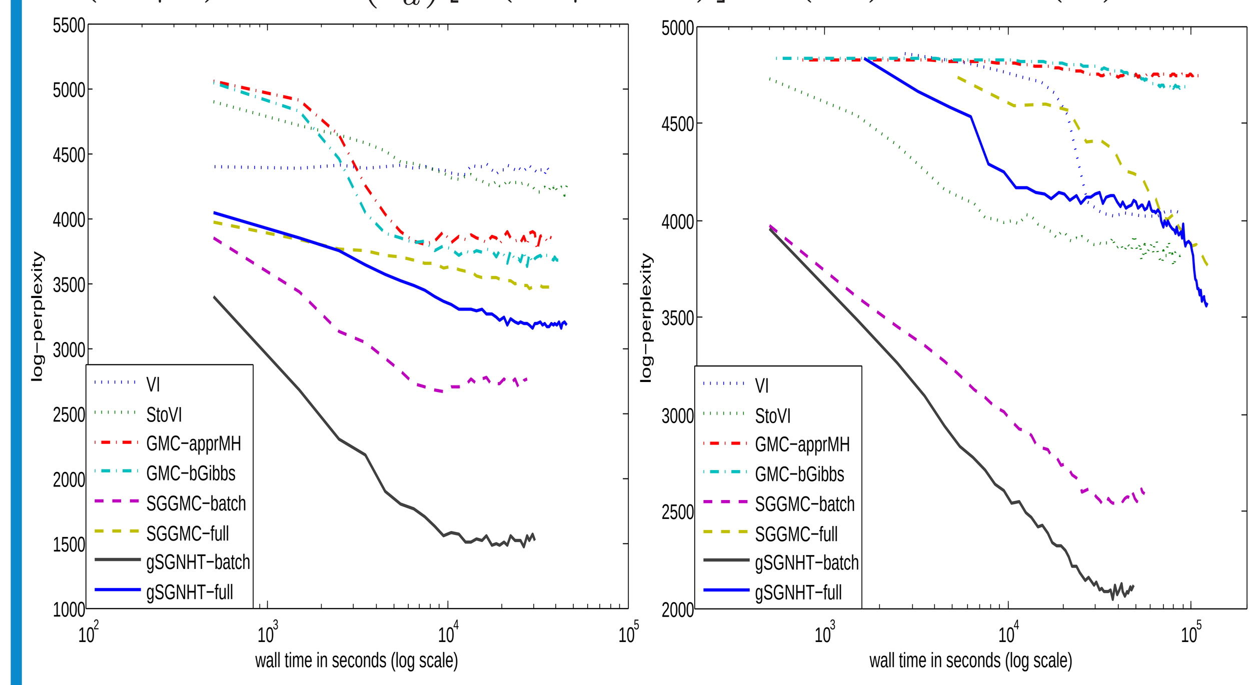
Real-world Experiments



Spherical Admixture Model (Reisinger et al., 2014): a topic model for *spherical* data $v_d \in \mathbb{S}^{V-1}$. $\beta, \mu \in \mathbb{S}^{V-1}$, $\theta \in \Delta^{K-1}$.

- Inference by SGGMC/gSGNHT: directly sample from the posterior $\pi(\beta|v)$.
  - μ can be collapsed analytically.
  - Monte-Carlo and mini-batch estimated (doubly stochastic) gradient: $-\nabla_\beta \log \pi(\beta|v)$ $= -\mathbb{E}_{\underbrace{\pi(\theta|\beta, v)}_{\text{drawn by GMC}}} \Big[\underbrace{\nabla_\beta \log \pi(v, \beta, \theta)}_{\text{closed-form known; estimated by mini-batch}} \Big]$
- Evaluation:

log-perplexity (the lower the better) along time: $-(1/|\mathcal{D}'|)\sum_{d \in \mathcal{D}'} \log\big[(1/M)\sum_{m=1}^{M} \pi(v_d|\beta^{(m)})\big]$, $\pi(v_d|\beta) = \mathbb{E}_{\pi(\theta_d)}[\pi(v_d|\beta, \theta_d)]$, $\pi(\theta_d) = \mathrm{Dir}(\alpha)$.



small dataset (1,666) — large dataset (150,000)

- Observations
  - SGGMC/gSGNHT: most accurate and fast; more salient on the larger dataset.
  - GMC-apprMH/GMC-bGibbs: more accurate than VI/StoVI on the small dataset but too slow on the large one.
  - VI/StoVI are blocked.

← Main paper

Appendix, codes, data →