Understanding MCMC Dynamics as Flows on the Wasserstein Space

Chang Liu, Jingwei Zhuo, Jun Zhu[1]

Department of Computer Science and Technology, Tsinghua University

chang-li14@mails.tsinghua.edu.cn

ICML 2019

[1]Corresponding author.

Introduction

What is known about dynamics-based MCMC:

- There are many instances, e.g. Langevin dynamics (LD) [30], Hamiltonian Monte Carlo (HMC) [12], stochastic gradient HMC (SGHMC) [8], etc.
- LD is recognized as the gradient flow of the KL divergence on the Wasserstein space [18].
    - Then its asymptotic [30] and non-asymptotic [13, 9] behaviors are clear.
    - Then its relation to existing particle-based variational inference methods (ParVIs) is clear [7, 21].

What remains unknown:

- Whether a general MCMC dynamics can be explained as an interpretable flow.

Introduction

Motivation

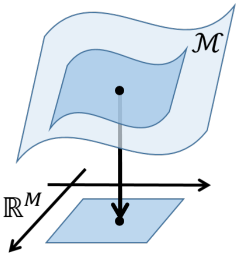
Explain a general MCMC dynamics as an interpretable flow.

- Then the behavior of general MCMC dynamics is clear.
- Then more MCMC dynamics than LD are connected to the ParVI family: ParVIs with more efficient MCMC dynamics, and MCMCs with more effective ParVI simulation.

Manifolds

Manifold $\mathcal{M}$:

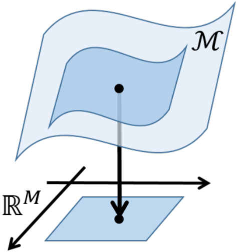
- Locally homeomorphic to an open subset of $\mathbb{R}^M$.
- We consider manifolds globally homeomorphic to $\mathbb{R}^M$ (global coordinate space).



Manifolds

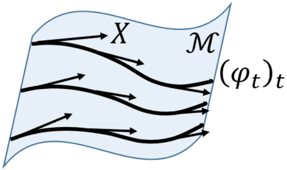
Manifold $\mathcal{M}$:

- Locally homeomorphic to an open subset of $\mathbb{R}^M$.
- We consider manifolds globally homeomorphic to $\mathbb{R}^M$ (global coordinate space).



Flows on $\mathcal{M}$:

- The set of curves $\{(\varphi_t)_t\}$ s.t. $\frac{\mathrm{d}\varphi_t}{\mathrm{d}t} = X(\varphi_t)$ given a vector field X.
- We use "vector fields" and "flows" interchangeably.



Manifolds

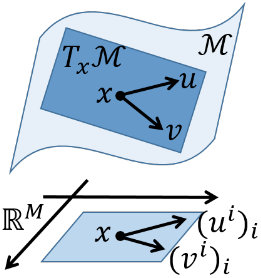
Riemannian structure on $\mathcal{M}$:

- An inner product in every tangent space $T_x\mathcal{M}$.
- Coordinate expression:

$$\langle u, v \rangle_{T_x\mathcal{M}} = g_{ij}(x) u^i v^j.$$

- Gradient: $\langle \operatorname{grad} f(x), v \rangle_{T_x\mathcal{M}} = v[f] := v^i \partial_i f(x)$,
  $\iff$ steepest ascending direction:
  $\operatorname{grad} f(x) = \max \cdot \operatorname{argmax}_{\|v\|_{T_x\mathcal{M}}=1} \frac{\mathrm{d}}{\mathrm{d}t} f(\varphi_t)$.
  Coordinate expression:

$$\big(\operatorname{grad} f(x)\big)^i = g^{ij}(x) \partial_j f(x).$$



Manifolds

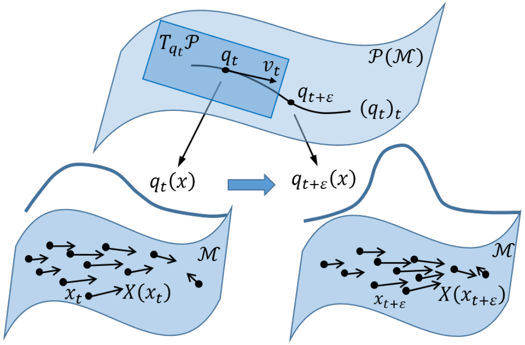
Wasserstein space $\mathcal{P}(\mathcal{M})$:

- Space of distributions on $\mathcal{M}$ (finite variance).
- Tangent vector $v \iff$ vector field X on $\mathcal{M}$.
- Tangent space at $q \in \mathcal{P}(\mathcal{M})$

$$T_q\mathcal{P}(\mathcal{M}) = \overline{\{\mathrm{grad}\, f \mid f \in \mathcal{C}_c^\infty(\mathcal{M})\}}^{\mathcal{L}_q^2(\mathcal{M})}$$

is a subspace of the Hilbert space

$$\mathcal{L}_q^2(\mathcal{M}) := \{X \mid \mathbb{E}_{q(x)}[\langle X(x), X(x)\rangle_{T_x\mathcal{M}}] < \infty\}.$$



Manifolds

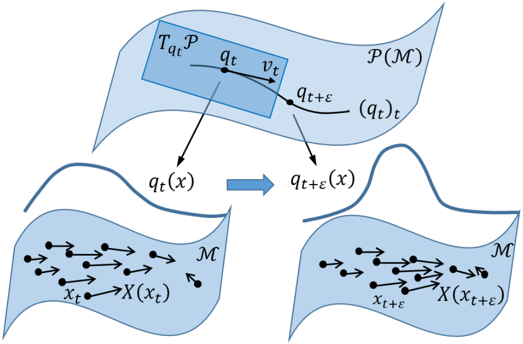
Wasserstein space $\mathcal{P}(\mathcal{M})$:

- Space of distributions on $\mathcal{M}$ (finite variance).
- Tangent vector $v \iff$ vector field X on $\mathcal{M}$.
- Tangent space at $q \in \mathcal{P}(\mathcal{M})$

  $$T_q\mathcal{P}(\mathcal{M}) = \overline{\{\operatorname{grad} f \mid f \in \mathcal{C}_c^\infty(\mathcal{M})\}}^{\mathcal{L}_q^2(\mathcal{M})}$$

  is a subspace of the Hilbert space

  $$\mathcal{L}_q^2(\mathcal{M}) := \{X \mid \mathbb{E}_{q(x)}[\langle X(x), X(x)\rangle_{T_x\mathcal{M}}] < \infty\}.$$



- Riemannian structure:
  $T_q\mathcal{P}$ inherits the inner product of $\mathcal{L}_q^2$.
  - Gradient of $\mathrm{KL}_p(q) := \int_\mathcal{M} \log(q/p)\,\mathrm{d}q$:

$$\operatorname{grad}\mathrm{KL}_p(q) = \operatorname{grad}\log(q/p),\ \big(\operatorname{grad}\mathrm{KL}_p(q)\big)^i(x) = g^{ij}(x)\,\partial_j \log(q(x)/p(x)).$$

LD as Gradient Flow

Equivalent dynamics:

- They produce the same distribution evolution rule.
- X and $\pi_q(X)$ are equivalent, where $\pi_q : \mathcal{L}_q^2 \to T_q\mathcal{P}$ is the orthogonal projection.

LD as Gradient Flow:

- The Langevin dynamics
$$\mathrm{d}x = \nabla \log p(x)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t(x)$$
is equivalent [7] to the deterministic dynamics:
$$\mathrm{d}x = \nabla \log(p(x)/q_t(x))\,\mathrm{d}t,$$
where q_t is the distribution of x at time t.

  It is the gradient flow of KL_p on $\mathcal{P}(\mathcal{M})$ for Euclidean $\mathcal{M}$!
- The gradient flow interpretation of LD is known earlier from another perspective [18].

Describe General MCMC Dynamics

The complete recipe [24] *(known knowledge)*:

- A general MCMC dynamics on $\mathbb{R}^M$ targeting p can be expressed as the diffusion process:

$$\begin{aligned} \mathrm{d}x &= V(x)\,\mathrm{d}t + \sqrt{2D(x)}\,\mathrm{d}B_t(x), \\ V^i(x) &= \frac{1}{p(x)}\partial_j\Big(p(x)\big(D^{ij}(x) + Q^{ij}(x)\big)\Big), \end{aligned} \tag{1}$$

  for some positive semi-definite matrix $D_{M\times M}$ (diffusion matrix) and some skew-symmetric matrix $Q_{M\times M}$ (curl matrix).

The First Reformulation

Lemma 1 (Equivalent deterministic MCMC dynamics)

MCMC dynamics Eq. (1) with symmetric D is equivalent to the deterministic dynamics:

$$\begin{gathered} \mathrm{d}x = W_t(x)\mathrm{d}t, \\ (W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x), \end{gathered} \tag{2}$$

where q_t is the distribution density of x at time t.

The First Reformulation

Lemma 1 (Equivalent deterministic MCMC dynamics)

MCMC dynamics Eq. (1) with symmetric D is equivalent to the deterministic dynamics:

$$\mathrm{d}x = W_t(x)\mathrm{d}t,$$
$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x), \tag{2}$$

where q_t is the distribution density of x at time t.

- $\Longrightarrow$ Barbour's generator [2] $\mathcal{A}f := \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{q_t}[f]\big|_{q_t=\delta_x} = \frac{1}{p}\partial_j\left[p\left(D^{ij} + Q^{ij}\right)(\partial_i f)\right]$ (c.f. [17]).

The First Reformulation

Lemma 1 (Equivalent deterministic MCMC dynamics)

MCMC dynamics Eq. (1) with symmetric D is equivalent to the deterministic dynamics:

$$\mathrm{d}x = W_t(x)\mathrm{d}t,$$
$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x), \tag{2}$$

where q_t is the distribution density of x at time t.

- $\Longrightarrow$ Barbour's generator [2] $\mathcal{A}f := \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{q_t}[f]\big|_{q_t=\delta_x} = \frac{1}{p}\partial_j\left[p\left(D^{ij}+Q^{ij}\right)(\partial_i f)\right]$ (c.f. [17]).

How to interpret $W_t(x)$?

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}(\mathcal{M})$.

- Euclidean $\mathcal{M}$ only allows $D = I$.

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}(\mathcal{M})$.

- Euclidean $\mathcal{M}$ only allows $D = I$.
- Hilbert $\mathcal{M}$ only allows constant and non-singular D.

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}(\mathcal{M})$.

- Euclidean $\mathcal{M}$ only allows $D = I$.
- Hilbert $\mathcal{M}$ only allows constant and non-singular D.
- Riemannian $\mathcal{M}$ allows position-dependent $D(x)$, but $D(x)$ needs to be non-singular.

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}(\mathcal{M})$.

- Euclidean $\mathcal{M}$ only allows $D = I$.
- Hilbert $\mathcal{M}$ only allows constant and non-singular D.
- Riemannian $\mathcal{M}$ allows position-dependent $D(x)$, but $D(x)$ needs to be non-singular.
- What kind of $\mathcal{M}$ allows position-dependent and positive *semi*-definite $D(x)$?

Interpret MCMC Dynamics

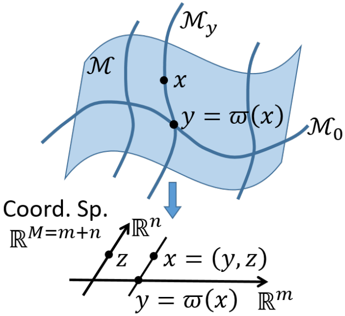
$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}(\mathcal{M})$.

- Fiber Bundle $\mathcal{M}$ (of dim. $M = m + n$) *(known knowledge)*:
  - $\mathcal{M}$ is locally $\mathcal{M}_0 \times \mathcal{F}$ ($\mathcal{M}_0$ of dim. m, $\mathcal{F}$ of dim. n) [27] in terms of a projection ϖ:

$$\varpi : \mathcal{M} \to \mathcal{M}_0 \overset{\text{locally}}{\Longleftrightarrow} \mathcal{M}_0 \times \mathcal{F} \to \mathcal{M}_0.$$

  - The *fiber* through $y \in \mathcal{M}_0$: $\mathcal{M}_y := \varpi^{-1}(y)$ (diffeom. to $\mathcal{F}$).
  - Coordinate decomposition: $x = (y, z)$, $y \in \mathbb{R}^m$: coord. of $\mathcal{M}_0$; $z \in \mathbb{R}^n$: coord. of $\mathcal{M}_y$.



Interpret MCMC Dynamics

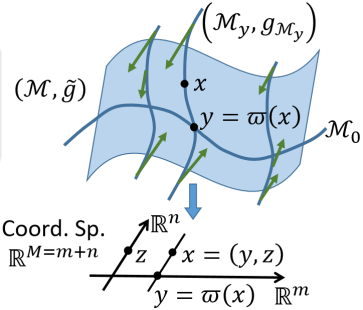
$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}(\mathcal{M})$.

- Fiber-Riemannian manifold $\mathcal{M}$:

Definition 3 (Fiber-Riemannian manifold)

$\mathcal{M}$ is a *fiber-Riemannian manifold* if it is a fiber bundle and there is a Riemannian structure $g_{\mathcal{M}_y}$ on each *fiber* $\mathcal{M}_y$.



Interpret MCMC Dynamics

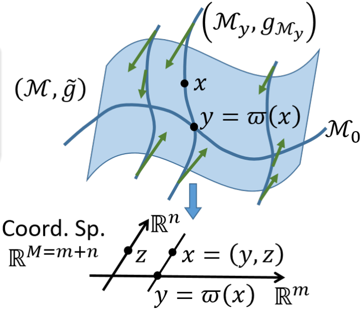
$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}(\mathcal{M})$.

- Fiber-Riemannian manifold $\mathcal{M}$:

Definition 3 (Fiber-Riemannian manifold)

$\mathcal{M}$ is a *fiber-Riemannian manifold* if it is a fiber bundle and there is a Riemannian structure $g_{\mathcal{M}_y}$ on each *fiber* $\mathcal{M}_y$.



- Gradient on fiber $\mathcal{M}_y$:

$$\left(\mathrm{grad}_{\mathcal{M}_y} f(y,z)\right)^a = (g_{\mathcal{M}_y}(z))^{ab}\,\partial_{z^b} f(y,z),$$
$$1 \le a, b \le n.$$

- Define *fiber-gradient* on $\mathcal{M}$ by taking union over y:

$$\left(\mathrm{grad}_{\mathrm{fib}} f(x)\right)_M := \left(0_m,\ \left(\mathrm{grad}_{\mathcal{M}_{\varpi(x)}} f(\varpi(x), z)\right)_n\right).$$

Interpret MCMC Dynamics

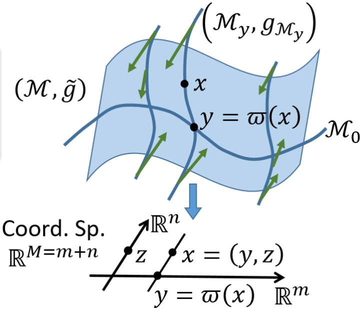
$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}(\mathcal{M})$.

- Fiber-Riemannian manifold $\mathcal{M}$:

Definition 3 (Fiber-Riemannian manifold)

$\mathcal{M}$ is a *fiber-Riemannian manifold* if it is a fiber bundle and there is a Riemannian structure $g_{\mathcal{M}_y}$ on each *fiber* $\mathcal{M}_y$.



- Alternatively, the fiber-gradient on $\mathcal{M}$ is:

$$\left(\mathrm{grad}_{\text{fib}}\, f(x)\right)^i = \tilde{g}^{ij}(x)\,\partial_j f(x), \quad 1 \le i, j \le M,$$

$$\left(\tilde{g}^{ij}(x)\right)_{M\times M} := \begin{pmatrix} 0_{m\times m} & 0_{m\times n} \\ 0_{n\times m} & \left((g_{\mathcal{M}_{\varpi(x)}}(z))^{ab}\right)_{n\times n} \end{pmatrix}. \quad (3)$$

We use $\tilde{g}$ to denote the fiber-Riemannian structure.

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}(\mathcal{M})$.

- Structures on $\mathcal{P}(\mathcal{M})$ with fiber-Riemannian $\mathcal{M}$.
  - Hard to decompose $\mathcal{P}(\mathcal{M})$.
  - Consider $\widetilde{\mathcal{P}}(\mathcal{M}) := \{q(z|y) \in \mathcal{P}(\mathcal{M}_y) \mid y \in \mathcal{M}_0\} \overset{\text{locally}}{\Longleftrightarrow} \mathcal{M}_0 \times \mathcal{P}(\mathcal{M}_y)$: fiber-Riemannian!
  - On $\mathcal{P}(\mathcal{M}_y)$,
$$\left(\operatorname{grad} \mathrm{KL}_{p(\cdot|y)}(q(\cdot|y))(z)\right)^a = (g_{\mathcal{M}_y}(z))^{ab}\,\partial_{z^b} \log \frac{q(z|y)}{p(z|y)}$$
$$= (g_{\mathcal{M}_y}(z))^{ab}\,\partial_{z^b} \log \frac{q(y,z)}{p(y,z)}, 1 \le a, b \le n.$$
  - Taking union over $y \in \mathcal{M}_0$, the fiber-gradient on $\widetilde{\mathcal{P}}(\mathcal{M})$ is:
$$\left(\operatorname{grad}_{\text{fib}} \mathrm{KL}_p(q)(x)\right)_M = \left(0_m, \left((g_{\mathcal{M}_{\varpi(x)}}(z))^{ab}\,\partial_{z^b} \log\left(q(x)/p(x)\right)\right)_n\right)$$
$$= \left(\tilde{g}^{ij}(x)\,\partial_j \log\left(q(x)/p(x)\right)\right)_M.$$

Project to make a tangent vector on $\mathcal{P}(\mathcal{M})$: $\pi_q(\operatorname{grad}_{\text{fib}} \mathrm{KL}_p(q)) \in T_q\mathcal{P}(\mathcal{M})$.

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}(\mathcal{M})$.

- $\left(\mathrm{grad}_{\mathrm{fib}}\,\mathrm{KL}_p(q)(x)\right)^i = \tilde{g}^{ij}(x)\,\partial_j \log\left(q(x)/p(x)\right)$, $(\tilde{g}^{ij}) = \begin{pmatrix} 0_{m\times m} & 0_{m\times n} \\ 0_{n\times m} & ({g_{\mathcal{M}_y}}^{ij})_{n\times n} \end{pmatrix}$.

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}(\mathcal{M})$.

- $\left(\mathrm{grad}_{\mathrm{fib}}\,\mathrm{KL}_p(q)(x)\right)^i = \tilde{g}^{ij}(x)\,\partial_j \log\left(q(x)/p(x)\right)$, $(\tilde{g}^{ij}) = \begin{pmatrix} 0_{m\times m} & 0_{m\times n} \\ 0_{n\times m} & ({g_{\mathcal{M}_y}}^{ij})_{n\times n} \end{pmatrix}$.

Assumption 4 (Regular MCMC dynamics (1/2))

(a) $D = C$ *or* $D = 0$ *or* $D = \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$, *for a symmetric positive definite* $C(x)$.

(b) $\ldots$

- Satisfied by existing MCMC instances.
- Could be relaxed by coordinate transformation.

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}(\mathcal{M})$.

- $\big(\mathrm{grad}_{\mathrm{fib}}\,\mathrm{KL}_p(q)(x)\big)^i = \tilde{g}^{ij}(x)\,\partial_j \log\big(q(x)/p(x)\big)$, $(\tilde{g}^{ij}) = \begin{pmatrix} 0_{m\times m} & 0_{m\times n} \\ 0_{n\times m} & ({g_{\mathcal{M}_y}}^{ij})_{n\times n} \end{pmatrix}$.

Assumption 4 (Regular MCMC dynamics (1/2))

(a) $D = C$ *or* $D = 0$ *or* $D = \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$, *for a symmetric positive definite* $C(x)$.

(b) $\ldots$

- Satisfied by existing MCMC instances.
- Could be relaxed by coordinate transformation.

- $D^{ij}\,\partial_j \log(p/q_t)$ is the fiber-gradient with fiber-Riemannian $(\mathcal{M}, \tilde{g})$ where $(\tilde{g}^{ij}) = D$.

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

2 $Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x)$ makes a Hamiltonian flow.

- The common Hamiltonian flow: $\mathcal{M} = \mathbb{R}^{2\ell}$, $Q = \begin{pmatrix} 0 & I_\ell \\ -I_\ell & 0 \end{pmatrix}$.

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

2 $Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x)$ makes a Hamiltonian flow.

- The common Hamiltonian flow: $\mathcal{M} = \mathbb{R}^{2\ell}$, $Q = \begin{pmatrix} 0 & I_\ell \\ -I_\ell & 0 \end{pmatrix}$.
- Symplectic manifold [10, 25]: $\mathcal{M}$ even-dim., Q non-singular.

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

2 $Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x)$ makes a Hamiltonian flow.

- The common Hamiltonian flow: $\mathcal{M} = \mathbb{R}^{2\ell}$, $Q = \begin{pmatrix} 0 & I_\ell \\ -I_\ell & 0 \end{pmatrix}$.
- Symplectic manifold [10, 25]: $\mathcal{M}$ even-dim., Q non-singular.
- What kind of structure can be more general, while being Hamiltonian (conserves a certain function)?

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

2 $Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x)$ makes a Hamiltonian flow.

- Poisson manifold $\mathcal{M}$ [14] *(known knowledge)*:
  - A Poisson structure on $\mathcal{M}$ can be represented by a *bivector field* β, whose coordinate expression $(\beta^{ij}(x))$ is *skew-symmetric* and satisfies:

$$\beta^{il}\partial_l\beta^{jk} + \beta^{jl}\partial_l\beta^{ki} + \beta^{kl}\partial_l\beta^{ij} = 0, \forall i, j, k. \quad (4)$$

  - A Poisson structure defines a *Hamiltonian flow* X_f given a smooth function f:

$$\big(X_f(x)\big)^i = \beta^{ij}(x)\,\partial_j f(x).$$

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

2 $Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x)$ makes a Hamiltonian flow.

- Poisson structure on $\mathcal{P}(\mathcal{M})$ [23, 1, 15] *(known knowledge)*:
  - The Hamiltonian flow of a function F on $\mathcal{P}(\mathcal{M})$ is

    $$\mathcal{X}_F(q) = \pi_q(X_f),$$

    where the function f on $\mathcal{M}$ relates to F by $\operatorname{grad}_q \mathbb{E}_q[f] = \operatorname{grad}_q F(q)$.
  - The Hamiltonian flow $\mathcal{X}_F$ conserves F: $\frac{\mathrm{d}}{\mathrm{d}t} F(q_t) = 0$.

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

2 $Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x)$ makes a Hamiltonian flow.

- Poisson structure on $\mathcal{P}(\mathcal{M})$ (new):

Lemma 2 (Hamiltonian flow of KL on $\mathcal{P}(\mathcal{M})$)

The Hamiltonian flow of KL_p *on* $\mathcal{P}(\mathcal{M})$ *is*

$$\mathcal{X}_{\mathrm{KL}_p}(q) = \pi_q(X_{\log(q/p)}), \text{ where } \big(X_{\log(q/p)}(x)\big)^i = \beta^{ij}(x)\,\partial_j \log(q(x)/p(x)).$$

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

2 $Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x)$ makes a Hamiltonian flow.

- $-\big(X_{\log(q/p)}(x)\big)^i = \beta^{ij}(x)\,\partial_j \log p(x) - \beta^{ij}(x)\,\partial_j \log q(x).$

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

2 $Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x)$ makes a Hamiltonian flow.

- $-\big(X_{\log(q/p)}(x)\big)^i = \beta^{ij}(x)\,\partial_j \log p(x) - \beta^{ij}(x)\,\partial_j \log q(x)$.

Assumption 4 (Regular MCMC dynamics (2/2))

(a) $D = C$ *or* $D = 0$ *or* $D = \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$, *for a symmetric positive definite* $C(x)$.

(b) $Q(x)$ *satisfies Eq.* (4): $Q^{il}\partial_l Q^{jk} + Q^{jl}\partial_l Q^{ki} + Q^{kl}\partial_l Q^{ij} = 0, \forall i, j, k.$

- Satisfied by MCMCs except for SGNHT-related methods [11, 34].
- Required to match Poisson structure; unnecessary for conservation of Hamiltonian.

Interpret MCMC Dynamics

$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

2 $Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x)$ makes a Hamiltonian flow.

- $-\big(X_{\log(q/p)}(x)\big)^i = \beta^{ij}(x)\,\partial_j \log p(x) - \beta^{ij}(x)\,\partial_j \log q(x)$.

Assumption 4 (Regular MCMC dynamics (2/2))

(a) $D = C$ *or* $D = 0$ *or* $D = \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$, *for a symmetric positive definite* $C(x)$.

(b) $Q(x)$ *satisfies Eq.* (4)*:* $Q^{il}\partial_l Q^{jk} + Q^{jl}\partial_l Q^{ki} + Q^{kl}\partial_l Q^{ij} = 0, \forall i, j, k.$

- Satisfied by MCMCs except for SGNHT-related methods [11, 34].
- Required to match Poisson structure; unnecessary for conservation of Hamiltonian.

$$Q^{ij}\,\partial_j \log p + \partial_j Q^{ij} \overset{?}{\Longleftrightarrow} Q^{ij}\,\partial_j \log p - Q^{ij}\,\partial_j \log q? \text{ Yes!}$$

Interpret MCMC Dynamics: Main Theorem

Theorem 5 (Equivalence between regular MCMC dynamics on $\mathbb{R}^M$ and fGH flows on $\mathcal{P}(\mathcal{M})$.)

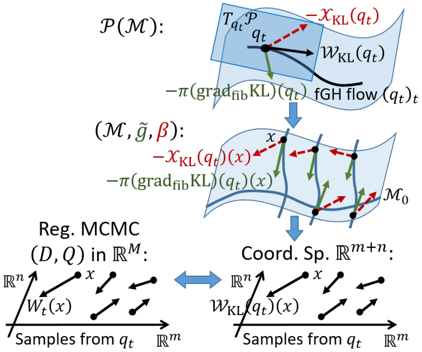
We call $(\mathcal{M}, \tilde{g}, \beta)$ a fiber-Riemannian Poisson (fRP) manifold, and define the fiber-gradient Hamiltonian (fGH) flow on $\mathcal{P}(\mathcal{M})$ as

$$\mathcal{W}_{\mathrm{KL}_p} := -\pi(\mathrm{grad}_{\mathrm{fib}}\, \mathrm{KL}_p) - \mathcal{X}_{\mathrm{KL}_p},$$
$$\left(\mathcal{W}_{\mathrm{KL}_p}(q)\right)^i = \pi_q\Big((\tilde{g}^{ij} + \beta^{ij})\, \partial_j \log(p/q) \Big). \tag{5}$$

Then:

Regular MCMC dynamics $\iff$ fGH flow with fRP $\mathcal{M}$,

$$(D, Q) \iff (\tilde{g}, \beta).$$



Interpret MCMC Dynamics: Case Study

Type 1: D is non-singular ($m = 0$ in Eq. (3)).

- $\mathcal{M}_0$ degenerates, $\mathcal{M}$ is the unique fiber.
- $\mathcal{M}$ is Riemannian, fiber gradient $\Longrightarrow$ gradient.
- The fGH flow: $\mathcal{W}_{\mathrm{KL}_p} = -\pi(\operatorname{grad} \mathrm{KL}_p) - \mathcal{X}_{\mathrm{KL}_p}$,
  - $-\pi(\operatorname{grad} \mathrm{KL}_p)$: minimizes KL_p steepestly on $\mathcal{P}(\mathcal{M})$.
  - $-\mathcal{X}_{\mathrm{KL}_p}$: conserves KL_p on $\mathcal{P}(\mathcal{M})$ and helps mixing/exploration.
- Converges to p uniquely (c.f. [24]).
- Robust to SG (c.f. [31, 32]).

Instances:

- LD [29] / SGLD [33]: $Q = 0$, $\mathcal{M}$ is Euclidean.
- RLD [16] / SGRLD [28]: $Q = 0$, $\mathcal{M}$ is the manifold under consideration.

Interpret MCMC Dynamics: Case Study

Type 2: $D = 0$ ($n = 0$ in Eq. (3)).

- $\mathcal{M}_0 = \mathcal{M}$, fibers degenerate.
- $\mathcal{M}$ has no (fiber-)Riemannian structures.
- The fGH flow: $\mathcal{W}_{\mathrm{KL}_p} = -\mathcal{X}_{\mathrm{KL}_p}$ conserves KL_p on $\mathcal{P}(\mathcal{M})$ and helps mixing/exploration.
- Fragile against SG: no stablizing forces (i.e. (fiber-)gradient flows) (c.f. [8, 3]).
- Hard to extend to ParVIs.

Instances (ℓ-dim. sample space $\mathcal{S}$):

- HMC [12, 26, 4]: $\mathcal{S} = \mathbb{R}^\ell$; $\mathcal{M}$ is $\mathbb{R}^{2\ell}$.
- HMC relies on *geometric ergodicity* for convergence [22, 4].
- RHMC [16] / LagrMC [19] / GMC [5]: manifold $\mathcal{S}$; $\mathcal{M}$ is $T^*\mathcal{S}$.

Interpret MCMC Dynamics: Case Study

Type 3: $D \neq 0$ and D is singular ($m, n \geq 1$ in Eq. (3)).

- Non-degenerate $\mathcal{M}_0$ and $\mathcal{M}_y$.
- $\mathcal{M}$ is a non-trivial fRP manifold.
- The fGH flow: $\mathcal{W}_{\mathrm{KL}_p} := -\pi(\mathrm{grad}_{\mathrm{fib}} \mathrm{KL}_p) - \mathcal{X}_{\mathrm{KL}_p}$,
  - $-\pi(\mathrm{grad}_{\mathrm{fib}} \mathrm{KL}_p)$: minimizes $\mathrm{KL}_{p(\cdot|y)}(q(\cdot|y))$ steepestly on each fiber $\mathcal{P}(\mathcal{M}_y)$.
  - $-\mathcal{X}_{\mathrm{KL}_p}$: conserves KL_p on $\mathcal{P}(\mathcal{M})$ and helps mixing/exploration.
- Robust to SG (SG appears on each fiber) (c.f. [8, 6]).

Instances (ℓ-dim. sample space $\mathcal{S}$):

- SGHMC [8] ($\mathcal{S} = \mathbb{R}^\ell$) and SGRHMC [24] / SGGMC [20] (manifold $\mathcal{S}$): $\mathcal{M}_0$ is $\mathcal{S}$ and $\mathcal{M}_\theta$ is $T^*_\theta \mathcal{S}$.
- SGNHT [11] ($\mathcal{S} = \mathbb{R}^\ell$) and gSGNHT [20] (manifold $\mathcal{S}$): $\mathcal{M}_0$ is $\mathcal{S}$ and $\mathcal{M}_\theta$ is $\mathbb{R} \times T^*_\theta \mathcal{S}$.

ParVI Simulation for SGHMC

Simulate the deterministic dynamics of SGHMC:

By Lemma 1 (Eq. (2)): $$\begin{cases} \frac{\mathrm{d}\theta}{\mathrm{d}t} = \Sigma^{-1} r, \\ \frac{\mathrm{d}r}{\mathrm{d}t} = \nabla_\theta \log p(\theta) - C\Sigma^{-1} r - C\nabla_r \log q(r). \end{cases}$$

By Theorem 5 (Eq. (5)): $$\begin{cases} \frac{\mathrm{d}\theta}{\mathrm{d}t} = \Sigma^{-1} r + \nabla_r \log q(r), \\ \frac{\mathrm{d}r}{\mathrm{d}t} = \nabla_\theta \log p(\theta) - C\Sigma^{-1} r - C\nabla_r \log q(r) - \nabla_\theta \log q(\theta). \end{cases}$$

- Problem: estimate $\nabla \log q$ with finite particles.

ParVI Simulation for SGHMC

Simulate the deterministic dynamics of SGHMC:

By Lemma 1 (Eq. (2)): $$\begin{cases} \frac{\mathrm{d}\theta}{\mathrm{d}t} = \Sigma^{-1} r, \\ \frac{\mathrm{d}r}{\mathrm{d}t} = \nabla_\theta \log p(\theta) - C\Sigma^{-1} r - C\nabla_r \log q(r). \end{cases}$$

By Theorem 5 (Eq. (5)): $$\begin{cases} \frac{\mathrm{d}\theta}{\mathrm{d}t} = \Sigma^{-1} r + \nabla_r \log q(r), \\ \frac{\mathrm{d}r}{\mathrm{d}t} = \nabla_\theta \log p(\theta) - C\Sigma^{-1} r - C\nabla_r \log q(r) - \nabla_\theta \log q(\theta). \end{cases}$$

- Problem: estimate $\nabla \log q$ with finite particles.
- Solution: use ParVI techniques [21], e.g. Blob [7]:

$$-\nabla_r \log q(r^{(i)}) \approx -\frac{\sum_k \nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(i,j)}} - \sum_k \frac{\nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(j,k)}},$$

where $K_r^{(i,j)} := K_r(r^{(i)}, r^{(j)})$.

ParVI Simulation for SGHMC

Simulate the deterministic dynamics of SGHMC:

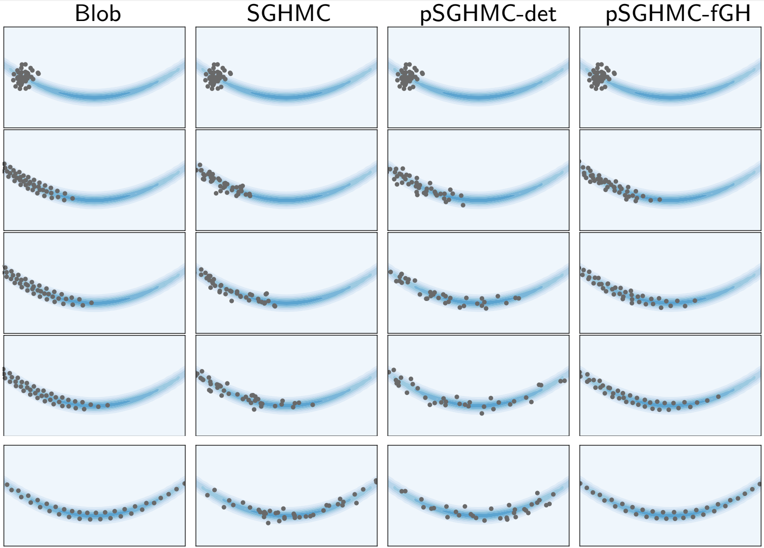
$$\text{pSGHMC-det: } \begin{cases} \frac{\Delta\theta^{(i)}}{\varepsilon} = \Sigma^{-1} r^{(i)}, \\ \frac{\Delta r^{(i)}}{\varepsilon} = \nabla_\theta \log p(\theta^{(i)}) - C\Sigma^{-1} r^{(i)} - C\left(\frac{\sum_k \nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(i,j)}} + \sum_k \frac{\nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(j,k)}} \right). \end{cases}$$

$$\text{pSGHMC-fGH: } \begin{cases} \frac{\Delta\theta^{(i)}}{\varepsilon} = \Sigma^{-1} r^{(i)} + \frac{\sum_k \nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(i,j)}} + \sum_k \frac{\nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(j,k)}}, \\ \frac{\Delta r^{(i)}}{\varepsilon} = \nabla_\theta \log p(\theta^{(i)}) - \left(\frac{\sum_k \nabla_{\theta^{(i)}} K_\theta^{(i,k)}}{\sum_j K_\theta^{(i,j)}} + \sum_k \frac{\nabla_{\theta^{(i)}} K_\theta^{(i,k)}}{\sum_j K_\theta^{(j,k)}} \right) \\ \qquad - C\Sigma^{-1} r^{(i)} - C\left(\frac{\sum_k \nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(i,j)}} + \sum_k \frac{\nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(j,k)}} \right). \end{cases}$$

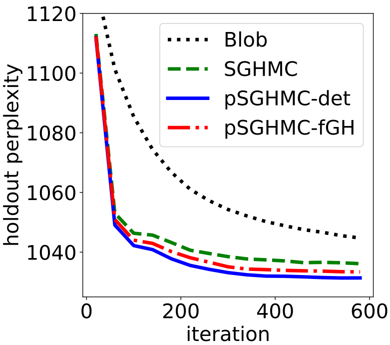
Advantages:

- Over SGHMC: particle-efficiency, ParVI techniques like HE [21].
- Over ParVIs: more efficient dynamics over LD.

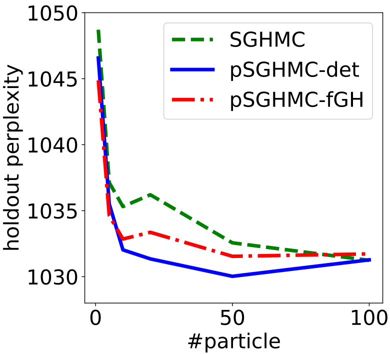
Synthetic Experiment



Latent Dirichlet Allocation (LDA)



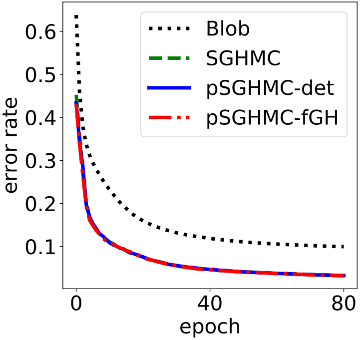
(a) Learning curve (20 ptcls)



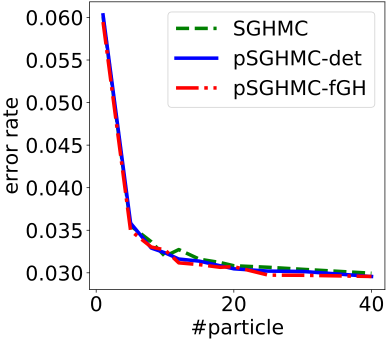
(b) Particle efficiency (iter 600)

Figure: Performance on LDA with the ICML data set.

Bayesian Neural Networks (BNNs)



(a) Learning curve (10 ptcls)



(b) Particle efficiency (epch 80)

Figure: Performance on BNN with MNIST data set.

Thank you!

Luigi Ambrosio and Wilfrid Gangbo.
Hamiltonian odes in the wasserstein space of probability measures.
Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 61(1):18–53, 2008.

Andrew D Barbour.
Stein's method for diffusion approximations.
Probability theory and related fields, 84(3):297–322, 1990.

Michael Betancourt.
The fundamental incompatibility of scalable hamiltonian monte carlo and naive data subsampling.
In *International Conference on Machine Learning*, pages 533–540, 2015.

Michael Betancourt.
A conceptual introduction to hamiltonian monte carlo.
arXiv preprint arXiv:1701.02434, 2017.

Simon Byrne and Mark Girolami.
Geodesic monte carlo on embedded manifolds.
Scandinavian Journal of Statistics, 40(4):825–845, 2013.

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 2278–2286, 2015.

Changyou Chen, Ruiyi Zhang, Wenlin Wang, Bai Li, and Liqun Chen.
A unified particle-optimization framework for scalable bayesian sampling.
arXiv preprint arXiv:1805.11659, 2018.

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.

Xiang Cheng and Peter Bartlett.
Convergence of langevin mcmc in kl-divergence.
arXiv preprint arXiv:1705.09048, 2017.

Ana Cannas Da Silva.
Lectures on symplectic geometry, volume 3575.
Springer, 2001.

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.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth.
Hybrid monte carlo.
Physics Letters B, 195(2):216–222, 1987.

Alain Durmus and Eric Moulines.
High-dimensional bayesian inference via the unadjusted langevin algorithm.
arXiv preprint arXiv:1605.01559, 2016.

Rui Loja Fernandes and Ioan Marcut.
Lectures on Poisson Geometry.
Springer, 2014.

Wilfrid Gangbo, Hwa Kil Kim, and Tommaso Pacini.
Differential forms on Wasserstein space and infinite-dimensional Hamiltonian systems.
American Mathematical Soc., 2010.

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.

Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, and Lester Mackey.
Measuring sample quality with diffusions.
arXiv preprint arXiv:1611.06972, 2016.

Richard Jordan, David Kinderlehrer, and Felix Otto.
The variational formulation of the fokker–planck equation.
SIAM journal on mathematical analysis, 29(1):1–17, 1998.

Shiwei Lan, Vasileios Stathopoulos, Babak Shahbaba, and Mark Girolami.
Markov chain monte carlo from lagrangian dynamics.
Journal of Computational and Graphical Statistics, 24(2):357–378, 2015.

Chang Liu, Jun Zhu, and Yang Song.
Stochastic gradient geodesic mcmc methods.
In *Advances In Neural Information Processing Systems*, pages 3009–3017, 2016.

Chang Liu, Jingwei Zhuo, Pengyu Cheng, Ruiyi Zhang, Jun Zhu, and Lawrence Carin.
Accelerated first-order methods on the wasserstein space for bayesian inference.

arXiv preprint arXiv:1807.01750, 2018.

Samuel Livingstone, Michael Betancourt, Simon Byrne, and Mark Girolami.
On the geometric ergodicity of hamiltonian monte carlo.
arXiv preprint arXiv:1601.08057, 2016.

John Lott.
Some geometric calculations on wasserstein space.
Communications in Mathematical Physics, 277(2):423–437, 2008.

Yi-An Ma, Tianqi Chen, and Emily Fox.
A complete recipe for stochastic gradient mcmc.
In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015.

Jerrold E Marsden and Tudor S Ratiu.
Introduction to mechanics and symmetry: a basic exposition of classical mechanical systems, volume 17.
Springer Science & Business Media, 2013.

Radford M Neal et al.
Mcmc using hamiltonian dynamics.
Handbook of Markov Chain Monte Carlo, 2(11), 2011.

Liviu I Nicolaescu.
Lectures on the Geometry of Manifolds.
World Scientific, 2007.

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.

Gareth O Roberts and Osnat Stramer.
Langevin diffusions and metropolis-hastings algorithms.
Methodology and computing in applied probability, 4(4):337–357, 2002.

Gareth O Roberts, Richard L Tweedie, et al.
Exponential convergence of langevin distributions and their discrete approximations.
Bernoulli, 2(4):341–363, 1996.

Issei Sato and Hiroshi Nakagawa.
Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process.
In *International Conference on Machine Learning*, pages 982–990, 2014.

Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer.
Consistency and fluctuations for stochastic gradient langevin dynamics.
The Journal of Machine Learning Research, 17(1):193–225, 2016.

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

Yizhe Zhang, Changyou Chen, Zhe Gan, Ricardo Henao, and Lawrence Carin.
Stochastic gradient monomial gamma sampler.
arXiv preprint arXiv:1706.01498, 2017.