Sampling Methods on Manifolds and Their View from Probability Manifolds

Chang Liu

Microsoft Research Asia

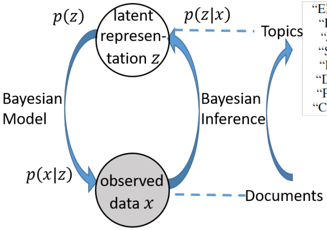
changliu@microsoft.com
Work done at Department of Computer Science and Technology, Tsinghua University

Nov. 20, 2019

Introduction: the Sampling Task

The need of drawing samples from a distribution:

- Bayesian inference: $p(z|x) = p(z)p(x|z)/p(x) \propto p(z)p(x|z)$:



"ENGINES"	speed	product	introduced	designs	fuel
"ROYAL"	britain	queen	sir	earl	died
"ARMY"	commander	forces	war	general	military
"STUDY"	analysis	space	program	user	research
"PARTY"	act	office	judge	justice	legal
"DESIGN"	size	glass	device	memory	engine
"PUBLIC"	report	health	community	industry	conference
"CHURCH"	prayers	communion	religious	faith	historical

- Generative model generation (e.g., MRF generation).
- Monte Carlo estimation (e.g., MRF likelihood gradient, doubly-stochastic gradient).

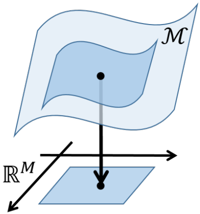
Introduction: Sampling Methods

Methods:

- Monte Carlo:
  - Directly draw i.i.d. samples.
  - Efficient but requires exact density.
- Markov Chain Monte Carlo (MCMC):
  - Draw samples by simulating a Markov chain with desired stationary distribution.
  - Admit unnormalized density but introduce autocorrelation.
- Particle-Based Variational Inference (ParVI):
  - Optimize a set of particles (i.e. samples) to drive the particle distribution towards the target distribution.
  - Admit unnormalized density but require assumption on the particle distribution (which affects performance).

Introduction: Manifold

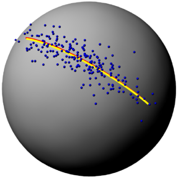
- Concept (M-dim manifold $\mathcal{M}$):
  topological space locally homeomporchic to an open subset of $\mathbb{R}^M$.
- Merits:
  - Inclusive concept: globally releases linearity.
  - Rich structures can be equipped: distance, gradient, distribution, dynamics, etc.
  - Fundamental view of geometry: parameterization-invariant.



Introduction: Sampling and Manifolds

Sampling from a distribution supported on a manifold.

- Spherical Admixture Model (SAM) [61]:
  topics on spheres for better representation.

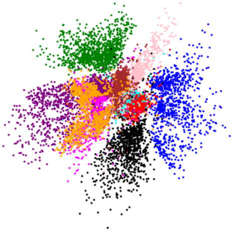


Model	Accuracy (%)				
	Overall	*prin.*	*war*	*cond.*	*Italy*
Bag-of-Words	57.9 ± 3.4	60.5	71.3	55.3	45.1
LDA	57.3 ± 3.0	59.4	63.9	58.1	34.9
movMF	49.6 ± 8.3	47.6	11.7	55.8	0.0
MH SAM $[\mathbb{S}_+]$	46.1 ± 6.9	46.5	31.8	54.4	8.3
MH SAM $[\mathbb{S}]$	59.4 ± 5.4	60.9	51.7	64.8	31.4
VEM SAM $[\mathbb{S}_+]$	58.7 ± 0.6	64.9	71.1	60.8	13.9
VEM SAM $[\mathbb{S}]$	$\mathbf{65.2 \pm 0.3}$	71.3	65.1	62.5	50.6

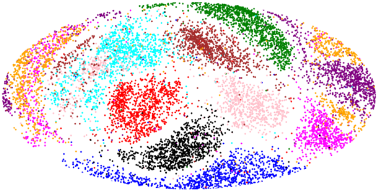
Introduction: Sampling and Manifolds

Sampling from a distribution supported on a manifold.

- Hyperspherical Variational Auto-Encoder [19, 30]: spherical latent space for uninformative prior.



(a) $\mathbb{R}^2$ latent space of the $\mathcal{N}$-VAE.

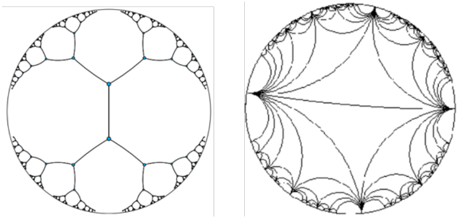


(b) Hammer projection of $\mathcal{S}^2$ latent space of the $\mathcal{S}$-VAE.

Introduction: Sampling and Manifolds

Sampling from a distribution supported on a manifold.

- Hyperbolic Variational Auto-Encoders [53, 30, 59, 55]:
  hyperbolic latent space (R) for the analogy to a tree structure (L).

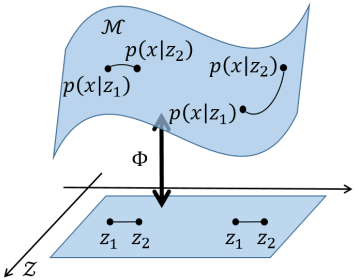


- Bayesian Matrix Factorization [64, 66, 73]:
  factor matrices on Stiefel manifold [68, 33]
  $\{M \in \mathbb{R}^{m \times n} \mid M^\top M = I_m\}$.

Introduction: Sampling and Manifolds

Sampling from a distribution supported on a manifold.

- Information Geometry [3, 4]:
  for Bayesian inference $p(z|x)$ for a Bayesian model $\{p(z), p(x|z)\}$:



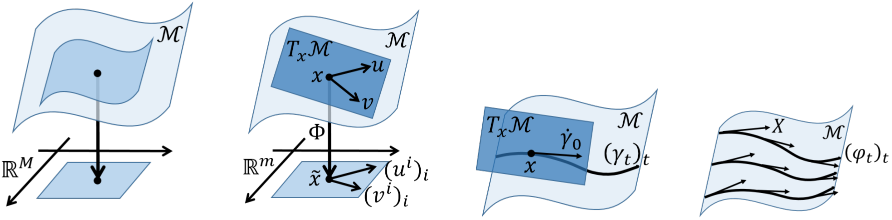
Introduction: Sampling and Manifolds

- Sampling from a distribution supported on a manifold:
  How to comply to the manifold geometry while being efficient?
- Viewing Sampling Methods on Probability Manifolds:
  - ParVIs have a natural optimization interpretation on a probability space. Can it be made concrete?
  - Do MCMCs have a similar interpretation?

Manifolds

M-dim. manifold $\mathcal{M}$ (a):
topological space locally homeomporchic to an open subset of $\mathbb{R}^M$.

- Tangent vector v at $x \in \mathcal{M}$ (b): linear function $C^\infty(\mathcal{M}) \to \mathbb{R}$ satifying the Leibniz rule (directional derivative).
  - A smooth curve γ_t through x defines a tangent vector (derivative along the curve) (c).
- Tangent space $T_x\mathcal{M}$ at x (b): M-dim. linear space.
- Flow of a vector field V (d): the set of curves $\{(\varphi_t)_t\}$ s.t. $\dot{\varphi}_t = V(\varphi_t)$ (exists at least locally).

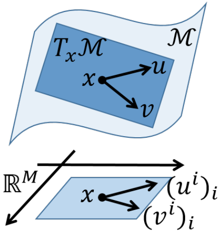


(a) (b) (c) (d)

Manifolds

Riemannian structure: 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 of f:
$\langle \mathrm{grad}\, f(x), v \rangle_{T_x\mathcal{M}} = v[f] := v^i \partial_i f(x).$
$\Longleftrightarrow$
Steepest ascending direction:
$\mathrm{grad}\, f(x) = \max \cdot \mathrm{argmax}_{\|v\|_{T_x\mathcal{M}}=1} \frac{\mathrm{d}}{\mathrm{d}t} f(\varphi_t).$
Coordinate expression:
$$\big(\mathrm{grad}\, f(x)\big)^i = g^{ij}(x) \partial_j f(x).$$



Manifolds

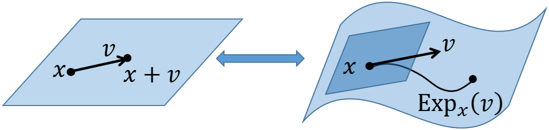
Riemannian structure: inner product in every tangent space $T_x\mathcal{M}$.

- Distance: $d(x,y) = \sqrt{\inf_{\gamma_t:\gamma_0=x,\gamma_1=y} \int_0^1 \langle \dot{\gamma}_t, \dot{\gamma}_t \rangle_{T_{\gamma_t}\mathcal{M}} \,\mathrm{d}t}$.
- Geodesic: the minimizing curve(s) when it exists (e.g., when $\mathcal{M}$ is complete as a metric space [32]).
  - More fundamental definition: auto-parallel curves under an affine connection (covariant derivative).
  - Generalization of straight lines.

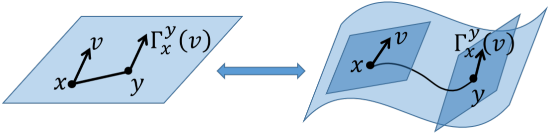
Manifolds

Riemannian structure: inner product in every tangent space $T_x\mathcal{M}$.

- Exponential map $\mathrm{Exp}_x(v)$: maps $v \in T_x\mathcal{M}$ to the end point of the geodesic tangent to v at x with length $\|v\|_{T_x\mathcal{M}}$.
    - Generalization of vector addition.



- Parallel transport $\Gamma_x^y(v)$: moves $v \in T_x\mathcal{M}$ to $T_y\mathcal{M}$ (in a certain sense of) parallelly, along the geodesic from x to y.
    - Generalization of conventional parallel transport.
    - Generally is path-dependent.
    - More fundamental def.: specified by an affine connection.



Manifolds

Measures on orientable manifolds can be expressed by volume forms:

- Volume form: alternative linear $(T_x\mathcal{M})^M \to \mathbb{R}$ for every x.
- Lebesgue measure of a coordinate space: $\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^M$.
- Riemannian volume form (Riemannian measure): coordinate invariant volume form $\sqrt{|G|}\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^M$.

MCMCs on Euclidean Space

Classical MCMCs: high autocorrelation.

- Metropolis-Hastings algorithm [54, 31].
- Gibbs sampling [27].

Dynamics-Based MCMCs: more effective move.

- Dynamics: continuous-time no-jump Markov process:
$$\mathrm{d}x = V(x)\,\mathrm{d}t + \sqrt{2D(x)}\,\mathrm{d}B_t(x).$$
- Key tool: the Fokker-Planck Equation:
$$\partial_t p_t = -\partial_i (p_t V^i) + \partial_i \partial_j (p_t D^{ij}).$$

MCMCs on Euclidean Space

Dynamics-Based MCMCs: more effective move.

- Langevin Dynamics (LD) [39] ([63, 62, 71]):
$$\mathrm{d}x = \Sigma^{-1}\nabla \log p \,\mathrm{d}t + \sqrt{2\Sigma^{-1}}\,\mathrm{d}B_t(x).$$
- Hamiltonian Dynamics (Hamitonian Monte Carlo (HMC) [21, 56, 10]):
$$\begin{cases} \mathrm{d}x = \Sigma^{-1} r \,\mathrm{d}t, \\ \mathrm{d}r = \nabla \log p \,\mathrm{d}t. \end{cases}$$
- Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) [37, 15]:
$$\begin{cases} \mathrm{d}x = \Sigma^{-1} r \,\mathrm{d}t, \\ \mathrm{d}r = \nabla \log p \,\mathrm{d}t - C r \,\mathrm{d}t + \sqrt{2C\Sigma}\,\mathrm{d}B_t(r). \end{cases}$$
- Stochastic Gradient Nosé-Hoover Thermostats (SGNHT) [20]:
$$\begin{cases} \mathrm{d}x = \Sigma^{-1} r \,\mathrm{d}t, \\ \mathrm{d}r = \nabla \log p \,\mathrm{d}t - \xi r \,\mathrm{d}t + \sqrt{2C\Sigma}\,\mathrm{d}B_t(r), \\ \mathrm{d}\xi = (\frac{1}{M} r^\top \Sigma^{-1} r - 1)\,\mathrm{d}t. \end{cases}$$

MCMCs on Euclidean Space

Dynamics-Based MCMCs: more effective move.

- The complete recipe [51] for the dynamics:

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

  for some pos. semi-def. $D_{M \times M}$ (diffusion matrix) and skew-symm. $Q_{M \times M}$ (curl matrix), keeps p invariant.
  - The inverse also holds.
  - If D is pos. def., then p is the unique stationary distribution.

MCMCs on Euclidean Space

Dynamics-Based MCMCs: more effective move.

- Stochastic Gradient MCMC: for Bayesian inference,

$$\nabla_z \log p(z|\{x^{(n)}\}_{n=1}^N) = \nabla_z \log p(z) + \sum_{n=1}^{N} \nabla_z \log p(x^{(n)}|z),$$

$$\tilde{\nabla}_z \log p(z|\{x^{(n)}\}_{n=1}^N) := \nabla_z \log p(z) + \frac{N}{|\mathcal{S}|} \sum_{n\in\mathcal{S}} \nabla_z \log p(x^{(n)}|z)$$

$$\approx \nabla_z \log p(z|\{x^{(n)}\}_{n=1}^N) + \mathcal{N}(0, A(z)).$$

Influence on the dynamics $\mathrm{d}x = V(x)\,\mathrm{d}t + \sqrt{2D(x)}\,\mathrm{d}B_t(x)$:

$$\mathrm{Var}(V(x)\,\mathrm{d}t) = \mathrm{Var}(V(x))\,\mathrm{d}t^2 = o(\mathrm{d}t),$$

$$\mathrm{Var}(\sqrt{2D(x)}\,\mathrm{d}B_t(x)) = 2D(x)\,\mathrm{d}t.$$

  - HMC cannot be simulated using stochastic gradient [15, 9].

MCMCs on Riemannian Manifolds

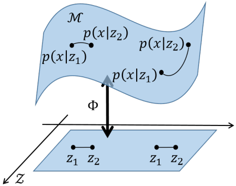
In the coordinate space (p is the density w.r.t. the Lebesgue meas.):

- Riemann Manifold Langevin Dynamics (RMLD) [28, 60]:
$$\mathrm{d}x = G^{-1}\nabla \log p \,\mathrm{d}t + \nabla \cdot G^{-1}\,\mathrm{d}t + \sqrt{2G^{-1}}\,\mathrm{d}B_t(x).$$
- Riemann Manifold Hamiltonian Monte Carlo (RMHMC) [28]:
$$\begin{cases} \mathrm{d}x = G^{-1} r\,\mathrm{d}t, \\ \mathrm{d}r = \nabla \log(p/\sqrt{|G|})\,\mathrm{d}t - \frac{1}{2}\nabla\left(r^\top G^{-1} r\right)\mathrm{d}t. \end{cases}$$
- Stochastic Gradient Riemann Hamiltonian Monte Carlo (SGRHMC) [51]:
$$\begin{cases} \mathrm{d}x = G^{-1/2} r\,\mathrm{d}t, \\ \mathrm{d}r = G^{-1/2}\nabla \log p\,\mathrm{d}t + \nabla \cdot G^{-1/2}\,\mathrm{d}t - G^{-1} r\,\mathrm{d}t + \sqrt{2G^{-1}}\,\mathrm{d}B_t(r). \end{cases}$$

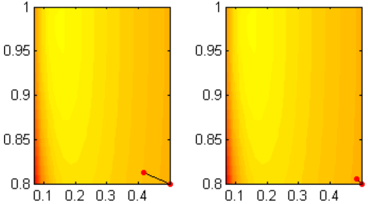
MCMCs on Riemannian Manifolds

In the coordinate space: application using the Fisher-Rao Metric (information geometry [3, 4]):

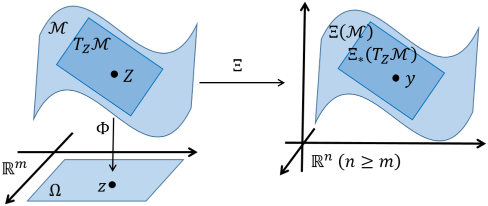
- Given a Bayesian model $p(z), p(x|z)$, z is a coordinate of the manifold $\{p(x|z) \mid z \in \mathcal{Z}\}$.



- Fisher-Rao metric:
  $G(z) := \mathbb{E}_{p(x|z)}[\nabla_z^\top \log p(x|z) \nabla_z \log p(x|z)]$.
  - Derived from the KL divergence.
  - Corresp. distance is the ($\sqrt{8}\times$) JS divergence.
  - Invariant under reparameterization of z.
- HMC (L) and RMHMC (R) [28]:



MCMCs on Riemannian Manifolds



Problems of coordinate space: a global one may not exist (e.g. hyperspheres $\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\}$).

- Cumbersome to switch between coordinate systems.
- G would be singular near the edge of a coordinate space.

Simulation in an embedded space $\Xi(\mathcal{M})$: homeo. injective $\Xi : \mathcal{M} \to \mathbb{R}^n$.

- Global representation.
- Common manifolds have a natural (isometric) embedding.
- Hausdorff meas. on $\Xi(\mathcal{M})$ (isom. emb.) is the Riem. meas. on $\mathcal{M}$.

MCMCs on Riemannian Manifolds

RMHMC in the embedded space:

- Constraint HMC (CHMC) [11].
- Geodesic Monte Carlo (GMC) [12].

Stochastic Gradient MCMCs in the embedded space [42]:

- Stochastic Gradient Geodesic Monte Carlo (SGGMC).
- Geodesic Stochastic Gradient Nosé-Hoover Thermostats (gSGNHT).

Stochastic Gradient MCMCs in the Embedded Space

Table: A summary of MCMCs on Riemannian Manifolds. –: 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 [13] and [40], respectively.

methods	stochastic gradient	no inner iteration	no global coordinates	order of integrator
LD [63, 62]	×	√	–	1st
HMC [56]	×	√	–	2nd
GMC [12]	×	√	√	2nd
RMLD [28]	×	√	×	1st
RMHMC [28]	×	×	×	2nd†
CHMC [11]	×	×	√	2nd†
SGLD [71]	√	√	–	1st
SGHMC [15] / SGNHT [20]	√	√	–	1st‡
SGRLD [60] / SGRHMC [51]	√	√	×	1st
SGGMC / gSGNHT [42]	√	√	√	2nd

Stochastic Gradient MCMCs in the Embedded Space

SGGMC dynamics (coordinate space):

- Augment with the momentum $r \in \mathbb{R}^m$ (more precisely, covector $\in T_x^* \mathcal{M}$).
- Augmented target distribution:

$$-\log p(z,r) = \underbrace{-\log p(z|x) + \frac{1}{2}\log|G(z)|}_{\text{potential energy}} \underbrace{+\frac{1}{2}r^\top G(z)^{-1} r}_{\text{kinetic energy}}.$$

- Let $\mathcal{M}$ isom. emb. in $\mathbb{R}^n$ via $y = \Xi(x)$. Define:

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

where $J_{n\times m} : J_{ai} = \frac{\partial y^a}{\partial x^i}$ ($J^\top J = G$).

Stochastic Gradient MCMCs in the Embedded Space

SGGMC dynamics (coordinate space):

$$\begin{cases} \mathrm{d}z = G^{-1} r \mathrm{d}t \\ \mathrm{d}r = \nabla_z \log p(z|x) \mathrm{d}t - \frac{1}{2} \nabla_z \log |G(z)| \mathrm{d}t \\ \qquad - J^\top C J G^{-1} r \, \mathrm{d}t - \frac{1}{2} \nabla_z \left[r^\top G^{-1} r \right] \mathrm{d}t \\ \qquad + \mathcal{N}(0, 2 J^\top C J \mathrm{d}t) \end{cases}$$

Stochastic Gradient MCMCs in the Embedded Space

SGGMC simulation (emb. sp.): Symmetric Splitting Integrator (SSI) [13].

- Split SGGMC dynamics (in the coordinate space):

$$\begin{cases} \mathrm{d}z = G^{-1} r \mathrm{d}t \\ \mathrm{d}r = \nabla_z \log p(z|x) \mathrm{d}t - \frac{1}{2} \nabla_z \log |G(z)| \mathrm{d}t \\ \qquad - J^\top C J G^{-1} r \, \mathrm{d}t - \frac{1}{2} \nabla_z \left[r^\top G^{-1} r \right] \mathrm{d}t \\ \qquad + \mathcal{N}(0, 2 J^\top C J \mathrm{d}t) \end{cases}$$

$$A: \begin{cases} \mathrm{d}z = G^{-1} r \mathrm{d}t \\ \mathrm{d}r = -\frac{1}{2} \nabla_z \left[r^\top G^{-1} r \right] \mathrm{d}t \end{cases} \Rightarrow (z_t, r_t) = \mathrm{GeodFlow}(z_0, r_0) \text{ [1, 12]}$$

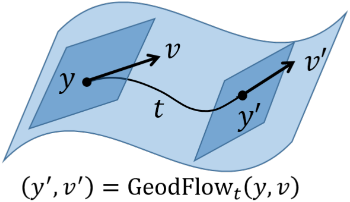
$$B: \begin{cases} \mathrm{d}z = 0 \\ \mathrm{d}r = -J^\top C J G^{-1} r \, \mathrm{d}t \end{cases} \Rightarrow \begin{cases} z_t = z_0 \\ r_t = J^\top \operatorname{expm}\{-Ct\} J G^{-1} r_0 \end{cases}$$

$$O: \begin{cases} \mathrm{d}z = 0 \\ \mathrm{d}r = \nabla_z \log p(z|x) \, \mathrm{d}t \\ \qquad - \frac{1}{2} \nabla_z \log |G(z)| \, \mathrm{d}t \\ \qquad + \mathcal{N}(0, 2 J^\top C J \mathrm{d}t) \end{cases} \Rightarrow \begin{cases} z_t = z_0 \\ r_t = \nabla_z \log p(z_0|x) t \\ \qquad - \frac{1}{2} \nabla_z \log |G(z_0)| \, t \\ \qquad + \mathcal{N}(0, 2 J^\top C J t) \end{cases}$$

Stochastic Gradient MCMCs in the Embedded Space

SGGMC simulation (emb. sp.): Symmetric Splitting Integrator (SSI) [13].

- Dynamics A in the embedded space: geodesic flow (i.e., exponential map + parallel transport).



Example 1 (Geodesic flow of hypersphere $\mathbb{S}^{n-1}$ in the embedded space)

$$\begin{cases} y(t) = y(0)\cos(\alpha t) + (v(0)/\alpha)\sin(\alpha t) \\ v(t) = -\alpha y(0)\sin(\alpha t) + v(0)\cos(\alpha t) \end{cases},$$

where $y \in \mathbb{S}^{n-1}$, $v = \dot{y} \in T_y(\mathbb{S}^{n-1})$, and $\alpha = \|v(0)\|$.

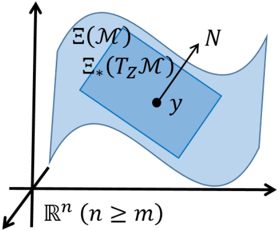
Stochastic Gradient MCMCs in the Embedded Space

SGGMC simulation (emb. sp.): Symmetric Splitting Integrator (SSI) [13].

- Dynamics B and O in the embedded space:

$$B : \begin{cases} y(t) = y(0) \\ v(t) = \Lambda\big(y(0)\big)\, \mathrm{expm}\{-Ct\} v(0) \end{cases}$$

$$O : \begin{cases} y(t) = y(0) \\ v(t) = v(0) + \Lambda\big(y(0)\big)\big[\nabla_y \log p_{\mathcal{H}}\big(y(0)|x\big)t \\ \qquad\qquad\qquad\qquad + \mathcal{N}(0, 2Ct)\big], \end{cases}$$



where: $p_{\mathcal{H}}$ is the density function w.r.t. the Hausdorff measure, and $\Lambda(y) = I_n - P(y)P(y)^\top$ is the projection onto $\Xi_*(T_z\mathcal{M})$.

Example 2 (The projection $\Lambda(y)$ for hypersphere in the embedded space)

$$\Lambda(y) = I_n - yy^\top.$$

Stochastic Gradient MCMCs in the Embedded Space

SGGMC simulation (emb. sp.): Symmetric Splitting Integrator (SSI) [13].

- Simulate following the sequence "ABOBA":

Algorithm 1 Sampling procedure of SGGMC

Sample a subset $\mathcal{S}$ for computing $\tilde{\nabla}_y \log p_{\mathcal{H}}(y)$. $(y_0, v_0) \leftarrow (y^{(n-1)}, v^{(n-1)})$.
for $l = 1, 2, \ldots, L$ **do**
  A: Update $(y^*, v^*) \leftarrow (y_{l-1}, v_{l-1})$ by the geodesic flow for time step $\frac{\varepsilon_n}{2}$.
  B: $v^* \leftarrow \exp\{-C\frac{\varepsilon_n}{2}\}v^*$.
  O: $v^* \leftarrow v^* + \Lambda(y^*) \cdot \left[\tilde{\nabla}_y \log p_{\mathcal{H}}(y^*)\varepsilon_n + \mathcal{N}\Big(0, (2C - \varepsilon_n V(y^*))\varepsilon_n\Big)\right]$.
  B: $v^* \leftarrow \exp\{-C\frac{\varepsilon_n}{2}\}v^*$.
  A: Update $(y_l, v_l) \leftarrow (y^*, v^*)$ by the geodesic flow for time step $\frac{\varepsilon_n}{2}$.
end for

- Second-order simulation: MSE $= O(L^{-2K/(2K+1)})$ [13].

Stochastic Gradient MCMCs in the Embedded Space

gSGNHT dynamics:

$$\begin{cases} \mathrm{d}z = G^{-1} r \, \mathrm{d}t, \\ \mathrm{d}r = \nabla_z \log p(z|x) \mathrm{d}t - \frac{1}{2} \nabla_z \log |G| \mathrm{d}t - \xi r \, \mathrm{d}t - \frac{1}{2} \nabla_z \left[r^\top G^{-1} r \right] \mathrm{d}t + \mathcal{N}(0, 2CG\mathrm{d}t) \\ \mathrm{d}\xi = (\frac{1}{m} r^\top G^{-1} r - 1) \, \mathrm{d}t. \end{cases}$$

gSGNHT simulation:

Algorithm 2 Sampling procedure of gSGNHT

A: Update $(y^*, v^*) \leftarrow (y_{l-1}, v_{l-1})$ by the geodesic flow for time step $\frac{\varepsilon_n}{2}$,
$\xi^* \leftarrow \xi_{l-1} + (\frac{1}{m} v_{l-1}^\top v_{l-1} - 1) \frac{\varepsilon_n}{2}$.

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

O: $v^* \leftarrow v^* + \Lambda(y^*) \cdot \left[\tilde{\nabla}_y \log p_{\mathcal{H}}(y^*) \varepsilon_n + \mathcal{N}\big(0, (2C - \varepsilon_n V(y^*)) \varepsilon_n\big) \right]$.

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

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

Stochastic Gradient MCMCs in the Embedded Space

Experimental results:

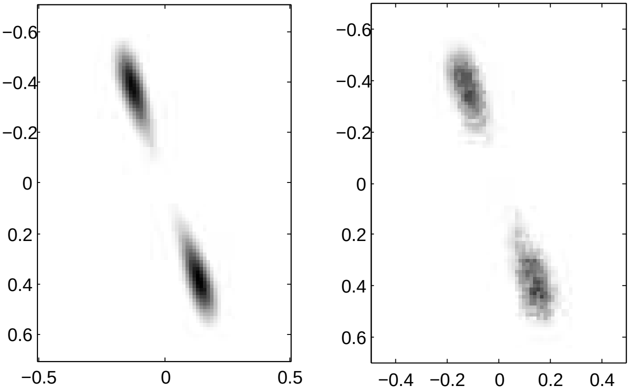
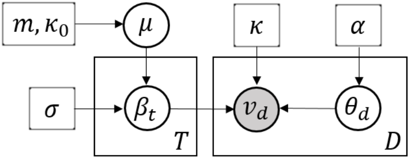


Figure: Joint posterior of z_1 and z_2 in gray scale. Left: true distribution; Right: empirical distribution by samples of SGGMC.

Stochastic Gradient MCMCs in the Embedded Space

Experimental results: inference for Spherical Admixture Model (SAM) [61]

- Model structure:



- Document v (e.g., normalized tf-idf), topic β, corpus mean μ: on hyperspheres.
- Posterior of interest: $p(\beta|v)$.

$$\nabla_\beta \log p(\beta|v) = \frac{1}{p(\beta|v)} \nabla_\beta \int p(\beta, \theta|v) \mathrm{d}\theta = \mathbb{E}_{p(\theta|\beta,v)} \left[\nabla_\beta \log p(\beta, \theta|v) \right].$$

Run another MCMC (GMC [12]) to sample from $p(\theta|\beta, v)$ (supported on simplex) to estimate the expectation.

Stochastic Gradient MCMCs in the Embedded Space

Experimental results: inference for Spherical Admixture Model (SAM) [61]

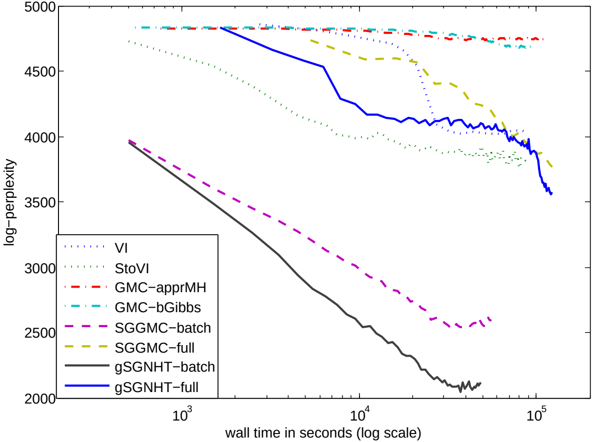


Figure: Results on the 150K Wikipedia subset (150K training and 1K test, 50 topics)

ParVIs on Euclidean Space

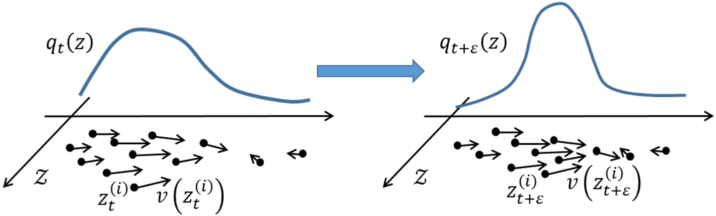
Particle-Based Variational Inference (ParVI):
optimize a set of particles (i.e. samples) to drive the particle distribution towards the target distribution.

- More flexible and accurate than classical (i.e., statistical-model-based) variational inference.
- Has a better convergence perspective than MCMCs.
- More particle-efficient than MCMCs.

ParVIs on Euclidean Space

Stein Variational Gradient Descent (SVGD) [46]:

- A deterministic dynamics $\dot{z}_t = v(z_t)$ on $\mathcal{M} = \mathbb{R}^m$ induces a continuously-evolving distribution $(q_t)_t$ on $\mathcal{M}$:



$\partial_t q_t = -\nabla \cdot (q_t v)$. (continuity equation / det. FPE)

ParVIs on Euclidean Space

Stein Variational Gradient Descent (SVGD) [46]:

- To drive $(q_t)_t$ towards p, let it minimize $\mathrm{KL}(q_t \| p)$:
  - Find the decreasing rate (directional derivative):

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t \| p) = \mathbb{E}_q[v \cdot \nabla \log p + \nabla \cdot v].$$

  - Find v maximizing the decreasing rate $v^* := \max \cdot \operatorname{argmax}_{\|v\|_{\mathfrak{X}}=1} -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t \| p)$ (functional gradient).
    - Taking $\mathfrak{X} = \mathcal{T}(\mathcal{M}) = \{\mathbb{R}^m \to \mathbb{R}^m \text{ functions}\}$: solution not estimable.
    - Taking $\mathfrak{X} = \mathcal{H}^m$ where $\mathcal{H}$ is the RKHS [67] of a kernel K:

$$v^*(x') = \mathbb{E}_{q(x)}[K(x, x')\nabla_x \log p(x) + \nabla_x K(x, x')].$$

    The expectation can be estimated directly by the particles!
  - Simulate the particles by applying the dynamics: $x^{(i)} \leftarrow x^{(i)} + \varepsilon v^*(x^{(i)})$.

ParVIs on Riemannian Manifolds

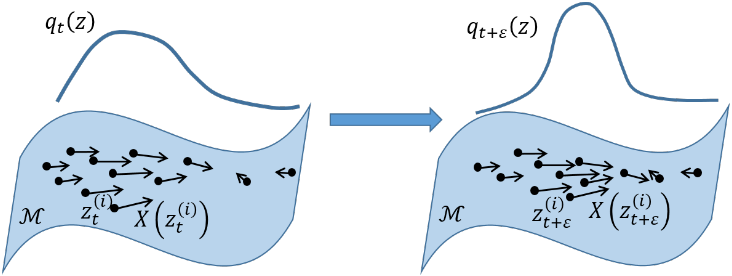
Riemannian SVGD [41]:

- Utilize information geometry to enhance efficiency (coordinate space).
- Enable ParVIs on manifolds like hyperspheres (embedded space).

ParVIs on Riemannian Manifolds

Dynamics on a Riemannian manifold:

- $\dot{z}_t = X(z_t)$, $(z_t)_t$ is a curve of the flow of X.



- Evolving distribution: let all densities be w.r.t. the Riem. meas.

Lemma 3 (Continuity Equation on Riemannian Manifold)

$$\begin{aligned}\partial_t q_t &= -\operatorname{div}(q_t X) = -X[q_t] - q_t \operatorname{div}(X) \\ &= -X^i \partial_i q_t - q_t \partial_i X^i - q_t X^i \partial_i \log \sqrt{|G|}.\end{aligned}$$

ParVIs on Riemannian Manifolds

Directional derivative:

Theorem 4 (Directional Derivative)

Let p be a fixed distribution. Then the directional derivative is

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t\|p)=\mathbb{E}_{q_t}[\mathrm{div}(pX)/p]=\mathbb{E}_{q_t}\big[X[\log p]+\mathrm{div}(X)\big].$$

- $X[q_t]$: the action of the vector field X on the smooth function q_t. In any coordinate system, $X[q_t] = X^i \partial_i q_t$.
- $\mathrm{div}(X)$: the divergence of vector field X. In any coordinate system, $\mathrm{div}(X) = \partial_i(\sqrt{|G|}X^i)/\sqrt{|G|}$.

ParVIs on Riemannian Manifolds

Functional gradient:

$$X^* := \max \cdot \underset{X \in \mathfrak{X}, \|X\|_{\mathfrak{X}}=1}{\operatorname{argmax}} \ \mathcal{J}(X) := \mathbb{E}_q \big[X[\log p] + \operatorname{div}(X) \big],$$

where $\mathfrak{X}$ is a subspace of vector fields on $\mathcal{M}$, such that:

1. X^* is a valid vector field on $\mathcal{M}$.

Example 5 (Nontriviality of a valid vector field)

Vector fields on an even-dimensional hypersphere must have one zero point (hairy ball theorem ([2], Thm 8.5.13)). The choice in SVGD $\mathfrak{X} = \mathcal{H}^m$ cannot guarantee this requirement.

2. X^* is coordinate invariant.
   - Concept: the expression in any coordinate system is the same.
   - Necessary for avoiding the arbitrariness of the solution.
   - The choice in SVGD $\mathfrak{X} = \mathcal{H}^m$ cannot guarantee this requirement.
3. X^* can be expressed in closed form.

ParVIs on Riemannian Manifolds

Functional gradient:

Our Solution

$\mathfrak{X} = \{\mathrm{grad}\, f \mid f \in \mathcal{H}\}$, where $\mathcal{H}$ is the RKHS of a kernel K.

The gradient a function is a valid, coordinate invariant vector field.

Lemma 6

For Gaussian RKHS, $\mathfrak{X}$ is isometrically isomorphic to $\mathcal{H}$.

Theorem 7 (Functional Gradient)

$$X^{*\prime} = \mathrm{grad}' f^{*\prime}, \; f^{*\prime} = \mathbb{E}_q\big[(\mathrm{grad}\, K)[\log p] + \Delta K\big],$$

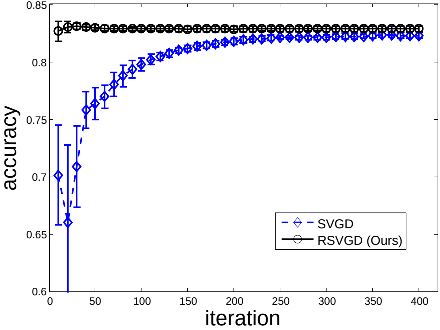
where "$\cdot'$" takes x' as argument, and $\Delta f := \mathrm{div}(\mathrm{grad}\, f)$.

$$X^{*\prime i} = g'^{ij} \partial'_j \mathbb{E}_q\Big[\big(g^{ab}\partial_a \log(p\sqrt{|G|}) + \partial_a g^{ab}\big)\partial_b K + g^{ab}\partial_a\partial_b K\Big].$$

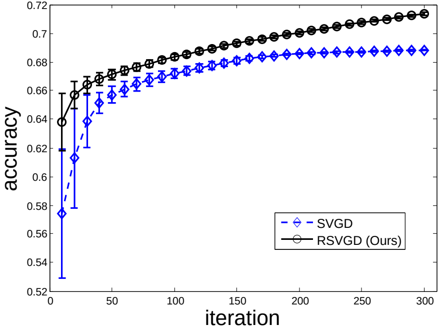
Simulate the dynamics: $z^{(s)} \leftarrow z^{(s)} + \varepsilon X^*(z^{(s)})$.

ParVIs on Riemannian Manifolds

Experimental Results (coordinate space):



(a) On Splice19 dataset



(b) On Covertype dataset

Figure: Test accuracy along iteration for BLR. Both methods are run 20 times on Splice19 and 10 times on Covertype.

ParVIs on Riemannian Manifolds

Functional gradient in the embedded space:

Proposition 8 (Functional Gradient in the Embedded Space)

Let m-dim $\mathcal{M}$ isometrically embedded in $\mathbb{R}^n$ (with orthonormal basis $\{y^\alpha\}_{\alpha=1}^n$) via $\Xi : \mathcal{M} \to \mathbb{R}^n$. Then $X^{*\prime} = (I_n - N' N'^\top) \nabla' f^{*\prime}$,

$$f^{*\prime} = \mathbb{E}_q \Big[\Big(\nabla \log \big(p \sqrt{|G|} \big) \Big)^\top \Big(I_n - P P^\top \Big) (\nabla K) + \nabla^\top \nabla K - \operatorname{tr} \Big(P^\top (\nabla \nabla^\top K) P \Big) + \Big((J^\top \nabla)^\top (G^{-1} J^\top) \Big) (\nabla K) \Big],$$

where $\nabla = (\partial_{y^1}, \dots, \partial_{y^n})^\top$, $J_{n \times m} : J_{ai} = \frac{\partial y^a}{\partial z^i}$, and $P \in \mathbb{R}^{n \times (n-m)}$ is the set of orthonormal basis of the orthogonal complement of $\Xi_*(T_z \mathcal{M})$.

Simulate the dynamics with exponential map:

$$y^{(s)} \leftarrow \mathrm{Exp}_{y^{(s)}}(\varepsilon X^*(y^{(s)})).$$

(Is a coordinate-independent expression possible?)

ParVIs on Riemannian Manifolds

Functional gradient on hyperspheres:

Proposition 9 (Functional Gradient for Embedded Hyperspheres)

For $\mathbb{S}^{n-1}$ isometrically embedded in $\mathbb{R}^n$ with orthonormal basis $\{y^\alpha\}_{\alpha=1}^n$, we have $X^{*\prime} = (I_n - y' y'^\top) \nabla' f^{*\prime}$, where $f^{*\prime} =$

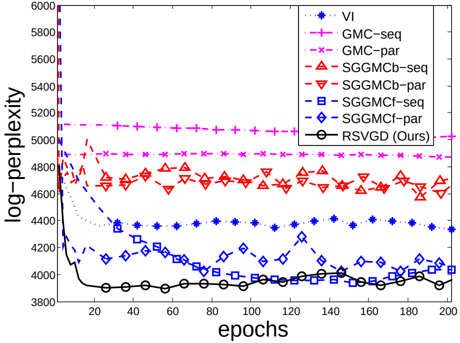
$$\mathbb{E}_q \Big[(\nabla \log p)^\top (\nabla K) + \nabla^\top \nabla K - y^\top (\nabla \nabla^\top K) y - (y^\top \nabla \log p + n - 1) y^\top \nabla K \Big].$$

Simulate the dynamics with exponential map on $\mathbb{S}^{n-1}$:

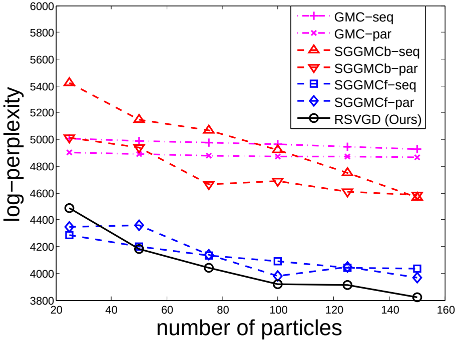
$$\mathrm{Exp}_y(v) = y \cos(\|v\|) + (v / \|v\|) \sin(\|v\|).$$

ParVIs on Riemannian Manifolds

Experimental Results (embedded space):



(a) Results with 100 particles



(b) Results at 200 epochs

Figure: Results on the SAM inference task on 20News-different dataset, in log-perplexity. SGGMCf: full batch; SGGMCb: mini-batch of size 50.

Questions on ParVIs and MCMCs

- ParVIs exhibit the intuition of minimizing $\mathrm{KL}_p(\cdot)$ on a probability space, along the speedest descending direction. Can this be made concrete?
  - Liu (2017) [45] conceives a probability manifold where SVGD simulates the gradient flow. But the validity of the artificial manifold is unknown.
- ParVIs do not assume a parametric statistical model, but need a kernel (or other treatment). Do they need an assumption / make an approximation?
- Do general MCMCs have a flow/optimization interpretation?

Things are made clear on the Wasserstein space.

The Wasserstein Space

For a metric space $(\mathcal{M}, d)$:
$\mathcal{P}_2(\mathcal{M}) := \{ q\text{: distribution on } \mathcal{M} \mid \exists x_0 \in \mathcal{M} \text{ s.t. } \mathbb{E}_q[d(x_0, x)^2] < +\infty \}$.

- $\mathcal{P}_2(\mathcal{M})$ is a metric space ([70], Def 6.4) with the Wasserstein distance:

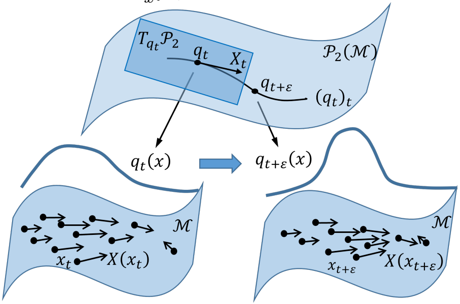
$$d_W(q, p) := \Big(\inf_{\pi \in \Pi(q,p)} \mathbb{E}_{\pi(x,y)}[d(x,y)^2] \Big)^{1/2},$$

where

$$\Pi(q,p) := \Big\{ \pi\text{: distribution on } \mathcal{M} \times \mathcal{M} \Big| \int_{\mathcal{M}} \pi(x,y) \, \mathrm{d}y = q(x), \int_{\mathcal{M}} \pi(x,y) \, \mathrm{d}x = p(y) \Big\}.$$

The Wasserstein Space: Riemannian Structure

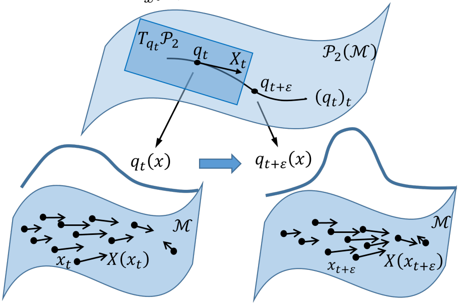
For a Riem. manif. $(\mathcal{M}, \langle\cdot,\cdot\rangle_{T_x\mathcal{M}})$, $\mathcal{P}_2(\mathcal{M})$ also has a Riem. str. [58, 70, 6]:



- Tangent vector $v \iff$ vector field X on $\mathcal{M}$.
- Tangent space at q: $T_q\mathcal{P}_2(\mathcal{M}) = \overline{\{\operatorname{grad} f \mid f \in \mathcal{C}_c^\infty(\mathcal{M})\}}^{\mathcal{L}_q^2(\mathcal{M})}$ is a subspace of $\mathcal{L}_q^2(\mathcal{M}) := \{X \mid \mathbb{E}_{q(x)}[\langle X(x), X(x)\rangle_{T_x\mathcal{M}}] < \infty\}$.
  ([70], Thm 13.8; [6], Thm 8.3.1, Def 8.4.1, Prop 8.4.5)

The Wasserstein Space: Riemannian Structure

For a Riem. manif. $(\mathcal{M}, \langle \cdot, \cdot \rangle_{T_x\mathcal{M}})$, $\mathcal{P}_2(\mathcal{M})$ also has a Riem. str. [58, 70, 6]:



- Riemannian structure: $T_q\mathcal{P}_2$ inherits the inner product of $\mathcal{L}_q^2$:

$$\langle X, Y \rangle_{T_q\mathcal{P}_2} = \mathbb{E}_{q(x)}\big[\langle X(x), Y(x) \rangle_{T_x\mathcal{M}}\big].$$

  It is consistent with d_W [8].

The Wasserstein Space: Riemannian Structure

- Gradient flow on $\mathcal{P}_2(\mathcal{M})$ for $\mathrm{KL}_p(q) := \mathbb{E}_q[\log(q/p)]$ (using Riem. meas):
  - $\mathcal{P}_2(\mathcal{M})$ as a Riemannian manifold:

$$V^{\mathsf{GF}} := -\operatorname{grad}\mathrm{KL}_p(q) = -\operatorname{grad}\left(\frac{\delta}{\delta q}\mathrm{KL}_p(q)\right) = \operatorname{grad}\log(p/q).$$

    ([70], Thm 23.18; [6], Example 11.1.2)
  - $\mathcal{P}_2(\mathcal{M})$ as a metric space: e.g., Minimizing Movement Scheme (MMS) ([6], Def. 2.0.6):

$$q_{t+\varepsilon} = \operatorname*{argmin}_{q \in \mathcal{P}_2(\mathcal{M})} \mathrm{KL}_p(q) + \frac{1}{2\varepsilon} d_W^2(q, q_t).$$

  They coincide under the Riemannian structure. ([70], Prop. 23.1, Rem. 23.4; [6], Thm. 11.1.6; [24], Lem. 2.7)
  Exponential convergence when p is log-concave. ([70], Thm 23.25, Thm 24.7; [6], Thm 11.1.4)

Langevin Dynamics as Wasserstein Gradient Flow

- The Langevin dynamics
$$\mathrm{d}x = \nabla \log p(x) \,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t(x)$$
produces the same [14] evolving distr. $(q_t)_t$ as:
$$\mathrm{d}x = \nabla \log(p(x)/q_t(x)) \,\mathrm{d}t,$$
which is the gradient flow of KL_p on $\mathcal{P}_2(\mathcal{M})$ for Euclidean $\mathcal{M}$.
- The gradient flow interpretation of LD is known earlier from the MMS perspective [34].

Understanding ParVIs on the Wasserstein Space

Understand and accelerate ParVIs from the Wasserstein gradient flow perspective [43].

- Consider Euclidean $\mathcal{M} = \mathbb{R}^m$ for brevity.

SVGD Approximates the Wasserstein Gradient Flow

Reformulate V^{GF} as:

$$V^{\mathsf{GF}} = \max_{V \in \mathcal{L}_q^2, \|V\|_{\mathcal{L}_q^2}=1} \cdot \operatorname{argmax} \langle V^{\mathsf{GF}}, V \rangle_{\mathcal{L}_q^2}. \quad (2)$$

We find:

Theorem 10 (V^{SVGD} approximates V^{GF})

$$V^{\mathsf{SVGD}} = \max_{V \in \mathcal{H}^D, \|V\|_{\mathcal{H}^D}=1} \cdot \operatorname{argmax} \langle V^{\mathsf{GF}}, V \rangle_{\mathcal{L}_q^2}.$$

- $\mathcal{H}^D$ is a subspace of $\mathcal{L}_q^2$, so V^{SVGD} is the projection of V^{GF} on $\mathcal{H}^D$.

ParVIs Approx. the Wass. Gradient Flow by Smoothing

Smoothing Functions

- SVGD restricts the optimization domain $\mathcal{L}_q^2$ to $\mathcal{H}^D$.

Theorem 11 ($\mathcal{H}^D$ smooths $\mathcal{L}_q^2$)

For $\mathcal{M} = \mathbb{R}^D$, a Gaussian kernel K on $\mathcal{M}$ and an absolutely continuous q, the vector-valued RKHS $\mathcal{H}^D$ of K is isometrically isomorphic to the closure $\mathcal{G} := \overline{\{\phi * K : \phi \in \mathcal{C}_c^\infty\}}^{\mathcal{L}_q^2}$.

$\overline{\mathcal{C}_c^\infty}^{\mathcal{L}_q^2} = \mathcal{L}_q^2$ ([36], Thm. 2.11) $\Longrightarrow$ $\mathcal{G}$ is roughly the kernel-smoothed $\mathcal{L}_q^2$.

Smoothing the Density

- The Blob method (w-SGLD-B) [14]: partially smooths the density.

$$V^{\text{GF}} = -\nabla\Big(\frac{\delta}{\delta q}\mathbb{E}_q[\log(q/p)]\Big) \Longrightarrow V^{\text{Blob}} = -\nabla\Big(\frac{\delta}{\delta q}\mathbb{E}_q[\log(\tilde{q}/p)]\Big),$$

where $\tilde{q} := q * K$ is the kernel-smoothed density.

ParVIs Approx. the Wass. Gradient Flow by Smoothing

- Equivalence:
  Smoothing-function objective $= \mathbb{E}_q[L(V)]$, $L : \mathcal{L}_q^2 \to L_q^2$ linear.
  $$\Longrightarrow \mathbb{E}_{\tilde{q}}[L(V)] = \mathbb{E}_{q*K}[L(V)] = \mathbb{E}_q[L(V) * K] = \mathbb{E}_q[L(V * K)].$$
- Necessity: $\operatorname{grad} \mathrm{KL}_p(q)$ undefined at $q = \hat{q} := \frac{1}{N} \sum_{i=1}^N \delta_{x^{(i)}}$.

Theorem 12 (Necessity of smoothing for SVGD)

For $q = \hat{q}$ and $V \in \mathcal{L}_p^2$, problem (2):

$$\max_{V \in \mathcal{L}_p^2, \|V\|_{\mathcal{L}_p^2} = 1} \left\langle V^{\mathsf{GF}}, V \right\rangle_{\mathcal{L}_{\hat{q}}^2},$$

has no optimal solution. In fact the supremum of the objective is infinite, indicating that a maximizing sequence of V tends to be ill-posed.

ParVIs rely on the smoothing assumption! No free lunch!

New ParVIs with Smoothing

- Gradient Flow with Smoothed Density (GFSD):
  Fully smooth the density:
  $$V^{\mathsf{GFSD}} := \nabla \log p - \nabla \log \tilde{q}.$$
- Gradient Flow with Smoothed test Functions (GFSF):
  $$V^{\mathsf{GF}} = \nabla \log p - \nabla \log q$$
  $$\implies V^{\mathsf{GF}} = \nabla \log p + \operatorname*{argmin}_{U \in \mathcal{L}^2} \max_{\substack{\phi \in \mathcal{C}_c^\infty, \\ \|\phi\|_{\mathcal{L}_q^2}=1}} \big(\mathbb{E}_q[\phi \cdot U - \nabla \cdot \phi]\big)^2.$$
  Smooth ϕ: take ϕ from $\mathcal{H}^D$:
  $$V^{\mathsf{GFSF}} := \nabla \log p + \operatorname*{argmin}_{U \in \mathcal{L}^2} \max_{\substack{\phi \in \mathcal{H}^D, \\ \|\phi\|_{\mathcal{H}^D}=1}} \big(\mathbb{E}_q[\phi \cdot U - \nabla \cdot \phi]\big)^2.$$
  Solution: $\hat{V}^{\mathsf{GFSF}} = \hat{V} + \hat{K}'\hat{K}^{-1}$. (Note $\hat{V}^{\mathsf{SVGD}} = \hat{V}^{\mathsf{GFSF}}\hat{K}$.)
  $\hat{V}_{:,i} = \nabla_{x^{(i)}} \log p(x^{(i)})$, $\hat{K}_{ij} = K(x^{(i)}, x^{(j)})$, $\hat{K}'_{:,i} = \sum_j \nabla_{x^{(j)}} K(x^{(j)}, x^{(i)})$.

Bandwidth Selection via the Heat Equation

Note

Under the dynamics $\mathrm{d}x = -\nabla \log q_t(x)\,\mathrm{d}t$, q_t evolves following the heat equation (HE): $\partial_t q_t(x) = \Delta q_t(x)$.

Smoothing the density: $q_t(x) \approx \tilde{q}(x) = \tilde{q}(x; \{x^{(i)}\}_{i=1}^N)$. Then for $q_{t+\varepsilon}(x)$,

- Due to HE, $q_{t+\varepsilon}(x) \approx \tilde{q}(x) + \varepsilon \Delta \tilde{q}(x)$.
- Due to the effect of the dynamics, updated particles $\{x^{(i)} - \varepsilon \nabla \log \tilde{q}(x^{(i)})\}_{i=1}^N$ approximate $q_{t+\varepsilon}$, so $q_{t+\varepsilon}(x) \approx \tilde{q}(x; \{x^{(i)} - \varepsilon \nabla \log \tilde{q}(x^{(i)})\}_{i=1}^N)$.

Objective: $\sum_k \left(\tilde{q}(x^{(k)}) + \varepsilon \Delta \tilde{q}(x^{(k)}) - \tilde{q}(x^{(k)}; \{x^{(i)} - \varepsilon \nabla \log \tilde{q}(x^{(i)})\}_{i=1}^N) \right)^2$.

Take $\varepsilon \to 0$, make the objective dimensionless (h/x^2 is dimensionless):

$\frac{1}{h^{D+2}} \sum_k \left[\Delta \tilde{q}(x^{(k)}; \{x^{(i)}\}_i) + \sum_j \nabla_{x^{(j)}} \tilde{q}(x^{(k)}; \{x^{(i)}\}_i) \cdot \nabla \log \tilde{q}(x^{(j)}; \{x^{(i)}\}_i) \right]^2$.

Also applicable to smoothing functions.

Bandwidth Selection via the Heat Equation

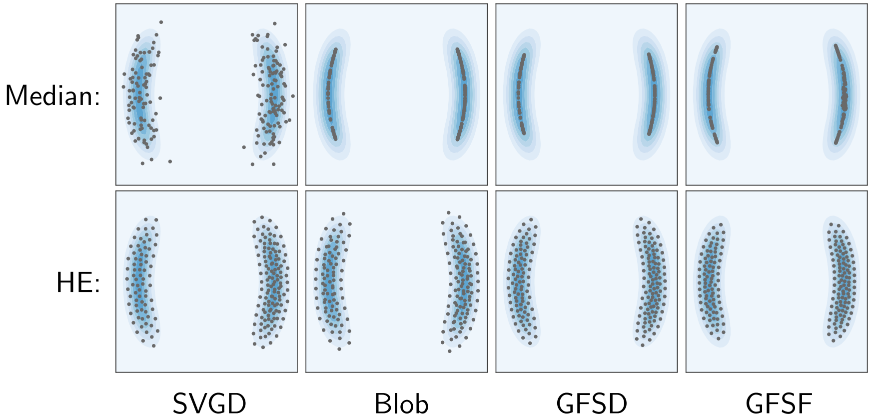


Figure: Comparison of HE (bottom row) with the median method (top row) for bandwidth selection.

Accelerated First-Order Methods on the Wasserstein Space

Nesterov's Acceleration Methods on Riemannian Manifolds:

$\rho_k \in \mathcal{P}_2(\mathcal{M})$: auxiliary variable. $V_k := -\operatorname{grad} \mathrm{KL}(\rho_k)$.

- Riemannian Accelerated Gradient (RAG) [47] (with simplification):

$$\begin{cases} q_k = \mathrm{Exp}_{\rho_{k-1}}(\varepsilon V_{k-1}), \\ \rho_k = \mathrm{Exp}_{q_k} \left[-\Gamma_{\rho_{k-1}}^{q_k} \left(\frac{k-1}{k} \mathrm{Exp}_{\rho_{k-1}}^{-1}(q_{k-1}) - \frac{k+\alpha-2}{k} \varepsilon V_{k-1} \right) \right]. \end{cases}$$

- Riemannian Nesterov's method (RNes) [74] (with simplification):

$$\begin{cases} q_k = \mathrm{Exp}_{\rho_{k-1}}(\varepsilon V_{k-1}), \\ \rho_k = \mathrm{Exp}_{q_k} \left\{ c_1 \mathrm{Exp}_{q_k}^{-1} \left[\mathrm{Exp}_{\rho_{k-1}} \left((1-c_2) \mathrm{Exp}_{\rho_{k-1}}^{-1}(q_{k-1}) + c_2 \mathrm{Exp}_{\rho_{k-1}}^{-1}(q_k) \right) \right] \right\}. \end{cases}$$

Required:

- Exponential map $\mathrm{Exp}_q : T_q\mathcal{P}_2(\mathcal{M}) \to \mathcal{P}_2(\mathcal{M})$ and its inverse.
- Parallel transport $\Gamma_q^\rho : T_q\mathcal{P}_2(\mathcal{M}) \to T_\rho\mathcal{P}_2(\mathcal{M})$.

Accelerated First-Order Methods on the Wasserstein Space

Leveraging the Riemannian Structure of $\mathcal{P}_2(\mathcal{M})$:

- Exponential map ([70], Coro. 7.22; [6], Prop. 8.4.6; [24], Prop. 2.1):
  $\mathrm{Exp}_q(V) = (\mathrm{id} + V)_\# q$, *i.e.*,
  $\{x^{(i)}\}_i \sim q \Rightarrow \{x^{(i)} + V(x^{(i)})\}_i \sim \mathrm{Exp}_q(V)$.
- Inverse exponential map: require the optimal transport map.
  - Sinkhorn methods [17, 72] appear costly and unstable.
  - Make approximations when $\{x^{(i)}\}_i$ and $\{y^{(i)}\}_i$ are pairwise close:
    $d(x^{(i)}, y^{(i)}) \ll \min\left\{ \min_{j \neq i} d(x^{(i)}, x^{(j)}), \min_{j \neq i} d(y^{(i)}, y^{(j)}) \right\}$.

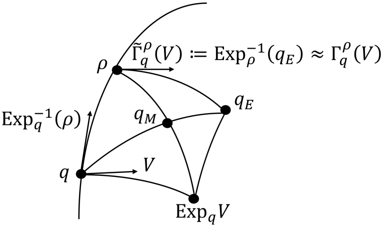
Proposition 13 (Inverse exponential map)

For pairwise close samples $\{x^{(i)}\}_i$ of q and $\{y^{(i)}\}_i$ of ρ, we have
$\big(\mathrm{Exp}_q^{-1}(\rho)\big)(x^{(i)}) \approx y^{(i)} - x^{(i)}$.

Accelerated First-Order Methods on the Wasserstein Space

Leveraging the Riemannian Structure of $\mathcal{P}_2(\mathcal{M})$:

- Parallel transport
  - Hard to implement analytical results [49, 50].
  - Use Schild's ladder method [23, 35] for approximation.



Proposition 14 (Parallel transport)

For pairwise close samples $\{x^{(i)}\}_i$ of q and $\{y^{(i)}\}_i$ of ρ, we have $\big(\Gamma_q^\rho(V)\big)(y^{(i)}) \approx V(x^{(i)})$, $\forall V \in T_q\mathcal{P}_2$.

Accelerated First-Order Methods on the Wasserstein Space

Algorithm 3 The acceleration framework with Wasserstein Accelerated Gradient (WAG) and Wasserstein Nesterov's method (WNes)

1: WAG: select acceleration factor $\alpha > 3$;
WNes: select or calculate $c_1, c_2 \in \mathbb{R}^+$;
2: Initialize $\{x_0^{(i)}\}_{i=1}^N$ distinctly; let $y_0^{(i)} = x_0^{(i)}$;
3: **for** $k = 1, 2, \cdots, k_{\max}$, **do**
4:     **for** $i = 1, \cdots, N$, **do**
5:         Find $V(y_{k-1}^{(i)})$ by SVGD/Blob/GFSD/GFSF;
6:         $x_k^{(i)} = y_{k-1}^{(i)} + \varepsilon V(y_{k-1}^{(i)})$;
7:         $y_k^{(i)} = x_k^{(i)} + \begin{cases} \text{WAG: } \frac{k-1}{k}(y_{k-1}^{(i)} - x_{k-1}^{(i)}) + \frac{k+\alpha-2}{k}\varepsilon V(y_{k-1}^{(i)}); \\ \text{WNes: } c_1(c_2-1)(x_k^{(i)} - x_{k-1}^{(i)}); \end{cases}$
8:     **end for**
9: **end for**
10: Return $\{x_{k_{\max}}^{(i)}\}_{i=1}^N$.

Accelerated First-Order Methods on the Wasserstein Space

Experimental results: Bayesian inference for Latent Dirichlet Allocation:

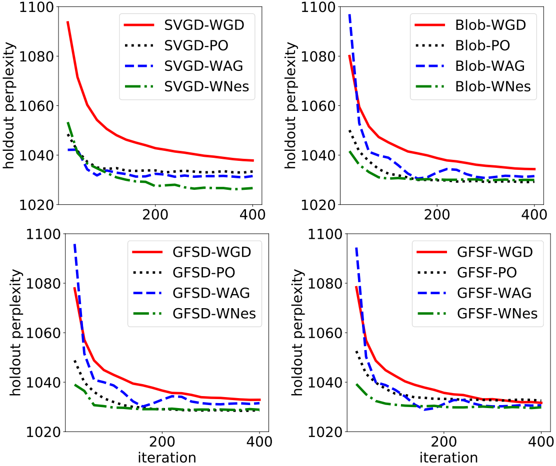


Figure: Acceleration effect of WAG and WNes on LDA (measured by hold-out perplexity).

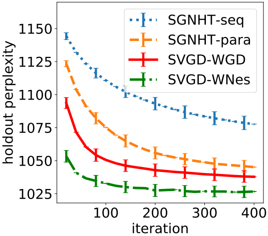


Figure: Comparison of SVGD and SGNHT on LDA, as representatives of ParVIs and MCMCs. Average over 10 runs.

Understanding MCMCs on the Wasserstein Space

Understanding MCMC dynamics as flows on the Wasserstein Space [44]:

- The Langevin dynamics (LD) is recognized as the Wasserstein gradient flow of the KL divergence [34].
  - Benefits its asymptotic [63] and non-asymptotic [22, 16] behaviors.
  - Relates it to ParVIs [14, 43].
- Does a general MCMC dynamics correspond to an interpretable flow on the Wasserstein space?

The First Reformulation

Lemma 15 (Equivalent deterministic MCMC dynamics)

A general MCMC dynamics specified by a symm. pos. semi-def. D *and skew-symm.* Q *via Eq.* (1) *produces the same distr. evolution as the deterministic dynamics:*

$$\mathrm{d}x = W_t(x)\,\mathrm{d}t,$$
$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x), \tag{3}$$

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

The First Reformulation

Lemma 15 (Equivalent deterministic MCMC dynamics)

A general MCMC dynamics specified by a symm. pos. semi-def. D *and skew-symm.* Q *via Eq.* (1) *produces the same distr. evolution as the deterministic dynamics:*

$$\mathrm{d}x = W_t(x)\,\mathrm{d}t,$$
$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x), \quad (3)$$

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

- $\Longrightarrow$ Barbour's generator [7] $\mathcal{A}f := \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{q_t}[f]\big|_{q_t=\delta_x} = \frac{1}{p}\partial_j\left[p\left(D^{ij}+Q^{ij}\right)(\partial_i f)\right]$ (c.f. [29]).

The First Reformulation

Lemma 15 (Equivalent deterministic MCMC dynamics)

A general MCMC dynamics specified by a symm. pos. semi-def. D and skew-symm. Q via Eq. (1) produces the same distr. evolution as the deterministic dynamics:

$$\mathrm{d}x = W_t(x)\,\mathrm{d}t,$$
$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x), \tag{3}$$

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

- $\Longrightarrow$ Barbour's generator [7] $\mathcal{A}f := \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{q_t}[f]\big|_{q_t=\delta_x} = \frac{1}{p}\partial_j\left[p\left(D^{ij}+Q^{ij}\right)(\partial_i f)\right]$ (c.f. [29]).

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 $\mathcal{P}_2(\mathcal{M})$.

- Euclidean $\mathcal{M}$: $D = I$.
- Hilbert $\mathcal{M}$: constant and non-singular D.
- Riemannian $\mathcal{M}$: non-singular $D(x)$.

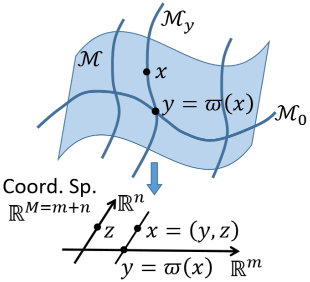
We need positive *semi*-definite $D(x)$.

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 $\mathcal{P}_2(\mathcal{M})$.

- Fiber Bundle $\mathcal{M}$ (of dim. $M = m + n$) *(known knowledge)*:
  - $\mathcal{M}$ is locally $\mathcal{M}_0 \times \mathcal{F}$ ($\dim(\mathcal{M}_0) = m$, $\dim(\mathcal{F}) = n$) [57] in terms of a projection ϖ:
    $$\varpi : \mathcal{M} \to \mathcal{M}_0 \overset{\text{locally}}{\iff} \mathcal{M}_0 \times \mathcal{F} \to \mathcal{M}_0.$$
  - 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

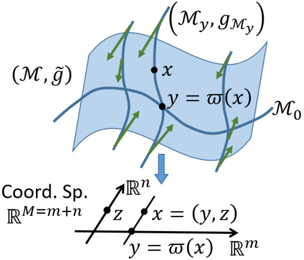
$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}_2(\mathcal{M})$.

- 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

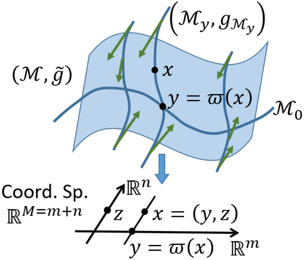
$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}_2(\mathcal{M})$.

- 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 $\mathcal{M}_y$:

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

- 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

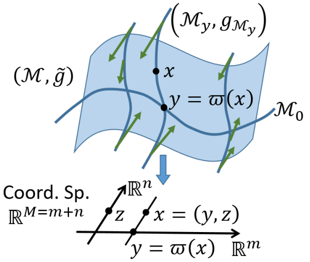
$$(W_t)^i(x) = D^{ij}(x)\,\partial_j \log(p(x)/q_t(x)) + Q^{ij}(x)\,\partial_j \log p(x) + \partial_j Q^{ij}(x).$$

1 $D^{ij}(x)\,\partial_j \log(p(x)/q_t(x))$ seems like a gradient flow on $\mathcal{P}_2(\mathcal{M})$.

- 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:

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

$$\left(\tilde{g}^{ij}(x)\right)_{M\times M} := \begin{pmatrix} 0_{m\times m} & 0_{m\times n} \\ 0_{n\times m} & \left((g_{\mathcal{M}_{\varpi(x)}}(z))^{ab}\right)_{n\times n} \end{pmatrix}. \quad (4)$$

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 $\mathcal{P}_2(\mathcal{M})$.

- Structures on $\mathcal{P}_2(\mathcal{M})$ with fiber-Riemannian $\mathcal{M}$.
  - Hard to decompose $\mathcal{P}_2(\mathcal{M})$.
  - $\widetilde{\mathcal{P}}_2(\mathcal{M}) := \{q(z|y) \in \mathcal{P}_2(\mathcal{M}_y) \mid y \in \mathcal{M}_0\} \overset{\text{locally}}{\Longleftrightarrow} \mathcal{M}_0 \times \mathcal{P}_2(\mathcal{M}_y)$: fiber-Riemannian!
  - On $\mathcal{P}_2(\mathcal{M}_y)$, $\left(\operatorname{grad} \mathrm{KL}_{p(\cdot|y)}(q(\cdot|y))(z)\right)^a$ $= (g_{\mathcal{M}_y}(z))^{ab}\,\partial_{z^b} \log \frac{q(z|y)}{p(z|y)} = (g_{\mathcal{M}_y}(z))^{ab}\,\partial_{z^b} \log \frac{q(y,z)}{p(y,z)}, 1 \leqslant a, b \leqslant n.$
  - Taking union over $y \in \mathcal{M}_0$, the fiber-gradient on $\widetilde{\mathcal{P}}_2(\mathcal{M})$ is:
$$\left(\operatorname{grad}_{\text{fib}} \mathrm{KL}_p(q)(x)\right)_M = \left(\tilde{g}^{ij}(x)\,\partial_j \log\left(q(x)/p(x)\right)\right)_M.$$

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 $\mathcal{P}_2(\mathcal{M})$.

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

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 symm. 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 support $(\mathcal{M}, \tilde{g})$ 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: $\mathcal{M} = \mathbb{R}^{2\ell}$, $Q = \begin{pmatrix} 0 & I_\ell \\ -I_\ell & 0 \end{pmatrix}$.
- Symplectic manifold [18, 52]: $\mathcal{M}$ even-dim., Q non-singular.
- Poisson manifold $\mathcal{M}$ [25]:
  - Poisson structure: bivector field $\beta = \beta^{ij}\partial_i \otimes \partial_j = \sum_{i<j} \beta^{ij}\partial_i \wedge \partial_j$ (anti-symm. 2nd-order contravariant tensor field; (β_{ij}) is *skew-symm.*) that satisfies the Jacobian identity:

$$\beta^{il}\partial_l\beta^{jk} + \beta^{jl}\partial_l\beta^{ki} + \beta^{kl}\partial_l\beta^{ij} = 0, \forall i, j, k. \tag{5}$$

  - Hamiltonian flow X_f of a smooth function f:

$$\big(X_f(x)\big)[h] := \big(\beta(\mathrm{d}f, \mathrm{d}h)\big)(x) = \beta^{ij}(x)\,\partial_i f(x)\,\partial_j h(x).$$

    Coordinate expression: $\big(X_f(x)\big)^i = \beta^{ij}(x)\,\partial_j f(x)$.
    X_f conserves f: $\frac{\mathrm{d}}{\mathrm{d}t} f(\varphi_t) = 0$.

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 $\mathcal{P}_2(\mathcal{M})$ [49, 5, 26] *(known knowledge)*:
  - Hamiltonian flow of a function F on $\mathcal{P}_2(\mathcal{M})$:

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

    where func. f on $\mathcal{M}$ relates to F via $\mathrm{grad}_q \mathbb{E}_q[f] = \mathrm{grad}_q F(q)$, and π_q is the orthogonal projection $\mathcal{L}_q^2(\mathcal{M}) \to T_q\mathcal{P}_2(\mathcal{M})$, which does not change distribution evolution.
- Hamiltonian flow of KL on $\mathcal{P}_2(\mathcal{M})$:

Lemma 2 (Hamiltonian flow of KL on $\mathcal{P}_2(\mathcal{M})$)

The Hamiltonian flow of KL_p on $\mathcal{P}_2(\mathcal{M})$ is:

$\mathcal{X}_{\mathrm{KL}_p}(q) = \pi_q(X_{\log(q/p)})$, *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.

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

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 symm. positive definite* $C(x)$.

(b) $Q(x)$ *satisfies Eq.* (5)*:* $Q^{il}\partial_l Q^{jk} + Q^{jl}\partial_l Q^{ki} + Q^{kl}\partial_l Q^{ij} = 0, \forall i, j, k.$

- Satisfied by MCMCs except for SGNHT-related methods [20, 75].
- 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 symm. positive definite* $C(x)$.

(b) $Q(x)$ *satisfies Eq.* (5): $Q^{il}\partial_l Q^{jk} + Q^{jl}\partial_l Q^{ki} + Q^{kl}\partial_l Q^{ij} = 0, \forall i, j, k.$

- Satisfied by MCMCs except for SGNHT-related methods [20, 75].
- Required to match Poisson structure; unnecessary for conservation of Hamiltonian.

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

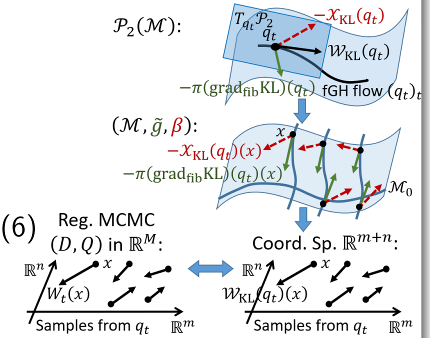
Interpret MCMC Dynamics: Main Theorem

Theorem 5 (Equivalence between regular MCMC dynamics on $\mathbb{R}^M$ and fGH flows on $\mathcal{P}_2(\mathcal{M})$.)

We call $(\mathcal{M}, \tilde{g}, \beta)$ a fiber-Riemannian Poisson (fRP) manifold, and define the fiber-gradient Hamiltonian (fGH) flow on $\mathcal{P}_2(\mathcal{M})$ as:

$$\mathcal{W}_{\mathrm{KL}_p} := -\pi(\mathrm{grad}_{\mathrm{fib}} \mathrm{KL}_p) - \mathcal{X}_{\mathrm{KL}_p},$$

$$\left(\mathcal{W}_{\mathrm{KL}_p}(q)\right)^i = \pi_q\left((\tilde{g}^{ij} + \beta^{ij})\partial_j \log(p/q)\right).$$



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. (4)).

- $\mathcal{M}_0$ 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 $\mathcal{P}_2(\mathcal{M})$.
  - $-\mathcal{X}_{\mathrm{KL}_p}$: conserves KL_p on $\mathcal{P}_2(\mathcal{M})$ and helps mixing/exploration.
- Converges to p uniquely (c.f. [51]).
- Robust to SG (c.f. [65, 69]).

Instances:

- LD [62] / SGLD [71]: $Q = 0$, $\mathcal{M}$ is Euclidean.
- RLD [28] / SGRLD [60]: $Q = 0$, $\mathcal{M}$ is the manifold under consideration.

Interpret MCMC Dynamics: Case Study

Type 2: $D = 0$ ($n = 0$ in Eq. (4)).

- $\mathcal{M}_0 = \mathcal{M}$, 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 $\mathcal{P}_2(\mathcal{M})$ and helps mixing/exploration.
- Fragile against SG: no stablizing forces (i.e. (fiber-)gradient flows) (c.f. [15, 9]).
- Hard to extend to ParVIs.

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

- HMC [21, 56, 10] ($\mathcal{S} = \mathbb{R}^\ell$): $\mathcal{M} = \mathbb{R}^{2\ell}$.
- HMC relies on *geometric ergodicity* for convergence [48, 10].
- RHMC [28] / LagrMC [38] / GMC [12] (manifold $\mathcal{S}$): $\mathcal{M} = T^*\mathcal{S}$.

Interpret MCMC Dynamics: Case Study

Type 3: $D \neq 0$ and D is singular ($m, n \geqslant 1$ in Eq. (4)).

- 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 $\mathrm{KL}_{p(\cdot|y)}(q(\cdot|y))$ steepest on each fiber $\mathcal{P}_2(\mathcal{M}_y)$.
  - $-\mathcal{X}_{\mathrm{KL}_p}$: conserves KL_p on $\mathcal{P}_2(\mathcal{M})$ and helps mixing/exploration.
- Robust to SG (SG appears on each fiber) (c.f. [15, 13]).

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

- SGHMC [15] ($\mathcal{S} = \mathbb{R}^\ell$), SGRHMC [51] / SGGMC [42] (manifold $\mathcal{S}$): $\mathcal{M}_0 = \mathcal{S}$, $\mathcal{M}_\theta = T^*_\theta \mathcal{S}$.
- SGNHT [20] ($\mathcal{S} = \mathbb{R}^\ell$), gSGNHT [42] (manifold $\mathcal{S}$): $\mathcal{M}_0 = \mathcal{S}$, $\mathcal{M}_\theta = \mathbb{R} \times T^*_\theta \mathcal{S}$.

ParVI Simulation for SGHMC

Simulate the deterministic dynamics of SGHMC: (take $C(x)$ as constant $C\Sigma$)

By Lemma 15 (Eq. (3)): $\begin{cases} \frac{\mathrm{d}\theta}{\mathrm{d}t} = \Sigma^{-1} r, \\ \frac{\mathrm{d}r}{\mathrm{d}t} = \nabla_\theta \log p(\theta) - Cr - C\Sigma \nabla_r \log q(r). \end{cases}$

By Theorem 5 (Eq. (6)): $\begin{cases} \frac{\mathrm{d}\theta}{\mathrm{d}t} = \Sigma^{-1} r + \nabla_r \log q(r), \\ \frac{\mathrm{d}r}{\mathrm{d}t} = \nabla_\theta \log p(\theta) - Cr - C\Sigma \nabla_r \log q(r) - \nabla_\theta \log q(\theta). \end{cases}$

To estimate $\nabla \log q$ with particles, use ParVI techniques [43], e.g. Blob [14]:

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

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

ParVI Simulation for SGHMC

Simulate the deterministic dynamics of SGHMC:

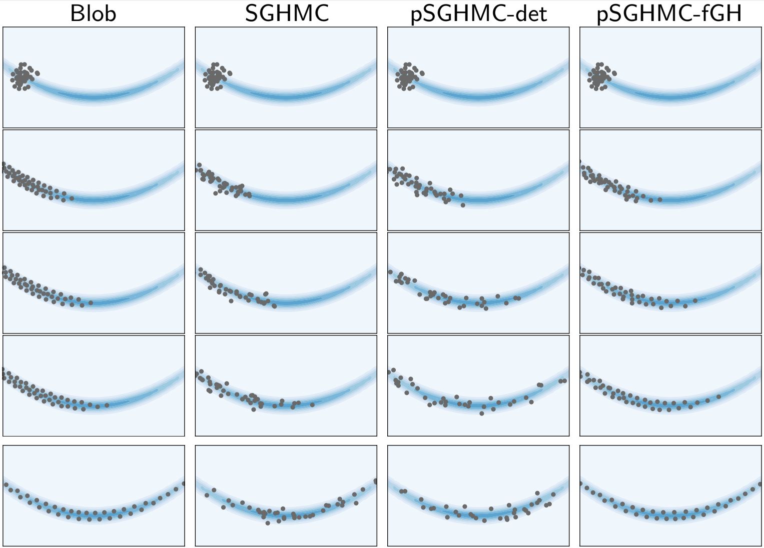
$$\text{pSGHMC-det:} \begin{cases} \frac{\Delta\theta^{(i)}}{\varepsilon} = \Sigma^{-1} r^{(i)}, \\ \frac{\Delta r^{(i)}}{\varepsilon} = \nabla_\theta \log p(\theta^{(i)}) - C r^{(i)} - C\Sigma\left(\frac{\sum_k \nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(i,j)}} + \sum_k \frac{\nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j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

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

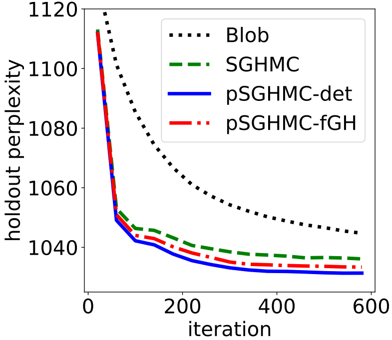
Advantages:

- Over SGHMC: particle-efficiency, ParVI techniques like HE [43].
- Over ParVIs: more efficient dynamics over LD.

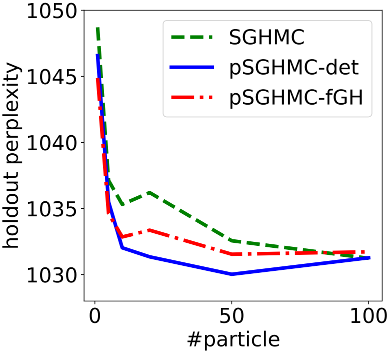
Experimental Results: Synthetic



Experimental Results: Latent Dirichlet Allocation (LDA)



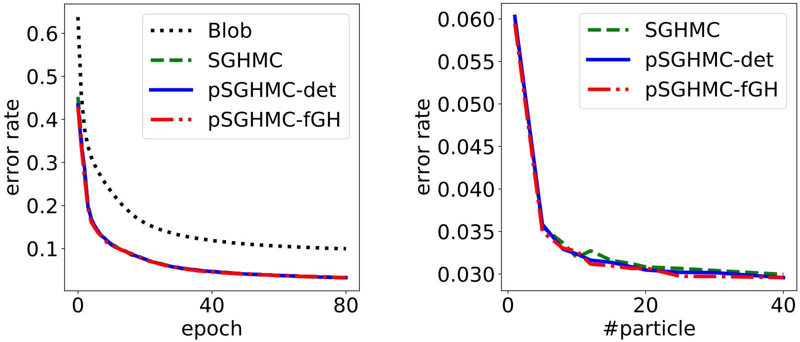
(a) Learning curve (20 ptcls)



(b) Particle efficiency (iter 600)

Figure: Performance on LDA with the ICML data set.

Experimental Results: Bayesian Neural Networks (BNNs)



(a) Learning curve (10 ptcls) (b) Particle efficiency (epch 80)

Figure: Performance on BNN with MNIST data set.

Thanks!
Questions?

Ralph Abraham, Jerrold E Marsden, and Jerrold E Marsden.
Foundations of mechanics.
Benjamin/Cummings Publishing Company Reading, Massachusetts, Providence, Rhode Island, 1978.

Ralph Abraham, Jerrold E Marsden, and Tudor Ratiu.
Manifolds, tensor analysis, and applications, volume 75.
Springer Science & Business Media, New York, 2012.

Shun-Ichi Amari.
Information geometry and its applications.
Springer, Tokyo, 2016.

Shun-Ichi Amari and Hiroshi Nagaoka.
Methods of information geometry, volume 191.
American Mathematical Soc., Providence, Rhode Island, 2007.

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.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré.
Gradient flows: in metric spaces and in the space of probability measures.
Springer Science & Business Media, Berlin, 2008.

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

Jean-David Benamou and Yann Brenier.
A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem.
Numerische Mathematik, 84(3):375–393, 2000.

Michael Betancourt.
The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling.
In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pages 533–540, Lille, France, 2015. IMLS.

Michael Betancourt.
A conceptual introduction to Hamiltonian Monte Carlo.
arXiv preprint arXiv:1701.02434, 2017.

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-12)*, pages 161–172, La Palma, Canary Islands, 2012. AISTATS Committee.

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 2269–2277, Montréal, Canada, 2015. NIPS Foundation.

Changyou Chen, Ruiyi Zhang, Wenlin Wang, Bai Li, and Liqun Chen.
A unified particle-optimization framework for scalable Bayesian sampling.
In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, Monterey, California USA, 2018. Association for Uncertainty in Artificial Intelligence.

Tianqi Chen, Emily Fox, and Carlos Guestrin.

Stochastic gradient Hamiltonian Monte Carlo.

In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pages 1683–1691, Beijing, China, 2014. IMLS.

Xiang Cheng and Peter Bartlett.

Convergence of Langevin MCMC in KL-divergence.

arXiv preprint arXiv:1705.09048, 2017.

Marco Cuturi.

Sinkhorn distances: Lightspeed computation of optimal transport.

In *Advances in Neural Information Processing Systems*, pages 2292–2300, Lake Tahoe, Nevada USA, 2013. NIPS Foundation.

Ana Cannas Da Silva.

Lectures on symplectic geometry, volume 3575.

Springer, Boston, 2001.

Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak.

Hyperspherical variational auto-encoders.

arXiv preprint arXiv:1804.00891, 2018.

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, Montréal, Canada, 2014. NIPS Foundation.

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.

J Ehlers, F Pirani, and A Schild.
The geometry of free fall and light propagation, in the book "General Relativity" (papers in honour of JL Synge), 63–84, 1972.

Matthias Erbar et al.
The heat equation on manifolds as a gradient flow in the Wasserstein space.
In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 46, pages 1–23, Paris, 2010. Institut Henri Poincaré.

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

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

Stuart Geman and Donald Geman.
Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.
In *Readings in Computer Vision*, pages 564–584. Elsevier, Los Altos, California USA, 1987.

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.

Daniele Grattarola, Lorenzo Livi, and Cesare Alippi.
Adversarial autoencoders with constant-curvature latent manifolds.
arXiv preprint arXiv:1812.04314, 2018.

W Keith Hastings.
Monte Carlo sampling methods using Markov chains and their applications.
Biometrika, 57(1):97–109, 1970.

Heinz Hopf and Willi Rinow.
Über den begriff der vollständigen differential geometrischen fläche.
Commentarii Mathematici Helvetici, 3(1):209–225, 1931.

I. M. James.
The topology of Stiefel manifolds, volume 24.
Cambridge University Press, New York, 1976.

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.

Arkady Kheyfets, Warner A Miller, and Gregory A Newton.
Schild's ladder parallel transport procedure for an arbitrary connection.

International Journal of Theoretical Physics, 39(12):2891–2898, 2000.

Ondrej Kováčik and Jiří Rákosník.
On spaces $L^p(x)$ and $W^{k,p}(x)$.
Czechoslovak Mathematical Journal, 41(4):592–618, 1991.

Hendrik Anthony Kramers.
Brownian motion in a field of force and the diffusion model of chemical reactions.
Physica, 7(4):284–304, 1940.

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.

Paul Langevin.
Sur la théorie du mouvement Brownien.
Compt. Rendus, 146:530–533, 1908.

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.

Chang Liu and Jun Zhu.

Riemannian Stein variational gradient descent for Bayesian inference.

In *The 32nd AAAI Conference on Artificial Intelligence*, pages 3627–3634, New Orleans, Louisiana USA, 2018. AAAI press.

Chang Liu, Jun Zhu, and Yang Song.

Stochastic gradient geodesic MCMC methods.

In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3009–3017. Curran Associates, Inc., Barcelona, Spain, 2016.

Chang Liu, Jingwei Zhuo, Pengyu Cheng, Ruiyi Zhang, Jun Zhu, and Lawrence Carin.

Understanding and accelerating particle-based variational inference.

In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4082–4092, Long Beach, California USA, 09–15 Jun 2019. IMLS, PMLR.

Chang Liu, Jingwei Zhuo, and Jun Zhu.

Understanding MCMC dynamics as flows on the Wasserstein space.

In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4093–4103, Long Beach, California USA, 09–15 Jun 2019. IMLS, PMLR.

Qiang Liu.
Stein variational gradient descent as gradient flow.
In *Advances in Neural Information Processing Systems*, pages 3118–3126, Long Beach, California USA, 2017. NIPS Foundation.

Qiang Liu and Dilin Wang.
Stein variational gradient descent: A general purpose Bayesian inference algorithm.
In *Advances in Neural Information Processing Systems*, pages 2370–2378, Barcelona, Spain, 2016. NIPS Foundation.

Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, and Licheng Jiao.
Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds.
In *Advances in Neural Information Processing Systems*, pages 4875–4884, Long Beach, California USA, 2017. NIPS Foundation.

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.

John Lott.
An intrinsic parallel transport in Wasserstein space.
Proceedings of the American Mathematical Society, 145(12):5329–5340, 2017.

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, Montréal, Canada, 2015. NIPS Foundation.

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, Berlin, 2013.

Emile Mathieu, Charline Le Lan, Chris J Maddison, Ryota Tomioka, and Yee Whye Teh.
Hierarchical representations with Poincaré variational auto-encoders.

arXiv preprint arXiv:1901.06033, 2019.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller.
Equation of state calculations by fast computing machines.
The journal of chemical physics, 21(6):1087–1092, 1953.

Yoshihiro Nagano, Shoichiro Yamaguchi, Yasuhiro Fujita, and Masanori Koyama.
A differentiable Gaussian-like distribution on hyperbolic space for gradient-based learning.
arXiv preprint arXiv:1902.02992, 2019.

Radford M. Neal.
MCMC using Hamiltonian dynamics.
Handbook of Markov Chain Monte Carlo, 2, 2011.

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

Felix Otto.
The geometry of dissipative evolution equations: the porous medium equation.
2001.

Ivan Ovinnikov.
Poincaré Wasserstein autoencoder.
arXiv preprint arXiv:1901.01427, 2019.

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, Lake Tahoe, Nevada USA, 2013. NIPS Foundation.

Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J. Mooney.
Spherical topic models.
In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pages 903–910, Haifa, Israel, 2010. IMLS.

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.

Ruslan Salakhutdinov and Andriy Mnih.
Bayesian probablistic matrix factorization using Markov chain Monte Carlo.
In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, pages 880–887, Helsinki, Finland, 2008. IMLS, Omnipress.

Issei Sato and Hiroshi Nakagawa.
Approximation analysis of stochastic gradient Langevin dynamics by using Fokker-Planck equation and Ito process.
In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pages 982–990, Beijing, China, 2014. IMLS.

Yang Song and Jun Zhu.
Bayesian matrix completion via adaptive relaxed spectral regularization.
In *The 30th AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 2044–2050, Phoenix, Arizona USA, 2016. AAAI press.

Ingo Steinwart and Andreas Christmann.
Support vector machines.
Springer Science & Business Media, New York, 2008.

Eduard L. Stiefel.
Richtungsfelder und fernparallelismus in n-dimensionalen mannigfaltigkeiten.

Commentarii Mathematici Helvetici, 8(1):305–353, 1935.

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.

Cédric Villani.
Optimal transport: old and new, volume 338.
Springer Science & Business Media, Berlin, 2008.

Max Welling and Yee Whye Teh.
Bayesian learning via stochastic gradient Langevin dynamics.
In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pages 681–688, Bellevue, Washington USA, 2011. IMLS.

Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha.
A fast proximal point method for computing Wasserstein distance.
arXiv preprint arXiv:1802.04307, 2018.

Viktor Yanush and Dmitry Kropotov.
Hamiltonian Monte-Carlo for orthogonal matrices.
arXiv preprint arXiv:1901.08045, 2019.

Hongyi Zhang and Suvrit Sra.
An estimate sequence for geodesically convex optimization.
In *Proceedings of the 31st Annual Conference on Learning Theory (COLT 2018)*, pages 1703–1723, Stockholm, Sweden, 2018. IMLS.

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