

Understanding MCMC Dynamics as Flows on the Wasserstein Space

Chang Liu, Jingwei Zhuo, Jun Zhu¹

Department of Computer Science and Technology, Tsinghua University

chang-li14@mails.tsinghua.edu.cn

ICML 2019

¹Corresponding author.

Introduction

- Langevin dynamics (LD) \iff gradient flow on the Wasserstein space of a Euclidean space [11].
- Does a general MCMC dynamics have such an explanation?

Introduction

- Langevin dynamics (LD) \iff gradient flow on the Wasserstein space of a Euclidean space [11].
- Does a general MCMC dynamics have such an explanation?

In this work:

- General MCMC dynamics \iff fiber-Gradient Hamiltonian (fGH) flow on the Wasserstein space of a fiber-Riemannian Poisson (fRP) manifold.
- “fGH flow = min-KL flow + const-KL flow” explains the behavior of MCMCs.
- The connection to particle-based variational inference (ParVI) inspires new methods.

First Reformulation

Describe a general MCMC dynamics targeting p [15]:

$$dx = V(x) dt + \sqrt{2D(x)} dB_t(x),$$
$$V^i(x) = \frac{1}{p(x)} \partial_j \left(p(x) (D^{ij}(x) + Q^{ij}(x)) \right),$$

for some pos. semi-def. D and skew-symm. Q .

First Reformulation

Describe a general MCMC dynamics targeting p [15]:

$$dx = V(x) dt + \sqrt{2D(x)} dB_t(x),$$

$$V^i(x) = \frac{1}{p(x)} \partial_j \left(p(x) (D^{ij}(x) + Q^{ij}(x)) \right),$$

for some pos. semi-def. D and skew-symm. Q .

Lemma 1 (Equivalent deterministic MCMC dynamics)

$$dx = W_t(x) dt,$$

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

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

Gradient flow of KL_p on $\mathcal{P}(\mathcal{M})$ with Riemannian (\mathcal{M}, g) :

$$-\text{grad}_{\mathcal{P}(\mathcal{M})} \text{KL}_p(q) = -\text{grad}_{\mathcal{M}} \log(q/p) = g^{ij}(x) \partial_j \log(p(x)/q(x)).$$

(g^{ij}) : symm. pos. def.

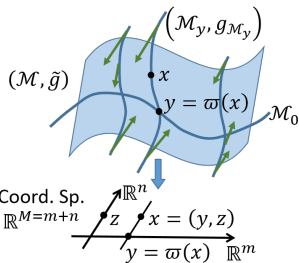
Interpret MCMC Dynamics

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

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

Definition 3 (Fiber-Riemannian manifold)

Fiber-Riemannian manifold: a fiber bundle with a Riem. strc. $g_{\mathcal{M}_y}$ on each fiber \mathcal{M}_y .



- Fiber-gradient: union of grad. over fibers

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

$$(\tilde{g}^{ij}(x))_{M \times M} := \begin{pmatrix} 0_{m \times m} & 0_{m \times n} \\ 0_{n \times m} & ((g_{\mathcal{M}_{\varpi(x)}}(z))^{ab})_{n \times n} \end{pmatrix}. \quad (1)$$

- On $\tilde{\mathcal{P}}(\mathcal{M})$: $(\text{grad}_{\text{fib}} \text{KL}_p(q)(x))_M = (\tilde{g}^{ij}(x) \partial_j \log(q(x)/p(x)))_M$.

Interpret MCMC Dynamics

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

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

Consider a Poisson manifold (\mathcal{M}, β) [8].

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

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

$\mathcal{X}_{\text{KL}_p}$ conserves KL_p on $\mathcal{P}(\mathcal{M})$ [1, 9].

Interpret MCMC Dynamics: Main Theorem

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

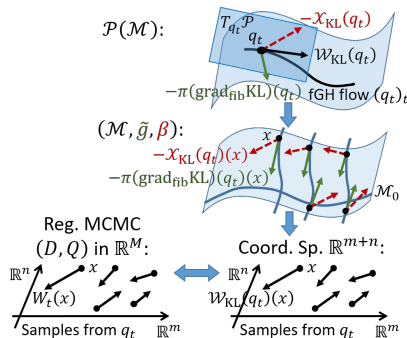
We call $(\mathcal{M}, \tilde{g}, \beta)$ a fiber-Riemannian Poisson (fRP) manifold, and define the fiber-gradient Hamiltonian (fGH) flow on $\mathcal{P}(\mathcal{M})$ as:

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

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

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. (1)).

- fGH flow $\mathcal{W}_{\text{KL}_p} = -\pi(\text{grad } \text{KL}_p) - \mathcal{X}_{\text{KL}_p}$,
 - $-\pi(\text{grad } \text{KL}_p)$: minimizes KL_p on $\mathcal{P}(\mathcal{M})$.
 - $-\mathcal{X}_{\text{KL}_p}$: conserves KL_p on $\mathcal{P}(\mathcal{M})$, helps mixing/exploration.
- LD [18] / SGLD [19], RLD [10] / SGRLD [17].

Type 2: $D = 0$ ($n = 0$ in Eq. (1)).

- fGH flow $\mathcal{W}_{\text{KL}_p} = -\mathcal{X}_{\text{KL}_p}$ conserves KL_p on $\mathcal{P}(\mathcal{M})$.
- Fragile against SG: no stabilizing forces (i.e. (fiber-)gradient flows).
- HMC [7, 16, 2], RHMC [10] / LagrMC [12] / GMC [3].

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

- fGH flow $\mathcal{W}_{\text{KL}_p} = -\pi(\text{grad}_{\text{fib}} \text{KL}_p) - \mathcal{X}_{\text{KL}_p}$,
 - $-\pi(\text{grad}_{\text{fib}} \text{KL}_p)$: minimizes $\text{KL}_{p(\cdot|y)}(q(\cdot|y))$ on each fiber $\mathcal{P}(\mathcal{M}_y)$.
 - $-\mathcal{X}_{\text{KL}_p}$: conserves KL_p on $\mathcal{P}(\mathcal{M})$, helps mixing/exploration.
- Robust to SG (SG appears on each fiber).
- SGHMC [5], SGRHMC [15]/SGGMC [13], SGNHT [6]/gSGNHT [13].

ParVI Simulation for SGHMC

Deterministic dynamics of SGHMC [5]:

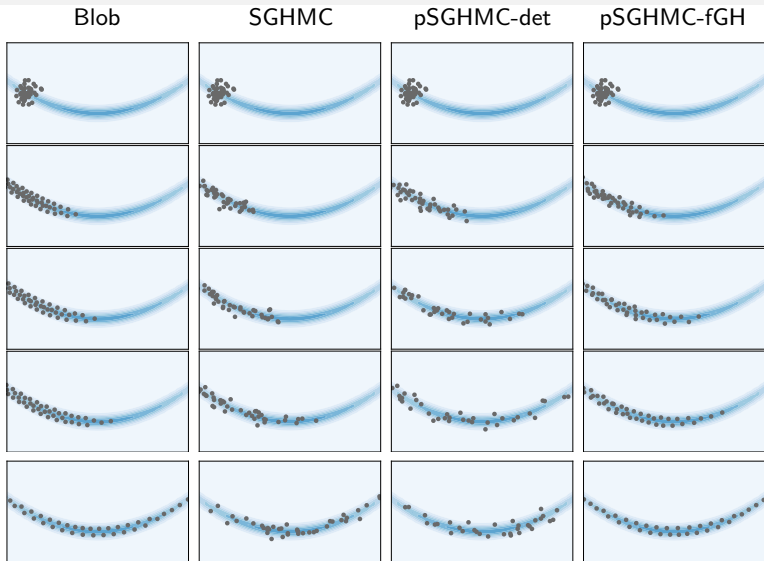
$$\text{By Lemma 1: } \left\{ \begin{array}{l} \frac{d\theta}{dt} = \Sigma^{-1}r, \\ \text{pSGHMC-det } \frac{dr}{dt} = \nabla_{\theta} \log p(\theta) - C\Sigma^{-1}r - C\nabla_r \log q(r). \end{array} \right.$$

$$\text{By Theorem 5: } \left\{ \begin{array}{l} \frac{d\theta}{dt} = \Sigma^{-1}r + \nabla_r \log q(r), \\ \text{pSGHMC-fGH } \frac{dr}{dt} = \nabla_{\theta} \log p(\theta) - C\Sigma^{-1}r - C\nabla_r \log q(r) - \nabla_{\theta} \log q(\theta). \end{array} \right.$$

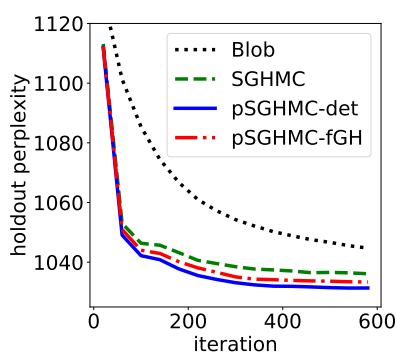
Estimate $\nabla \log q$ using ParVI techniques [14], e.g. Blob [4].

- Over SGHMC: particle-efficient.
- Over ParVIs: more efficient dynamics than LD.

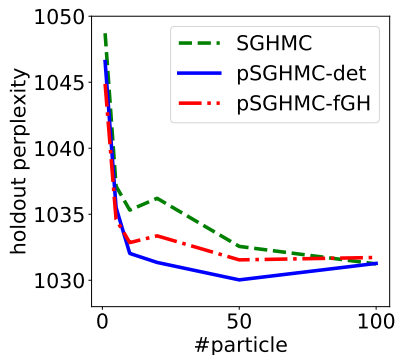
Synthetic Experiment



Latent Dirichlet Allocation (LDA)



(a) Learning curve (20 ptcls)



(b) Particle efficiency (iter 600)

Figure: Performance on LDA with the ICML data set.



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.



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, 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.

-  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.
-  Rui Loja Fernandes and Ioan Marcuț.
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.

-  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.
-  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.
-  Radford M Neal et al.

Mcmc using hamiltonian dynamics.

Handbook of Markov Chain Monte Carlo, 2(11), 2011.



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.



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.