Chapter 10

Geometry in sampling methods: A review on manifold MCMC and particle-based variational inference methods

Chang Liu[a] **and Jun Zhu**[b,*]

[a]*Microsoft Research Asia, Beijing, China*

[b]*Department of Computer Science and Technology, Beijing National Center for Information Science and Technology, Tsinghua-Bosch Joint Center for ML, Tsinghua University, Beijing, China*

[*]*Corresponding author: e-mail: dcszj@mail.tsinghua.edu.cn*

Abstract

Sampling methods are an indispensable tool for Bayesian inference, as they provide a flexible and asymptotically exact approximation to the intractable posterior in an out-of-the-box way. These methods generate or update the samples by simulating a dynamical process, which is a construct on a space with certain geometry. Non-Euclidean geometry has long been incorporated in Bayesian inference and continues to generate impact. It is considered either (1) directly due to that the target distribution is defined on a non-Euclidean manifold, or (2) for a proper dynamics that respects the geometry of a distribution space. In this chapter, we review the background and some recent progress on the interplay between geometry and sampling methods. We consider two major classes of sampling methods: Markov chain Monte Carlo (MCMC) and particle-based variational inference (ParVI). For MCMC, we cover some dynamics on manifolds and their simulation for both cases (1) and (2). For ParVI, we describe its geometric interpretation under the view of case (2), and introduce the variants that the interpretation inspires, including those for case (1).

Keywords: Bayesian inference, Riemannian geometry, Markov chain Monte Carlo, Particle-based variational inference

1 Geometry consideration in sampling: Why bother?

Bayesian inference is the central task in Bayesian modeling. Given a prior distribution $p_0(x)$ of a latent variable x and a likelihood distribution $p(o|x)$ that

Handbook of Statistics, Vol. 47. https://doi.org/10.1016/bs.host.2022.07.004

links x to an observation/data variable o, the task is to approximate the posterior distribution $p(x|o)$ of the latent variable x given an observation value o. By Bayes' rule, we know that $p(x|o) = \frac{p_0(x)p(o|x)}{p(o)} = \frac{p_0(x)p(o|x)}{\int p_0(x)p(o|x)\ \mathrm{d}x}$, but the task is still highly intractable due to the commonly high-dimensional integral. Sampling methods are a natural fit to Bayesian inference. On one hand, the most common expectation on a posterior approximator is to estimate an expectation of a function under the posterior. With samples of the posterior, this can be conveniently estimated using the average of the function values on the samples. On the other hand, sampling methods often allow an *unnormalized* density function of the target distribution, which is the case in Bayesian inference: $p(x|o) \propto p_0(x)p(o|x)$, so there is an easily accessible unnormalized density.

Commonly the latent variable is represented as a vector in a Euclidean space. But in many applications, the latent variable may be more appropriate to be defined in a non-Euclidean ("curly"/"nonflat") space, often formalized as a manifold. This is the first case for considering non-Euclidean geometry in sampling methods. For example, in some applications we only care about the direction of a data vector while its magnitude does not convey useful information. The vector is thus normalized and lies in a *hypersphere* $\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n | x^\top x = 1\}$. This includes the common "term frequency–inverse document frequency" (tf–idf) feature representing a document by a weighted term (i.e., word) distribution in it, which is normalized. More instances can be found in geology and bioinformatics, which leads to the study of directional/orientational statistics. Modeling such data often requires the latent variable also be in the hypersphere,[a] leading to the sampling task on a hypersphere. For tf–idf feature, spherical admixture model (Reisinger et al., 2010) confines the "topic" latent vectors also on the hypersphere. Hyperspherical latent vector is also used in variational autoencoder (Davidson et al., 2018), which improves the performance for hyperspherical data and also enables an uninformative prior that helps better clustering for the usual Euclidean data. Lan et al. (2014) considered sampling from a general norm-constrained space by converting the space to a hypersphere.

Another example is latent variables on a *simplex* $\Delta^{n-1} := \{x \in (\mathbb{R}^+)^n | \sum_{i=1}^n x^i = 1\}$,[b] which is common since it is where the probability parameter of a categorical likelihood lives. Although the simplex is flat, people often desire a parameterization using $(n-1)$ free/unconstrained parameters, which is a nonlinear representation of the space (Beck and Teboulle, 2003; Patterson and Teh, 2013). For more examples, in matrix completion (e.g., recommender system), the matrix is often seen as being generated from a singular-value decomposition form to handle the rank constraint (Salakhutdinov and Mnih, 2008; Song and

[a]This can be promoted from the fact that hyperspheres are not homeomorphic to Euclidean spaces.
[b]The superscript i here represents the contravariant index of a vector x, but not the exponent.

Zhu, 2016; Yanush and Kropotov, 2019). Under a Bayesian treatment, this leads to inferring latent variable on the set of orthogonal matrices (may not be in a squared shape), called the *Stiefel manifold* (James, 1976; Stiefel, 1935). Some works on variational autoencoders (Grattarola et al., 2018; Mathieu et al., 2019; Nagano et al., 2019; Ovinnikov, 2019) take a *hyperbolic* latent space for tree-structured data, due to their similar geometry that the volume (resp. number of nodes) grows exponentially with the distance to the origin (resp. the distance to the root node). Some recent works also consider endowing the latent space with the pulled-back metric from the data space so that the induced geometry objects such as gradient, geodesic (Arvanitidis et al., 2018, 2019; Chen et al., 2018b), exponential map (Shao et al., 2018), and isotropic Gaussian (Kalatzis et al., 2020) in the latent space follow the data-manifold geometry, which make operations such as interpolation in the latent space represent a semantic meaning.

The manifold perspective is more common for Bayesian methods than one may expect, as the concept of *information geometry* (Amari, 1998, 2016; Amari and Nagaoka, 2007) introduces a "natural" metric to the latent space (even for Euclidean latent space). Noting that each value of the latent variable x defines a likelihood distribution $p(o|x)$ on the observation variable o, the difference between x values may be more naturally measured by the difference between the distributions they define (see Fig. 1). If the infinitesimal distribution difference is measured by the KL divergence, the induced latent space metric is the Fisher–Rao metric, which is invariant to reparameterization. The well-known natural gradient (Amari, 1998; Khan and Nielsen, 2018) is the gradient under this metric, which is the fastest ascending direction of a *distribution* objective function, and remains the same under any parameterization of the distribution. Using natural gradient in optimization often achieves a much faster convergence.

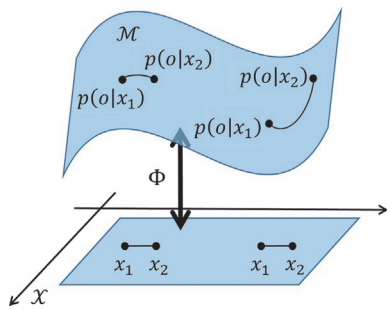

FIG. 1 Information geometry as an example of the second case for considering geometry in sampling methods: to respect the geometry of a distribution space. The difference between latent variable values x_1, x_2 is more appropriately measured as the difference between the likelihood distributions $p(o|x_1)$, $p(o|x_2)$ they define. Equally separated x_1, x_2 pairs in the latent space $\mathcal{X}$ may be differently separated in the distribution space $\mathcal{M}$.

Information geometry is the first example of the second case for considering geometry in sampling methods, i.e., to respect the geometry of a distribution space. Such a consideration is also the underlying principle of particle-based variational inference (ParVI) methods, which are a relatively new and fast-developing branch of sampling methods. In contrast to conventional MCMC methods, ParVI methods use a *deterministic* dynamics to iteratively update a *fixed-sized* set of particles (i.e., samples) toward the target distribution, and can often achieve a better sample efficiency. The dynamics is chosen to fastest descending the difference between the particle distribution and the target distribution under some metric, which is formally the gradient flow of the difference function on a certain space of distributions. Studying the geometry of the distribution space as a manifold then reveals the assumptions and convergence analysis of ParVI, and also draws the link to general MCMC dynamics. The geometric understanding also inspires ParVI variants that converge faster, produce more accurate approximations, and handles non-Euclidean latent space.

This review is organized as follows. We first introduce basic concepts of manifold in Section 2. We then describe general MCMC dynamics in Section 3.1, followed by MCMC instances on manifolds in Section 3.2 which can be simulated in the coordinate space of the manifold. For some manifolds, simulation in their embedded space is advantageous, as shown in Section 3.3. Next, we introduce ParVI methods starting with perhaps the most popular instance [Stein variational gradient descent (SVGD)] in Section 4.1. After introducing some background knowledge in Section 4.2, we present in Section 4.3 the geometric interpretation as the gradient flow on certain distribution spaces, which reveals the assumptions and convergence analysis of ParVI. This interpretation also draws the link to general MCMC dynamics as we show in Section 4.4, and inspires ParVI variants that converge faster, produce more accurate approximations, and handles non-Euclidean latent space as shown in Section 4.5.

2 Manifold and related concepts

In this part we introduce some basic concepts pertaining to manifold. The concept of manifold is a generalization of vector spaces, which allows a "curly" intuition and spacial heterogeneity. The so-called Riemannian manifold introduces a light structure but which then induces almost all counterparts of common concepts in an inner-product space, which enables tractable computation. We focus on the scheme and intuition of the concepts and include relations to linear space when possible. See formal textbooks (e.g., Abraham et al., 2012; Do Carmo, 1992; Nicolaescu, 2007; Romano, 2007) for a complete introduction.

2.1 Manifold

A common intuitive description of an m-dimensional *manifold* $\mathcal{M}$ is that it is a space that *locally* looks like the m-dimensional Euclidean space $\mathbb{R}^m$. It is

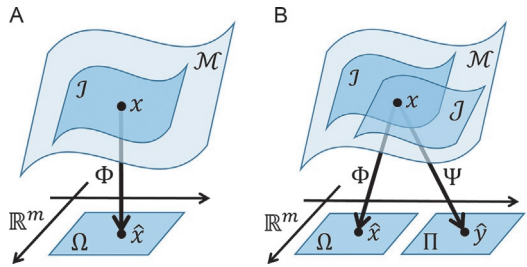

FIG. 2 Illustration of the concepts of manifold and coordinate system. (A) Concept of manifold and local coordinate system. (B) Intersecting coordinates for characterizing a smooth manifold.

formally defined as a topological space,[c] any point on which has a neighborhood $\mathcal{I}$ homeomorphic to a Euclidean open subset $\Omega \subseteq \mathbb{R}^m$, meaning that there is a continuous bijection $\Phi : \mathcal{I} \to \Omega$ whose inverse is also continuous (such Φ is called a homeomorphism) (see Fig. 2A). By definition, around any point the manifold can be locally represented using an m-dimensional coordinate $\hat{x} = \Phi(x) \in \mathbb{R}^m$, so $(\mathcal{I}, \Phi)$ is called a local coordinate system, and Ω a coordinate space. A function $f : \mathcal{M} \to \mathbb{R}$ on the manifold can also be concretized as a usual multivariate function $f \circ \Phi^{-1} : \mathbb{R}^m \to \mathbb{R}$, which holds the same continuity as f. This is what formalizes the intuition "locally looks like the Euclidean space." Note that for some manifolds there is no global coordinate system, e.g., the hypersphere $\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n | x^\top x = 1\}$, since it is not homeomorphic to any Euclidean space.

The definition for now covers the continuity of the manifold, but we often also care about differentiability and smoothness. A smooth manifold is expected to define smooth functions on it. For a function f, its smoothness around a point x can be characterized by that of its coordinate version $f \circ \Phi^{-1} : \mathbb{R}^m \to \mathbb{R}$. To make a consistent characterization independent of the choice of coordinate system, under any other coordinate system $(\mathcal{J}, \Psi)$ containing x (Fig. 2B), the multivariate function $f \circ \Psi^{-1}$ should have the same smoothness as $f \circ \Phi^{-1}$. This requires the coordinate conversions $\Psi \circ \Phi^{-1}$ and $\Phi \circ \Psi^{-1}$ be smooth as $\mathbb{R}^m \to \mathbb{R}^m$ functions. Such coordinate systems are called compatible, and the manifold $\mathcal{M}$ is called smooth if there is a set of compatible coordinate systems that covers $\mathcal{M}$. The differentiability/smoothness of a function f or a curve $\gamma_\cdot : [a, b] \to \mathcal{M}$ on a smooth manifold can be consistently determined by that of $f \circ \Phi^{-1} : \mathbb{R}^m \to \mathbb{R}^m$ or $t \mapsto \Phi(\gamma_t) : \mathbb{R} \to \mathbb{R}^m$.

[c] A topological space is a set with a *topology*, which is roughly a set of abstract open subsets (containing $\emptyset$ and the entire set, closed under finite intersection and countable union), enabling the definitions of neighborhood, limit, and continuity. Commonly for defining a manifold, the topological space requires second-countable (the topology can be generated from a countable set of open subsets) and second-separable (any two points have nonintersecting neighborhoods).

Denote the set of smooth functions on $\mathcal{M}$ as $C^\infty(\mathcal{M})$. Since common manifolds are smooth and common objects of interest are defined on smooth manifolds, we focus on smooth manifolds and smooth functions hereafter, and still call them "manifold" and "function" for brevity.

2.2 Tangent vector and vector field

(See Fig. 3 for the relations among the introduced concepts in this and the next subsection.) Fundamental geometric descriptions of the manifold and algorithmically concerned dynamics (e.g., gradient descent and Langevin dynamics analogies) calls for the concept of tangent vector. In $\mathbb{R}^m$ the tangent vector at a point γ_0 on a smooth curve $(\gamma_t)_t$ (Fig. 4A) is the limiting vector $v := \frac{\mathrm{d}\gamma_t}{\mathrm{d}t}\Big|_{t=0} = \lim_{h\to 0} \frac{1}{h}(\gamma_h - \gamma_0)$, which holds the meaning of the velocity of a particle moving along the curve. Unfortunately on manifolds there is no vector subtraction (yet). For another characterization, note that the curve induces a directional derivative of a function f at γ_0 as $\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t))\big|_{t=0}$, which is, by the chain rule, $\partial_i f(\gamma_0)\frac{\mathrm{d}\gamma_t^i}{\mathrm{d}t}\Big|_{t=0} = v^i(\gamma_0)\partial_i f(\gamma_0)$, where $\partial_i f$ is short for $\frac{\partial}{\partial x^i} f(x^1, \ldots, x^m)$, and we have used *Einstein's summation convention* that the index repeated in both a subscript and a superscript (e.g., i here) is summed over automatically

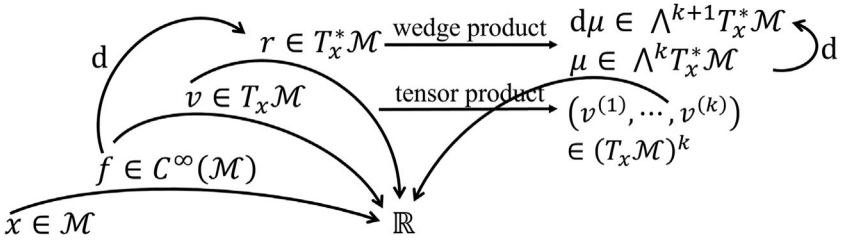

FIG. 3 Relations among some concepts on a manifold $\mathcal{M}$. A symbol just above a *curly arrow* represents an instance of the map. See Sections 2.2 and 2.3 for details.

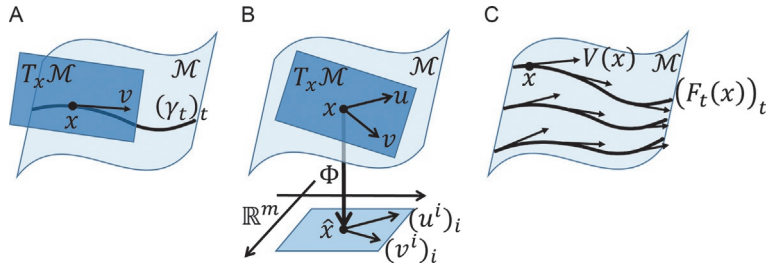

FIG. 4 Illustration of tangent vector, tangent space, vector field, and flow on a manifold $\mathcal{M}$. See Section 2.2 for details. (A) Tangent vector on a curve. (B) Tangent vector and tangent space on a manifold. (C) Vector field and its flow.

(i.e., the summation symbol "$\sum_{i=1}^{m}$" is omitted). So the tangent vector has another appearance $v = v^i\partial_i$ as a *directional derivative operator* on functions, and its action on a function $v[f]$ is the directional derivative of f along v.

This perspective can be extended to manifolds, since the directional derivative $\frac{\mathrm{d}}{\mathrm{d}t}f(\gamma_t)\big|_{t=0}$ is still well defined (note $t \mapsto f(\gamma_t)$ is a $\mathbb{R} \to \mathbb{R}$ function). Using any coordinate system $(\mathcal{I}, \Phi)$ around γ_0, it can be expressed as $\frac{\mathrm{d}}{\mathrm{d}t}(f \circ \Phi^{-1})(\Phi(\gamma_t))\big|_{t=0} = v^i\partial_i f$ in the sense that $v^i := \frac{\mathrm{d}}{\mathrm{d}t}\Phi^i(\gamma_t)\big|_{t=0}$ and $\partial_i f := \frac{\partial}{\partial \hat{x}^i} f(\Phi^{-1}(\hat{x}))\Big|_{\hat{x}=\Phi(\gamma_0)}$ (Fig. 4A). Covering all possible smooth curves passing through x, a general *tangent vector* v at $x \in \mathcal{M}$ as a directional derivative operator, is defined as a linear function on $C^\infty(\mathcal{M})$[d] satisfying the Leibniz rule: $v[fh] = f(x)v[h] + h(x)v[f]$. All tangent vectors at x form an m-dimensional linear space called the *tangent space* $T_x\mathcal{M}$ at x (Fig. 4B), and under any coordinate system containing x, $\{\partial_i\}_{i=1}^m$ is a basis, which is defined as $\partial_i f := \frac{\partial}{\partial \hat{x}^i} f(\Phi^{-1}(\hat{x}))\Big|_{\hat{x}=\Phi(x)}$. Any $v \in T_x\mathcal{M}$ can be expressed as $v = v^i\partial_i$ where $v^i = v[\Phi^i]$.[e] Under the change of coordinate system, the basis transforms as $\tilde{\partial}_\alpha = \frac{\partial \hat{x}^i}{\partial \hat{y}^\alpha}\partial_i$, where $\tilde{\partial}_\alpha f := \frac{\partial}{\partial \hat{y}^\alpha} f(\Psi^{-1}(\hat{y}))\Big|_{\hat{y}=\Psi(x)}$ forms the basis under the new coordinate system $(\mathcal{J}, \Psi)$, and $\frac{\partial \hat{x}^i}{\partial \hat{y}^\alpha} := \frac{\partial}{\partial \hat{y}^\alpha}\Phi^i \circ \Psi^{-1}(\hat{y})\Big|_{\hat{y}=\Psi(x)}$. For a given tangent vector v, its coordinate transforms similarly: $\tilde{v}^\alpha = \frac{\partial \hat{y}^\alpha}{\partial \hat{x}^i}v^i$. Note that the coordinate expression $v = v^i\partial_i = \tilde{v}^\alpha\tilde{\partial}_\alpha$ is invariant under coordinate change.

If $\mathcal{M}$ is a linear space, then it is isomorphic to the tangent space $T_x\mathcal{M}$ for any $x \in \mathcal{M}$: $y \in \mathcal{M} \mapsto v_y \in T_x\mathcal{M}, v_y[f] := \dfrac{\mathrm{d}}{\mathrm{d}t}f(x+ty)\Big|_{t=0}$. But generally tangent spaces at different points are different linear spaces.

A *vector field* V (Fig. 4C) defines a tangent vector $V(x)$ at every point x on the manifold, and $V(x)$ depends on x smoothly (e.g., $V^i \circ \Phi^{-1}$ for each i is smooth in any coordinate system). Denote the set of vector fields on $\mathcal{M}$ as $\mathcal{T}(\mathcal{M})$. A vector field defines a *dynamics* (dynamical system) on the manifold: $\frac{\mathrm{d}x_t}{\mathrm{d}t} = V(x_t)$, which describes how a particle moves on the manifold as time proceeds. Its solution is called a *flow*, which is a set of curves $\{(\phi_t(x))_t | x \in \mathcal{M}\}$ such that $\phi_0(x) = x$ and $\frac{\mathrm{d}}{\mathrm{d}t}\phi_t(x)\big|_{t=0} = V(x)$. The flow of any vector field V exists at least locally (due to Picard–Lindelöf theorem).

For a concise and conventional notation, in the following, symbols with indices (e.g., x^i, v_i) represent the coordinates of the same objects (e.g., x, v) in some coordinate system. Particularly we use x^i to represent the same thing as $\hat{x}^i$, i.e., the coordinates of a manifold point x.

---

[d] More precisely, instead of $C^\infty(\mathcal{M})$, it is sufficient to only consider functions that are smooth in a neighborhood of x.

[e] This can be seen through $v[\Phi^j] = v^i\partial_i\Phi^j = v^i\frac{\partial}{\partial \hat{x}^i}\Phi^j(\Phi^{-1}(\hat{x}^1, \ldots, \hat{x}^m))\big|_{\hat{x}=\Phi(x)} = v^i\dfrac{\partial \hat{x}^j}{\partial \hat{x}^i}\Big|_{\hat{x}=\Phi(x)} = v^i\delta_i^j = v^j$.

2.3 Cotangent vector and differential form

The cotangent space $T_x^*\mathcal{M}$ at x is the dual space of the tangent space $T_x\mathcal{M}$. Under a coordinate system, the basis $\{\partial_i\}_{i=1}^m$ of $T_x\mathcal{M}$ induces a *dual basis* $\{\partial^{*i}\}_{i=1}^m$ for $T_x^*\mathcal{M}$: $\partial^{*i}[v] := v^i = v[\Phi^i]$, which satisfies $\partial^{*i}[\partial_j] = \delta_j^i$.[f] Under this perspective, the directional derivative along vector v, $v[f] = v^i\partial_i f = (\partial_i f\partial^{*i})[v^j\partial_j]$ can be viewed as the action of the covector $\partial_i f\partial^{*i}$ on the vector v. We define this covector as the *differential* of f: $\mathrm{d}f \in T_x^*\mathcal{M}$, $\mathrm{d}f[v] := v[f], \forall v \in T_x\mathcal{M}$. We then recognize that $\mathrm{d}\Phi^i[\partial_j] = \partial_j[\Phi^i] = \frac{\partial}{\partial\hat{x}^j}\Phi^i(\Phi^{-1}(\hat{x})) = \frac{\partial\hat{x}^i}{\partial\hat{x}^j} = \delta_j^i$, so ∂^{*i} is essentially $\mathrm{d}\Phi^i$. Conventionally $\mathrm{d}\Phi^i$ is denoted using the symbol of the corresponding coordinates as $\mathrm{d}x^i$. The covector is then expressed as $\mathrm{d}f = \partial_i f\,\mathrm{d}x^i$, which is the form of the usual differential in calculus.

To describe a k-dimensional volume element on a manifold and more, we need the concept of k-form. To draw the intuition, consider the volume of the k-dimensional parallelepiped formed by k vectors. We know that the volume responds linearly to each vector, and switching two vectors alters the orientation of the parallelepiped, so the volume changes its sign. This tells us that the volume is an antisymmetric linear function on the k vectors. Indeed, in $\mathbb{R}^k$, the volume formed by k vectors is the $k \times k$ determinant of the vector-stacking matrix, which is an antisymmetric linear function.

On a manifold, a k-dimensional infinitesimal volume element at a point x is formed by k tangent vectors from $T_x\mathcal{M}$. We call the map from the k vectors to the volume value as a k-differential form, or just k-form. Formally, a k-form $\mu : (T_x\mathcal{M})^k \to \mathbb{R}$ is an antisymmetric k-multilinear function on $T_x\mathcal{M}$. Denote the space of k-forms as $\wedge^k T_x^*\mathcal{M}$, which is a subspace of $(T_x^*\mathcal{M})^k := \otimes^k T_x^*\mathcal{M}$, i.e. the space of k-multilinear functions. In this sense, a covector from $T_x^*\mathcal{M}$ is also recognized as a 1-form. Due to antisymmetry, if there are two identical vectors in the k input vectors, the k-form outputs zero. Due to linearity, this case can be extended to k linearly dependent vectors. Since when $k > m$, any k vectors are linearly dependent, and we know that all k-forms are trivially zero for $k > m$. In other cases, due to antisymmetry, $\wedge^k T_x^*\mathcal{M}$ is $\binom{m}{k}$-dimensional.

To construct a k-form using k covectors $\{r^{(1)}, \ldots, r^{(k)}\}$ from $T_x^*\mathcal{M}$ (i.e., using k 1-forms), we introduce the *wedge product*, which is essentially the antisymmetrized tensor product: $r^{(1)} \wedge \cdots \wedge r^{(k)} := \sum_\sigma (-1)^\sigma r^{(\sigma_1)} \otimes \cdots \otimes r^{(\sigma_k)}$, where σ traverses over all the permutations of $\{1, \ldots, k\}$ and $(-1)^\sigma$ is its sign.[g]

---

[f] The symbol δ_j^i is the Kronecker delta tensor, i.e., $\delta_j^i := 1$ if $i = j$ and $\delta_j^i := 0$ otherwise. Similarly, δ_{ij}, δ^{ij} are defined for the corresponding types of tensors.

[g] The notation as the product of a binary operator $\wedge$ (instead of an operator on k covectors altogether) is valid since $(r^{(1)} \wedge \cdots \wedge r^{(i)}) \wedge (r^{(i+1)} \wedge \cdots \wedge r^{(k)}) = r^{(1)} \wedge \cdots \wedge r^{(k)}$ for any $1 \le i < k$.

Using this notion, the $\binom{m}{k}$ k-forms $\{\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}\}_{1 \le i_1 < \cdots < i_k \le m}$ build a basis of $\wedge^k T_x^* \mathcal{M}$. As a tensor in $(T_x^* \mathcal{M})^k$, any k-form can be expanded as $\mu = \mu_{i_1 \ldots i_k} \mathrm{d}x^{i_1} \otimes \cdots \otimes \mathrm{d}x^{i_k}$, where $\mu_{i_1 \ldots i_k} = \mu[\partial_{i_1}, \ldots, \partial_{i_k}]$. Due to antisymmetry, permutationally equivalent index tuples can be grouped together: $\mu = \sum_{1 \le i_1 < \cdots < i_k \le m} \sum_{\sigma} \mu_{i_{\sigma_1} \ldots i_{\sigma_k}} \mathrm{d}x^{i_{\sigma_1}} \otimes \cdots \otimes \mathrm{d}x^{i_{\sigma_k}} = \sum_{1 \le i_1 < \cdots < i_k \le m} \mu_{i_1 \cdots i_k} \sum_{\sigma} (-1)^{\sigma} \mathrm{d}x^{i_{\sigma_1}} \otimes \cdots \otimes \mathrm{d}x^{i_{\sigma_k}} = \sum_{1 \le i_1 < \cdots < i_k \le m} \mu_{i_1 \cdots i_k} \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k} = \frac{1}{k!} \mu_{i_1 \ldots i_k} \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}$.

The differential operator d can be extended as *exterior derivative* to act on a k-form and give a $(k + 1)$-form: $\mathrm{d}\mu := \sum_{1 \le i_1 < \cdots < i_k \le m} \mathrm{d}\mu_{i_1 \ldots i_k} \wedge \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k} = \frac{1}{k!} \partial_i \mu_{i_1 \ldots i_k} \mathrm{d}x^{i} \wedge \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}$. This definition is independent of the choice of coordinate system; particularly it has an alternative definition: df is the differential of function for a 0-form f (i.e., a function), and $\mathrm{d}(\mu \wedge \nu) = \mathrm{d}\mu \wedge \nu + (-1)^k \mu \wedge \mathrm{d}\nu$ for k-form μ, and $\mathrm{d} \circ \mathrm{d} = 0$.

2.4 Riemannian manifold

A *Riemannian manifold* is a manifold $\mathcal{M}$ in any of whose tangent space $T_x \mathcal{M}$ there is an inner product $\langle \cdot, \cdot \rangle_{T_x \mathcal{M}}$ defined, and that it depends on x smoothly. This $\langle \cdot, \cdot \rangle_{T_x \mathcal{M}}$ is called a Riemannian structure or metric. In any coordinate system it can be expressed as a positive definite matrix, $\langle u, v \rangle_{T_x \mathcal{M}} = g_{ij}(x) u^i v^i$ where $g_{ij}(x) := \langle \partial_i, \partial_j \rangle_{T_x \mathcal{M}}$. It also induces a norm in each tangent space, $\|v\|_{T_x \mathcal{M}} := \sqrt{\langle v, v \rangle_{T_x \mathcal{M}}}$. This simple structure endows the manifold with many useful concepts that make computation possible.

The *gradient* of a function $f \in C^{\infty}(\mathcal{M})$ at $x \in \mathcal{M}$ is defined as the tangent vector $\mathrm{grad}\, f(x) \in T_x \mathcal{M}$ that satisfies $\langle \mathrm{grad} f(x), v \rangle_{T_x \mathcal{M}} = \mathrm{d}f[v] = v[f]$, $\forall v \in T_x \mathcal{M}$. It has the following coordinate expression:

$$\mathrm{grad}\, f(x) = g^{ij}(x) \partial_j f(x) \partial_i = G(x)^{-1} \nabla f(x), \tag{1}$$

where $(g^{ij}) = G^{-1}$ is the inverse matrix of the Riemann metric tensor $G := (g_{ij})$. The gradient can be defined all over the manifold, and the flow of the negative of this vector field is called a gradient flow. This abstract definition of gradient meets the common intuition of a fastest ascending direction for f: $\max \cdot \mathrm{argmax}_{v \in T_x \mathcal{M}: \|v\| = 1} v[f] = \max \cdot \mathrm{argmax}_{v \in T_x \mathcal{M}: \|v\| = 1} \langle \mathrm{grad} f(x), v \rangle_{T_x \mathcal{M}} = \mathrm{grad} f(x)$, where "max · argmax" denotes the scalar product of the maximum to the maximizing vector.

The *length* of a curve $\gamma : [a, b] \to \mathcal{M}$ can be defined using the Riemannian structure as: $L(\gamma) := \sqrt{\int_a^b \|\mathrm{d}\gamma_t / \mathrm{d}t\|^2_{T_{\gamma_t} \mathcal{M}} \, \mathrm{d}t}$, which leads to the *distance* between two points on the manifold:

$$d_{\mathcal{M}}(x, y) := \inf_{(\gamma_t)_{t \in [0,1]} : \gamma_0 = x, \gamma_1 = y} L(\gamma). \tag{2}$$

If the manifold is complete under this distance, the distance-minimizing curve exists [Hopf–Rinow theorem (Hopf and Rinow, 1931)], which is called a

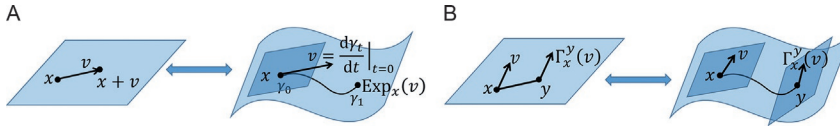

FIG. 5 Illustration of concepts on a Riemannian manifold, with analogy to a linear space. See Section 2.4 for more details. (A) Vector addition in linear space (*left*) and exponential map on Riemannian manifold (*right*). (B) Parallel transport in linear space (*left*) and on Riemannian manifold (*right*).

geodesic. It is the counterpart of straight line in Euclidean spaces. As illustrated in Fig. 5A, in analogy to vector addition $x + v$ in linear space that moves a point x along the straight line in the direction of v, define the *exponential map* $\mathrm{Exp}_x : T_x\mathcal{M} \to \mathcal{M}, v \mapsto \gamma_1$ as moving a point x along the geodesic $(\gamma_t)_{t\in[0,\,1]}$ tangent to v at $x : \gamma_0 = x, \left.\frac{\mathrm{d}\gamma_t}{\mathrm{d}t}\right|_{t=0} = v$, which exists uniquely. As mentioned in Section 2.2, tangent space of a linear space is everywhere the same, but is not for a general manifold. Fortunately the Riemannian structure induces a link between tangent spaces at different points, which is the *parallel transport* $\Gamma_x^y : T_x\mathcal{M} \to T_y\mathcal{M}$ (Fig. 5B). It transports a tangent vector at x to a tangent vector at y along the geodesic[h] $(\gamma_t)_{t\in[0,1]}$ from x to y in a certain way that is regarded "parallel"; particularly $\left\langle \frac{\mathrm{d}\gamma_t}{\mathrm{d}t}, v_t \right\rangle_{T_{\gamma_t}\mathcal{M}}$ does not change with t, where $v_0 = v$, $v_1 = \Gamma_x^y(v)$.[i]

2.5 Measure

A measure on an m-dimensional manifold[j] is represented by an m-form ω, which measures an m-dimensional infinitesimal volume element everywhere on the manifold. Since the space of m-forms is one-dimensional, ω is represented by $\omega_{1\ldots m}\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^m$, or simply $\omega\,\mathrm{d}x$, where $\mathrm{d}x := \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^m$ represents the usual Lebesgue measure of the Euclidean coordinate space. As a measure, $\omega_{1\ldots m}$ is required to be a nonnegative function. Such an m-form is also called a volume form. Under coordinate system change, the coordinate

[h] A noteworthy distinction from the linear case is that transporting the vector in the parallel way but along different paths would generally yield different results, and the difference is related to the curvature of the manifold. If there is no difference, the manifold is seen as flat (though unnecessarily linear).

[i] We would like to mention that the general definitions of geodesic, exponential map and parallel transport are built on an independent manifold structure called affine connection, and the definitions here correspond to the special version under the Levi-Civita connection, which is an affine connection induced from the Riemannian structure.

[j] To define a measure, the manifold is required to be *orientable*: there exists a set of coordinate systems covering the manifold and in the intersection of any two coordinate systems, the Jacobian determinant of the coordinate transformation $\left|\left(\frac{\partial \tilde{y}^{\alpha}}{\partial x^{\beta}}\right)\right|$ is positive.

expression transforms with the Jacobian determinant of the coordinate transformation: $\tilde{\omega}(y) = \omega(x)/\left|\left(\frac{\partial \tilde{y}^{\alpha}}{\partial x^{i}}\right)\right|$.

On a Riemannian manifold, there is a special measure called the Riemannian measure, which has the coordinate expression $\omega_g = \sqrt{|G|}\,\mathrm{d}x$, where $|G|$ is the determinant of G. This expression is coordinate invariant, meaning that in another coordinate system, it becomes $\tilde{\omega}_g = \sqrt{|\tilde{G}|}\,\mathrm{d}y$.

Integral on the manifold can be defined under a measure. Particularly the density function of a distribution/measure η on the manifold can be defined: for any measurable subset $\mathcal{I} \subseteq \mathcal{M}$, $\eta(\mathcal{I}) = \int_{\Phi(\mathcal{I})} p_{\mathrm{L}}\,\mathrm{d}x = \int_{\Phi(\mathcal{I})} p_{\mathrm{R}}\,\mathrm{d}\omega_g$, where p_{L} is the density w.r.t. the Lebesgue measure $\mathrm{d}x$ in the coordinate space, and p_{R} is the density function w.r.t. the Riemannian measure ω_g. They are related by:

$$p_{\mathrm{L}} = p_{\mathrm{R}}\sqrt{|G|}.$$

Finally we mention the well-known Stokes theorem: for a proper manifold region $\mathcal{I}$ with boundary $\partial\mathcal{I}$, and an $(m-1)$-form η, we have $\int_{\mathcal{I}} \mathrm{d}\eta = \int_{\partial\mathcal{I}} \eta$.

2.6 Divergence and Laplacian

In Euclidean space, the divergence of a vector field V is defined as $(\nabla\cdot V)(x) := \partial_i V^i(x)$. It is considered in a type of integral with compactly supported function f: $\int_{\mathbb{R}^m} V\cdot\nabla f\,\mathrm{d}x = \int_{\mathbb{R}^m} V^i(\partial_i f)\,\mathrm{d}x = -\int_{\mathbb{R}^m} f\partial_i V^i\,\mathrm{d}x = -\int_{\mathbb{R}^m} f\nabla\cdot V\,\mathrm{d}x$, where the second equality is due to integration by parts and that the term $\int_{\mathbb{R}^m}\partial_i(fV^i)\,\mathrm{d}x = \int_{\partial\mathbb{R}^m} fV^i\,\mathrm{d}S_i$ (due to Stokes theorem; $(\mathrm{d}S_i)_i$ is the infinitesimal surface element with out-pointing normal direction on the infinitely large sphere $\partial\mathbb{R}^m$, such that $\mathrm{d}\mathrm{d}S = \mathrm{d}x$) vanishes since $f = 0$ on $\partial\mathbb{R}^m$. On a Riemannian manifold, the concept of divergence can be extended similarly under this characterization. For a coordinate-invariant definition, integrals are considered under the Riemannian measure:

$$\mathrm{div}:\mathcal{T}(\mathcal{M})\to C^{\infty}(\mathcal{M}),\quad \text{s.t.}\int_{\mathcal{M}} V[f]\,\mathrm{d}\omega_g = -\int_{\mathcal{M}} f\,\mathrm{div}\,V\,\mathrm{d}\omega_g,\ \forall f\in C_{\mathrm{c}}^{\infty}(\mathcal{M}), \tag{3}$$

where $C_{\mathrm{c}}^{\infty}(\mathcal{M})$ denotes the set of all compactly supported functions on manifold $\mathcal{M}$. Note by the definition of gradient, the l.h.s. can also be written as $\int\langle \mathrm{grad} f(x), V(x)\rangle_{T_x\mathcal{M}}\,\mathrm{d}\omega_g(x)$. In any coordinate space, the l.h.s. is $\int V^j(\partial_j f)\sqrt{|G|}\,\mathrm{d}x = -\int f\partial_j(V^j\sqrt{|G|})\,\mathrm{d}x$, which equals to the r.h.s $-\int f\,\mathrm{div}\,V\sqrt{|G|}\,\mathrm{d}x$ by the definition. This then leads to the coordinate expression of the divergence:

$$\mathrm{div}\ V = \partial_i(\sqrt{|G|}V^i)/\sqrt{|G|} = \partial_i V^i + V^i\partial_i\log\sqrt{|G|}.$$

In Euclidean space, the Laplacian of a function $\nabla^2 f := \sum_{i=1}^{m} \partial_i \partial_i f = \nabla \cdot \nabla f$ can be seen as the divergence of the gradient of f. This can be extended to Riemannian manifold:

$$\mathrm{Lap}\, f := \mathrm{div}(\mathrm{grad}\, f) = \partial_i(\sqrt{|G|} g^{ij} \partial_j f)/\sqrt{|G|}.$$

2.7 Manifold embedding

Many manifolds are defined as a subset of a Euclidean space $\mathbb{R}^n$, e.g., the hypersphere $\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n | x^\top x = 1\}$. This is often how we imagine or picture a manifold (even in our illustrative figures), but is it possible for any manifold defined in the abstract way described in Section 2.1? This is formally described as the *embedding* of a manifold $\mathcal{M}$, which is a smooth injection $\Xi : \mathcal{M} \to \mathbb{R}^n$ to a Euclidean space so that we can understand the manifold as a subset $\Xi(\mathcal{M})$ of $\mathbb{R}^n$ (Fig. 6). Whitney embedding theorem (Persson, 2014; Whitney, 1944) shows that an m-dimensional manifold can always be embedded into $\mathbb{R}^{2m}$.

For a Riemannian manifold, the embedding also pulls back the Euclidean metric $\delta_{\alpha\beta}$ to the manifold as $\tilde{g}_{ij} = \delta_{\alpha\beta} \frac{\partial \hat{y}^\alpha}{\partial x^i} \frac{\partial \hat{y}^\beta}{\partial x^j}$ for $\hat{y} = \Xi(\Phi^{-1}(\hat{x}))$. If it coincides with the original metric g_{ij}, the embedding is said *isometric*. Nash embedding theorem (Nash, 1956) shows that any Riemannian manifold can be isometrically embedded into a Euclidean space.

In the embedded space $\mathbb{R}^n$, the restriction of the n-dimensional Lebesgue measure onto the subset $\Xi(\mathcal{M}) \subseteq \mathbb{R}^n$ induces a measure on $\Xi(\mathcal{M})$, which is called the Hausdorff measure. A distribution on manifold then also admits a density function $p_{\mathrm{H}} : \Xi(\mathcal{M}) \to \mathbb{R}_{\geq 0}$ w.r.t. this measure. If the manifold is Riemannian and the embedding is isometric, then the Hausdorff measure coincides with the Riemannian measure, and p_{R} coincides with p_{H} in the sense that $p_{\mathrm{R}} = p_{\mathrm{H}} \circ \Xi \circ \Phi^{-1}$.

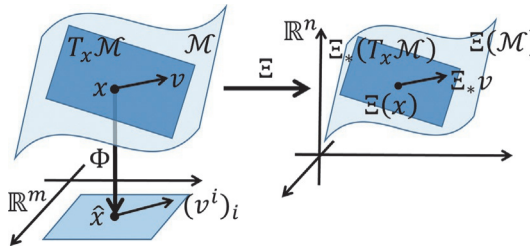

FIG. 6 Illustration of the embedding of a manifold.

3 Markov chain Monte Carlo on Riemannian manifolds

Markov chain Monte Carlo (MCMC) methods have long been a workhorse for Bayesian inference. When for inferring latent variables on a manifold, classical MCMC methods such as Gibbs sampling and Metropolis–Hastings (MH) can be adapted in some cases. For example, for sampling from a hypersphere $\mathbb{S}^{n-1}$ using MH, one can draw a proposal from an isotropic Gaussian centered at the current sample and project (normalize) the sample onto $\mathbb{S}^{n-1}$, then compute the acceptance rate using the (unnormalized) target distribution density (the proposal density ratio is 1 since the Gaussian is isotropic thus symmetric after projection) (Reisinger et al., 2010). But for a general manifold there does not seem to be a systematic way to construct such variants. More importantly, same as in the usual Euclidean case, these classical methods are not sufficiently efficient as their Markov chains mix slowly due to the uninformative and local transition.

Dynamics-based MCMC methods are more efficient and lead to the recent trend. They carry out the sampling process by simulating a dynamics, or formally a diffusion process or a continuous-time no-jump Markov process [see e.g., Särkkä and Solin (2019) for a comprehensive introduction]. The dynamics typically leverages the gradient of the target distribution log-density so the move is more informative and leads to distant transition. This lowers autocorrelation and increases the effective sample size and sampling efficiency. Many dynamics show nice convergence properties (Eberle et al., 2019; Mangoubi and Smith, 2017; Roberts et al., 1996; Seiler et al., 2014), which gift an exponential convergence guarantee to many algorithms under proper conditions (Cheng and Bartlett, 2017; Cheng et al., 2018; Dalalyan, 2017; Durmus and Moulines, 2016; Durmus et al., 2017; Livingstone et al., 2019; Roberts et al., 1996; Seiler et al., 2014). Particularly for Bayesian inference, the gradient formulation also allows using *stochastic gradient* in the simulation of many dynamics (Chen et al., 2014; Welling and Teh, 2011) to scale to large datasets. Given a set of independently and identically distributed (IID) data points $\{o^{(i)}\}_{i=1}^N$, the stochastic gradient estimates the gradient of the posterior log-density $\nabla_x \log p(x|\{o^{(i)}\}_{i=1}^N) = \nabla_x \log p_0(x) + \sum_{i=1}^N \log p(o^{(i)}|x)$ using a uniformly randomly chosen sub-dataset $\mathcal{S}$ (resampled for each call to the gradient) of a fixed size $|\mathcal{S}|$ as:

$$\text{(Stochastic gradient)} \qquad \nabla_x \log p_0(x) + \frac{N}{|\mathcal{S}|} \sum_{o \in \mathcal{S}} \log p(o|x). \tag{4}$$

It is a stochastic but unbiased and cheap estimate of the true gradient so that the method is scalable to large datasets. Moreover, dynamics are constructed using geometric objects, so can be naturally extended to Riemannian manifolds. We thus focus on dynamics-based MCMC methods in the following.

3.1 Technical description of general MCMC dynamics

In the Euclidean case, a dynamics is described by the so-called stochastic differential equation (SDE):

$$(\text{Dynamics}) \quad \mathrm{d}x = V_t(x)\,\mathrm{d}t + \sqrt{2D_t(x)}\,\mathrm{d}B_t(x), \tag{5}$$

where $V_t(x)$ is a vector or velocity field (may be time dependent) describing the deterministic *drift* of particles, $D_t(x)$ is a symmetric positive semidefinite matrix called the *diffusion matrix* (also may be time dependent) describing the strength of the stochastic diffusion, and $B_t(x)$ is the standard Brownian motion (or called the Wiener process), which can be seen as adding a sample of infinitesimal scale from the normal distribution $\mathcal{N}(x,\mathrm{d}t)$ independently ateach time instance t for advancing an infinitesimal time period $\mathrm{d}t$. The square root $\sqrt{2D_t}$ represents the positive semidefinite matrix such that $\sqrt{2D_t}\sqrt{2D_t}^{\top} = 2D_t$. A dynamics induces the evolution of the distribution of particles moving under the dynamics (see Fig. 7). Denoting the evolving distribution as $(q_t)_t$ in terms of the density function w.r.t. the Lebesgue measure, the evolution rule is explicitly given by the Fokker–Planck equation (FPE):

$$(\text{FPE}) \quad \partial_t q_t = -\nabla \cdot (q_t V_t) + \nabla\nabla^{\top} : (q_t D_t), \tag{6}$$

where $\nabla\nabla^{\top} : (q_t D_t) := \partial_i \partial_j (q_t D_t^{ij}) = \mathrm{tr}(\nabla\nabla^{\top}(q_t D_t^{\top}))$ is the double-dot product of two matrices of the same size (see e.g., Villani (2008, pp. 27–30), for the time-dependent, weak derivative, and Riemannian manifold extension).

The FPE plays a central role in developing proper dynamics for MCMC. Particularly, the target distribution p needs to be a stationary distribution of the dynamics, which leads to the requirement on the dynamics that $0 = -\nabla\cdot(pV_t) + \nabla\nabla^{\top} : (pD_t)$. Based on this, Ma et al. (2015) developed a

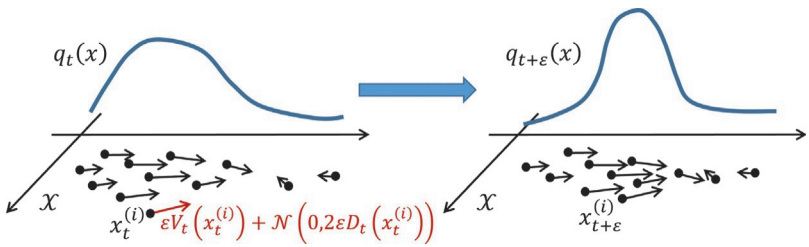

FIG. 7 Illustration of the evolving distribution induced by a dynamics equation (5). Given particles $\{x_t^{(i)}\}_i$ at time t that distribute obeying distribution q_t, if they move according to the dynamics equation (5), then after a short period of time ε their positions undergo a displacement (shown as the *red arrow*) and their new positions $\{x_{t+\varepsilon}^{(i)}\}_i$ distribute obeying a new distribution $q_{t+\varepsilon}$. When ε approaches to zero, the distribution q_t evolves continuously in time. The rule of the evolution is given by the Fokker–Planck equation (6).

complete recipe for all the dynamics that has p as a stationary distribution (and with time-independent V and D)[k]:

$$(p\text{-stationary dynamics}) \quad \mathrm{d}x = V_p(x)\,\mathrm{d}t + \sqrt{2D(x)}\,\mathrm{d}B_t(x), \tag{7}$$

$$\begin{aligned} \text{where} \quad & V_p(x) := \nabla \cdot (p(x)D(x) + p(x)Q(x))/p(x) \\ \text{which means} \quad & V_p^i(x) := \partial_j\big(p(x)D^{ij}(x) + p(x)Q^{ij}(x)\big)/p(x), \end{aligned} \tag{8}$$

and $Q(x)$ is any antisymmetric matrix ($Q^\top = -Q$) called a *curl matrix*. When D is positive (strictly) definite (i.e., it is nonsingular), p is the only stationary distribution of this dynamics. By the FPE (6), the dynamics equation (7) is equivalent to the following *deterministic* dynamics (Liu et al., 2019b, Lem. 1, Thm. 5) in the sense of the induced evolving distribution:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = D(x)\nabla \log\left(p(x)/q_t(x)\right) + Q(x)\nabla \log p(x) + \nabla \cdot Q(x), \tag{9}$$

$$\text{or,} \quad \frac{\mathrm{d}x}{\mathrm{d}t} = D(x)\nabla \log\left(p(x)/q_t(x)\right) + Q(x)\nabla \log\left(p(x)/q_t(x)\right), \tag{10}$$

where q_t is the distribution of x_t. This formulation will be explored in Section 4.4 for a geometric interpretation of a general MCMC dynamics. The deterministic equivalent of a diffusion process is also leveraged in diffusion-based generative models (Song et al., 2021) ("probability flow").

In early versions, dynamics simulation is followed by an MH step to correct discretization error. But in many cases the MH step is intractable or costly, so it is omitted when using small enough discretization.

3.2 Riemannian MCMC in coordinate space

By definition (Section 2.1), a manifold can naturally be described in each coordinate space, and we can then carry out Euclidean space operations there. This is particularly convenient for manifolds that has a global coordinate system.

3.2.1 Langevin dynamics

Perhaps the first dynamics-based MCMC method that got extended is the Langevin dynamics (Langevin, 1908; Roberts et al., 1996). Its plain form in the Euclidean space $\mathbb{R}^m$ is described by the SDE:

$$(\mathrm{LD}) \quad \mathrm{d}x = \nabla \log p_{\mathrm{L}}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \tag{11}$$

[k]The "matrix-divergence" operation $\partial_j M^{ij}$ may be better denoted as $(\nabla \cdot M^\top)^\top$. Here we just use $\nabla \cdot M$ for a concise notation.

where p_{L} is the density of the target distribution in the usual sense, i.e., w.r.t. the Lebesgue measure in $\mathbb{R}^m$ (see Section 2.5). Its manifold extension is then developed following the same form:

$$\text{(RLD)} \quad \mathrm{d}x = \operatorname{grad} \log p_{\mathrm{R}} \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}\tilde{B}_t \tag{12}$$

$$= G^{-1} \nabla \log p_{\mathrm{R}} \, \mathrm{d}t + \nabla \cdot \left(\sqrt{|G|} G^{-1} \right) / \sqrt{|G|} \, \mathrm{d}t + \sqrt{2G^{-1}} \, \mathrm{d}B_t \tag{13}$$

$$= G^{-1} \nabla \log p_{\mathrm{L}} \, \mathrm{d}t + \nabla \cdot G^{-1} \, \mathrm{d}t + \sqrt{2G^{-1}} \, \mathrm{d}B_t. \tag{14}$$

Here, $\tilde{B}_t$ is the standard Brownian motion on the Riemannian manifold. This extension can be developed from their effects in evolving distribution. In the Euclidean space the standard Brownian motion ($V \equiv 0$, $D \equiv \frac{1}{2} I_m$ in Eq. 5) leads to the diffusion equation or heat equation $\partial_t q_t = \frac{1}{2} \nabla^2 q_t$ by the FPE (6), so on the Riemannian manifold the standard Brownian motion should lead to the distribution evolution $\partial_t q_t = \frac{1}{2} \mathrm{Lap} \, q_t$ (see Section 2.6). This relates the manifold motion to the Euclidean motion in the coordinate space, which leads to its coordinate expression (Kent, 1978) and subsequently Eq. (13).

Algorithmic developments are then pursued. Roberts and Stramer (2002) developed sampling algorithms using the expression (14). Following the common practice of plain Langevin dynamics, the Euler–Maruyama integrator is used due to its simplicity:

$$x_{k+1} = x_k + \varepsilon G^{-1}(x_k) \nabla \log p_{\mathrm{L}}(x_k) + \varepsilon \nabla \cdot G^{-1}(x_k) + \mathcal{N}(0, 2G^{-1}(x_k)\varepsilon), \tag{15}$$

where "$+ \mathcal{N}(0, 2G^{-1}(x_k)\varepsilon)$" means adding a random sample from the specified normal distribution for this step of update. As the counterpart of the Euler integrator for ordinary differential equations, the Euler–Maruyama integrator is also of first order, i.e., the discretization error is proportional to the step size ε. An MH step is also called after each update to correct discretization error. This is possible since the density function of the proposal distribution by Eq. (15) can be computed. Compared to random-walk MH samplers, using Langevin dynamics makes more effective move as there is a driving force toward the nearby high-probability region (while the Brownian motion keeps the samples a reasonable dispersion and helps exploring all the high-probability regions). Moreover, if discretization error can be omitted, then the MH acceptance ratio is 1. The work also includes a convergence analysis on the simulation. Girolami and Calderhead (2011) applied the algorithm to Bayesian inference under the information geometry perspective, i.e., using the Fisher information matrix as the Riemann metric tensor: $G(x) := \mathbb{E}_{p(o|x)}[\nabla_x \log p(o|x) \nabla_x^{\top} \log p(o|x)]$. Improved convergence rate is shown empirically over several Bayesian models including Bayesian logistic regression. The same spirit is applied to Bayesian neural networks (Li et al., 2016), where a different metric is designed based on adaptive-gradient

optimization methods, as computing the Fisher information matrix is unaffordably costly in this case. Patterson and Teh (2013) made the extension to using stochastic gradient and applied to the inference task of latent Dirichlet allocation (LDA) (Blei et al., 2003) on large datasets. In the task the latent variable lies on simplexes, which is represented by nonlinear coordinate systems.

A literature remark is that the dynamics formulated in Roberts and Stramer (2002) and in Girolami and Calderhead (2011) and Patterson and Teh (2013) turn out to be[1] :

$$\mathrm{d}x = G^{-1}\nabla \log p_{\mathrm{L}}\ \mathrm{d}t + 2\nabla \cdot G^{-1}\ \mathrm{d}t + \sqrt{2G^{-1}}\ \mathrm{d}B_t, \text{ and} \tag{16}$$

$$\mathrm{d}x = G^{-1}\nabla \log p_{\mathrm{L}}\ \mathrm{d}t + 2\nabla \cdot \left(\sqrt{|G|}G^{-1}\right)/\sqrt{|G|}\ \mathrm{d}t + \sqrt{2G^{-1}}\ \mathrm{d}B_t, \tag{17}$$

respectively. The respective differences from Eqs. (14) and (13) are pointed out by Xifara et al. (2014) and are seen as a transcription error. However, Xifara et al. (2014) also showed that Eq. (17) recovers the correct dynamics by noting:

$$M^{-1}\nabla \log |M| + \nabla \cdot M^{-1} = 0, \quad \text{if } M_{ij} = M_{ji}, \text{ and } \partial_k M_{ij} = \partial_i M_{kj}, \tag{18}$$

which is satisfied if G is defined as the Hessian of a convex function (e.g., in mirror descent).

3.2.2 Hamiltonian dynamics

The Hamiltonian dynamics is a reformulation of Newton's law in classical mechanics. It describes the state of a system with the position x and also the momentum r. The augmented (x, r) space is called phase space. A nice property is that, for a target distribution $p(x)$ on the positional space, and any constant symmetric positive definite matrix Σ, the dynamics under the potential energy $-\log p(x)$,

$$(\mathrm{HMC}) \quad \mathrm{d}x = \Sigma^{-1}r\ \mathrm{d}t, \quad \mathrm{d}r = \nabla \log p(x)\ \mathrm{d}t, \tag{19}$$

keeps the phase-space distribution $p(x)\mathcal{N}(r|0,\Sigma)$ stationary. This can be seen from the FPE (6) when taking x as (x, r) on the phase space. A physical understanding can be seen by noting that the log-density of the distribution is the energy of the state (x, r) (hence Σ holds the physical meaning of a mass matrix), which the dynamics conserves. Moreover the dynamics is so-called volume preserving (Liouville theorem), i.e., the Jacobian determinant of an infinitesimal transformation by the dynamics (ratio of the infinitesimal volume change) is 1, so the distribution is kept stationary. By this property, Duane et al. (1987) proposed an MCMC algorithm called hybrid Monte Carlo, which is subsequently called Hamiltonian Monte Carlo (HMC) as the common name.

[1]We use a half time scale here: Eqs. (16) and (17) are the results after replacing t with $2t$ in the formulations in the references.

However, the simulation trajectory only covers a subset of the positional space (particularly, due to energy conservation, the trajectory cannot go beyond regions with higher potential energy than the total energy). So for drawing each sample/proposal, the momentum variable r is resampled from $\mathcal{N}(r|0,\Sigma)$. Moreover, as the volume-preserving property is a key ingredient to keep the target stationary, the simulation method (integrator) is also expected so, and to make this property well defined, the integrator is also required invertible (the continuous-time dynamics is naturally invertible). Otherwise, the volume change would accumulate for long simulations and strongly deviate the stationary distribution, and the MH acceptance probability drops, making ineffective move. Due to this consideration, the Störmer–Verlet or called the leapfrog integrator is used, which updates x and r separately on an interleaving time-discretization scheme and is both invertible and volume preserving. In all, for drawing a sample using L inner simulation steps, the process of HMC is:

$$\begin{aligned} &x^{(k,0)} := x^{(k-1)}, r^{(k,0)} \sim \mathcal{N}(0,\Sigma), r^{(k,\frac{1}{2})} := r^{(k,0)} + \frac{\varepsilon}{2}\nabla \log p(x^{(k,0)});\\ &x^{(k.l)} := x^{(k,l-1)} + \varepsilon\Sigma^{-1} r^{(k,l-\frac{1}{2})}, r^{(k,l+\frac{1}{2})} := r^{(k,l-\frac{1}{2})} + \varepsilon\nabla \log p(x^{(k,l)}),\\ &\quad l = 1,\ldots,L; x^{(k)} := x^{(k,L)}. \end{aligned} \tag{20}$$

Although the leapfrog integrator does not require second-order quantities (e.g., Hessian), it is a second-order method. An MH step is appended (where the ratio of proposal transition is 1 due to volume preservation). When $L = 1$, HMC recovers the Langevin dynamics (with $\frac{\varepsilon}{2}$ as the step size), so HMC walks much farther, making a distant thus less correlated proposal still with a high MH acceptance rate. See e.g., Neal (2011) and Betancourt (2017) for more detailed descriptions. The leapfrog integrator can be seen to split the dynamics equation (19) into $\mathrm{d}x = \Sigma^{-1} r\ \mathrm{d}t$, $\mathrm{d}r = 0$ and $\mathrm{d}x = 0$, $\mathrm{d}r = \nabla \log p(x)\ \mathrm{d}t$, and alternately simulate each symmetrically, in closed form. This pattern can be generalized to allow multiple splitting and simulating SDE (Chen et al., 2015; Hairer et al., 2006), which is called symmetric splitting integrator (SSI). The discretization error is of a higher order in ε than the Euler–Maruyama integrator. Interestingly, the interleaving structure for invertibility is of the same pattern as the coupling layer in flow-based generative models (Dinh et al., 2017; Kingma and Dhariwal, 2018), and some other works more directly use the Hamiltonian dynamics for generative modeling (Caterini et al., 2018; Dockhorn et al., 2021; Toth et al., 2020).

Riemannian HMC is developed also by Girolami and Calderhead (2011). Hamiltonian dynamics on a Riemannian manifold is given by the coordinate expression

$$\text{(RHMC)}\quad \mathrm{d}x = G(x)^{-1} r\ \mathrm{d}t, \quad \mathrm{d}r = \nabla \log p_{\mathrm{R}}(x)\ \mathrm{d}t - \frac{1}{2}\nabla_x \left(r^{\top} G(x)^{-1} r \right)\ \mathrm{d}t. \tag{21}$$

The algorithm redraws momentum from $r \sim \mathcal{N}(r|0, G(x))$, and a generalized leapfrog integrator is used in simulation. Invertibility and volume preservation are again guaranteed, but each update step requires solving two equations, which is done by fixed point iteration. To bypass this costly inner iteration, Lan et al. (2015) reformulated the dynamics using velocity instead of momentum, i.e., in the form of Lagrangian dynamics. The new form avoids one equation, and the rest equation can also be eliminated by an approximation which slightly violates volume-preservation. Lee and Vempala (2018) developed nonasymptotic convergence rate of Riemannian HMC and applied to sampling on polytopes.

3.2.3 Stochastic gradient Hamiltonian dynamics

Assuming large IID data, the stochastic gradient equation (4) can be seen as the true gradient corrupted by a Gaussian noise with zero mean and some covariance matrix $A(x)$, due to the central limit theorem. Directly using stochastic gradient in Langevin dynamics equation (11) simulation (Welling and Teh, 2011) does not introduce much problem. The change in its discretization $x_{k+1} = x_k + \varepsilon \nabla \log p_{\mathrm{L}}(x_k) + \mathcal{N}(0, 2\varepsilon)$ is to add a noise obeying $\mathcal{N}(0, A(x_k)\varepsilon^2)$ to the r.h.s. The variance of this new noise is a higher-order infinitesimal of that of the originally included noise (Chen et al., 2015).

But it is not the case for HMC which is found fragile to gradient noise (Betancourt, 2015; Chen et al., 2014), since the original dynamics is deterministic. In another way, the dynamics preserves energy so the energy from the gradient noise gets accumulated all the way, driving the distribution toward uniformity (Chen et al., 2014). To counteract the noise, Chen et al. (2014) introduced a friction term into the dynamics that dissipates the accumulated energy. The resulting stochastic-gradient HMC (SGHMC) dynamics is known as the second-order (underdamped) Langevin dynamics (Wang and Uhlenbeck, 1945) in physics[m]:

$$\text{(SGHMC)} \quad \mathrm{d}x = \Sigma^{-1} r \, \mathrm{d}t, \quad \mathrm{d}r = \nabla \log p(x) \, \mathrm{d}t - Cr \, \mathrm{d}t + \sqrt{2C\Sigma} \, \mathrm{d}B_t(r), \tag{22}$$

where C is a constant positive definite matrix controlling the friction intensity, and the Brownian motion dominates over the gradient noise as is in the Langevin dynamics case. To adaptively adjust C to match the gradient noise, Ding et al. (2014) proposed stochastic gradient Nosé–Hoover thermostats (SGNHT) which introduces a thermostat variable $\xi \in \mathbb{R}$ and composes the dynamics as[n]:

[m]Compared to Eq. (13) in Chen et al. (2014), we replace its M with Σ and its C with $C\Sigma$.
[n]Compared to Eqs. (5,6) in Ding et al. (2014), we allow a variance in the momentum distribution $\mathcal{N}(r|0, \Sigma)$ and take the diffusion factor A as $C\Sigma$.

$$\text{(SGNHT)} \quad \mathrm{d}x = \Sigma^{-1} r \, \mathrm{d}t, \quad \mathrm{d}r = \nabla \log p(x) \, \mathrm{d}t - \xi r \, \mathrm{d}t + \sqrt{2C\Sigma} \, \mathrm{d}B_t(r),$$
$$\mathrm{d}\xi = \left(\frac{1}{m} r^\top \Sigma^{-1} r - 1\right) \mathrm{d}t. \tag{23}$$

Their extensions to Riemannian manifolds are also developed. Ma et al. (2015) designed the following SGRHMC dynamics using the complete recipe (Eq. 8) to extend SGHMC:

$$\text{(SGRHMC)} \quad \mathrm{d}x = \sqrt{G(x)^{-1}} r \, \mathrm{d}t,$$
$$\mathrm{d}r = \sqrt{G(x)^{-1}} \nabla \log p_{\mathrm{L}}(x) \, \mathrm{d}t + \nabla \cdot \sqrt{G(x)^{-1}} \, \mathrm{d}t - G(x)^{-1} r \, \mathrm{d}t + \sqrt{2G(x)^{-1}} \, \mathrm{d}B_t(r).$$

Although it converges correctly, it does not seem to have a physical or geometric interpretation. Particularly, it is unknown if the dynamics can be simulated in an embedded space of the manifold, which is required in many applications as will be discussed next. To tackle this problem, Liu et al. (2016a) developed the following dynamics:

$$\text{(SGGMC)} \quad \begin{cases} \mathrm{d}x = G(x)^{-1} r \, \mathrm{d}t, \\ \mathrm{d}r = \nabla_x \log p_{\mathrm{R}}(x) \, \mathrm{d}t - \frac{1}{2} \nabla_x \left(r^\top G(x)^{-1} r\right) \mathrm{d}t - J(x)^\top C J(x) G(x)^{-1} r \, \mathrm{d}t \\ \qquad + \sqrt{2J(x)^\top C J(x)} \, \mathrm{d}B_t(r), \end{cases} \tag{24}$$

where $J(x) := \sqrt{G(x)}^\top$. If C commutes with $J(x)$ (e.g., when C is a scalar matrix), the last two terms become $-Cr \, \mathrm{d}t + \sqrt{2CG(x)} \, \mathrm{d}B_t(r)$. It recovers the RHMC dynamics equation (21) when no friction $C = 0$, and recovers the SGHMC dynamics equation (22) on a Hilbert space $G(x) \equiv \Sigma$ (note in which case $\nabla \log p_{\mathrm{R}} = \nabla \log p_{\mathrm{L}}$). Liu et al. (2016a) also proposed a manifold version of SGNHT:

$$\text{(gSGNHT)} \quad \begin{cases} \mathrm{d}x = G(x)^{-1} r \, \mathrm{d}t, \\ \mathrm{d}r = \nabla_x \log p_{\mathrm{R}}(x) \, \mathrm{d}t - \frac{1}{2} \nabla_x \left(r^\top G(x)^{-1} r\right) \mathrm{d}t - \xi r \, \mathrm{d}t + \sqrt{2CG(x)} \, \mathrm{d}B_t(r), \\ \mathrm{d}\xi = \left(\frac{1}{m} r^\top G(x)^{-1} r - 1\right) \mathrm{d}t, \end{cases} \tag{25}$$

which recovers the SGNHT dynamics equation (23) when $G(x) \equiv \Sigma$.

3.3 Riemannian MCMC in embedded space

The coordinate expressions of these dynamics allow straightforward simulation in the coordinate space, as if in the usual Euclidean space. However, this may also be problematic in some cases. When the manifold does not have a global coordinate system, such as hyperspheres, coordinate simulation requires

changing local coordinate system when the sample is about to move out of the current one. In addition to this inconvenience, the metric tensor $G(x)$ often goes to singularity near the boundary of the coordinate system, which may cause numerical problems. On the other hand, common manifolds are defined as a subset of a higher-dimensional Euclidean space, which also endows the manifold a Riemannian metric by pulling-back the Euclidean metric via the inclusion map. This naturally gives an isometric embedding of the manifold as a result (Section 2.7). In the embedded space the manifold is described globally. Distributions on the manifold are also commonly defined in the embedded space in the form of the density function p_{H} w.r.t. the Hausdorff measure on $\Xi(\mathcal{M})$, such as the von Mises-Fisher distribution on hyperspheres. It is thus more attractive to simulate the dynamics in the embedded space.

Brubaker et al. (2012) considered general manifold defined as a subset of $\mathbb{R}^n$ via a constraint, and generalized HMC on such manifold. Simulation is basically done using the usual leapfrog process in the embedded space subject to the constraint for the sample and the induced constraint on the momentum. This generalized leapfrog is called RATTLE and is still invertible, volume preserving, and of the second order. But each simulation step requires solving several nonlinear equations, which is done using Newton's method in the work.

Byrne and Girolami (2013) developed a simulation method in the embedded space for HMC, called geodesic Monte Carlo (GMC). The name comes from the use of explicit geodesic flow solutions for simulation on the embedded manifold, which is available for common manifolds such as the simplex, hypersphere, and the Stiefel manifold. It avoids the costly inner iteration for solving nonlinear equations in each update step, in contrast to, e.g., Girolami and Calderhead (2011) and Brubaker et al. (2012). Adopting the leapfrog integrator pattern (or, SSI), GMC splits the Riemannian Hamiltonian dynamics equation (21) as:

$$\begin{cases} \mathrm{d}x = G(x)^{-1} r \, \mathrm{d}t, \\ \mathrm{d}r = -\frac{1}{2} \nabla_x \left(r^\top G(x)^{-1} r \right) \mathrm{d}t, \end{cases} + \begin{cases} \mathrm{d}x = 0, \\ \mathrm{d}r = \nabla \log p_{\mathrm{R}}(x) \, \mathrm{d}t. \end{cases}$$

The first dynamics turns out to be the geodesic equation, which describes the free motion on the manifold, in analogy to the uniform linear motion in a Euclidean space. Its solution is just the geodesic flow, and for some common manifolds there is a closed-form expression, e.g., this is the rotation along the great circle on the hypersphere $\mathbb{S}^{n-1}$:

$$y(t) = y(0) \cos(\alpha t) + (v(0)/\alpha) \sin(\alpha t), \quad v(t) = -\alpha y(0) \sin(\alpha t) + v(0) \cos(\alpha t),$$

where $y \in \mathbb{S}^{n-1}$ and $v \in T_y(\mathbb{S}^{n-1})$ are the embedded version of x and r, and $\alpha = \| v(0) \|$. For the second dynamics, since x does not change with time, the solution is $r(t) = r(0) + t \nabla \log p_{\mathrm{R}}(x)$, or $y(t) = y(0) + t \Lambda(y(0)) \nabla \log p_{\mathrm{H}}(y)$ in

the embedded space, where $\Lambda(y)$ is the projection from the tangent space of $\mathbb{R}^n$ to the tangent space of $\Xi(\mathcal{M})$ at y. For $\mathbb{S}^{n-1}$, $\Lambda(y) = I_n - yy^\top$. For drawing the momentum variable, sampling $r \sim \mathcal{N}(0, G(x))$ in the coordinate space is equivalent to drawing $r \in \mathcal{N}(0, I_n)$ in the isometrically embedded space $\mathbb{R}^n$ and projecting onto the tangent space using Λ. The final process is to draw r from $\mathcal{N}(0, G(x))$ and alternately simulating the two dynamics on an interleaving time scheme, similar to the leapfrog integrator equation (20).

Byrne and Girolami (2013) also showed closed-form expressions of the geodesic flow and tangent space projection for the Stiefel manifold, which involves computationally expensive matrix exponential. Yanush and Kropotov (2019) then made the computation cheaper by replacing the exact operations with a retraction operation, which is explored in the field of optimization on Riemannian manifolds, and is shown to be sufficient to simulate the dynamics.

Following this spirit, Liu et al. (2016a) developed simulation methods in the embedded space for their proposed dynamics. The SGGMC dynamics equation (24) is split as:

$$\begin{cases} \mathrm{d}x = G(x)^{-1} r \, \mathrm{d}t, \\ \mathrm{d}r = -\frac{1}{2} \nabla_x \left(r^\top G(x)^{-1} r \right) \mathrm{d}t, \end{cases} + \begin{cases} \mathrm{d}x = 0, \\ \mathrm{d}r = -J(x)^\top C J(x) G(x)^{-1} r \, \mathrm{d}t, \end{cases}$$

$$+ \begin{cases} \mathrm{d}x = 0, \\ \mathrm{d}r = \nabla_x \log p_{\mathrm{R}}(x) \, \mathrm{d}t + \sqrt{2 J(x)^\top C J(x)} \, \mathrm{d}B_t(r), \end{cases}$$

where the first dynamics is the same. The second dynamics has solution $r(t) = J^\top \mathrm{expm}(-Ct) J G^{-1} r(0)$ or $v(t) = \Lambda \, \mathrm{expm}(-Ct) v(0)$ in the embedded space, where expm is matrix exponential and the quantities are evaluated at $x(0)$ or $y(0)$ if not specified. For infinitesimal time interval ε, the solution is approximately $v(\varepsilon) = v(0) - \varepsilon \Lambda C v(0)$. The third dynamics can be simulated by $v(\epsilon) = v(0) + \Lambda(\varepsilon \nabla \log p_{\mathrm{H}} + \mathcal{N}(0, 2C\varepsilon))$. The gSGNHT dynamics equation (25) can be modified similarly. The two methods show remarkable accuracy and scalability for the hypersphere inference task of the spherical admixture model (Reisinger et al., 2010) on large datasets.

4 Particle-based variational inference methods

Although MCMC methods are perhaps the most popular sampling method for their wide applicability and numerous success in various domains, there are also complaints. The central drawback is the (auto)correlation among samples since they are in a Markov chain. It downgrades the effective sample size, and a long run is required for a reasonable approximation. This can be imagined that if the target distribution has two equally high separated modes, the

correlation tends to put the next sample in the same mode, so such set of samples are not as representative as the same amount of IID samples, which distribute in the two modes roughly equally. A considerable number of simulation steps are required to get a chance for the sampler to find the other mode. Besides the cost of the long run itself, evaluating the expectation of a function also requires quite a lot evaluations on these large set of samples.

Recently there emerged another class of sampling methods that try to approximate the target distribution using a *fixed* number of samples, or called particles. They iteratively update the particles deterministically to minimize the difference (typically the KL divergence) of the distribution they represent to the target distribution. This is the spirit of variational inference, hence the name particle-based variational inference (ParVI). The use of particles is more flexible thus more accurate than parametric approximations that classical variational inference uses. Since the particles need to approximate the target distribution jointly as a whole, they communicate/interact with each other, and appear negatively correlated. So sample efficiency is promoted.

In the following we start with the groundbreaking version, SVGD (Liu and Wang, 2016). The algorithm is then viewed from a geometric perspective as approximate simulation to the gradient flow of the KL divergence on the Wasserstein space and its variants, which can be seen as Riemannian manifolds of distributions. We then show the consequences of the geometric perspective, including variant algorithms, approximation techniques, and convergence analysis.

4.1 Stein variational gradient descent

SVGD (Liu and Wang, 2016) updates the particles $\{x^{(i)}\}_{i=1}^N$ using a deterministic dynamics $\mathrm{d}x = V_t(x)\,\mathrm{d}t$ such that the evolving distribution $(q_t)_t$ decreases the KL divergence to the target distribution p. The dynamics is chosen such that the KL divergence $\mathrm{KL}_p(q_t) := \mathrm{KL}(q_t \parallel p) := \mathbb{E}_{q_t}[\log(q_t/p)]$ is minimized steepest. To derive the expression for $V_t(x)$, one needs to connect the decreasing rate to the dynamics:

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}_p(q_t) = \mathbb{E}_{q_t}[\nabla \cdot (pV_t)/p] = \mathbb{E}_{q_t}[V_t \cdot \nabla \log p + \nabla \cdot V_t]. \tag{26}$$

The steepest descent is achieved if V_t maximizes Eq. (26). As the objective is linear in V_t, it can be made arbitrarily large by increasing the norm of V_t. So the maximization makes sense only if the norm is fixed, i.e., finding the steepest descending direction, and the minimum represents the steepest rate of decrease which is used as the magnitude in that direction:

$$V_t^{\mathrm{opt}} := \max_{V_t \in \mathfrak{X}, \|V_t\|_{\mathfrak{X}}=1} \cdot \arg\max \; -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}_p(q_t). \tag{27}$$

This leads to the next subtlety that which normed space $\mathfrak{X}$ should be used for V_t. If $\mathfrak{X}$ is taken as the Hilbert space $\mathcal{L}^2_{q_t}(\mathbb{R}^m) := \{V : \mathbb{R}^m \to \mathbb{R}^m | \ \|V\|_{\mathcal{L}^2_{q_t}} < \infty\}$ with inner product $\langle U, V\rangle_{\mathcal{L}^2_q} := \int U(x) \cdot V(x) q(x)\, \mathrm{d}x$, we have:

$$V_t^{\mathrm{L2}} = \nabla \log (p/q_t). \tag{28}$$

But this result is not easily estimable since it requires $\nabla \log q_t$ while we only have samples/particles of q_t. Liu and Wang (2016) then used $\mathfrak{X} = \mathcal{H}^m$, where $\mathcal{H}$ is the reproducing kernel Hilbert space (RKHS) of a kernel function $K : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ (Steinwart and Christmann, 2008, Ch. 4), which can be seen as the linear space of functions $\left\{\sum_l \alpha_l K(\,\cdot\,, x^{(l)})\right\}$ (summation over finite set or countably infinite set in the sense of pointwise convergence) with inner product $\left\langle \sum_l \alpha_l K(\,\cdot\,, x^{(l)}), \sum_{l'} \beta_{l'} K(\,\cdot\,, y^{(l')})\right\rangle_{\mathcal{H}} := \sum_{l,l'} \alpha_l \beta_{l'} K(x^{(l)}, y^{(l')})$. It is named after its reproducing property $\langle f(\,\cdot\,), K(\,\cdot\,, x)\rangle_{\mathcal{H}} = f(x), \forall f \in \mathcal{H}$ and that it is a Hilbert space. The solution is then:

$$V_t^{\mathrm{SVGD}}(x) := \mathbb{E}_{q_t(x')}[K(x, x')\nabla_{x'} \log p(x') + \nabla_{x'} K(x, x')]. \tag{29}$$

The expression depends on q_t only in terms of expectation, which can be estimated by averaging over the particles. The particles are then updated by simulating the dynamics:

$$x_{k+1}^{(i)} = x_k^{(i)} + \varepsilon \frac{1}{N} \sum_{j=1}^{N} \left(\mathbf{K}_k^{(ij)} \nabla \log p\left(x_k^{(j)}\right) + \nabla_{x_k^{(j)}} \mathbf{K}_k^{(ij)}\right), \tag{30}$$

where $\mathbf{K}_k^{(ij)} := K\left(x_k^{(i)}, x_k^{(j)}\right)$. Since $\mathbf{K}^{(ij)}$ increases as $x^{(i)}$ and $x^{(j)}$ move closer, the gradient $\nabla_{x^{(j)}} \mathbf{K}^{(ij)}$ points toward $x^{(i)}$ at $x^{(j)}$, and points away from $x^{(j)}$ at $x^{(i)}$. So the second term in the update pushes $x^{(i)}$ away from all other particles, which presents a repulsion force and incurs a negative correlation among the particles. Another attractive property is that it degenerates to gradient descent targeting the mode of $p(x)$ if there is only one particle, since $\nabla_{x'} K(x, x')|_{x'=x} = 0$, so it naturally transits to the maximum a posteriori (MAP) estimate.

4.2 The Wasserstein space

As seen above, the derivation of SVGD involves geometric considerations such as steepest descending direction, which is very much alike a gradient. Can this geometric perspective be formalized so that Eq. (27) is the gradient of KL_p on some space of distributions? The Wasserstein space gives an elegant answer.

Particularly we consider the 2-Wasserstein space $\mathcal{P}_2(\mathcal{M})$ on a complete metric space $\mathcal{M}$ ($= \mathbb{R}^m$ for common ParVIs), defined as the set of all (unnecessarily absolutely continuous) distributions on $\mathcal{M}$ with finite variance: $\exists x_0 \in \mathcal{M}$ s.t. $\mathbb{E}_q[d_{\mathcal{M}}(x, x_0)^2] < \infty$. It is usually equipped with the 2-Wasserstein

distance and seen as a metric space (Villani, 2008, Def. 6.4; Ambrosio et al., 2008, Ch. 7), which explains the name. The 2-*Wasserstein distance* (Villani, 2008, Def. 6.1; Ambrosio et al., 2008, Eq. (7.1.1)) between two distributions $p,q \in \mathcal{P}_2(\mathcal{M})$ arises from the *optimal transport* problem, and is defined as the minimal cost for transporting mass on $\mathcal{M}$ distributed as $p(x)$ to distribute as $q(x')$ by all possible (may be stochastic) transport plans $\pi(x'|x)$:

$$d_{\mathcal{P}_2}(p,q) := \left(\inf_{\pi(x'|x):\ \int_{\mathcal{M}} \pi(x'|x)\, \mathrm{d}p(x) = q(x')} \int_{\mathcal{M}\times\mathcal{M}} d_{\mathcal{M}}(x,x')^2\, \mathrm{d}(\pi(x'|x)p(x)) \right)^{1/2}. \tag{31}$$

We first note that quite a range of geometric structures can be extended to a general metric space (Ambrosio et al., 2008, Part I). For future reference, we mention a particular example of the definition of the gradient flow $(q_t)_t$ of a function F on $\mathcal{P}_2(\mathcal{M})$ as a metric space in terms of the minimizing movement scheme (Ambrosio et al., 2008, Def. 2.0.6): $(q_t)_t$ is the limiting curve when $\tau \to 0$ of the piecewise constant curve (Ambrosio et al., 2008, Def. 2.0.2):

$$q_{t+\varepsilon} := \operatorname*{argmin}_{q\in\mathcal{P}_2(\mathcal{M})} \frac{1}{2\tau} d^2_{\mathcal{P}_2}(q,q_t) + F(q), \quad \forall \varepsilon \in (0,\tau]. \tag{32}$$

It extends the implicit Euler discretization.[o] Nevertheless, it is more attractive to consider the space as a Riemannian manifold (if possible) for explicit calculation and simulation. This is possible for $\mathcal{P}_2(\mathcal{M})$. We start with characterizing tangent vectors on it. See Fig. 8 for an illustration.

Similar to the process in defining tangent vector for a general manifold in Section 2.2, consider a curve $(q_t)_t$ on $\mathcal{P}_2$ which is an evolving distribution on $\mathcal{M}$. As mentioned in Section 3.1, a dynamics induces an evolving distribution, and the connection is given by the FPE (6). If we only consider deterministic dynamics ($D \equiv 0$), the FPE becomes what is commonly called the *continuity*

[o]Other examples of concepts mentioned in Ambrosio et al. (2008) for a general complete metric space: the speed (metric derivative) at a point on a curve (Thm. 1.1.2), curve absolute continuity (Def. 1.1.1) and length (Lem. 1.1.4), geodesic (Def. 2.4.2), gradient of a function in terms of norm/modulus (strong (Def. 1.2.1) and weak (Def. 1.2.2) upper gradient; local (also Villani, 2008, Prop. 23.1(ii)), and global slopes (Def. 1.2.4) are weak, and strong (for lower semicontinuous functions) upper gradients (Thm. 1.2.5); relaxed slope (Section 2.3) is the local slope under some conditions (Rem. 2.3.2)), and gradient flow [curve of maximal slope (Def. 1.3.2) in terms of the evolution variational inequality, which indicates energy identity (also Villani, 2008, Prop.23.1 (ii)) in some cases (Rem. 1.3.3); (generalized) minimizing movement scheme (Def. 2.0.6), whose convergence (Cor. 2.2.2) and connection to curve of maximal slope (Thm. 2.3.1) and energy identity (Thm. 2.3.3) are made; curve with dispersion from geodesic bounded by directional derivative along the geodesic (Villani, 2008, Prop. 23.1(iv), Def. 23.7); they all coincide with the Riemannian gradient flow (Villani, 2008, Prop.23.1, Rem. 23.4; Ambrosio et al., 2008, Thm. 11.1.6)].

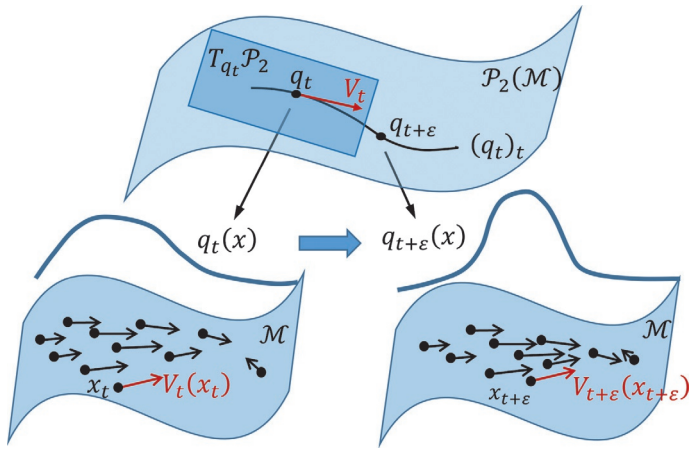

FIG. 8 Illustration of the support-space vector-field representation of a Wasserstein tangent vector. Similar to the illustration in Fig. 7, when particles $\{x_t^{(i)}\}_i$ distribute obeying q_t at time t and move under a deterministic dynamics $\mathrm{d}x_t = V_t(x_t)\,\mathrm{d}t$ (*bottom two red arrows*), i.e., each moving with velocity $V_t(x_t^{(i)})$, they form a new configuration at time $t+\varepsilon$ which defines a new distribution $q_{t+\varepsilon}$. As $\varepsilon \to 0$, the dynamics induces a continuously evolving distribution $(q_t)_t$, which is a curve on the Wasserstein space $\mathcal{P}_2(\mathcal{M})$. The tangent vector along the curve at q_t is the instantaneous evolution of the distribution at t, which is given by the continuity equation (33) (or, the FPE (6) without diffusion), which is in turn determined by V_t. So the vector field V_t represents a Wasserstein tangent vector (the *red arrow* at the *top*).

equation or the *conservation-of-mass formula*, which takes the following form on manifold $\mathcal{M}$:

$$\partial_t q_t = -\mathrm{div}(q_t V_t), \tag{33}$$

where q_t is the density under the Riemannian measure [(Villani, 2008, p. 28); see e.g., Liu and Zhu (2018, Appx. A1) for deriving the manifold extension]. The conclusion equation (33) can be extended to distributions that do not have a density (i.e., not absolutely continuous) under the Riemannian measure in the sense of *weak derivatives* (or called in the sense of distributions) (Ambrosio et al., 2008, Eq. (8.1.3), Rem. 8.1.1):

$$-\int_{\mathbb{R}}\int_{\mathcal{M}} \partial_t f_t(x)\,\mathrm{d}q_t(x)\mathrm{d}t = \int_{\mathbb{R}}\int_{\mathcal{M}} \langle \mathrm{grad}\, f_t(x), V_t(x)\rangle_{T_x\mathcal{M}}\,\mathrm{d}q_t(x)\mathrm{d}t, \quad \forall f \in C_c^\infty(\mathbb{R}\times\mathcal{M}).$$

(Note Eq. 3 for understanding this form of Eq. 33.) Conversely, it is shown (Ambrosio et al., 2008, Thm. 8.3.1; Villani, 2008, Thm. 13.8; Erbar et al., 2010, Prop. 2.5) that for any curve $(q_t)_t$ on $\mathcal{P}_2(\mathcal{M})$, at point q_t there exists a vector field V_t on $\mathcal{M}$ such that the continuity equation (33) holds in the weak sense, and the *unique existence* (up to a set of $\mathrm{d}q_t(x)\mathrm{d}t$-measure zero) is attained in a subspace of vector fields,

$$\overline{\{\operatorname{grad}\varphi | \varphi \in C_c^\infty(\mathcal{M})\}}^{\mathcal{L}^2_{q_t}(\mathcal{M})}, \tag{34}$$

where the overline means the closure as a subset of the Hilbert space $\mathcal{L}^2_{q_t}(\mathcal{M})$.[p,q] This correspondence suggests using a vector field on $\mathcal{M}$ to represent a tangent vector on $\mathcal{P}_2(\mathcal{M})$, and using the above subspace of vector fields equation (34) as the tangent space $T_{q_t}\mathcal{P}_2(\mathcal{M})$ (Ambrosio et al., 2008, Def. 8.4.1, Prop. 8.4.5; Erbar et al., 2010, Def. 2.3).

What is particularly insightful of this correspondence is that the vector field V_t defines a dynamics on $\mathcal{M}$ which determines how the Wasserstein curve $(q_t)_t$ moves along the corresponding Wasserstein tangent vector as an evolving distribution. So we can carry out simulation of the Wasserstein curve by simulating the dynamics on samples/particles in $\mathcal{M}$. Specifically, $\operatorname{Exp}_{\cdot}(\varepsilon V_t(\cdot))_\#[q_t]$ is a first-order approximation of $q_{t+\varepsilon}$ in terms of the Wasserstein distance $d_{\mathcal{P}_2}$ (Ambrosio et al., 2008, Prop. 8.4.6), where $\operatorname{Exp}_{\cdot}(\varepsilon V_t(\cdot)) : \mathcal{M} \to \mathcal{M}, x \mapsto \operatorname{Exp}_x(\varepsilon V_t(x))$ is the exponential map on the support space $\mathcal{M}$ as a function of position, and $\phi_\#[q]$ is the pushed-forward distribution of q by measurable map $\phi : \mathcal{M} \to \mathcal{M}$, defined by $\phi_\#[q](\mathcal{I}) := q(\phi^{-1}(\mathcal{I}))$ for any measurable subset $\mathcal{I} \in \mathcal{M}$ (Billingsley, 2012, p. 196). More concretely, this means that if $\{x^{(i)}\}$ is a set of particles of q_t, then $\{\operatorname{Exp}_{x^{(i)}}(\varepsilon V_t(x^{(i)}))\}$ is approximately a set of particles of $q_{t+\varepsilon}$.

For a Riemannian structure, a natural inner product in the tangent space $T_{q_t}\mathcal{P}_2(\mathcal{M})$ is the one inherited from $\mathcal{L}^2_{q_t}(\mathcal{M})$. What makes a coincidence is that the induced distance in the Riemannian sense (Eq. 2) is exactly the 2-Wasserstein distance (Eq. 31), which is revealed by the Benamou–Brenier formula (Benamou and Brenier, 2000; Otto, 2001; Villani, 2008, Ch. 15). So this Riemannian structure can be seen as a finer characterization of the Wasserstein space as a metric space. Various geometric objects can be then induced. Of particular interest is the gradient of a function F on $\mathcal{P}_2(\mathcal{M})$. The general formulation is (Villani, 2008, Ex. 15.10; Ambrosio et al., 2008, Lem. 10.4.1):

---

[p]For this conclusion, Ambrosio et al. (2008, Thm. 8.3.1, Def. 5.1.11) and Erbar et al. (2010, Prop. 2.5) require the curve $(q_t)_t$ to be locally absolutely continuous in a metric-space sense (Ambrosio et al., 2008, Def. 1.1.1; Erbar et al., 2010, Def. 2.2), while Villani (2008, Thm. 13.8) further restricts $(q_t)_t$ to be Lipschitz-continuous which enlarges the subspace by allowing $\varphi \in C_c^1(\mathcal{M})$.

[q]The space defined in Eq. (34) can be seen as the quotient space of $\mathcal{L}^2_{q_t}(\mathcal{M})$ under the equivalent relation $U \sim V : \operatorname{div}(q_t U) = \operatorname{div}(q_t V)$, i.e., they induce the same evolving distribution via the continuity equation (33). More precisely, the space (34) is the orthogonal complement of $\{V \in \mathcal{L}^2_{q_t}(\mathcal{M}) \mid \operatorname{div}(q_t V) = 0\}$ (Erbar et al., 2010, Lem. 2.4), as can be seen from Eq. (3): $\operatorname{div}(q_t V) = 0 \iff \int_{\mathcal{M}} \varphi \operatorname{div}(q_t V)\, \mathrm{d}\omega_g = -\int_{\mathcal{M}} \langle \operatorname{grad}\varphi, V\rangle_{T,\mathcal{M}} q_t \,\mathrm{d}\omega_g = -\langle \operatorname{grad}\varphi, V\rangle_{\mathcal{L}^2_{q_t}(\mathcal{M})} = 0, \forall \varphi \in C_c^\infty(\mathcal{M})$. This space (34) is also equivalent to the set of vector fields that achieve a certain evolving distribution with the minimum $\mathcal{L}^2_{q_t}(\mathcal{M})$ norm: $\{\operatorname{argmin}_{V \in \mathcal{L}^2_{q_t}(\mathcal{M}), \operatorname{div}(q_t V) = f} \| V \|_{\mathcal{L}^2_{q_t}(\mathcal{M})} \mid f \in L^1_{\omega_g}(\mathcal{M}) \text{ s.t. } \int_{\mathcal{M}} f \,\mathrm{d}\omega_g = 0\}$ (Erbar et al., 2010, Lem. 2.4).

$$\operatorname{grad} F(q) = \operatorname{grad} \frac{\delta F}{\delta q}(q), \tag{35}$$

where the function $\frac{\delta F}{\delta q}(q)$ on $\mathcal{M}$ is the variation of F at q as a functional of distribution q, characterized by $\frac{\mathrm{d}}{\mathrm{d}t}F(q_t) = \int_{\mathcal{M}} \frac{\delta F}{\delta q}(q_t)\, \mathrm{d}\partial_t q_t$ in the sense of weak derivative for any evolving distribution $(q_t)_t$. Note that the l.h.s. is a gradient vector at q on the Wasserstein space $\mathcal{P}_2(\mathcal{M})$, while the r.h.s is a gradient function as a vector field on the supporting manifold $\mathcal{M}$. Particularly for the KL divergence, its gradient is given by (Villani, 2008, Formula 15.2, Thm. 23.18)[r]:

$$\operatorname{grad} \mathrm{KL}_p(q) = \operatorname{grad} \log(q/p). \tag{36}$$

Note that this coincides with $-V_t^{\mathrm{L2}}$ (Eq. 28) on Euclidean spaces. For an analogy to the Euclidean case, the gradient flow converges exponentially if the Wasserstein function is geodesically strongly convex on the Wasserstein space (Villani, 2008, Thm. 23.25, Thm. 24.7; Ambrosio et al., 2008, Ex. 11.1.2). For the KL divergence, this requires $\log p$ to be strongly log-concave for $\mathcal{M} = \mathbb{R}^m$ (Ambrosio et al., 2008, Def. 9.4.9, Lem. 9.4.7) and a similar but involved requirement for a general manifold $\mathcal{M}$ (Villani, 2008, Thm. 17.15). Gradient flow can then be defined and simulated in the way described in the previous paragraph.

As a literature remark, Villani (2008) also made detailed description on the optimal transport problem, dedicated analysis on the curvature of $\mathcal{P}_2(\mathcal{M})$ induced from that of the base manifold $\mathcal{M}$, and on the convexity of functions on $\mathcal{P}_2(\mathcal{M})$. Ambrosio et al. (2008) extended and formalized the concept of gradient flow and other geometric objects for general metric spaces. For the Wasserstein space, they only considered the Euclidean support but made analysis in parallel on the p-Wasserstein space $\mathcal{P}_p(\mathbb{R}^m)$ for $p \neq 2$, which is no longer a Riemannian manifold but a metric space. Erbar et al. (2010) and Santambrogio (2017) gave concise summaries of the two books. Lott (2008) presents explicit calculations of more Riemannian manifold objects for $\mathcal{P}_2(\mathcal{M})$.

In the spirit of information geometry (see Fig. 1), Chen and Li (2018) and Wang and Li (2022) also considered using the Wasserstein distance (resp. KL divergence) to measure the infinitesimal difference between two likelihood distributions, and induced a parametric Wasserstein metric (resp. Fisher–Rao metric). More variants are also considered (e.g., the Stein geometry to be introduced in Section 4.3.2).

4.3 Geometric view of particle-based variational inference methods

As mentioned in Section 4.1, the development of SVGD resembles the process of defining the gradient as a steepest ascending direction on an abstract

[r]To make the expression well defined, q needs to be absolutely continuous w.r.t. p.

space. We now review some formal descriptions of this intuition from the view on the Wasserstein space and a variant space. In this section we restrict our attention to the Euclidean case $\mathcal{M} = \mathbb{R}^m$.

4.3.1 View from the Wasserstein space

As seen from Section 4.2, theory on the Wasserstein space formalizes the intuition of the gradient of a Wasserstein function, and links it to a particle dynamics, so it is a natural choice (Chen et al., 2018a; Liu et al., 2019a). Indeed, as mentioned right below Eq. (27), if the space $\mathfrak{X}$ is taken as $\mathcal{L}^2_{q_t}(\mathbb{R}^m)$, the optimal vector field V_t^{L2} is just $-\mathrm{grad}\ \mathrm{KL}_p$ by Eq. (36). To relate it to the SVGD vector field equation (29), Liu et al. (2019a, Thm. 2) gave the following relation:

$$V_t^{\mathrm{SVGD}} = \max_{V_t \in \mathcal{H}^m, \|V_t\|_{\mathcal{H}^m}=1} \cdot \operatorname{argmax} \langle -\mathrm{grad}\ \mathrm{KL}_p(q_t), V_t \rangle_{\mathcal{L}^2_{q_t}(\mathbb{R}^m)},$$

which suggests V_t^{SVGD} is the projection of the Wasserstein gradient of KL_p onto the vector-valued RKHS $\mathcal{H}^m$.

To further understand this approximation, note that if the space of optimization is taken as $\mathcal{L}^2_{q_t}(\mathbb{R}^m)$, then $-\mathrm{grad}\ \mathrm{KL}_p$ is recovered. For a Gaussian kernel, the replacement from $\mathcal{L}^2_{q_t}(\mathbb{R}^m)$ to $\mathcal{H}^m$ is roughly doing a kernel smoothing: $\mathcal{H}^m$ is isometrically isomorphic to $\overline{\{\phi * K \mid \phi \in \mathcal{L}^2_{q_t}\}}^{\mathcal{L}^2_{q_t}}$ (Liu et al., 2019a, Thm. 3), where $(\phi * K)(x) := \int_{\mathbb{R}^m} \phi(x') K(x, x')\, \mathrm{d}x'$ is the convolution of ϕ with the kernel K. Furthermore, for each individual ϕ, this kernel smoothing reduces its "sharpness": elementwise, $\phi^i * K$ is $\sqrt{\frac{2}{\pi}}\frac{B}{\sigma}$-Lipschitz, where $B := \sup_{x \in \mathbb{R}^m} |\phi^i(x)|$ bounds the original function ϕ^i, and σ is the bandwidth of the Gaussian kernel K (He et al., 2022, Thm. 2). This smoothing operation over vector fields is shown equivalent to smoothing the density itself, i.e., using $\tilde{q}_K := q * K$ in place of q so that the empirical distribution $\hat{q}(x) := \frac{1}{N}\sum_{i=1}^N \delta_{x^{(i)}}(x)$ can be plugged in for q, where $\delta_{x^{(i)}}(x)$ denotes the Dirac delta measure that puts all the probability mass only at the point $x^{(i)}$. The smoothing operation (in either way) is regarded mandatory for a well-defined vector field (Liu et al., 2019a), as $\mathrm{KL}_p(q)$ is defined as infinity if q is not absolutely continuous w.r.t. p, which is the case when $q = \hat{q}$.

4.3.2 View from the Stein geometry

Under the view of Wasserstein space, the kernel enters only as a smoothing approximation to geometric objects. This may be insufficient to analyze the SVGD dynamics precisely, and a geometric structure involving the kernel is expected. Liu (2017) made an attempt where the tangent space, i.e., a space

of vector fields, is taken as the vector-valued RKHS $\mathcal{H}^m$ with the same inner product. Under this choice, the gradient of KL_p recovers the SVGD vector field equation (29) indeed. Nevertheless, this is only a formulation. For a complete geometric development, it remains to show the manifold as a set and the one-to-one correspondence between an element in the defined tangent space and a tangent vector of a curve on the manifold. In this spirit, the theory is improved by Duncan et al. (2019), who called it Stein geometry.[s]

Specifically, the manifold as a set is taken as the collection of fully supported, absolutely continuous distributions that make $K(x, x)$ have finite expectation, denoted $\mathcal{P}_K$. In contrast to using the entire $\mathcal{H}^m$, they refine the tangent space at $q \in \mathcal{P}_K$ as (Duncan et al., 2019, Def. 5):

$$T_q\mathcal{P}_K := \overline{\{\mathcal{K}_q \nabla\varphi | \varphi \in C_c^\infty(\mathbb{R}^m)\}}^{\mathcal{H}^m},$$

$$\text{where} \qquad \mathcal{K}_q : L_q^2(\mathbb{R}^m) \to \mathcal{H}, \quad (\mathcal{K}_q f)(x) := \int_{\mathbb{R}^m} K(x,x')f(x')q(x')\,\mathrm{d}x'$$

is the integral operator of kernel K and is applied elementwise if the operand is a vector.[t] This is similar to Eq. (34) except that kernel smoothing is applied to each vector field, so it guarantees a unique representation of a tangent vector of $\mathcal{P}_K$ (Duncan et al., 2019, Lem. 7(2)) (see Footnote 4.2). Under the same inner product of $\mathcal{H}^m$, the gradient of a function F on $\mathcal{P}_K$ is (Duncan et al., 2019, Lem. 9) $\mathrm{grad}_{\mathcal{P}_K} F(q) = \mathcal{K}_q \nabla \frac{\delta F}{\delta q}(q)$, which is the kernel-smoothed Wasserstein gradient equation (35). This leads to the SVGD vector field

$$\mathrm{grad}_{\mathcal{P}_K} \mathrm{KL}_p(q_t) = -V_t^{\mathrm{SVGD}}, \tag{37}$$

since $V_t^{\mathrm{SVGD}} = \mathcal{K}_{q_t} V_t^{\mathrm{L2}} = -\mathcal{K}_{q_t} \mathrm{grad}\ \mathrm{KL}_p(q_t)$ by definition (Eq. 29) (and integration by parts). More geometric structures such as the corresponding distance (Duncan et al., 2019, Def. 13) and geodesic (Duncan et al., 2019, Prop. 18) are developed. Note that if $K(x, x') = \delta_x(x')/q(x')$, then $\mathcal{K}_q$ is the identity operator, and Stein geometry reduces to the Wasserstein geometry.

---

[s]Nevertheless, this name may not well reflect the exact characteristics of the theory. Back to the origin (Liu and Wang, 2016), the label "Stein" is introduced via the connection of the objective in Eq. (27) to Stein's method (Stein, 1972) for constructing probability metric (Stein discrepancy (Gorham and Mackey, 2015)), where $\mathcal{S}_p : V \mapsto -\nabla \cdot (pV)/p$ is called Stein operator. It describes how the distribution evolves and is also used in the Wasserstein space case. What is special in the new geometric structure is the incorporation of kernels. So it may be better called "kernel-smoothed geometry" or "kernelized-Stein geometry."

[t]Duncan et al. (2019) originally defined the tangent space as the corresponding space of $\partial_t q_t$ via the continuity equation (33), i.e., applying $V \mapsto -\nabla \cdot (qV)$ to the definition here. The results of the two definitions are isometrically isomorphic thus both are Hilbert spaces (Duncan et al., 2019, Lem. 7).

Duncan et al. (2019) also introduced some concepts and results that are particularly useful for analyzing the SVGD dynamics. They defined the Hessian of KL_p (Lem. 22), local geodesic strong convexity of KL_p in terms of the Hessian (Lem. 25), and the consequent exponential convergence of KL_p near the optimal solution p (Thm. 20). Furthermore, they found an equivalent criterion for the strong convexity of KL_p near p with strength $\lambda > 0$, called the *Stein–Poincaré inequality* (Lem. 32):

$$\begin{aligned} &\langle \varphi, \mathcal{A}_{p,K}\varphi\rangle_{L^2_p(\mathbb{R}^m)} \geq \lambda \langle \varphi, \varphi\rangle_{L^2_p(\mathbb{R}^m)}, \quad \forall \varphi \in C_c^\infty(\mathbb{R}^m) \text{ s.t. } \int_{\mathbb{R}^m} \varphi \, \mathrm{d}p = 0, \\ \text{where} \quad &\mathcal{A}_{p,K}\varphi := -\nabla \cdot (p\mathcal{K}_p \nabla \varphi)/p, \quad \forall \varphi \in C_c^\infty(\mathbb{R}^m) \end{aligned} \tag{38}$$

is the kernelized Barbour's generator of the Stein gradient flow of KL_p, which relates to the kernelized Stein operator by $\mathcal{A}_{p,K}\varphi = \mathcal{S}_{p,K}\nabla\varphi$.[u] This criterion makes convexity identification easier, since $\mathcal{A}_{p,K}$ is a self-adjoint and positive semidefinite operator on $L^2_q(\mathbb{R}^m)$[v] which leads to the linear algebra problem of spectral gap.

The convexity can also be described in other ways. Recall that a nice property of a convex function is the exponential convergence along its gradient flow. For KL_p, we can impose a simple condition to achieve exponential convergence: $\mathrm{KL}_p(q_t) \leq -\frac{1}{2\lambda}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}_p(q_t)$, where $(q_t)_t$ is the gradient flow of KL_p on $\mathcal{P}_K$. Substituting Eq. (37) into Eq. (26), we get the *Stein-log-Sobolev inequality*:

$$\begin{aligned} &\mathrm{KL}_p(q) \leq \frac{1}{2\lambda} I_{p,K}(q), \\ &\text{where } I_{p,K}(q) := \mathbb{E}_p[\nabla(q/p) \cdot \mathcal{K}_p \nabla(q/p)] = \| \mathcal{K}_p \nabla \log(q/p) \|^2_{\mathcal{H}} \end{aligned} \tag{39}$$

is again a kernelized object called Stein–Fisher information. In fact, it is a probability metric constructed using Stein's method, called the *kernel Stein discrepancy* (after square-rooted) (Chwialkowski et al., 2016; Gorham and Mackey, 2017; Liu et al., 2016b), which generalizes the Fisher divergence. Stein-log-Sobolev inequality is a stronger condition than Stein–Poincaré inequality (Duncan et al., 2019, Eq. (61)). Note that this line of developing a convexity theory on a probability manifold is in parallel to that of Villani (2008) for the Wasserstein space $\mathcal{P}_2(\mathcal{M})$ (Hessian formula (15.7), convexity

[u] Likewise, the vanilla Barbour's generator (Barbour, 1990) of the Wasserstein gradient flow of KL_p (which is the overdamped Langevin dynamics) is $\mathcal{A}_p\varphi := -\nabla \cdot (p\nabla\varphi)/p$, which holds the same relation to the Stein operator $\mathcal{S}_p(V) := -\nabla \cdot (pV)/p$ (see also Footnote s).

[v] The domain in the definition (38) can be extended to $L^2_q(\mathbb{R}^m)$ in the sense of weak derivative. Also note $\overline{C_c^\infty(\mathbb{R}^m)}^{L^2_q(\mathbb{R}^m)} = L^2_q(\mathbb{R}^m)$ if q is absolutely continuous (Kováčik and Rákosník, 1991, Thm. 2.11).

(Def. 16.5), Poincaré inequality (Def. 21.17), log-Sobolev inequality (Def. 21.1) and the sufficiency for the former (Thm. 22.17)). The Wasserstein theory can be recovered by choosing $K(x, x') = \delta_x(x')/q(x')$.

This kernelized extension is more suitable for analyzing SVGD. Korba et al. (2020) made a nonasymptotic analysis. In the infinite-particle regime, the difference of KL_p between two adjacent updates is bounded by $I_{p,K}$ at the former update (Prop. 5; also Liu, 2017, 3.3) which serves as a descent lemma in optimization theory, and the cumulative moving average of $I_{p,K}$ converges to zero inversely linearly (Cor. 6). For bounded $\log p$ and Lipschitz kernel with Lipschitz gradients, they also found a bound on the expected squared Wasserstein distance between the empirical distribution of finite particles and the true particle distribution, which decays inversely to the square-root of the number of particles. Chewi et al. (2020) reformulated $V_t^{\mathrm{SVGD}} = -\mathcal{K}_{q_t} \nabla \log (q_t/p)$ as $-\mathcal{K}_{q_t}((p/q_t)\nabla(q_t/p)) = -\mathcal{K}_p \nabla(q_t/p)$, which is the Wasserstein gradient of the chi-squared divergence $\chi^2_p(q_t) := \int (q_t^2/p)\, \mathrm{d}x - 1$ by Eq. (35) (up to a factor of 2). The integral operator $\mathcal{K}_p$ does not depend on time anymore which inspires a characterization for a desired kernel and a new ParVI method. They also proved exponential convergence of KL_p along the (unkernelized) χ^2_p gradient flow given Poincaré inequality, and the inverse quadratic, or exponential convergence of χ^2_p itself given Poincaré inequality, or the log-Sobolev inequality (implied by log-concavity of p).

4.4 Geometric view of MCMC dynamics and relation to ParVI methods

As an interlude, one may also wonder if a similar geometric view for MCMC dynamics is possible. As MCMC dynamics do not need kernel smoothing, the Wasserstein-space view is more relevant.

4.4.1 Langevin dynamics

To begin with, it was found decades ago by Jordan et al. (1998) that the Langevin dynamics equation (11) induces evolving distributions that are the gradient flow of KL_p on the Wasserstein space $\mathcal{P}_2(\mathbb{R}^m)$. The argument was made where the Wasserstein gradient flow is defined in the metric-space sense of the minimizing movement scheme (Eq. 32) (so it is also called the JKO scheme following the authors' initials). With the Riemannian structure introduced in Section 4.2, this can be seen more directly: the dynamics $\mathrm{d}x_t = \nabla \log (p/q_t)\, \mathrm{d}t$ (in the sense of weak derivative if p or q_t as a distribution is not absolutely continuous) defined by the Wasserstein gradient of KL_p (Eqs. 36 and 28) and the Langevin dynamics equation (11) induce the same evolving distribution due to FPE (6).

This coincidence also holds for their Riemannian versions, $\mathrm{d}x_t = G^{-1}\nabla \log(p_{\mathrm{L}}/q_{\mathrm{L}t})\,\mathrm{d}t$ in coordinate space from the gradient-flow side (Eq. 36), and Eq. (14) in coordinate space from the Riemannian Langevin dynamics side (Eq. 12).[w] Alternatively, we can assume this coincidence in the first place, and derive the coordinate expression of the Riemannian Brownian motion via FPE (6):

$$\sqrt{2}\mathrm{d}\tilde{B}_t = \nabla \cdot (\sqrt{|G|}G^{-1})/\sqrt{|G|}\,\mathrm{d}t + \sqrt{2G^{-1}}\,\mathrm{d}B_t,$$

which bridges Eq. (12) to Eq. (13) or Eq. (14). This could serve as an alternative way to generalize Brownian motion to Riemannian manifolds, in addition to using the generalization of the heat equation as introduced in Section 3.2 (Eq. 14).

Compared to the general form of MCMC dynamics equation (8), the Wasserstein gradient flow of KL_p, i.e., the Riemannian Langevin dynamics equation (14), covers the case when $Q = 0$ and D is nondegenerate.

4.4.2 Hamiltonian dynamics

Another instance that has a geometric interpretation is the Hamiltonian dynamics used in HMC. Since it is a classical physical process, its analysis has a longer history. Typically it is interpreted as the Hamiltonian flow/vector field of a function called Hamiltonian, which is $\log p(x,r)$ in this context, on a symplectic manifold,[x] or more commonly on the cotangent bundle $T^*\mathcal{B}$ of a manifold $\mathcal{B}$ with its canonical symplectic structure (Betancourt et al., 2017).[y] A key feature of the Hamiltonian flow is that it keeps Hamiltonian constant.

---

[w]The dynamics expressions in coordinate space all use densities w.r.t. the Lebesgue measure per requirement of the FPE (6).

[x]A symplectic structure is a nondegenerate ($\varsigma_x \neq 0, \forall x \in \mathcal{M}$) and closed ($\mathrm{d}\varsigma = 0$) 2-form (Section 2.3) on the manifold $\mathcal{M}$. The Hamiltonian vector field X_f of a function f, which is called the Hamiltonian, is characterized by $\varsigma(X_f, V) = V[f], \forall V \in \mathcal{T}(\mathcal{M})$. Such X_f is unique: $\varsigma(X_f - X_f', V) = \varsigma(X_f, V) - \varsigma(X_f', V) = V[f] - V[f] = 0$ indicates $X_f - X_f' = 0$ since ς is nondegenerate. It keeps Hamiltonian invariant: $X_f[f] = \varsigma(X_f, X_f)$ which is 0 since ς is antisymmetric. Its coordinate expression is $\varsigma = \sum_{1 \leq i < j \leq m} \varsigma_{ij}\,\mathrm{d}x^i \wedge \mathrm{d}x^j$, where the $m \times m$ matrix (ς_{ij}) is antisymmetric. Note that $\det[(\varsigma_{ij})] = \det[(\varsigma_{ij})^\top] = \det[-(\varsigma_{ij})] = (-1)^m \det[(\varsigma_{ij})]$, so the nondegeneracy requires m to be even.

[y]Let $\{x^i\}_{i=1}^m$ be a coordinate system of $\mathcal{B}$. Recalling Section 2.3, at each point $x \in \mathcal{B}$, $\{\mathrm{d}x^i\}_{i=1}^m$ is a basis of $T_x^*\mathcal{B}$, and any 1-form χ has a coordinate expression $\chi = r_i\,\mathrm{d}x^i$. So $\{x^i, r^i\}_{i=1}^m$ is a coordinate system of $T^*\mathcal{B}$. The canonical symplectic structure is then defined as $\varsigma = \sum_{i=1}^m \mathrm{d}x^i \wedge \mathrm{d}r^i$, which can be shown to be coordinate independent. See e.g., Da Silva (2001) for more information. Under this construction, the coordinate expression of ς is S_m defined below.

Similar to the Riemannian structure case, a symplectic structure can be constructed to the Wasserstein space $\mathcal{P}_2(\mathcal{M})$ using that of $\mathcal{M}$ (Ambrosio and Gangbo, 2008; Gangbo et al., 2010, Ch. 6). For the cotangent bundle $T^*\mathcal{B}$ of a base Riemannian manifold $\mathcal{B}$ with its canonical symplectic structure, the resulting Hamiltonian vector field of a function H on the Wasserstein space $\mathcal{P}_2(T^*\mathcal{B})$ is:

$$X_F(q) = \pi_q\left(-S_m \mathrm{grad}_{x,r} \frac{\delta H}{\delta q}\right),$$

where $S_m := \begin{pmatrix} 0 & -I_m \\ I_m & 0 \end{pmatrix}$, and π_q is the projection from $\mathcal{L}^2_q(T^*\mathcal{B})$ to $T_q\mathcal{P}_2(T^*\mathcal{B})$.[z] Particularly for KL_p where $p(x,r) := p(x)\mathcal{N}(r|0,\Sigma)$ in the Euclidean case, this Hamiltonian vector field recovers the Hamiltonian dynamics equation (19).[aa]

This corresponds to the general form of MCMC dynamics equation (8) when $D = 0$ and $Q = S$. Although a more general Q can be interpreted as a general symplectic structure, a symplectic structure always requires the dimension of the manifold to be even, following the nondegeneracy requirement on its 2-form representation (see Footnote x).

4.4.3 General MCMC dynamics

To enclose the gap to the general form of MCMC dynamics equation (8), new geometric constructions are expected. For this, Liu et al. (2019b) introduced the so-called fiber-Riemannian Poisson (fRP) structure that generalizes the Riemannian and symplectic structures, and the fiber-gradient Hamiltonian (fGH) flow that generalizes the gradient and Hamiltonian flows.

4.4.3.1 Fiber-Riemannian structure and fiber-gradient flow

For a positive semidefinite D, it can be transformed to $D(x) = \begin{pmatrix} C(x) & 0 \\ 0 & 0 \end{pmatrix}$ by choosing a proper coordinate system $(x^1,\ldots,x^m) = (y^1,\ldots,y^\ell,z^1,\ldots,z^{m-\ell})$, where the $\ell \times \ell$ matrix $C(x)$ is everywhere positive definite (thus nondegenerate). The manifold $\mathcal{M}$ can then be split into two parts: the space $\mathcal{F}$ with coordinates $(y^1,\ldots,y^\ell)$ that correspond to the matrix C, and the space $\mathcal{M}_0$ with the rest coordinates $(z^1,\ldots,z^{m-\ell})$. On the first part, $C(x) = C(y, z)$ can be treated as the inverse Riemannian metric $G^{-1}(y|z)$ on that space, which matches the Wasserstein gradient flow. But such a metric in general also depends on the position z in the second part $\mathcal{M}_0$, which is out of the $\mathcal{F}$ space itself. To

[z] That is, subtracting a vector field V that satisfies $\mathrm{div}(qV) = 0$ to make the vector field achieve minimal $\mathcal{L}^2_q$ norm.

[aa] Up to a minus sign; note that the additional part $V = \begin{pmatrix} -\nabla_r \log q \\ \nabla_x \log q \end{pmatrix}$ has $\nabla\cdot(qV) = 0$.

describe such dependency, the construct of fiber bundle is introduced, which is roughly a generalization of the product space $\mathcal{M} = \mathcal{F} \otimes \mathcal{M}_0$, where both factor spaces $\mathcal{F}$ and $\mathcal{M}_0$ can be manifolds, and at each point of the second space $z \in \mathcal{M}_0$, the first space is allowed to have a z-specific structure (hence denoted $\mathcal{F}_z$) while is still diffeomorphic to $\mathcal{F}$. The above $G^{-1}(y|z)$ can then be interpreted to define a Riemannian structure on the corresponding fiber space $\mathcal{F}_z$. This partial Riemannian structure defined by $D(x)$ is thus called a fiber-Riemannian structure by Liu et al. (2019b). Correspondingly, the gradient now can only be defined on each fiber space, and the union of gradients over all fibers $\mathcal{F}_z, \forall z \in \mathcal{M}_0$ is defined as a fiber gradient:

$$\mathrm{grad}_{\mathrm{fib}} f(x) = D^{ij}(x)\partial_j f(x)\partial_i. \tag{40}$$

However, it is hard to develop a fiber-Riemannian structure on the Wasserstein space $\mathcal{P}_2(\mathcal{M})$, so Liu et al. (2019b) turn to consider $\widetilde{\mathcal{P}}_2(\mathcal{M}) := \{q(\cdot\,|z) \in \mathcal{P}_2(\mathcal{F}_z)|z \in \mathcal{M}_0\}$. It naturally has a fiber-Riemannian structure. The fiber gradient of KL_p then recovers the first part of the MCMC dynamics equation (10) exactly:

$$-\pi_q(\mathrm{grad}_{\mathrm{fib}}\mathrm{KL}_p(q)) = D\nabla \log(p/q). \tag{41}$$

4.4.3.2 Poisson structure and Hamiltonian flow

For a general antisymmetric Q, it can be viewed as a Poisson structure on the manifold $\mathcal{M}$. A Poisson structure refers to a Poisson bracket $\{\cdot, \cdot\}$, which has its origin from classical mechanics describing the evolution of a mechanical quantity. Formally, $\{\cdot, \cdot\}$ is a Lie bracket on $C^\infty(\mathcal{M})$, i.e., an antisymmetric bilinear $C^\infty(\mathcal{M}) \times C^\infty(\mathcal{M}) \to C^\infty(\mathcal{M})$ map with Jacobi identity $\{f, \{g, h\}\} + \{g, \{h, f\}\} + \{h, \{f, g\}\} = 0$, that satisfies the Leibniz rule $\{f, gh\} = g\{f, h\} + \{f, g\}h$. Since the Poisson bracket is linear and has a differentiation behavior, it can be linearly represented using tangent-vector basis for any given coordinate system: $\{f, h\} = \{x^i, x^j\}\partial_i f \partial_j h = \sum_{1\le i<j\le m}\{x^i, x^j\}(\partial_i f\partial_j h - \partial_j f\partial_i h) = \sum_{1\le i<j\le m} \{x^i, x^j\}(\partial_i \otimes \partial_j - \partial_j \otimes \partial_i)(\mathrm{d}f \otimes \mathrm{d}h) = \sum_{1\le i<j\le m}\{x^i, x^j\}(\partial_i \wedge \partial_j)(\mathrm{d}f \otimes \mathrm{d}h)$. This suggests another representation of the Poisson structure as a bivector (or 2-vector) field[ab] $\beta(x) = \sum_{1\le i<j\le m}\beta^{ij}(x)\partial_i \wedge \partial_j$, where $(\beta^{ij})(x)$ is everywhere an antisymmetric matrix (dual to the symplectic structure; see Footnote x) and satisfies the corresponding differential Jacobi identity $\beta^{il}\partial_l\beta^{jk} + \beta^{jl}\partial_l\beta^{ki} + \beta^{kl}\partial_l\beta^{ij} = 0, \forall i, j, k$.

[ab]Formally, dual to a 2-form (Section 2.3), a bivector field is everywhere an antisymmetric bilinear function on $(T_x^*\mathcal{M})^2$ (where $\mathrm{d}f \otimes \mathrm{d}h$ lives), or alternatively a combination of the wedge products (antisymmetrized tensor product) of vector pairs (hence the name).

The correspondence between the two representations is given by $\{f, h\} = \beta(\mathrm{d}f, \mathrm{d}h)$. As both the Poisson structure (a 2-vector) and the symplectic structure (a 2-form) bear the antisymmetric bilinearity, they have a one-to-one correspondence via the linear space duality if the Poisson structure is nondegenerate (see Footnote x; Fernandes and Marcut, 2014, Lem. 1.28). So a Poisson structure covers more than a symplectic structure, as it does not have to be nondegenerate; particularly, it allows an odd-dimensional manifold.

Similar to the case on a symplectic manifold, a Hamiltonian flow can also be defined under a Poisson structure. The corresponding Hamiltonian vector field is defined as $X_h(\cdot) := \{\cdot, h\}$, which keeps the Hamiltonian invariant $X_h(h) = 0$, as expected. It holds the following coordinate expression from definition:

$$X_h(x) = \beta^{ij}(x)\partial_j h(x)\partial_i. \tag{42}$$

A Poisson structure $\{\cdot, \cdot\}_{\mathcal{P}_2(\mathcal{M})}$ on the Wasserstein space $\mathcal{P}_2(\mathcal{M})$ can also be developed using that $\{\cdot, \cdot\}_{\mathcal{M}}$ of $\mathcal{M}$ (Lott, 2008, Sec. 6; Gangbo et al., 2010, Sec. 7.2). For functions $F(q)$, $H(q)$ on $\mathcal{P}_2(\mathcal{M})$, define their Poisson bracket as $\{F,H\}_{\mathcal{P}_2(\mathcal{M})}(q) := \int_{\mathcal{M}} \left\{\frac{\delta F}{\delta q}(q), \frac{\delta H}{\delta q}(q)\right\}_{\mathcal{M}} \mathrm{d}q$ (note $\frac{\delta F}{\delta q}(q)$ and $\frac{\delta H}{\delta q}(q)$ are functions on $\mathcal{M}$). The corresponding Hamiltonian flow on the Wasserstein space $\mathcal{P}_2(\mathcal{M})$ is given per point q as $X_H(q) = \pi_q(X_{\frac{\delta H}{\delta q}(q)})$, where the projection π_q is used to map the dynamics $X_{\frac{\delta H}{\delta q}(q)}$ to the Wasserstein tangent space $T_q\mathcal{P}_2(\mathcal{M})$ (Eq. 34) while achieving the same evolving distribution. Particularly, for KL_p on $\mathcal{P}_2(\mathcal{M})$, the Hamiltonian flow is (Liu et al., 2019b, Lem. 2):

$$X_{\mathrm{KL}_p}(q) = \pi_q(\beta^{ij}\partial_j \log(q/p)\partial_i). \tag{43}$$

This form exactly matches the rest part of their reformulation of a general MCMC dynamics equation (10) if $(\beta^{ij}(x))$ is taken as $-Q(x)$ (Liu et al., 2019b, Thm. 5).[ac]

4.4.3.3 fRP structure and fGH flow

To wrap up, if a manifold $\mathcal{M}$ is equipped with both structures, it is called a fiber-Riemannian Poisson (fRP) manifold, and the combination of both

[ac]Common instances of Q satisfy the differential Jacobi identity, required by a Poisson structure. Exceptions include SGNHT-related dynamics. Nevertheless, the identity does not seem fundamental for this geometric construction (even unnecessary for the conservation of Hamiltonian along Hamiltonian flow).

flows Eqs. (40) and (42)) of a function is called the fiber-gradient Hamiltonian (fGH) flow. The geometric view of a general MCMC dynamics targeting p can be then stated as the fGH flow of KL_p (Eqs. 41 and 43) on the Wasserstein space $\mathcal{P}_2(\mathcal{M})$ of an fRP manifold $\mathcal{M}$ (Liu et al., 2019b, Thm. 5).

To draw more intuitions from this geometric view, the fiber-gradient flow decreases $\mathrm{KL}_{p(\cdot|z)}$ on each fiber and drives each $q(\cdot|z)$ toward $p(\cdot|z)$, while the Hamiltonian flow keeps the target distribution $p(x) = p(y, z)$ invariant while makes more exploration in the sample space. The Langevin dynamics (Eqs. 11 and 14) defines a usual Riemannian structure (nondegenerate D, all fibers $= \mathcal{M}$, $\mathcal{M}_0$ degenerates) and null Poisson structure ($Q = 0$). The evolving distribution is driven to the target distribution p by the gradient flow. On the other extreme, the Hamiltonian dynamics (Eqs. 19 and 21) defines a null fiber-Riemannian structure ($D = 0$, fiber space degenerates, $\mathcal{M}_0 = \mathcal{M}$) and a nondegenerate Poisson structure (nondegenerate Q). It allows wilder while eligible exploration hence appears more efficient than the Langevin dynamics. Nevertheless it is fragile to gradient noise (Betancourt, 2015; Chen et al., 2014), as it only guarantees p is a stationary point but lacks a driving force toward p under perturbation. The SGHMC dynamics (22) interpolates between the two extremes. Its half-ranked D matrix defines a Riemannian structure only in each cotangent space of a base manifold $\mathcal{B}$, so the fiber bundle $\mathcal{M}$ is the cotangent bundle $T^*\mathcal{B}$. The symplectic structure of $T^*\mathcal{B}$ defines the Hamiltonian flow, and the fiber-gradient flow stabilizes the process in each fiber space, where the gradient noise comes.

4.4.3.4 Inspiration for more general ParVI methods

As mentioned in Section 4.3.1, the classical ParVI method, SVGD, can be seen as a deterministic simulation of the Wasserstein gradient flow of KL_p. Under the general geometric view here, there are now more options. Liu et al. (2019b) introduced ParVI methods that simulate the SGHMC dynamics (22). The development is done by reformulating the dynamics as equivalent deterministic ones using Eq. (9) or Eq. (10),

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = \Sigma^{-1} r, \\ \dfrac{\mathrm{d}r}{\mathrm{d}t} = \nabla_x \log p(x) - Cr - C\Sigma\nabla_r \log q_t(r), \end{cases} \qquad \begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = \Sigma^{-1} r + \nabla_r \log q_t(r), \\ \dfrac{\mathrm{d}r}{\mathrm{d}t} = \nabla_x \log p(x) - Cr - C\Sigma\nabla_r \log q_t(r) - \nabla_x \log q_t(x), \end{cases}$$

and then leveraging ParVI techniques (including SVGD) to estimate all the $\nabla \log q_t$ with corresponding particles. For example, the Blob method (Chen et al., 2018a) (see Section 4.5.1) leads to the following particle updates:

$$\begin{cases} \frac{1}{\varepsilon}(x_{k+1}^{(i)} - x_k^{(i)}) = \Sigma^{-1} r_k^{(i)}, \\ \frac{1}{\varepsilon}(r_{k+1}^{(i)} - r_k^{(i)}) = \nabla_{x_k^{(i)}} \log p(x_k^{(i)}) - C r_k^{(i)} \\ \qquad -C\Sigma\left(\frac{\sum_l \nabla_{r_k^{(i)}} \mathbf{K}_{r,k}^{(i,l)}}{\sum_j \mathbf{K}_{r,k}^{(i,j)}} + \sum_l \frac{\nabla_{r_k^{(i)}} \mathbf{K}_{r,k}^{(i,l)}}{\sum_j \mathbf{K}_{r,k}^{(j,l)}}\right), \end{cases} \quad \begin{cases} \frac{1}{\varepsilon}(x_{k+1}^{(i)} - x_k^{(i)}) = \Sigma^{-1} r_k^{(i)} + \frac{\sum_l \nabla_{r_k^{(i)}} \mathbf{K}_{r,k}^{(i,l)}}{\sum_j \mathbf{K}_{r,k}^{(i,j)}} + \sum_l \frac{\nabla_{r_k^{(i)}} \mathbf{K}_{r,k}^{(i,l)}}{\sum_j \mathbf{K}_{r,k}^{(j,l)}}, \\ \frac{1}{\varepsilon}(r_{k+1}^{(i)} - r_k^{(i)}) = \nabla_{x_k^{(i)}} \log p(x_k^{(i)}) - \left(\frac{\sum_l \nabla_{x_k^{(i)}} \mathbf{K}_{x_k^{(i)}}^{(i,l)}}{\sum_j \mathbf{K}_{x_k^{(i)}}^{(i,j)}} + \sum_l \frac{\nabla_{x_k^{(i)}} \mathbf{K}_{x_k^{(i)}}^{(i,l)}}{\sum_j \mathbf{K}_{x_k^{(i)}}^{(j,l)}}\right) \\ \qquad -C r_k^{(i)} - C\Sigma\left(\frac{\sum_l \nabla_{r_k^{(i)}} \mathbf{K}_{r,k}^{(i,l)}}{\sum_j \mathbf{K}_{r,k}^{(i,j)}} + \sum_l \frac{\nabla_{r_k^{(i)}} \mathbf{K}_{r,k}^{(i,l)}}{\sum_j \mathbf{K}_{r,k}^{(j,l)}}\right), \end{cases} \tag{44}$$

where $\mathbf{K}_{r,k}^{(i,j)} := \mathbf{K}_r(r_k^{(i)}, r_k^{(j)})$ for a kernel $\mathbf{K}_r$ for the momentum r, and similarly for $\mathbf{K}_{x,k}^{(i,j)}$. The new methods inherit the faster exploration from the Hamiltonian flow.

4.5 Variants and Techniques Inspired by the Geometric View

4.5.1 Other methods for Wasserstein gradient flow simulation

The key to the deterministic simulation of the Wasserstein gradient flow (36) is the estimation of the set of vectors $\mathbf{U}^* : \mathbf{U}^{*:,i} := \nabla \log q(x^{(i)})$ using particles $\{x^{(i)}\}_{i=1}^N$ of q. This is also where a smoothing operation is required as discussed in Section 4.3.1. Besides the way that SVGD estimates it, other methods are developed.

Perhaps the most straightforward treatment is the kernel density estimator $q(x) \approx \tilde{q}_K(x; \{x^{(i)}\}_i) := \frac{1}{N}\sum_{i=1}^N K(x, x^{(i)})$, which forms a ParVI method named gradient flow with smoothed density (GFSD) (Liu et al., 2019a). The corresponding approximation is $\mathbf{U}^{\mathrm{GFSD}:,i} = \frac{\sum_k \nabla_{x^{(i)}} \mathbf{K}^{(i,k)}}{\sum_j \mathbf{K}^{(i,j)}}$, where $\mathbf{K}^{(i,j)} := K(x^{(i)}, x^{(j)})$. It shows less diverse particles than SVGD.

This spirit of smoothing density with kernel can be extended. Noting that $\nabla \log q = \nabla(\frac{\delta}{\delta q}\mathbb{E}_q[\log q])$, it only requires smoothing the q in $\log q$ for allowing the use of the empirical distribution $\hat{q}(x)$ hence particles. This can be done by: $U^{\mathrm{Blob}} := \nabla(\frac{\delta}{\delta q}\mathbb{E}_q[\log (q * K)]) = \nabla \log (q * K) + \nabla(\frac{q}{q*K} * K)$, or in terms of particles, $\mathbf{U}^{\mathrm{Blob}:,i} = \frac{\sum_k \nabla_{x^{(i)}} \mathbf{K}^{(i,k)}}{\sum_j \mathbf{K}^{(i,j)}} + \sum_k \frac{\nabla_{x^{(i)}} \mathbf{K}^{(i,k)}}{\sum_j \mathbf{K}^{(j,k)}}$ (Chen et al., 2018a). The method is called Blob due to its origin in fluid mechanics. Note that it adds a term to $\mathbf{U}^{\mathrm{GFSD}}$, which perhaps explains its more spread particles than GFSD.

A more involved method is proposed by Liu et al. (2019a) under the perspective of doing kernel smoothing on vector fields (vs on densities), hence called gradient flow with smoothed function (GFSF). This is done by noting that $\nabla \log q = -\mathrm{argmin}_{U \in \mathcal{L}^2} \max_{\phi \in \mathcal{C}_c^\infty, \|\phi\|_{\mathcal{L}_q^2}=1} (\mathbb{E}_q[\phi \cdot U - \nabla \cdot \phi])^2$, and smoothing the vector field can be done by replacing $\mathcal{L}_q^2$ by the RKHS $\mathcal{H}^m$ as done by SVGD: $U^{\mathrm{GFSF}} := -\mathrm{argmin}_{U \in \mathcal{L}^2} \max_{\phi \in \mathcal{H}^m, \|\phi\|_{\mathcal{H}^m}=1} (\mathbb{E}_q[\phi \cdot U - \nabla \cdot \phi])^2$. The solution using particles is given by $\mathbf{U}^{\mathrm{GFSF}} = -\mathbf{J}\mathbf{K}^{-1}$, where $\mathbf{J}^{:,i} := \sum_j \nabla_{x^{(j)}} \mathbf{K}^{(i,j)}$. An interesting connection to SVGD is that, if denoting $\mathbf{P}^{:,i} := \nabla_{x^{(i)}} \log p(x^{(i)})$, then $\mathbf{V}^{\mathrm{GFSF}} = \mathbf{P} + \mathbf{J}\mathbf{K}^{-1}$ while $\mathbf{V}^{\mathrm{SVGD}} = \mathbf{P}\mathbf{K} + \mathbf{J}$, so $\mathbf{V}^{\mathrm{GFSF}} = \mathbf{V}^{\mathrm{SVGD}}\mathbf{K}^{-1}$. Another noticeable connection is to the research direction of gradient estimation for implicit (meaning no density; only samples) generative models. Particularly, $\mathbf{U}^{\mathrm{GFSF}}$ coincides with the result by Li and Turner (2018) developed from another intuition. Moreover, there are other techniques (e.g., Shi et al., 2018) in the direction that are worth a trial for ParVI.

Besides simulation using the Wasserstein gradient under a Riemannian perspective, there are also works that explored gradient-flow simulation in the metric-space sense. Chen et al. (2018a) leveraged the minimizing movement scheme (Eq. 32) for defining the gradient flow (Ambrosio et al., 2008, Def. 2.0.2). Due to the definition of the Wasserstein distance equation (31),

each simulation step amounts to a regularized discrete optimal transport problem, which is solved approximately and analytically with preset Lagrange multipliers for both marginal-distribution constraints. For other variants, Ranganath et al. (2016) solved Eq. (27) over the parameter space of a neural network model, but this introduces additional optimization cost in each update. The smoothing effect is implicitly controlled by the Lipschitzness of the neural network.

4.5.2 Riemannian-manifold support space

4.5.2.1 Riemannian SVGD

Liu and Zhu (2018) made the first attempt to extend SVGD to a Riemannian-manifold support/particle space $(\mathcal{M},g)$, called RSVGD. They started with the Riemannian version of Eq. (26) based on the Riemannian continuity equation (33) (Liu and Zhu, 2018, Lem. 1):

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}_p(q_t) = \mathbb{E}_{q_t}[\mathrm{div}(pV_t)/p] = \mathbb{E}_{q_t}[V_t[\log p] + \mathrm{div}\ V_t], \tag{45}$$

where the densities are w.r.t. the Riemannian measure ω_g on the manifold $(\mathcal{M},g)$ (Liu and Zhu, 2018, Thm. 2). To find a vector field V_t maximizing the decreasing rate by solving Eq. (27), requirements on the optimization domain $\mathfrak{X}$ are introduced: the analytic solution should be a valid vector field on $\mathcal{M}$ and is coordinate independent. The requirements appear very natural, but are not easily satisfied. For example, every valid vector field on an even-dimensional hypersphere must have a zero point [hairy ball theorem (Abraham et al., 2012, Thm. 8.5.13)], which cannot be guaranteed by the choice in SVGD $\mathfrak{X}=\mathcal{H}^m$ for a particular coordinate system, nor can coordinate independency. Liu and Zhu (2018) then chose the space $\mathfrak{X}=\{\mathrm{grad}\,\varphi\,|\,\varphi\in\mathcal{H}\}$, where $\mathcal{H}$ is the RKHS of a kernel K defined on $\mathcal{M}$. It naturally guarantees the requirements since the gradient is always a valid and coordinate-independent vector field. For common kernels (e.g., Gaussian kernel), such $\mathfrak{X}$ inherits an inner product from $\mathcal{H}$, which makes $\mathfrak{X}$ a Hilbert space (Liu and Zhu, 2018, Lem. 3) and leads to the analytic solution:

$$\begin{aligned} V_t^{\mathrm{RSVGD}}(x') &= \mathrm{grad}_{x'}\mathbb{E}_{q_t(x)}[(\mathrm{grad}_x K)[\log p(x)] + \mathrm{Lap}_x K] \\ &= g(x')^{i'j'}\partial_{x'}\mathbb{E}_{q_t(x)}\left[\left(g(x)^{ij}\partial_{x^i}\log p_{\mathrm{L}}(x) + \partial_{x^i}g(x)^{ij}\right)\partial_{x^j}K + g(x)^{ij}\partial_{x^i}\partial_{x^j}K\right]\partial_{x^{i'}} \\ &= G(x')^{-1}\nabla_{x'}\mathbb{E}_{q_t(x)}\left[\left(G(x)^{-1}\nabla\log p_{\mathrm{L}}(x) + \nabla\cdot G(x)^{-1}\right)\nabla_x K + G(x)^{-1}:\nabla_x\nabla_x^{\top}K\right], \end{aligned} \tag{46}$$

where K is evaluated at (x, x'), and $p_{\mathrm{L}}(x) = p(x)\sqrt{|G(x)|}$ (Liu and Zhu, 2018, Thm. 4). This expression can also be estimated using particles since q_t only appears in expectations. When used with the Fisher–Rao metric, RSVGD achieves faster convergence than SVGD for Bayesian logistic regression. For applications on manifolds without global coordinate system, e.g., hyperspheres, simulation in the embedded space $\Xi(\mathcal{M})\subseteq\mathbb{R}^n$ of the manifold is

preferred due to the argument in Section 3.3. Liu and Zhu (2018) also derived such an expression:

$$V_t^{\mathrm{RSVGD,emb}}(y') = \Lambda(y')\nabla_{y'}\mathbb{E}_{q_t(y)}\Big[\nabla \log\left(p(y)\sqrt{|G(y)|}\right)^\top \Lambda(y)\nabla_y K + \nabla_y^2 K - \mathrm{tr}\Big(P(y)^\top(\nabla_y\nabla_y^\top K)P(y)\Big) + (J(y)^\top\nabla)^\top(G(y)^{-1}J(y)^\top)\nabla_y K\Big],$$

where $P(y) \in \mathbb{R}^{n\times(n-m)}$ is a set of orthonormal basis of the orthogonal complement of the embedded tangent space $\Xi_*(T_x\mathcal{M})$. Particularly for the hypersphere $\mathbb{S}^{n-1}$, this reduces to:

$$V_t^{\mathrm{RSVGD,sph}}(y') = \left(I_n - y'y'^\top\right)\nabla_{y'}\mathbb{E}_{q_t(y)}\Big[\nabla \log p(y)^\top\nabla_y K + \nabla_y^2 K - y^\top\left(\nabla_y\nabla_y^\top K\right)y - (y^\top\nabla \log p(y) + n - 1)y^\top\nabla_y K\Big].$$

It achieves a faster convergence and better particle efficiency than embedded-space MCMC methods introduced in Section 3.3 for the spherical admixture model (Reisinger et al., 2010).

4.5.2.2 Mirrored SVGD

More recently, Shi et al. (2022) considered an extension of SVGD to leverage the mirror descent technique (Beck and Teboulle, 2003), which is an approach in optimization to handle constrained domain and Riemannian geometry. For minimizing a function f on a possibly constrained Euclidean domain $\mathcal{X} \subseteq \mathbb{R}^m$, the method follows the minimizing movement scheme (Eq. 32), $x_{k+1} = \mathrm{argmin}_{x\in\mathcal{X}}\frac{1}{\varepsilon}d_\psi(x, x_k) + \nabla f(x_k)^\top x$, using the so-called Bregman divergence (may not be a distance) $d_\psi(x, x_k) := \psi(x) - \psi(x_k) - \nabla\psi(x_k)^\top(x - x_k)$ defined by a strongly convex smooth function ψ. The solution has an expression $x_{k+1} = \nabla\psi^*(\nabla\psi(x_k) - \varepsilon\nabla f(x_k))$, where $\psi^*(y) := \sup_{x\in\mathcal{X}} y^\top x - \psi(x)$ is the convex conjugate (Legendre transformation) of ψ, and the strong convexity indicates $(\psi^*)^* = \psi$ and that $y = \nabla\psi(x)$ is a bijection to the mirrored space $\mathbb{Y} := \nabla\psi(\mathcal{X})$ with inverse $x = \nabla\psi^*(y)$. The expression is interpreted as first mirroring x_k to the mirrored space $y_k = \nabla\psi(x_k)$ and conducting gradient descent there $y_{k+1} = y_k - \varepsilon\nabla f(x_k)$, then mirroring back to the original space $x_{k+1} = \nabla\psi^*(y_{k+1})$. This explains the name. For an example to handle a constrained optimization domain, consider the simplex $\Delta^m := \{x \in (\mathbb{R}_+)^m \mid \sum_{i=1}^m x^i < 1\}$ and $\psi(x) := \sum_{i=1}^m x^i \log x^i + x^{m+1}(x)\log x^{m+1}(x)$ where $x^{m+1}(x) := 1 - \sum_{i=1}^m x^i$. The mirror map is $\nabla\psi(x) = \log(x/x^{m+1}(x))$ elementwise, which leads to an unconstrained mirror space $\nabla\psi(\Delta^m) = \mathbb{R}^m$.

For developing mirrored SVGD, a helpful insight is that mirror descent can be seen as a Riemannian gradient descent with $G = \nabla\nabla^\top\psi$ as the Riemannian metric, since $\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}(x_{k+1} - x_k) = -(\nabla\nabla^\top\psi(x_k))^{-1}\nabla f(x_k)$ matches Eq. (1). Particularly, the above simplex example also shows the usage of mirror descent

for information geometry: the Bregman divergence $d_\psi(x, x')$ recovers the KL divergence $\mathrm{KL}(\mathrm{Cat}(\cdot|x) \,\|\, \mathrm{Cat}(\cdot|x'))$ if the simplex points x, x' are treated as the parameter of a categorical distribution (on $m+1$ categories), and the Riemannian metric $\nabla\nabla^\top\psi(x)$ recovers the Fisher–Rao metric (Fisher information matrix) of $\mathrm{Cat}(\cdot|x)$.

Under this insight, mirrored SVGD may be seen as a special case of RSVGD, but there are more subtleties. The tasks are on Euclidean spaces, meaning a natural, global coordinate system of the manifold, so vector-field validity and coordinate-invariance are not concerned. On the other hand, as shown in the simplex example, Δ^m is a constrained space so updating particles there is quite involved. But it is much easier to update particles in the unconstrained mirrored space, which exactly guarantees the constraint in the original space. So Shi et al. (2022) considered solving a dynamics in the mirrored space, $\mathrm{d}y_t = U_t(x_t)\,\mathrm{d}t$, similar to the form of mirror descent in the mirror space. This is equivalent to formulating the dynamics in the original space as $\mathrm{d}x_t = (\nabla\nabla^\top\psi(x_t))^{-1}U_t(x_t)\,\mathrm{d}t = G(x_t)^{-1}U_t(x_t)\,\mathrm{d}t$. Under this formulation, the Riemannian version of KL decreasing rate equation (45)becomes[ad] :

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}_p(q_t) = \mathbb{E}_{q_t}[\mathcal{S}_p^\psi U_t],$$

$$\text{where}\quad \mathcal{S}_p^\psi U_t := U_t^\top(\nabla\nabla^\top\psi)^{-1}\nabla\log p + \nabla\cdot\left((\nabla\nabla^\top\psi)^{-1}U_t\right)$$

$$= U_t^\top(\nabla\nabla^\top\psi)^{-1}\nabla\log\left(p/|\nabla\nabla^\top\psi|\right) + (\nabla\nabla^\top\psi)^{-1} : (\nabla U_t^\top),$$

is called the mirrored Stein operator, and p is the density under the Lebesgue measure of $\mathcal{X}$ (different from the case in Eq. 45). Using this expression, Shi et al. (2022) solved Eq. (27) for maximizing the decreasing rate in the RKHS of a matrix-valued kernel $\boldsymbol{K}$ on $\mathcal{X}$:

$$U_t^{\mathrm{MSVGD}}(x') := \mathbb{E}_{q_t(x)}[\mathcal{S}_p^\psi \boldsymbol{K}(x,x')],$$

where $\mathcal{S}_p^\psi$ operates on each row of $\boldsymbol{K}(x, x')$ as a vector-valued function of x (Shi et al., 2022, Thm. 3). Similar to mirror descent, dynamics simulation for mirrored SVGD is done by first mirroring all x particles to y particles using $\nabla\psi$, updating y particles using U^{MSVGD}, and then mirroring back to x particles.

The simplest choice of kernel is $\boldsymbol{K}(x,x') = K(x,x')I_m$ where K is a scalar-valued kernel on $\mathcal{X}$. But Shi et al. (2022) found with one particle, the algorithm does not reduce to mirror descent targeting the mode of $p(x)$. They thus introduced a geometry-aware, position-dependent kernel:

$$\boldsymbol{K}_t^\psi(x,x') := \mathbb{E}_{q_t(x'')}\left[K_t^{1/2}(x, x'')\nabla\nabla^\top\psi(x'')K_t^{1/2}(x'', x')\right],$$

[ad]Eq. (18) is used in the last equality.

where $K_t^{1/2}(x,x') := \sum_\alpha \lambda_{t,\alpha}^{1/2} u_{t,\alpha}(x) u_{t,\alpha}(x')$ is defined under the spectral expansion/Mercer representation of the scalar-valued kernel, $K(x, x') = \sum_\alpha \lambda_{t,\alpha} u_{t,\alpha}(x) u_{t,\alpha}(x')$, where each $(\lambda_{t,\alpha}, u_{t,\alpha})$ solves the eigenvalue problem $\mathbb{E}_{q_t(x')}[K(x,x')u_t(x')] = \lambda_t u_t(x)$. It drives to the mode of $p(x)$ thus transits to MAP estimate when using one particle (Shi et al., 2022, Prop. 5), as desired. Shi et al. (2022) also used a general Riemannian metric in place of $\nabla\nabla^\top\psi$ (and using the geometry-aware kernel) for leveraging information geometry in general Euclidean/unconstrained inference tasks, which recovers natural gradient descent using one particle.

4.5.3 Accelerated gradient flow

Inspired by the geometric view of ParVI methods as approximations to the gradient descent of KL_p on the Wasserstein space (see Sections 4.3.1 and 4.5.1), Liu et al. (2019a) developed new ParVI methods corresponding to accelerated first-order optimization on the Wasserstein space. Acceleration of gradient descent in the Euclidean space is done by the well-known Nesterov's acceleration method (Nesterov, 1983), and it has been extended to Riemannian manifolds, including Riemannian accelerated gradient (RAG) (Liu et al., 2017) and Riemannian Nesterov's method (RNes) Zhang and Sra (2018). When written for the Wasserstein space $\mathcal{P}_2(\mathbb{R}^m)$, these accelerated methods introduce an auxiliary distribution variable $\rho \in \mathcal{P}_2(\mathbb{R}^m)$ in addition to the optimized distribution variable q, and the optimization updates are given by [with slight simplification by Liu et al. (2019a)]:

$$\begin{cases} q_k = \mathrm{Exp}_{\rho_{k-1}}(\varepsilon V_{k-1}), \\ \rho_k = \begin{cases} \mathrm{Exp}_{q_k}\left[-\Gamma_{\rho_{k-1}}^{q_k}\left(\frac{k-1}{k}\mathrm{Exp}_{\rho_{k-1}}^{-1}(q_{k-1}) - \frac{k+\alpha-2}{k}\varepsilon V_{k-1}\right)\right], & \text{(RAG)} \\ \mathrm{Exp}_{q_k}\left\{c_1 \mathrm{Exp}_{q_k}^{-1}\left[\mathrm{Exp}_{\rho_{k-1}}\left((1-c_2)\mathrm{Exp}_{\rho_{k-1}}^{-1}(q_{k-1}) + c_2 \mathrm{Exp}_{\rho_{k-1}}^{-1}(q_k)\right)\right]\right\}, & \text{(RNes)} \end{cases} \end{cases}$$

where $V_k := -\mathrm{grad}\ \mathrm{KL}(\rho_k)$, and $\alpha > 3$ and $c_1, c_2 > 0$ are hyperparameters. Implementing the algorithms requires estimating the exponential map $\mathrm{Exp}_q(V)$, its inverse $\mathrm{Exp}_q^{-1}(\rho)$, and the parallel transport Γ_q^ρ using particles $\{x^{(i)}\}_i$ of q and particles $\{y^{(j)}\}_j$ of ρ. Recall from Section 4.3.1 (Ambrosio et al., 2008, Prop. 8.4.6), since the geodesic defining the exponential map is obviously tangent to V at q, the Wasserstein-space exponential map $\mathrm{Exp}_q(V)$ can be estimated by:

$$\mathrm{Exp}_q(V) = \mathrm{Exp}_{\cdot}(V(\cdot))_\# [q] = (\mathrm{id} + V)_\# [q], \tag{47}$$

where the second Exp is the exponential map of the support space, and the last equality holds on Euclidean support spaces. This means that $\{x^{(i)} + V(x^{(i)})\}_i$ is a set of particles of $\mathrm{Exp}_q(V)$. The inverse exponential map $\mathrm{Exp}_q^{-1}(\rho)$ is given by $\mathcal{T}_q^\rho - \mathrm{id}$ on Euclidean support space $\mathbb{R}^m$, where $\mathcal{T}_q^\rho$ is the optimal transport map from q to ρ (Ambrosio et al., 2008, Thm. 7.2.2; Villani, 2008, Cor. 7.22). Solving the optimal transport problem directly is costly, but Liu et al. (2019a)

noticed that $\mathrm{Exp}_q^{-1}(\rho)$ is invoked only for ρ as infinitesimally updated q, which means the two sets of particles possibly hold the pairwise close assumption: $d(x^{(i)}, y^{(i)}) \ll \min\{\min_{j\neq i} d(x^{(i)}, x^{(j)}), \min_{j\neq i} d(y^{(i)}, y^{(j)})\}$. The optimal transport evaluated at $x^{(i)}$ can then be approximated as $y^{(i)} - x^{(i)}$. As for the parallel transport, Liu et al. (2019a) leveraged the Schild's ladder (Ehlers et al., 1972; Kheyfets et al., 2000) approximation method that only requires the exponential map and its inverse. The resulting accelerated ParVIs are named Wasserstein accelerated gradient (WAG) and Wasserstein Nesterov's method (WNes) that update particles as:

$$\begin{cases} x_k^{(i)} = y_{k-1}^{(i)} + \varepsilon V\left(y_{k-1}^{(i)}\right), \\ y_k^{(i)} = x_k^{(i)} + \begin{cases} \dfrac{k-1}{k}\left(y_{k-1}^{(i)} - x_{k-1}^{(i)}\right) + \dfrac{k+\alpha-2}{k}\varepsilon V\left(y_{k-1}^{(i)}\right), & \text{(WAG)} \\ c_1(c_2-1)\left(x_k^{(i)} - x_{k-1}^{(i)}\right), & \text{(WNes)} \end{cases} \end{cases}$$

where each $V\ (y^{(i)})_k$ is estimated using any existing ParVI techniques (e.g., SVGD, variants in Section 4.5.1).

Taghvaei and Mehta (2019) considered accelerating ParVI dynamics using a recent powerful unifying framework for optimization dynamics (Wibisono et al., 2016). The resulting algorithm is in a similar form of the ParVI counterpart of SGHMC, Eq. (44) introduced in Section 4.4.3. The framework has also been leveraged to accelerate MCMC dynamics (Ma et al., 2019; Wibisono, 2018).

Another way to accelerate optimization is Newton's method, which steers the gradient with the inverse Hessian matrix of the objective. Detommaso et al. (2018) developed such generalization under the early version of the Stein geometry, i.e., taking the vector-valued RKHS $\mathcal{H}^m$ as the tangent space (Liu, 2017). Specifically, the Hessian of KL_p at q is defined as a function of two tangent vectors V, W: $\mathrm{Hess}\,\mathrm{KL}_p(q)(V,W) := \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}(V[\mathrm{KL}_p(\mathrm{Exp}_q(\varepsilon W))] - V[\mathrm{KL}_p(q)])$, where $\mathrm{Exp}_q(\varepsilon W) = (\mathrm{id} + \varepsilon W)_{\#}q$ is the exponential map (see Eq. 47), and $V[F(q)] = \langle V(q), \mathrm{grad}_{\mathcal{P}_{\mathcal{H}^m}} F(q)\rangle_{\mathcal{H}^m} = \langle V(q), \nabla\frac{\delta F}{\delta q}(q)\rangle_{\mathcal{H}^m}$ is the action of the tangent vector V on F at q (i.e., the directional derivative along V; see Section 2.2). Its explicit expression is $\mathrm{Hess}\,\mathrm{KL}_p(q)(V,W) = -\langle \mathbb{E}_{q(x)}[H_{q,p}(x,\cdot)W(x)], V(\cdot)\rangle_{\mathcal{H}^m}$, where $H_{q,p}(x,x') := K(x,x')\nabla\nabla^\top \log p(x) + \nabla_x\nabla_x^\top K(x,x') + \nabla_x K(x,x')\nabla \log q(x)^\top$. In the Euclidean case, the Newtonian descending direction w is $-(\nabla\nabla^\top f)^{-1}\nabla f$, or $v^\top(\nabla\nabla^\top f)w = -v^\top\nabla f$ for all vector v. So the Newtonian descending direction W for KL_p at q is given by: $\mathrm{Hess}\,\mathrm{KL}_p(q)(V, W) = -V\,[\mathrm{KL}_p(q)]$ for all tangent vector $V \in \mathcal{H}^m$. This amounts to solving $\mathbb{E}_{q(x)}[H_{q,p}(x,x')W(x)] = V^{\mathrm{SVGD}}(x') = \mathbb{E}_{q(x)}[K(x,x')\nabla \log p(x) + \nabla_x K(x,x')]$, which is very costly $(\Omega(N^3d^3))$. Detommaso et al. (2018) made a relaxation of the problem as $\mathbb{E}_{q(x)}[\tilde{H}_{q,p}(x,x')]$ $W(x') = V^{\mathrm{SVGD}}(x')$, where the interaction of W between particles is decoupled,

and $H_{q,p}$ is simplified as $\tilde{H}_{q,p}(x,x') := \tilde{N}_p(x)K(x,x') + \tilde{N}_K(x,x')$ where the Hessians of $\log p$ and K are replaced with their Gauss–Newton approximations, and the third term is omitted due to the difficulty of estimation by particles.[ae] They also consider using a kernel with preconditioning, $K(x,x') := \exp(-(x-x')^\top \mathbb{E}_q[-\nabla\nabla^\top \log p](x-x')/(2m))$ for faster convergence. Although this method also leverages Fisher information (through $\tilde{N}_p$) and the kernel Hessian $\nabla_x \nabla_x^\top K$ as RSVGD (coordinate-space version (46)) does, the two methods are different in formulation, and the connection is yet to be studied.

Quasi-Newton methods for ParVI are also considered to reduce the costly second-order derivative computation. Zhu et al. (2020) leveraged a Riemannian quasi-Newton method (Kasai et al., 2018) that also allows stochastic gradient and variance reduction, and applied it to the Wasserstein space to develop a quasi-Newton ParVI method. Geometric constructions are implemented using the techniques by Liu et al. (2019a) introduced in Section 4.5.3. They also leveraged this general approach to develop variance-reduced ParVI methods, based on Riemannian stochastic variance reduction gradient (Zhang et al., 2016) and Riemannian stochastic path integrated differential estimator (Zhou et al., 2019).

4.5.4 Treatment of the kernel

As mentioned in Section 4.3.1, under the perspective of ParVI as KL_p minimization on the Wasserstein space, a smoothing operation is mandatory. The most popular way for this is using kernels (either using kernel density estimation or using RKHS for a function class; see Section 4.5.1), due to its elegant theoretical properties and efficient implementation. Nevertheless, it also faces challenges.

It is well known that kernel methods do not work as well in high dimensions. Zhuo et al. (2018) studied the implication for SVGD, and found the particles tend to collapse and underestimate the marginal variances as dimension increases. This is due to the vanishing of the repulsive term (second term in Eq. 30) which is responsible for diversity. Specifically, for a Gaussian kernel, a Gaussian target distribution, and a Gaussian or compactly supported particle distribution, the maximal scale of the repulsion was shown to decrease inversely sublinearly to dimension. To work around this limitation, they proposed a message-passing SVGD, which leverages the conditional independency in the high-dimensional random variable under the target distribution, and remove the corresponding noninteracting dimension pairs in the kernel. This breaks down a high-dimensional problem into several low-dimensional problems, and better performance was seen. A similar method is concurrently

[ae] A better approximation may be omitting the last two terms together, since they cancel each other under the expectation with $q(x)$ under mild boundary conditions: $\int q(x)\nabla_x K(x,x')\nabla \log q(x)^\top \,\mathrm{d}x = \int \nabla_x K(x,x')\nabla q(x)^\top \,\mathrm{d}x = -\int \nabla_x \nabla_x^\top K(x,x')q(x)\,\mathrm{d}x$.

developed by Wang et al. (2018) which is called graphical SVGD. Ba et al. (2019) made some further analysis. They showed that for a Gaussian kernel and Gaussian target distribution, the marginal variance of SVGD-stationary particles is proportional to the ratio of the number of particles over dimension, when both are sufficiently large and the ratio is less than 1. They also emphasized the collapse is due to both the randomness of estimating the first term of Eq. (30) using finite particles, and the deterministic update. They then proposed variants to reduce the randomness and enable stochastic particle resampling, though impractical for real applications. Some other works considered general approaches to reducing the dimensionality of the kernel to combat the variance collapse problem. Chen and Ghattas (2020) find a dominating subspace support of the target Bayesian posterior and run SVGD there. Gong et al. (2021) introduced a sliced kernelized Stein discrepancy [in a similar spirit as the sliced Wasserstein distance (Kolouri et al., 2016, 2019)], which can be unbiased estimated by a one-dimensional kernel calculation on a randomly/properly selected direction, which largely reduces the dimensionality. The resulting ParVI is also shown to outperform SVGD. Liu et al. (2022) took a step further to make a k-dimensional slice ($1 \leq k \leq m$). The k-dimensional subspace is defined by an optimization problem, which is solved by simulating a diffusion process on the space of k-dimensional subspaces, called the Grassmann manifold. The Grassmann manifold can be seen as the quotient space of the Stiefel manifold (all $m \times k$ orthogonal matrices) over the orthogonal group of order k.

There are works that analyzed the impact of kernels on the approximation power of SVGD convergent particles. Gorham and Mackey (2017) studied the discriminative power of kernel Stein discrepancy $I_{p,K}(q)$ defined in Eq. (39), which indicates properties of SVGD convergent particles since $I_{p,K}(q)$ converges to zero along SVGD updates (Korba et al., 2020, Cor. 6). They found $I_{p,K}(q)$ detects convergence (i.e., it converges to zero for a q sequence weakly converges to p) if K is twice continuously differentiable with uniformly bounded second-order cross derivatives and $\nabla \log p$ is Lipschitz with finite $\mathcal{L}_p^2$ norm. However for the converse, they found a counterexample that light-tailed kernels[af] fail to detect nonconvergence to standard Gaussian for dimension $m \geq 3$. This failure unfortunately applies to common kernels such as the Gaussian kernel, Matérn kernel, compactly supported kernels, and inverse multiquadric (IMQ) kernel $K^{\mathrm{IMQ}}(x,y) := (c^2 + \| x-y\|_2^2)^{-\beta}$ with $\beta \in (1, m/(m-2))$. A choice that guarantees the success of detecting nonconvergence is an IMQ kernel with $\beta \in (0,1)$. Liu and Wang (2018) analyzed the condition for a set of particles to be SVGD convergent, and found the convergent particles give the exact expectation under p for Stein-operator-transformed RKHS functions. Implications of this result include that linear

[af] Formally, this means $\sup\{\max\{|K(x,y)|, \| \nabla_x K(x,y)\|_2, |\nabla_x^\top \nabla_y K(x,y)|\}; | \ \| x-y\|_2 \geq r\} = o\left(r^{-1/\left(\frac{1}{2}-\frac{1}{m}\right)}\right)$.

kernels $K(x, x') := x^\top x' + c^2$ lead to convergent particles that exactly estimate the mean and variance of Gaussian distributions, and that a kernel combining randomly chosen feature maps of another kernel yields $O(1/\sqrt{N})$ -close convergent particles in terms of kernel Stein discrepancy to p.

Liu et al. (2019a) studied the impact of kernel under a dynamics perspective. Under their view of approximate Wasserstein-gradient-flow simulation (Section 4.3.1), the kernel is introduced for the mandatory smoothing operation for the intractable gradient $\nabla \log q(x)$ (see Section 4.5.1 for variants besides SVGD). Due to the equivalence between smoothing densities and smoothing functions (Section 4.3.1), they took GFSD (Section 4.5.1) as an example, which uses kernel density estimation for the particle density, $q(x) \approx \tilde{q}_K(x; \{x^{(i)}\}_{i=1}^N) := \frac{1}{N}\sum_{i=1}^N K(x, x^{(i)})$. What matters of this kernel smoothing is that the density update from the resulting particle update matches that from the exact dynamics $\mathrm{d}x = -\nabla \log q(x)\, \mathrm{d}t$. The updated particles after a time step ε are $\{x^{(i)} - \varepsilon \nabla \log \tilde{q}(x^{(i)}; \{x^{(j)}\}_j)\}_i$, so the resulting density update is $q_{t+\varepsilon}(x) \approx \tilde{q}(x; \{x^{(i)} - \varepsilon \nabla \log \tilde{q}(x^{(i)}; \{x^{(j)}\}_j)\}_i)$. On the other hand, the exact dynamics leads to the density update $\partial_t q_t(x) = \nabla^2 q_t(x)$ from FPE (6), so $q_{t+\varepsilon}(x) \approx \tilde{q}(x; \{x^{(i)}\}_i) + \varepsilon \nabla^2 \tilde{q}(x; \{x^{(i)}\}_i)$. The principle then translates to the requirement $\tilde{q}(x; \{x^{(i)} - \varepsilon \nabla \log \tilde{q}(x^{(i)}; \{x^{(j)}\}_j)\}_i) = \tilde{q}(x; \{x^{(i)}\}_i) + \varepsilon \nabla^2 \tilde{q}(x; \{x^{(i)}\}_i)$, which can be enforced by minimizing the averaged squared difference over the particles. In the limit $\varepsilon \to 0$, the objective can be formulated as $\frac{1}{N}\sum_l \left[\nabla^2 \tilde{q}(x^{(l)}; \{x^{(i)}\}_i) + \sum_j \nabla_{x^{(j)}} \tilde{q}(x^{(l)}; \{x^{(i)}\}_i) \cdot \nabla \log \tilde{q}(x^{(j)}; \{x^{(i)}\}_i)\right]^2$. Liu et al. (2019a) applied this objective to select the bandwidth parameter σ of a Gaussian kernel,[ag] which achieves an attractive well-aligned pattern that ameliorates particle collapse.

Wang et al. (2019) made a generalization of SVGD that uses a matrix-valued kernel $\boldsymbol{K}(\cdot, \cdot)$, which is a symmetric matrix-valued function $\boldsymbol{K}(x, x') = \boldsymbol{K}(x', x)^\top$ that is positive-definite everywhere. They solved Eq. (27) for the optimal vector field in the associated vector-valued RKHS $\mathcal{H}_{\boldsymbol{K}}$ of the kernel, which is correspondingly defined as the linear space of vector-valued functions $\left\{\sum_l \boldsymbol{K}(\cdot, x^{(l)})\boldsymbol{\alpha}^{(l)}\right\}$ (every $\boldsymbol{\alpha}^{(l)}$ is a vector; summation over a countable index set) with inner product $\left\langle \sum_l \boldsymbol{K}(\cdot, x^{(l)})\boldsymbol{\alpha}^{(l)}, \sum_{l'} \boldsymbol{K}(\cdot, y^{(l')})\boldsymbol{\beta}^{(l')} \right\rangle_{\mathcal{H}_{\boldsymbol{K}}} := \sum_{l,l'} \boldsymbol{\alpha}^{(l)\top} \boldsymbol{K}(x^{(l)}, x^{(l')})\boldsymbol{\beta}^{(l')}$ (*c.f.* Section 4.1). The corresponding reproducing property is $\langle \boldsymbol{f}(\cdot), \boldsymbol{K}(\cdot, x)\boldsymbol{\alpha}\rangle_{\mathcal{H}_{\boldsymbol{K}}} = \boldsymbol{\alpha}^\top \boldsymbol{f}(x)$. The resulting matrix-valued SVGD vector field is:

$$V_t^{\mathrm{SVGD},\boldsymbol{K}}(x) := \mathbb{E}_{q_t(x')}[\boldsymbol{K}(x, x')\nabla_{x'} \log p(x') + \nabla_{x'} \cdot \boldsymbol{K}(x, x')].$$

[ag]For optimizing bandwidth σ, they divided the objective by σ^{2m+4} to make a dimensionless objective.

Apparently, $\boldsymbol{K}(\cdot, \cdot) = K(\cdot, \cdot)I_m$ recovers the previous vector-valued RKHS $\mathcal{H}_{\boldsymbol{K}} = \mathcal{H}^m$ and the SVGD vector field equation (29). The general formulation includes message-passing/graphical SVGD (Wang et al., 2018; Zhuo et al., 2018) and mirrored SVGD (Shi et al., 2022) that came afterward. They made a discussion on the choice of $\boldsymbol{K}$ by using the induced kernel under a geometric transformation, and proposed a second-order method that is cheaper than Newton SVGD (Detommaso et al., 2018).

5 Conclusion

In this chapter we have reviewed the geometry consideration in sampling methods, including MCMC and ParVI methods, which are among the major tools for Bayesian inference. This is directly required when latent variables are defined on a manifold to better reflect the data structure/semantics. We have shown how MCMC and ParVI methods on manifolds are developed and simulated to solve the inference task of such latent variables. These methods also enable leveraging information geometry, which endows the latent space with a metric on the likelihood distribution, and leads to a faster convergence. Particularly for ParVI methods as a formulation of variational inference, a geometric interpretation as the gradient flow of the KL divergence on certain abstract distribution manifolds can be coined. This inspires various analysis and methods of ParVI, and also connects ParVI with MCMC which benefits one with techniques of the other. Nevertheless, computation involving non-Euclidean geometry is often more costly. Dynamics simulation in the coordinate space requires matrix inversion or higher-order derivatives, and only a few manifolds have a closed-form expression of geodesic and exponential map in the embedded space. ParVIs need to find a tractable approximation to the gradient flow, and the currently prevailing kernel method is not as effective in high dimensions. Further progress addressing these issues would enable people to enjoy the benefits of these geometric methods with less cost.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2021ZD0110502, 2017YFA0700904), NSFC Projects (Nos. 62061136001, 61621136008, 62106121, U19B2034, U19A2081, U1811461), the major key project of PCL (No. PCL2021A12), and Tsinghua Guo Qiang Institute, and the High Performance Computing Center, Tsinghua University.

References

Abraham, R., Marsden, J.E., Ratiu, T., 2012. Manifolds, Tensor Analysis, and Applications. vol. 75, Springer Science & Business Media, New York.

Amari, S.-I., 1998. Natural gradient works efficiently in learning. Neural Comput. 10 (2), 251–276.

Amari, S.-I., 2016. Information Geometry and Its Applications. Springer, Tokyo.
Amari, S.-I., Nagaoka, H., 2007. Methods of Information Geometry. vol. 191, American Mathematical Soc., Providence, Rhode Island.
Ambrosio, L., Gangbo, W., 2008. Hamiltonian ODEs in the Wasserstein space of probability measures. Commun. Pure Appl. Math. 61 (1), 18–53.
Ambrosio, L., Gigli, N., Savaré, G., 2008. Gradient Flows: In Metric Spaces and in the Space of Probability Measures. Springer Science & Business Media, Berlin.
Arvanitidis, G., Hansen, L.K., Hauberg, S., 2018. Latent space oddity: on the curvature of deep generative models. In: International Conference on Learning Representations.
Arvanitidis, G., Hauberg, S., Hennig, P., Schober, M., 2019. Fast and robust shortest paths on manifolds learned from data. In: The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, pp. 1506–1515.
Ba, J., Erdogdu, M.A., Ghassemi, M., Suzuki, T., Wu, D., Sun, S., Zhang, T., 2019. Towards characterizing the high-dimensional bias of kernel-based particle inference algorithms. In: The 2nd Symposium on Advances in Approximate Bayesian Inference.
Barbour, A.D., 1990. Stein's method for diffusion approximations. Probab. Theory Relat. Fields 84 (3), 297–322.
Beck, A., Teboulle, M., 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. 31 (3), 167–175.
Benamou, J.-D., Brenier, Y., 2000. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numer. Math. 84 (3), 375–393.
Betancourt, M., 2015. The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling. In: Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), IMLS, Lille, France, pp. 533–540.
Betancourt, M., 2017. A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434.
Betancourt, M., Byrne, S., Livingstone, S., Girolami, M., et al., 2017. The geometric foundations of Hamiltonian Monte Carlo. Bernoulli 23 (4A), 2257–2298.
Billingsley, P., 2012. Probability and Measure. John Wiley & Sons, New Jersey, ISBN: 978-1-118-12237-2.
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
Brubaker, M.A., Salzmann, M., Urtasun, R., 2012. A family of MCMC methods on implicitly defined manifolds. In: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS-12), AISTATS Committee, La Palma, Canary Islands, pp. 161–172.
Byrne, S., Girolami, M., 2013. Geodesic Monte Carlo on embedded manifolds. Scand. J. Stat. 40 (4), 825–845.
Caterini, A.L., Doucet, A., Sejdinovic, D., 2018. Hamiltonian variational auto-encoder. In: Advances in Neural Information Processing Systems, vol. 31.
Chen, P., Ghattas, O., 2020. Projected Stein variational gradient descent. Adv. Neural Inf. Proces. Syst. 33, 1947–1958.
Chen, Y., Li, W., 2018. Natural gradient in Wasserstein statistical manifold. arXiv:1805.08380.
Chen, T., Fox, E., Guestrin, C., 2014. Stochastic gradient Hamiltonian Monte Carlo. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014), IMLS, Beijing, China, pp. 1683–1691.
Chen, C., Ding, N., Carin, L., 2015. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In: Advances in Neural Information Processing Systems, NIPS Foundation, Montréal, Canada, pp. 2269–2277.

Chen, C., Zhang, R., Wang, W., Li, B., Chen, L., 2018a. A unified particle-optimization framework for scalable Bayesian sampling. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2018), Association for Uncertainty in Artificial Intelligence, Monterey, California USA.

Chen, N., Klushyn, A., Kurle, R., Jiang, X., Bayer, J., Smagt, P., 2018b. Metrics for deep generative models. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp. 1540–1550.

Cheng, X., Bartlett, P., 2017. Convergence of Langevin MCMC in KL-divergence. arXiv:1705.09048.

Cheng, X., Chatterji, N.S., Bartlett, P.L., Jordan, M.I., 2018. Underdamped Langevin MCMC: a non-asymptotic analysis. In: Conference on Learning Theory, PMLR, pp. 300–323.

Chewi, S., Le Gouic, T., Lu, C., Maunu, T., Rigollet, P., 2020. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. In: Advances in Neural Information Processing Systems, vol. 33, pp. 2098–2109.

Chwialkowski, K., Strathmann, H., Gretton, A., 2016. A kernel test of goodness of fit. In: Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), IMLS, New York, New York USA, pp. 2606–2615.

Da Silva, A.C., 2001. Lectures on Symplectic Geometry. vol. 3575 Springer, Boston.

Dalalyan, A.S., 2017. Theoretical guarantees for approximate sampling from smooth and log-concave densities. J. R. Stat. Soc. B (Stat. Methodol.) 79 (3), 651–676.

Davidson, T.R., Falorsi, L., De Cao, N., Kipf, T., Tomczak, J.M., 2018. Hyperspherical variational auto-encoders. arXiv:1804.00891.

Detommaso, G., Cui, T., Marzouk, Y., Spantini, A., Scheichl, R., 2018. A Stein variational Newton method. In: Advances in Neural Information Processing Systems, NIPS Foundation, Montréal, Canada, pp. 9187–9197.

Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R.D., Neven, H., 2014. Bayesian sampling using stochastic gradient thermostats. In: Advances in Neural Information Processing Systems, NIPS Foundation, Montréal, Canada, pp. 3203–3211.

Dinh, L., Sohl-Dickstein, J., Bengio, S., 2017. Density estimation using real NVP. In: Proceedings of the International Conference on Learning Representations (ICLR 2017).

Do Carmo, M.P., 1992. Riemannian Geometry. Birkhäuser, Boston.

Dockhorn, T., Vahdat, A., Kreis, K., 2021. Score-based generative modeling with critically-damped Langevin diffusion. In: Proceedings of the International Conference on Learning Representations (ICLR 2021).

Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D., 1987. Hybrid Monte Carlo. Phys. Lett. B 195 (2), 216–222.

Duncan, A., Nüsken, N., Szpruch, L., 2019. On the geometry of Stein variational gradient descent. arXiv:1912.00894.

Durmus, A., Moulines, E., 2016. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. arXiv:1605.01559.

Durmus, A., Moulines, E., Saksman, E., 2017. On the convergence of Hamiltonian Monte Carlo. arXiv:1705.00166.

Eberle, A., Guillin, A., Zimmer, R., 2019. Couplings and quantitative contraction rates for Langevin dynamics. Ann. Probab. 47 (4), 1982–2010.

Ehlers, J., Pirani, F., Schild, A., 1972. The geometry of free fall and light propagation. In: General Relativity, Clarendon Press, Oxford, pp. 63–84 (papers in honour of JL Synge).

Erbar, M., et al., 2010. The heat equation on manifolds as a gradient flow in the Wasserstein space. Ann. Inst. H. Poincaré Probab. Stat. 46 (1), 1–23.

Fernandes, R.L., Marcut, I., 2014. Lectures on Poisson Geometry. Springer, Basel.
Gangbo, W., Kim, H.K., Pacini, T., 2010. Differential Forms on Wasserstein Space and Infinite-Dimensional Hamiltonian Systems. American Mathematical Soc., Providence, Rhode Island.
Girolami, M., Calderhead, B., 2011. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J. R. Stat. Soc. B (Stat. Methodol.) 73 (2), 123–214.
Gong, W., Li, Y., Hernández-Lobato, J.M., 2021. Sliced kernelized Stein discrepancy. In: Proceedings of the International Conference on Learning Representations (ICLR 2021).
Gorham, J., Mackey, L., 2015. Measuring sample quality with Stein's method. In: Advances in Neural Information Processing Systems, NIPS Foundation, Montréal, Canada, pp. 226–234.
Gorham, J., Mackey, L., 2017. Measuring sample quality with kernels. arXiv:1703.01717.
Grattarola, D., Livi, L., Alippi, C., 2018. Adversarial autoencoders with constant-curvature latent manifolds. arXiv:1812.04314.
Hairer, E., Lubich, C., Wanner, G., 2006. Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations. vol. 31 Springer Science & Business Media.
He, D., Shi, W., Li, S., Gao, X., Zhang, J., Bian, J., Wang, L., Liu, T.-Y., 2022. Learning physics-informed neural networks without stacked back-propagation. arXiv:2202.09340.
Hopf, H., Rinow, W., 1931. Über den begriff der vollständigen differential geometrischen fläche. Comment. Math. Helv. 3 (1), 209–225.
James, I.M., 1976. The Topology of Stiefel Manifolds. vol. 24 Cambridge University Press, New York.
Jordan, R., Kinderlehrer, D., Otto, F., 1998. The variational formulation of the Fokker-Planck equation. SIAM J. Math. Anal. 29 (1), 1–17.
Kalatzis, D., Eklund, D., Arvanitidis, G., Hauberg, S., 2020. Variational autoencoders with Riemannian Brownian motion priors. In: International Conference on Machine Learning, PMLR, pp. 5053–5066.
Kasai, H., Sato, H., Mishra, B., 2018. Riemannian stochastic quasi-Newton algorithm with variance reduction and its convergence analysis. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp. 269–278.
Kent, J., 1978. Time-reversible diffusions. Adv. Appl. Probab. 10 (4), 819–835.
Khan, M.E., Nielsen, D., 2018. Fast yet simple natural-gradient descent for variational inference in complex models. In: 2018 International Symposium on Information Theory and Its Applications (ISITA), IEEE, Singapore, pp. 31–35.
Kheyfets, A., Miller, W.A., Newton, G.A., 2000. Schild's ladder parallel transport procedure for an arbitrary connection. Int. J. Theor. Phys. 39 (12), 2891–2898.
Kingma, D.P., Dhariwal, P., 2018. Glow: generative flow with invertible 1×1 convolutions. In: Advances in Neural Information Processing Systems, vol. 31.
Kolouri, S., Zou, Y., Rohde, G.K., 2016. Sliced Wasserstein kernels for probability distributions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5258–5267.
Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., Rohde, G., 2019. Generalized sliced Wasserstein distances. In: Advances in Neural Information Processing Systems, vol. 32.
Korba, A., Salim, A., Arbel, M., Luise, G., Gretton, A., 2020. A non-asymptotic analysis for Stein variational gradient descent. In: Advances in Neural Information Processing Systems, vol. 33, pp. 4672–4682.
Kováčik, O., Rákosník, J., 1991. On spaces $L^p(x)$ and $W^{k,p}(x)$. Czechoslov. Math. J. 41 (4), 592–618.
Lan, S., Zhou, B., Shahbaba, B., 2014. Spherical Hamiltonian Monte Carlo for constrained target distributions. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014), IMLS, Beijing, China, pp. 629–637.

Lan, S., Stathopoulos, V., Shahbaba, B., Girolami, M., 2015. Markov chain Monte Carlo from Lagrangian dynamics. J. Comput. Graph. Stat. 24 (2), 357–378.

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

Lee, Y.T., Vempala, S.S., 2018. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, pp. 1115–1121.

Li, Y., Turner, R.E., 2018. Gradient estimators for implicit models. In: Proceedings of the International Conference on Learning Representations (ICLR 2018), ICLR Committee, Vancouver, Canada. https://openreview.net/forum?id=SJi9WOeRb.

Li, C., Chen, C., Carlson, D.E., Carin, L., 2016. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In: The 30th AAAI Conference on Artificial Intelligence (AAAI-16), AAAI Press, Phoenix, Arizona USA, vol. 2, pp. 1788–1794. 3.

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

Liu, Q., Wang, D., 2016. Stein variational gradient descent: a general purpose Bayesian inference algorithm. In: Advances in Neural Information Processing Systems, NIPS Foundation, Barcelona, Spain, pp. 2370–2378.

Liu, Q., Wang, D., 2018. Stein variational gradient descent as moment matching. In: Advances in Neural Information Processing Systems, vol. 31.

Liu, C., Zhu, J., 2018. Riemannian Stein variational gradient descent for Bayesian inference. In: The 32nd AAAI Conference on Artificial Intelligence, AAAI Press, New Orleans, Louisiana USA, pp. 3627–3634.

Liu, C., Zhu, J., Song, Y., 2016a. Stochastic gradient geodesic MCMC methods. In: Advances in Neural Information Processing Systems 29, NIPS Foundation, Barcelona, Spain, pp. 3009–3017.

Liu, Q., Lee, J.D., Jordan, M.I., 2016b. A kernelized Stein discrepancy for goodness-of-fit tests. In: Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), IMLS, New York, New York USA.

Liu, Y., Shang, F., Cheng, J., Cheng, H., Jiao, L., 2017. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In: Advances in Neural Information Processing Systems, NIPS Foundation, Long Beach, California USA, pp. 4875–4884.

Liu, C., Zhuo, J., Cheng, P., Zhang, R., Zhu, J., Carin, L., 2019a. Understanding and accelerating particle-based variational inference. In: Proceedings of the 36th International Conference on Machine Learning, IMLS, Long Beach, California USA, vol. 97, pp. 4082–4092.

Liu, C., Zhuo, J., Zhu, J., 2019b. Understanding MCMC dynamics as flows on the Wasserstein space. In: Proceedings of the 36th International Conference on Machine Learning, IMLS, Long Beach, California USA, vol. 97, pp. 4093–4103.

Liu, X., Zhu, H., Ton, J.-F., Wynne, G., Duncan, A., 2022. Grassmann Stein variational gradient descent. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp. 2002–2021.

Livingstone, S., Betancourt, M., Byrne, S., Girolami, M., 2019. On the geometric ergodicity of Hamiltonian Monte Carlo. Bernoulli 25 (4A), 3109–3138.

Lott, J., 2008. Some geometric calculations on Wasserstein space. Commun. Math. Phys. 277 (2), 423–437.

Ma, Y.-A., Chen, T., Fox, E., 2015. A complete recipe for stochastic gradient MCMC. In: Advances in Neural Information Processing Systems, NIPS Foundation, Montréal, Canada, pp. 2899–2907.

Ma, Y.-A., Chatterji, N., Cheng, X., Flammarion, N., Bartlett, P., Jordan, M.I., 2019. Is there an analog of Nesterov acceleration for MCMC? arXiv:1902.00996.

Mangoubi, O., Smith, A., 2017. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. arXiv:1708.07114.

Mathieu, E., Lan, C.L., Maddison, C.J., Tomioka, R., Teh, Y.W., 2019. Hierarchical representations with Poincaré variational auto-encoders. arXiv:1901.06033.

Nagano, Y., Yamaguchi, S., Fujita, Y., Koyama, M., 2019. A differentiable Gaussian-like distribution on hyperbolic space for gradient-based learning. arXiv:1902.02992.

Nash, J., 1956. The imbedding problem for Riemannian manifolds. Ann. Math. 63 (1), 20–63.

Neal, R.M., 2011. MCMC using Hamiltonian dynamics. In: Handbook of Markov Chain Monte Carlo, vol. 2.

Nesterov, Y., 1983. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Math. Doklady 27 (2), 372–376.

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

Otto, F., 2001. The geometry of dissipative evolution equations: the porous medium equation. Commun. Partial Differ. Equ. 26 (1), 101–174.

Ovinnikov, I., 2019. Poincaré Wasserstein autoencoder. arXiv:1901.01427.

Patterson, S., Teh, Y.W., 2013. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In: Advances in Neural Information Processing Systems, NIPS Foundation, Lake Tahoe, Nevada USA, pp. 3102–3110.

Persson, M., 2014. The Whitney Embedding Theorem. Umeå University, Umeå, Sweden.

Ranganath, R., Tran, D., Altosaar, J., Blei, D., 2016. Operator variational inference. In: Advances in Neural Information Processing Systems, NIPS Foundation, Barcelona, Spain, pp. 496–504.

Reisinger, J., Waters, A., Silverthorn, B., Mooney, R.J., 2010. Spherical topic models. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), IMLS, Haifa, Israel, pp. 903–910.

Roberts, G.O., Stramer, O., 2002. Langevin diffusions and Metropolis-Hastings algorithms. Methodol. Comput. Appl. Probab. 4 (4), 337–357.

Roberts, G.O., Tweedie, R.L., et al., 1996. Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli 2 (4), 341–363.

Romano, G., 2007. Continuum mechanics on manifolds. In: Lecture notes University of Naples Federico II, Naples, Italy, pp. 1–695.

Salakhutdinov, R., Mnih, A., 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In: Proceedings of the 25th International Conference on Machine Learning (ICML 2008), IMLS, Helsinki, Finland, Omnipress, pp. 880–887.

Santambrogio, F., 2017. Euclidean, metric, and Wasserstein gradient flows: an overview. Bull. Math. Sci. 7 (1), 87–154.

Särkkä, S., Solin, A., 2019. Applied Stochastic Differential Equations. vol. 10 Cambridge University Press.

Seiler, C., Rubinstein-Salzedo, S., Holmes, S., 2014. Positive curvature and Hamiltonian Monte Carlo. In: Advances in Neural Information Processing Systems, vol. 27.

Shao, H., Kumar, A., Thomas Fletcher, P., 2018. The Riemannian geometry of deep generative models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 315–323.

Shi, J., Sun, S., Zhu, J., 2018. A spectral approach to gradient estimation for implicit distributions. In: Proceedings of the 35th International Conference on Machine Learning (ICML 2018), IMLS, Stockholm, Sweden, pp. 4651–4660.

Shi, J., Liu, C., Mackey, L., 2022. Sampling with mirrored Stein operators. In: Proceedings of the International Conference on Learning Representations (ICLR 2022).

Song, Y., Zhu, J., 2016. Bayesian matrix completion via adaptive relaxed spectral regularization. In: The 30th AAAI Conference on Artificial Intelligence (AAAI-16), AAAI Press, Phoenix, Arizona USA, pp. 2044–2050.

Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B., 2021. Score-based generative modeling through stochastic differential equations. In: Proceedings of the International Conference on Learning Representations (ICLR 2021).

Stein, C., 1972. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory, The Regents of the University of California, Oakland.

Steinwart, I., Christmann, A., 2008. Support Vector Machines. Springer Science & Business Media, New York.

Stiefel, E.L., 1935. Richtungsfelder und fernparallelismus in n-dimensionalen mannigfaltigkeiten. Comment. Math. Helv. 8 (1), 305–353.

Taghvaei, A., Mehta, P.G., 2019. Accelerated gradient flow for probability distributions. In: Proceedings of the 36th International Conference on Machine Learning (ICML 2019), IMLS, Long Beach, California USA.

Toth, P., Rezende, D.J., Jaegle, A., Racanière, S., Botev, A., Higgins, I., 2020. Hamiltonian generative networks. In: Proceedings of the International Conference on Learning Representations (ICLR 2020).

Villani, C., 2008. Optimal Transport: Old and New. vol. 338 Springer Science & Business Media, Berlin.

Wang, Y., Li, W., 2022. Accelerated information gradient flow. J. Sci. Comput. 90 (1), 1–47.

Wang, M.C., Uhlenbeck, G.E., 1945. On the theory of the Brownian motion II. Rev. Mod. Phys. 17 (2–3), 323.

Wang, D., Zeng, Z., Liu, Q., 2018. Stein variational message passing for continuous graphical models. In: International Conference on Machine Learning, PMLR, pp. 5219–5227.

Wang, D., Tang, Z., Bajaj, C., Liu, Q., 2019. Stein variational gradient descent with matrix-valued kernels. In: Advances in Neural Information Processing Systems, vol. 32.

Welling, M., Teh, Y.W., 2011. Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), IMLS, Bellevue, Washington USA, pp. 681–688.

Whitney, H., 1944. The self-intersections of a smooth n-manifold in 2n-space. Ann. Math. 45 (220-446), 180.

Wibisono, A., 2018. Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem. arXiv:1802.08089.

Wibisono, A., Wilson, A.C., Jordan, M.I., 2016. A variational perspective on accelerated methods in optimization. Proc. Natl. Acad. Sci. 113 (47), E7351–E7358.

Xifara, T., Sherlock, C., Livingstone, S., Byrne, S., Girolami, M., 2014. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. Stat. Probab. Lett. 91, 14–19.

Yanush, V., Kropotov, D., 2019. Hamiltonian Monte-Carlo for orthogonal matrices. arXiv:1901.08045.

Zhang, H., Sra, S., 2018. An estimate sequence for geodesically convex optimization. In: Proceedings of the 31st Annual Conference on Learning Theory (COLT 2018), IMLS, Stockholm, Sweden, pp. 1703–1723.

Zhang, H., Reddi, S.J., Sra, S., 2016. Riemannian SVRG: fast stochastic optimization on Riemannian manifolds. In: Advances in Neural Information Processing Systems, NIPS Foundation, Barcelona, Spain, pp. 4592–4600.

Zhou, P., Yuan, X.-T., Feng, J., 2019. Faster first-order methods for stochastic non-convex optimization on Riemannian manifolds. In: The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, pp. 138–147.

Zhu, M., Liu, C., Zhu, J., 2020. Variance reduction and quasi-Newton for particle-based variational inference. In: Proceedings of the 37th International Conference on Machine Learning, Virtual, vol. 119, pp. 11576–11587.

Zhuo, J., Liu, C., Shi, J., Zhu, J., Chen, N., Zhang, B., 2018. Message passing Stein variational gradient descent. In: Proceedings of the 35th International Conference on Machine Learning, Stockholmsmässan, Stockholm Sweden, vol. 80, pp. 6018–6027.