Appendix

A. Proofs

A.1. Proof of Lemma 1

Given the dynamics (1), the distribution curve $(q_t)_t$ is governed by the Fokker-Planck equation (*e.g.*, Risken (1996)):

$$\partial_t q_t = -\partial_i(q_t V^i) + \partial_i\partial_j(q_t D^{ij}),$$

which reduces to:

$$\begin{aligned}
\partial_t q_t =& -(\partial_i q_t)V^i - q_t(\partial_i V^i) \\
&+ q_t(\partial_i\partial_j D^{ij}) + (\partial_i\partial_j q_t)D^{ij} \\
&+ (\partial_i q_t)(\partial_j D^{ij}) + (\partial_j q_t)(\partial_i D^{ij}) \\
=& -(\partial_i q_t)(\partial_j D^{ij} + \partial_j Q^{ij}) - (\partial_i q_t)(D^{ij} + Q^{ij})\frac{\partial_j p}{p} \\
&- q_t\partial_i\partial_j(D^{ij} + Q^{ij}) - q_t(\partial_i D^{ij} + \partial_i Q^{ij})\frac{\partial_j p}{p} \\
&- q_t(D^{ij} + Q^{ij})(\frac{\partial_i\partial_j p}{p} - \frac{(\partial_i p)(\partial_j p)}{p^2}) \\
&+ q_t(\partial_i\partial_j D^{ij}) + (\partial_i\partial_j q_t)D^{ij} \\
&+ (\partial_i q_t)(\partial_j D^{ij}) + (\partial_j q_t)(\partial_i D^{ij}) \\
=& \quad (\partial_i q_t - \frac{q_t}{p}\partial_i p)(\partial_j D^{ij} - \partial_j Q^{ij}) \\
&- \frac{1}{p}(\partial_i q_t)(\partial_j p)(D^{ij} + Q^{ij}) \\
&- \frac{q_t}{p}(\partial_i\partial_j p)D^{ij} + \frac{q_t}{p^2}(\partial_i p)(\partial_j p)D^{ij} + (\partial_i\partial_j q_t)D^{ij},
\end{aligned}$$

where we have used the symmetry of D and skew-symmetry of Q in the last equality: $(\partial_j p)(\partial_i D^{ij}) = (\partial_i p)(\partial_j D^{ji}) = (\partial_i p)(\partial_j D^{ij})$ and similarly $(\partial_j p)(\partial_i Q^{ij}) = -(\partial_i p)(\partial_j Q^{ij})$; $\partial_i\partial_j Q^{ij} = \partial_j\partial_i Q^{ji} = -\partial_i\partial_j Q^{ij}$ so $\partial_i\partial_j Q^{ij} = 0$ and similarly $(\partial_i p)(\partial_j p)Q^{ij} = 0$, $(\partial_i\partial_j p)Q^{ij} = 0$.

The deterministic dynamics in the theorem $\mathrm{d}x = W_t(x)\,\mathrm{d}t$ with $W_t(x)$ defined in Eq. (7) induces the curve:

$$\begin{aligned}
\partial_t q_t =& -\partial_i(q_t(W_t)^i) \\
=& -(\partial_i q_t)(W_t)^i - q_t(\partial_i(W_t)^i) \\
=& -(\partial_i q_t)D^{ij}(\frac{\partial_j p}{p} - \frac{\partial_j q_t}{q_t}) \\
&- (\partial_i q_t)Q^{ij}(\frac{\partial_j p}{p}) - (\partial_i q_t)(\partial_j Q^{ij}) \\
&- q_t(\partial_i D^{ij})(\frac{\partial_j p}{p} - \frac{\partial_j q_t}{q_t}) \\
&- q_t D^{ij}(\frac{\partial_i\partial_j p}{p} - \frac{(\partial_j p)(\partial_i p)}{p^2} - \frac{\partial_i\partial_j q_t}{q_t} + \frac{(\partial_j q_t)(\partial_i q_t)}{q_t^2}) \\
&- q_t(\partial_i Q^{ij})\frac{\partial_j p}{p} - q_t Q^{ij}(\frac{\partial_i\partial_j p}{p} - \frac{(\partial_j p)(\partial_i p)}{p^2}) \\
&- q_t(\partial_i\partial_j Q^{ij})
\end{aligned}$$

$$\begin{aligned}
=& \quad (\partial_i q_t - \frac{q_t}{p}\partial_i p)(\partial_j D^{ij} - \partial_j Q^{ij}) \\
&- \frac{1}{p}(\partial_i q_t)(\partial_j p)(D^{ij} + Q^{ij}) \\
&- \frac{q_t}{p}(\partial_i\partial_j p)D^{ij} + \frac{q_t}{p^2}(\partial_i p)(\partial_j p)D^{ij} + (\partial_i\partial_j q_t)D^{ij},
\end{aligned}$$

where we have also applied aforementioned properties in the last equality. Now we see that the two dynamics induce the same distribution curve thus they are equivalent.

A.2. Derivation of Eq. (8)

Barbour's generator is understood as the directional derivative $(\mathcal{A}f)(x) = \frac{\mathrm{d}}{\mathrm{d}t}F_f(q_t)\Big|_{\substack{q_0=\delta_x\\ t=0}}$ on $\mathcal{P}(\mathbb{R}^M)$. Due to the definition of gradient, this can be written as $(\mathcal{A}f)(x) = \langle \operatorname{grad} F_f, \pi_{q_0}(W_0)\rangle_{T_{q_0}\mathcal{P}} = \langle \operatorname{grad} F_f, W_0\rangle_{\mathcal{L}^2_{q_0}}$, where $\pi_{q_0}(W_0)$ is the tangent vector of the distribution curve $(q_t)_t$ at time 0 due to Lemma 1, and the last equality holds due to that π_q is the orthogonal projection from $\mathcal{L}^2_q$ to $T_q\mathcal{P}$ and $\operatorname{grad} F_f \in T_{q_0}\mathcal{P}$ (see Section 2.2.1).

Before going on, we first introduce the notion of *weak derivative* (*e.g.*, Nicolaescu (2007), Def. 10.2.1) of a distribution. For a distribution with smooth density function q and a smooth function $f \in \mathcal{C}_c^\infty(\mathbb{R}^M)$, the rule of integration by parts tells us:

$$\begin{aligned}
\int_{\mathbb{R}^M} f(x)(\partial_i q(x))\,\mathrm{d}x = & \int_{\mathbb{R}^M} \partial_i(f(x)q(x))\,\mathrm{d}x \\
& - \int_{\mathbb{R}^M} (\partial_i f(x))q(x)\,\mathrm{d}x.
\end{aligned}$$

Due to Gauss's theorem (*e.g.*, Abraham et al. (2012), Thm. 8.2.9), $\int_{\mathbb{R}^M} \partial_i(f(x)q(x))\,\mathrm{d}x = \lim_{R\to+\infty}\int_{\mathbb{S}^{M-1}(R)}(f(y)q(y))v_i(y)\,\mathrm{d}y$, where $\mathbb{S}^{M-1}(R)$ is the $(M-1)$-dimensional sphere in $\mathbb{R}^M$ with radius R, $y \in \mathbb{S}^{M-1}$, and v_i is the i-th component of the unit normal vector v (pointing outwards) on $\mathbb{S}^{M-1}(R)$. Since f is compactly supported and $\lim_{\|x\|\to+\infty} q(x) = 0$, after a sufficiently large R, $f(y)q(y) = 0$, so the integral vanishes, and we have:

$$\begin{aligned}
\int_{\mathbb{R}^M} f(x)(\partial_i q(x))\,\mathrm{d}x = -\int_{\mathbb{R}^M} (\partial_i f(x))q(x)\,\mathrm{d}x, \\
\forall f \in \mathcal{C}_c^\infty(\mathbb{R}^M).
\end{aligned}$$

We can use this property as the definition of $\partial_i q$ for non-absolutely-continuous distributions, like the Dirac measure δ_{x_0}:

$$\begin{aligned}
\int_{\mathbb{R}^M} f(x)(\partial_i\delta_{x_0}(x))\,\mathrm{d}x :=& \int_{\mathbb{R}^M} (\partial_i f(x))\delta_{x_0}(x)\,\mathrm{d}x \\
=& \partial_i f(x_0).
\end{aligned}$$

Now we begin the derivation. Using the form in Eq. (7) and noting $q_0 = \delta_{x_0}$, we have:

$$
\begin{aligned}
&(\mathcal{A}f)(x_0) = \langle \operatorname{grad} F_f, W_0\rangle_{\mathcal{L}^2_{q_0}} \\
=&\mathbb{E}_{q_0(x)}[\langle \operatorname{grad} f(x), W_0(x)\rangle_{\mathbb{R}^M}] = \mathbb{E}_{q_0}[(\partial_i f) W_0^i] \\
=&\mathbb{E}_{q_0}\big[D^{ij}(\partial_i f)\big(\partial_j \log(p/q_0)\big) + Q^{ij}(\partial_i f)(\partial_j \log p) \\
&\qquad + (\partial_j Q^{ij})(\partial_i f)\big] \\
=&\quad \big[D^{ij}(\partial_i f)(\partial_j \log p)\big](x_0) \\
&\quad - \int_{\mathbb{R}^M} \big(D^{ij}(\partial_i f)\big)(x)(\partial_j q_0)(x)\,\mathrm{d}x \\
&\quad + \big[Q^{ij}(\partial_i f)(\partial_j \log p) + (\partial_j Q^{ij})(\partial_i f)\big](x_0) \\
=&\left[D^{ij}(\partial_i f)(\partial_j \log p) + \frac{1}{p}\partial_j(pQ^{ij})(\partial_i f)\right](x_0) \\
&\quad + \int_{\mathbb{R}^M} \partial_j\big(D^{ij}(\partial_i f)\big)(x) q_0(x)\,\mathrm{d}x \\
=&\left[D^{ij}(\partial_i f)(\partial_j \log p) + \frac{1}{p}\partial_j(pQ^{ij})(\partial_i f)\right](x_0) \\
&\quad + \big[\partial_j\big(D^{ij}(\partial_i f)\big)\big](x_0) \\
=&\Big[D^{ij}(\partial_i f)(\partial_j \log p) + \frac{1}{p}\partial_j(pQ^{ij})(\partial_i f) \\
&\qquad + (\partial_j D^{ij})(\partial_i f) + D^{ij}(\partial_i \partial_j f)\Big](x_0) \\
=&\left[\frac{1}{p}\partial_j\big(p(D^{ij} + Q^{ij})\big)(\partial_i f) + D^{ij}(\partial_i\partial_j f)\right](x_0) \\
=&\left[\frac{1}{p}\partial_j\big(p(D^{ij} + Q^{ij})\big)(\partial_i f) + (D^{ij} + Q^{ij})(\partial_i\partial_j f)\right](x_0) \\
=&\left[\frac{1}{p}\partial_j\left[p\left(D^{ij} + Q^{ij}\right)(\partial_i f)\right]\right](x_0),
\end{aligned}
$$

where the second last equality holds due to $Q^{ij}(\partial_i \partial_j f) = 0$ from the skew-symmetry of Q. This completes the derivation.

A.3. Proof of Lemma 2

Noting that the KL divergence $\mathrm{KL}_p(q) = \int_{\mathcal{M}} \log(q/p)\,\mathrm{d}q$ is a non-linear function on $\mathcal{P}(\mathcal{M})$, we need to first find its linearization. We fix a point $q_0 \in \mathcal{P}(\mathcal{M})$. Eq. (2) gives its gradient at q_0: $\operatorname{grad} \mathrm{KL}_p(q_0) = \operatorname{grad} \log(q_0/p)$. Consider the linear function on $\mathcal{P}(\mathcal{M})$:

$$F : q \mapsto \int_{\mathcal{M}} \log(q_0/p)\,\mathrm{d}q.$$

According to existing knowledge (*e.g.*, Villani (2008), Ex. 15.10; Ambrosio et al. (2008), Lem. 10.4.1; Santambrogio (2017), Eq. 4.10), its gradient at q_0 is given by:

$$\big(\operatorname{grad} F\big)(q_0) = \operatorname{grad}\left(\left.\frac{\delta F}{\delta q}\right|_{q=q_0}\right),$$

where $\frac{\delta F}{\delta q}$ is the first functional variation of F, which is $\log(q_0/p)$ at $q = q_0$. Now we find that $\operatorname{grad} F(q_0) = \operatorname{grad} \log(q_0/p) = \operatorname{grad} \mathrm{KL}_p(q_0)$, so $F(q)$ is the linearization of $\mathrm{KL}_p(q)$ at $q = q_0$ and the corresponding $f \in \mathcal{C}_c^\infty(\mathcal{M})$ in Eq. (6) is $\log(q_0/p)$. Then we have:

$$\mathcal{X}_{\mathrm{KL}_p}(q_0) = \pi_{q_0}(X_{\log(q_0/p)}).$$

Referring to Eq. (4), $X_{\log(q_0/p)} = \beta^{ij}\partial_j \log(q_0/p)\partial_i$. Due to the generality of q_0, this completes the proof.

A.4. Proof of Theorem 5

For a fixed $q \in \mathcal{P}(\mathcal{M})$, two vector fields on $\mathcal{M}$ produce the same distribution curve if they have the same projection on $T_q\mathcal{P}(\mathcal{M})$, so showing $\pi_q(W) = \mathcal{W}_{\mathrm{KL}_p}(q)$ is sufficient for showing the equivalence of the two dynamics. This in turn is equivalent to show $\pi_q(W - \mathcal{W}_{\mathrm{KL}_p}(q)) = 0_{\mathcal{L}^2_q}$, or $\operatorname{div}\big(q(W - \mathcal{W}_{\mathrm{KL}_p}(q))\big) = \operatorname{div}(q0_{\mathcal{L}^2_q}) = 0$ (see Section 2.2.1).

We first consider case (b): given an fRP manifold $(\mathcal{M}, \tilde{g}, \beta)$, we define an MCMC dynamics whose diffusion matrix D and curl matrix Q are the coordinate expressions of the fiber-Riemannian structure $(\tilde{g}^{ij})$ and the Poisson structure (β^{ij}), respectively. It is regular, as Assumption 4 is satisfied due to properties of $(\tilde{g}^{ij})$ (see Eq. (9)) and (β^{ij}) (see Section 2.2.2). Its equivalent deterministic dynamics at q (see Lemma 1) is given by:

$$W^i = \tilde{g}^{ij}\partial_j \log(p/q) + \beta^{ij}\partial_j \log p + \partial_j \beta^{ij}.$$

So we have:

$$
\begin{aligned}
&\operatorname{div}\big(q(W - \mathcal{W}_{\mathrm{KL}_p}(q))\big) \\
=&\operatorname{div}\Big(q\big(\tilde{g}^{ij}\partial_j \log(p/q) + \beta^{ij}\partial_j \log p + \partial_j \beta^{ij} \\
&\qquad - (\tilde{g}^{ij} + \beta^{ij})\partial_j \log(p/q)\big)\,\partial_i\Big) \\
=&\operatorname{div}\big(q(\partial_j\beta^{ij} + \beta^{ij}\partial_j \log q)\partial_i\big) \\
=&\operatorname{div}\big((q\partial_j\beta^{ij} + \beta^{ij}\partial_j q)\partial_i\big) \\
=&\operatorname{div}\big(\partial_j(q\beta^{ij})\partial_i\big) \\
=&\partial_i\partial_j(q\beta^{ij}) \\
=&0,
\end{aligned}
$$

where the last equality holds due to the skew-symmetry of (β^{ij}). This shows that the constructed regular MCMC dynamics is equivalent to the fiber-gradient Hamiltonian flow $\mathcal{W}_{\mathrm{KL}_p}$ on $\mathcal{M}$.

For case (a), given any regular MCMC dynamics whose matrices (D, Q) satisfy Assumption 4, we can define an fRP manifold $(\mathcal{M}, \tilde{g}, \beta)$ whose structures are defined in the coordinate space by the matrices: $\tilde{g}^{ij} := D^{ij}$, $\beta^{ij} := Q^{ij}$. Assumption 4 guarantees that such $\tilde{g}$ is a valid fiber-Riemannian structure and β a valid Poisson structure. On this constructed manifold, we follow the above procedure to construct a regular MCMC dynamics equivalent to the fGH flow $\mathcal{W}_{\mathrm{KL}_p}$ on it, whose equivalent deterministic dynamics is:

$$W^i = D^{ij}\partial_j \log(p/q) + Q^{ij}\partial_j \log p + \partial_j Q^{ij},$$

which is exactly the one of the original MCMC dynamics.

This shows that the original regular MCMC dynamics is equivalent to the fGH flow $\mathcal{W}_{\mathrm{KL}_p}$ on the constructed fRP manifold.

Finally, statement (c) is verified in both cases by the introduced construction. This completes the proof.

B. Details on Flow Simulation of SGHMC Dynamics

We first introduce more details on the Blob method, referring to the works of Chen et al. (2018a) and Liu et al. (2018). The key problem in simulating a general flow on the Wasserstein space is to estimate the gradient $u(x) := -\nabla \log q(x)$ where $q(x)$ is the distribution corresponding to the current configuration of the particles. The gradient has to be estimated using the finite particles $\{x^{(i)}\}_{i=1}^N$ distributed obeying $q(x)$. The analysis of Liu et al. (2018) finds that an estimate method has to make a smoothing treatment, in the form of either smoothing density or smoothing functions. The Blob method (Chen et al., 2018a) first reformulates $u(x)$ in a variation form:

$$u(x) = \nabla \left(-\frac{\delta}{\delta q} \mathbb{E}_q[\log q] \right),$$

then with a kernel function K, it replaces the density in the $\log q$ term with a smoothed one:

$$\begin{aligned} u(x) \approx& \nabla \left(-\frac{\delta}{\delta q} \mathbb{E}_q[\log(q * K)] \right) \\ =& -\nabla \log(q * K) - \nabla \left(\frac{q}{(q * K)} * K \right), \end{aligned}$$

where "*" denotes convolution. This form enjoys the benefit of enabling the usage of the empirical distribution: take $q(x) = \hat{q}(x) := \frac{1}{N} \sum_{i=1}^N \delta_{x^{(i)}}(x)$, with $\delta_{x^{(i)}}(x)$ denoting the Dirac measure at $x^{(i)}$. The above formulation then becomes:

$$\begin{aligned} u(x^{(i)}) =& -\nabla_x \log q(x^{(i)}) \\ \approx& -\frac{\sum_k \nabla_{x^{(i)}} K^{(i,k)}}{\sum_j K^{(i,j)}} - \sum_k \frac{\nabla_{x^{(i)}} K^{(i,k)}}{\sum_j K^{(j,k)}}, \end{aligned}$$

where $K^{(i,j)} := K(x^{(i)}, x^{(j)})$. This coincides with Eq. (15).

The vanilla SGHMC dynamics replaces the dynamics $\mathrm{d}r = -C \nabla_r \log q(r) \, \mathrm{d}t$ in Eq. (13) with $\mathrm{d}r = 2C \, \mathrm{d}B_t$ or more intuitively $\mathrm{d}r = \mathcal{N}(0, 2C \, \mathrm{d}t)$, where B_t denotes the standard Brownian motion. The equivalence between these two dynamics can also be directly derived from the Fokker-Planck equation: the first one produces a curve by $\partial_t q_t = -\partial_i \big(q_t (-C^{ij} \partial_j \log q_t) \big) = \partial_i (C^{ij} \partial_j q_t)$, and the second one by $\partial_t q_t = \partial_i \partial_j (q_t C^{ij}) = \partial_i (C^{ij} \partial_j q_t)$ for a constant C, so the two curves coincides. But dynamics (14) cannot be simulated in a stochastic way, since $-\nabla_r \log q(r)$ and $-\nabla_\theta \log q(\theta)$ are used to update θ and r, respectively, that is, the correspondence of gradients and variables is switched. In this case, estimating the gradient cannot be avoided.

Finally, we write the explicit update rule of the proposed methods using Blob with particles $\{(\theta, r)^{(i)}\}_{i=1}^N$. Let K_θ, K_r be the kernel functions for θ and r, and ε be a step size. The update rule for pSGHMC-det in Eq. (13) becomes:

$$\begin{cases} \theta^{(i)} \leftarrow \theta^{(i)} + \varepsilon \Sigma^{-1} r^{(i)}, \\ r^{(i)} \leftarrow r^{(i)} + \varepsilon \nabla_\theta \log p(\theta^{(i)}) \\ \qquad - \varepsilon C \Big(\Sigma^{-1} r^{(i)} + \frac{\sum_k \nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(i,j)}} + \sum_k \frac{\nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(j,k)}} \Big), \end{cases}$$

and for pSGHMC-fGH in Eq. (14):

$$\begin{cases} \theta^{(i)} \leftarrow \theta^{(i)} + \varepsilon \Big(\Sigma^{-1} r^{(i)} + \frac{\sum_k \nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(i,j)}} + \sum_k \frac{\nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(j,k)}} \Big), \\ r^{(i)} \leftarrow r^{(i)} + \varepsilon \nabla_\theta \log p(\theta^{(i)}) \\ \qquad - \varepsilon \Big(\frac{\sum_k \nabla_{\theta^{(i)}} K_\theta^{(i,k)}}{\sum_j K_\theta^{(i,j)}} + \sum_k \frac{\nabla_{\theta^{(i)}} K_\theta^{(i,k)}}{\sum_j K_\theta^{(j,k)}} \Big) \\ \qquad - \varepsilon C \Big(\Sigma^{-1} r^{(i)} + \frac{\sum_k \nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(i,j)}} + \sum_k \frac{\nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(j,k)}} \Big), \end{cases}$$

where $K_\theta^{(i,j)} := K_\theta(\theta^{(i)}, \theta^{(j)})$ and similarly for $K_r^{(i,j)}$.

C. Detailed Settings of Experiments

C.1. Detailed Settings of the Synthetic Experiment

For the random variable $x = (x_1, x_2)$, the target distribution density $p(x)$ is defined by:

$$\begin{aligned} \log p(x) =& -0.01 \times \left(\frac{1}{2}(x_1^2 + x_2^2) + \frac{0.8}{2}(25 x_1 + x_2^2)^2 \right) \\ &+ \text{const}, \end{aligned}$$

which is inspired by the target distribution used in the work of Girolami & Calderhead (2011). We use the exact gradient of the log density instead of stochastic gradient. Fifty particles are used, which are initialized by $\mathcal{N}\big((-2, -7), 0.5^2 I\big)$. The window range is $(-7, 3)$ horizontally and $(-9, 9)$ vertically. See the caption of Fig. 3 for other settings.

C.2. Detailed Settings of the LDA Experiment

We follow the same settings as Ding et al. (2014), which is also adopted in Liu et al. (2018). The data set is the ICML data set[2] developed by Ding et al. (2014). We use 90% words in each document to train the topic proportion of the document and the left 10% words for evaluation. A random 80%-20% train-test split of the data set is conducted in each run.

For the LDA model, parameters of the Dirichlet prior of topics is $\alpha = 0.1$. The mean and standard deviation of the Gaussian prior on the topic proportions is $\beta = 0.1$ and $\sigma = 1.0$. Number of topics is 30 and batch size is fixed as 100. The number of Gibbs sampling in each stochastic gradient evaluation is 50.

[2] https://cse.buffalo.edu/~changyou/code/SGNHT.zip

All the inference methods share the same step size $\varepsilon = 1 \times 10^{-3}$. SGHMC-related methods (SGHMC, pSGHMC-det and pSGHMC-fGH) share the same parameters $\Sigma^{-1} = 300$ and $C = 0.1$. ParVI methods (Blob, pSGHMC-det and pSGHMC-fGH) use the HE method for kernel bandwidth selection (Liu et al., 2018). To match the fashion of ParVI methods, SGHMC is run with parallel chains and the last samples of each chain are collected.

C.3. Detailed Settings of the BNN Experiment

We use a 784-100-10 feedforward neural network with sigmoid activation function. The batch size is 500. SGHMC, pSGHMC-det and pSGHMC-fGH share the same parameters $\varepsilon = 5 \times 10^{-5}$, $\Sigma^{-1} = 1.0$ and $C = 1.0$, while Blob uses $\varepsilon = 5 \times 10^{-8}$ (larger ε leads to diverged result). For the ParVI methods, we find the median method and the HE method for bandwidth selection perform similarly, and we adopt the median method for faster implementation.