Appendix

A. Proofs

A.1: Proof of Theorem 2

For $v \in \mathcal{H}^D$, the objective can be expressed as:

$$
\begin{aligned}
&\left\langle v^{\mathrm{GF}}, v\right\rangle_{\mathcal{L}_q^2} \\
=&\mathbb{E}_q[(\nabla \log p - \nabla \log q)\cdot v] \\
=&\mathbb{E}_q[\nabla \log p \cdot v] - \int_{\mathcal{X}} \nabla q \cdot v \,\mathrm{d}x \\
\overset{(*)}{=}&\mathbb{E}_q[\nabla \log p \cdot v] + \int_{\mathcal{X}} q \nabla \cdot v \,\mathrm{d}x \\
=&\mathbb{E}_{q(x)}\left[\sum_{\alpha=1}^{D}\Big(\partial_\alpha \log p(x) v_\alpha(x) + \partial_\alpha v_\alpha(x)\Big)\right] \\
\overset{(\#)}{=}&\mathbb{E}_{q(x)}\left[\sum_{\alpha=1}^{D}\Big(\partial_\alpha \log p(x)\langle K(x,\cdot), v_\alpha(\cdot)\rangle_{\mathcal{H}}\right. \\
&\left.+ \langle \partial_\alpha K(x,\cdot), v_\alpha(\cdot)\rangle_{\mathcal{H}}\Big)\right] \\
=&\mathbb{E}_{q(x)}\left[\langle K(x,\cdot)\nabla \log p(x), v(\cdot)\rangle_{\mathcal{H}^D} + \langle \nabla K(x,\cdot), v(\cdot)\rangle_{\mathcal{H}^D}\right] \\
=&\mathbb{E}_{q(x)}\left[\langle K(x,\cdot)\nabla \log p(x) + \nabla K(x,\cdot), v(\cdot)\rangle_{\mathcal{H}^D}\right] \\
=&\left\langle \mathbb{E}_{q(x)}\left[K(x,\cdot)\nabla \log p(x) + \nabla K(x,\cdot)\right], v(\cdot)\right\rangle_{\mathcal{H}^D} \\
=&\left\langle v^{\mathrm{SVGD}}, v\right\rangle_{\mathcal{H}^D},
\end{aligned}
$$

where the equality $(*)$ holds due to the definition of weak derivative of distributions, and equality $(\#)$ holds due to the reproducing property for any function f in the reproducing kernel Hilbert space $\mathcal{H}$ of kernel K: $\langle K(x,\cdot), f(\cdot)\rangle_{\mathcal{H}} = f(x)$ (Steinwart & Christmann (2008), Chapter 4), and $\langle \partial_{x_\alpha} K(x,\cdot), f(\cdot)\rangle_{\mathcal{H}} = \partial_{x_\alpha} f(x)$ (Zhou, 2008).

A.2: Proof of Theorem 3

When q is absolutely continuous with respect to the Lebesgue measure of $\mathcal{X} = \mathbb{R}^D$, $\mathcal{L}_q^2$ has the same topological properties as $\mathcal{L}^2$, so conclusions we cite below can be adapted from $\mathcal{L}^2$ to $\mathcal{L}_q^2$. Note that the map $\phi \mapsto \phi * K, \mathcal{L}^2 \to \mathcal{L}^2$ is continuous, so $\mathcal{G} := \overline{\{\phi * K : \phi \in \mathcal{C}_c^\infty\}}^{\mathcal{L}_q^2} = \{\phi * K : \phi \in \overline{\mathcal{C}_c^\infty}^{\mathcal{L}^2}\} = \{\phi * K : \phi \in \mathcal{L}^2\} = \{\phi * K : \phi \in L^2\}^D$, where the second last equality holds due to *e.g.*, Theorem 2.11 of (Kováčik & Rákosník, 1991). On the other hand, due to Proposition 4.46 and Theorem 4.47 of (Steinwart & Christmann, 2008), the map $\phi \mapsto \phi * K$ is an isometric isomorphism between $\{\phi * K : \phi \in L^2\}$ and $\mathcal{H}$, the reproducing kernel Hilbert space of K. This indicates that $\mathcal{G}$ is isometrically isomorphic to $\mathcal{H}^D$.

A.3: Proof of Theorem 4

We will redefine some notations in this proof. According to the deduction in Appendix A.1, the objective of the optimization problem Eq. (5) $\langle v^{\mathrm{GF}}, v\rangle_{\mathcal{L}_q^2}$ can be cast as $\mathbb{E}_q[\nabla \log p \cdot v + \nabla \cdot v]$. With $q = \hat{q}$ and $v \in \mathcal{L}_p^2$, we write the optimization problem as:

$$
\sup_{v \in \mathcal{L}_p^2, \|v\|=1} \sum_{i=1}^{N}\Big(\nabla \log p(x^{(i)}) \cdot v(x^{(i)}) + \nabla \cdot v(x^{(i)})\Big), \tag{7}
$$

We will find a sequence of functions $\{v_n\}$ satisfying conditions in Eq. (7) while the objective goes to infinity.

We assume that there exists $r_0 > 0$ such that $p(x) > 0$ for any $\left\|x - x^{(i)}\right\|_\infty < r_0$, $i = 1, 2, \cdots, N$, which is reasonable because it is almost impossible to sample $x^{(i)}$ with $p(x)$ vanishes in every neighborhood of $x^{(i)}$.

Denoting $v(x) = (v_1(x), \cdots, v_D(x))^\top$ for any D-dimensional vector function v and $\nabla f(x) = (\partial_1 f(x), \cdots, \partial_D f(x))^\top$ for any real-valued function f, the objective can be written as:

$$
\begin{aligned}
\mathcal{L}_v =& \sum_{i=1}^{N}\Big(\nabla \log p(x^{(i)}) \cdot v(x^{(i)}) + \nabla \cdot v(x^{(i)})\Big) \\
=& \sum_{i=1}^{N}\Big(\sum_{\alpha=1}^{D}\partial_\alpha[\log p(x^{(i)})] v_\alpha(x^{(i)}) + \sum_{\alpha=1}^{D}\partial_\alpha[v_\alpha(x^{(i)})]\Big) \\
=& \sum_{\alpha=1}^{D}\sum_{i=1}^{N}\Big(\partial_\alpha[\log p(x^{(i)})] v_\alpha(x^{(i)}) + \partial_\alpha[v_\alpha(x^{(i)})]\Big).
\end{aligned} \tag{8}
$$

For every $v \in \mathcal{L}_p^2, \|v\| = 1$, we can define a function $\phi = (\phi_1, \cdots, \phi_D)^\top \in \mathcal{L}^2$ correspondingly, such that $\phi(x) = p(x)^{\frac{1}{2}} v(x)$, which means $\phi_\alpha(x) = p(x)^{\frac{1}{2}} v_\alpha(x)$, and

$$
\begin{aligned}
\|\phi\|_2^2 =& \int_{\mathbb{R}^D} \phi^2 \,\mathrm{d}x = \int_{\mathbb{R}^D} \sum_{\alpha=1}^{D} (\phi_\alpha(x))^2 \,\mathrm{d}x \\
=& \int_{\mathbb{R}^D} \sum_{\alpha=1}^{D} (v_\alpha(x))^2 p(x) \,\mathrm{d}x = \|v\|^2 = 1.
\end{aligned}
$$

Rewrite Eq. (8) in term of ϕ, we have:

$$
\begin{aligned}
\mathcal{L}_\phi =& \sum_{\alpha=1}^{D}\sum_{i=1}^{N}\Big(\partial_\alpha[\log p(x^{(i)})] v_\alpha(x^{(i)}) + \partial_\alpha[v_\alpha(x^{(i)})]\Big) \\
=& \sum_{\alpha=1}^{D}\sum_{i=1}^{N}\Big(\partial_\alpha[\log p(x^{(i)})] \phi_\alpha(x^{(i)}) p(x^{(i)})^{-\frac{1}{2}} \\
&+ \partial_\alpha[\phi_\alpha(x^{(i)}) p(x^{(i)})^{-\frac{1}{2}}]\Big)
\end{aligned} \tag{9}
$$

$$= \sum_{\alpha=1}^{D}\sum_{i=1}^{N}\Big(\frac{1}{2}p(x^{(i)})^{-\frac{3}{2}}\partial_\alpha[p(x^{(i)})]\phi_\alpha(x^{(i)}) + p(x^{(i)})^{-\frac{1}{2}}\partial_\alpha[\phi_\alpha(x^{(i)})]\Big)$$
$$= \sum_{\alpha=1}^{D}\sum_{i=1}^{N}\Big(A_\alpha^{(i)}\phi_\alpha(x^{(i)}) + B^{(i)}\partial_\alpha[\phi_\alpha(x^{(i)})]\Big),$$

where $A_\alpha^{(i)} = \frac{1}{2}p(x^{(i)})^{-\frac{3}{2}}\partial_\alpha[p(x^{(i)})]$ and $B^{(i)} = p(x^{(i)})^{-\frac{1}{2}} > 0$. We will now construct a sequence $\{\phi_n\}$ to show the following problem:

$$\inf_{\phi\in\mathcal{L}^2,\|\phi\|=1}\sum_{\alpha=1}^{D}\sum_{i=1}^{N}\Big(A_\alpha^{(i)}\phi_\alpha(x^{(i)}) + B^{(i)}\partial_\alpha[\phi_\alpha(x^{(i)})]\Big) \tag{10}$$

has no solution, then induce a sequence $\{v_n\}$ by $\{\phi_n\}$ for problem Eq. (7).

Define a sequence of functions

$$\chi_n(x) = \begin{cases} I_n^{-1/2}(1-x^2)^{n/2}, & \text{for } x\in[-1,1],\\ 0, & \text{otherwise.}\end{cases}$$

We have $\int_{\mathbb{R}}\chi_n(x)^2\,\mathrm{d}x = 1$ with $I_n = \int_{-1}^{1}(1-x^2)^n\,\mathrm{d}x = \sqrt{\pi}\frac{\Gamma(n+1)}{\Gamma(n+3/2)}$, where $\Gamma(\cdot)$ is the Gamma function. Note that when $x = -1/\sqrt{n}$,

$$\begin{aligned}\chi_n'(x) =& -nI_n^{-\frac{1}{2}}x(1-x^2)^{\frac{n-2}{2}}\\ =&\pi^{-\frac{1}{4}}\sqrt{\frac{\Gamma(n+\frac{3}{2})}{\Gamma(n+1)}}\sqrt{n}(1-\frac{1}{n})^{\frac{n-2}{2}} \qquad (x=-\frac{1}{\sqrt{n}})\\ >&\pi^{-\frac{1}{4}}\sqrt{n}(1-\frac{1}{n})^{\frac{n-2}{2}}, \qquad (\Gamma(n+\frac{3}{2})>\Gamma(n+1))\end{aligned}$$

therefore,

$$\lim_{n\to\infty}\chi_n'(-\frac{1}{\sqrt{n}})$$
$$> \lim_{n\to\infty}\pi^{-\frac{1}{4}}\sqrt{n}(1-\frac{1}{n})^{\frac{n-2}{2}} = \pi^{-\frac{1}{4}}e^{-\frac{1}{2}}\lim_{n\to\infty}\sqrt{n} = +\infty.$$

Denote $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_D^{(i)})^\top \in \mathbb{R}^D, i = 1,\cdots,N$ and

$$r_1 = \frac{1}{3}\min_{i\neq j}\left\|x^{(i)}-x^{(j)}\right\|_\infty = \frac{1}{3}\min_{\alpha\in\{1,\cdots,D\},i\neq j}\left|x_\alpha^{(i)}-x_\alpha^{(j)}\right|.$$

We extend χ_n to $\mathbb{R}^D$ as ξ_n with support $\mathrm{supp}(\xi_n) = [-r,r]^D$,

$$\xi_n(x_1,x_2,\cdots,x_D) = r^{-D/2}\prod_{\alpha=1}^{D}\chi_n(\frac{x_i}{r}),$$

where $r = \min\{r_0, r_1\}$. It is easy to show that $\int_{\mathbb{R}^D}\xi_n(x)^2\,\mathrm{d}x = 1$, and

$$\lim_{n\to\infty}\partial_\alpha\xi_n(-\epsilon_n) = +\infty,\ \alpha = 1,2,\cdots,D,$$

with $\epsilon_n = \frac{r}{\sqrt{n}}(1,1,\cdots,1)^\top$.

We choose $\phi_\alpha(x) = \frac{1}{ND}\sum_{i=1}^{N}\psi_\alpha^{(i)}$, where $\psi_\alpha^{(i)}$ is defined by:

$$\psi_\alpha^{(i)}(x) = \begin{cases}\xi_n(x-x^{(i)}-\epsilon_n), & \text{if } A_\alpha^{(i)} >= 0,\\ -\xi_n(x-x^{(i)}+\epsilon_n), & \text{if } A_\alpha^{(i)} < 0.\end{cases}$$

With $\int_{\mathbb{R}^D}\psi_\alpha^{(i)}(x)\psi_\alpha^{(j)}(x)\,\mathrm{d}x = 0, \forall i\neq j$, we know ϕ_n satisfies conditions in Eq. (10). Noting that $\forall i,j$, $A_\alpha^{(i)}\psi_\alpha^{(j)}(x^{(i)}) \geq 0$, and

$$\partial_\alpha\psi_\alpha^{(j)}(x^{(i)}) = \begin{cases}+\infty, & \text{when } n\to\infty, \text{if } i=j,\\ 0, & \text{if } i\neq j,\end{cases}$$

we see $\mathcal{L}_{\phi_n}\to+\infty$ in Eq. (9) when $n\to\infty$.

Since $\mathrm{supp}(\phi_n)\subset\mathrm{supp}(p)$, we can induce a sequence of $\{v_n\}$ from $\{\phi_n\}$ as $v_n = \phi_n/\sqrt{p(x)}$, which satisfies restrictions in Eq. (7) and the objective $\mathcal{L}_{v_n}$ will go to infinity when $n\to\infty$. Note that any element in $\mathcal{L}_p^2$, as a function, cannot take infinite value. So the infinite supremum of the objective in Eq. (7) cannot be obtained by any element in $\mathcal{L}_p^2$, thus no optimal solution for the optimization problem.

A.4: Deduction of Proposition 5

We first derive the exact inverse exponential map on the Wasserstein space $\mathcal{P}_2(\mathcal{X})$, then develop finite-particle estimation for it. Given $q, r\in\mathcal{P}_2(\mathcal{X})$, $\mathrm{Exp}_q^{-1}(r)$ is defined as the tangent vector at q of the geodesic curve $q_{t\in[0,1]}$ from q to r. When q is absolutely continuous, the optimal transport map $\mathcal{T}_q^r$ from q to r exists (Villani (2008), Thm. 10.38). This condition fits the case of ParVIs since as our theory indicates, ParVIs have to do a smoothing treatment in any way, which is equivalent to assume an absolutely continuous distribution q. Under this case, the geodesic is given by $q_t = \big((1-t)\,\mathrm{id}+t\mathcal{T}_q^r\big)_\# q$ (Ambrosio et al. (2008), Thm. 7.2.2), and its tangent vector at q (*i.e.*, $t=0$) can be characterized by $\mathrm{Exp}_q^{-1}(r) = \lim_{t\to0}\frac{1}{t}(\mathcal{T}_q^{q_t}-\mathrm{id})$ (Ambrosio et al. (2008), Prop. 8.4.6). Due to the uniqueness of optimal transport map, we have $\mathcal{T}_q^{q_t} = (1-t)\,\mathrm{id}+t\mathcal{T}_q^r$, so we finally get $\mathrm{Exp}_q^{-1}(r) = \mathcal{T}_q^r - \mathrm{id}$.

To estimate it with a finite set of particles, we approximate the optimal transport map with the discrete one from particles $\{x^{(i)}\}_{i=1}^N$ of q to particles $\{y^{(i)}\}_{i=1}^N$ of r. As mentioned in the main context, it is a costly task, and the Sinkhorn approximations (for both the original version (Cuturi, 2013) and an improved version (Xie et al., 2018)) suffers from an

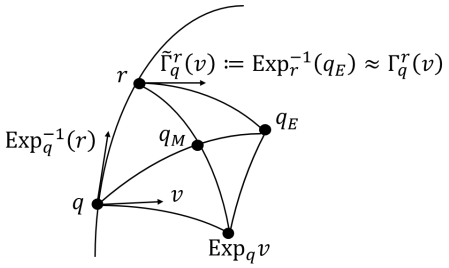


Figure 7. Illustration of the Schild's ladder method. Figure inspired by (Kheyfets et al., 2000).

unstable behavior in our experiments. We now utilize the pairwise-close condition and develop a light and stable approximation. The pairwise-close condition $d(x^{(i)}, y^{(i)}) \ll \min\{\min_{j\neq i} d(x^{(i)}, x^{(j)}), \min_{j\neq i} d(y^{(i)}, y^{(j)})\}$ indicates that $\frac{d(x^{(i)},x^{(j)})}{d(x^{(j)},y^{(j)})} \gg 1$, for any $i \neq j$. On the other hand, due to triangle inequality, we have $d(x^{(i)}, y^{(j)}) \geq \left|d(x^{(i)}, x^{(j)}) - d(x^{(j)}, y^{(j)})\right|$, or equivalently $\frac{d(x^{(i)},y^{(j)})}{d(x^{(j)},y^{(j)})} \geq \left|\frac{d(x^{(i)},x^{(j)})}{d(x^{(j)},y^{(j)})} - 1\right|$. Due to the above knowledge $\frac{d(x^{(i)},x^{(j)})}{d(x^{(j)},y^{(j)})} \gg 1$ by the pairwise-close condition, we have $\frac{d(x^{(i)},y^{(j)})}{d(x^{(j)},y^{(j)})} \gg 1$, or equivalently (by switching i and j) $d(x^{(i)}, y^{(i)}) \ll \min_{j\neq i} d(x^{(i)}, y^{(j)})$. This means that when transporting $\{x^{(i)}\}_i$ to $\{y^{(i)}\}_i$, the map $x^{(i)} \mapsto y^{(i)}$ for any i, has presumably the least cost. More formally, consider any amount of transportation from $x^{(i)}$ to $y^{(j)}$ other than $y^{(i)}$. It will introduce a change in the transportation cost that is proportional to $d(x^{(i)}, y^{(j)}) - d(x^{(i)}, y^{(i)}) + d(x^{(j)}, y^{(i)}) - d(x^{(j)}, y^{(j)})$, which is always positive due to our above recognition. Thus we can reasonably approximate the optimal transport map $\mathcal{T}_q^r$ by the discrete one $\mathcal{T}_q^r(x^{(i)}) \approx y^{(i)}$. With this approximation, we have $\left(\mathrm{Exp}_q^{-1}(r)\right)(x^{(i)}) = \mathcal{T}_q^r(x^{(i)}) - x^{(i)} \approx y^{(i)} - x^{(i)}$.

A.5: Deduction of Proposition 6

We derive the finite-particle estimation of the parallel transport on the Wasserstein space $\mathcal{P}_2(\mathcal{X})$. We follow the Schild's ladder method (Ehlers et al., 1972; Kheyfets et al., 2000) to parallel transport a tangent vector at q, $v \in T_q\mathcal{P}_2(\mathcal{X})$, to the tangent space at r, $T_r\mathcal{P}_2(\mathcal{X})$. As shown in Fig. 7, given q, r and $v \in T_q\mathcal{P}_2(\mathcal{X})$, the procedure to approximate $\Gamma_q^r(v)$ is

1. find the point $\mathrm{Exp}_q(v)$;
2. find the midpoint of the geodesic from r to $\mathrm{Exp}_q(v)$: $q_M := \mathrm{Exp}_q(\frac{1}{2}\mathrm{Exp}_q^{-1}(\mathrm{Exp}_q(v)))$;
3. extrapolate the geodesic from q to q_M by doubling the length to find $q_E := \mathrm{Exp}_q(2\,\mathrm{Exp}_q^{-1}(q_M))$;
4. take the approximator $\tilde{\Gamma}_q^r(v) := \mathrm{Exp}_r^{-1}(q_E) \approx \Gamma_q^r(v)$.

Note that the Schild's ladder method only requires the exponential map and its inverse. It provides a tractable first-order approximation $\tilde{\Gamma}_q^r$ of the parallel transport Γ_q^r under Levi-Civita connection, as needed.

Assume q and r are close in the sense of the Wasserstein distance, so that the Schild's ladder finds a good first-order approximation. In the following we consider transporting εv for small $\varepsilon > 0$ for the sake of the pairwise-close condition, and the result can be recovered by noting the linearity of the parallel transport: $\Gamma_q^r(\varepsilon v) = \varepsilon\Gamma_q^r(v)$. Let $\{x^{(i)}\}_{i=1}^N$ and $\{y^{(i)}\}_{i=1}^N$ be the sets of samples of q and r, respectively, and assume that they are pairwise close.

Now we follow the procedure.

1. The measure $\mathrm{Exp}_q(\varepsilon v)$ can be identified as $(\mathrm{id} + \varepsilon v)_{\#}q$ due to the knowledge on the exponential map on $\mathcal{P}_2(\mathcal{X})$ explained in Section 4, thus $\{x^{(i)} + \varepsilon v(x^{(i)})\}_{i=1}^N$ is a set of samples of $\mathrm{Exp}_q(\varepsilon v)$, and still pairwise close to $\{y^{(i)}\}_i$.

2. The optimal map $\mathcal{T}$ from r to $\mathrm{Exp}_q(\varepsilon v)$ can be approximated by $\mathcal{T}(y^{(i)}) = x^{(i)} + \varepsilon v(x^{(i)})$ since the two sets of samples are pairwise close. According to Theorem 7.2.2 of (Ambrosio et al., 2008), the geodesic from r to $\mathrm{Exp}_q(\varepsilon v)$ is $t \mapsto \left((1-t)\,\mathrm{id} + t\mathcal{T}\right)_{\#}r$. Thus a set of samples of q_M, *i.e.*, the midpoint of the geodesic, can be derived as $\left\{\frac{1}{2}\left(y^{(i)} + x^{(i)} + \varepsilon v(x^{(i)})\right)\right\}_i$.

3. Similarly, a set of samples of q_E is found as $\left\{(1-t)x^{(i)} + \frac{1}{2}t\left(y^{(i)} + x^{(i)} + \varepsilon v(x^{(i)})\right)\right\}_i\Big|_{t=2} = \left\{y^{(i)} + \varepsilon v(x^{(i)})\right\}_i$ and is pairwise close to $\{y^{(i)}\}_i$.

4. The approximated transported tangent vector $\mathrm{Exp}_r^{-1}(q_E)$ satisfies $\left(\mathrm{Exp}_r^{-1}(q_E)\right)(y^{(i)}) = \varepsilon v(x^{(i)})$.

Finally, we get the approximation $\left(\Gamma_q^r(v)\right)(y^{(i)}) \approx \left(\tilde{\Gamma}_q^r(v)\right)(y^{(i)}) = v(x^{(i)})$.

B. Derivations of GFSF Vector Field $\hat{u}^{\mathrm{GFSF}}$

B.1: Derivation with Vector-Valued Functions

The vector field $\hat{u}^{\mathrm{GFSF}}$ is identified by the optimization problem (6):

$$\min_{u\in\mathcal{L}^2} \max_{\substack{\phi\in\mathcal{H}^D,\\ \|\phi\|_{\mathcal{H}^D}=1}} \left(\sum_{i=1}^N \left(\phi(x^{(i)})\cdot u^{(i)} - \nabla\cdot\phi(x^{(i)})\right)\right)^2,$$

where $u^{(i)} := u(x^{(i)})$. For ϕ in $\mathcal{H}^D$, by using the reproducing property $\langle\phi_\alpha(\cdot), K(x,\cdot)\rangle_{\mathcal{H}} = \phi_\alpha(x)$ and $\langle\phi_\alpha(\cdot), \partial_{x_\beta}K(x,\cdot)\rangle_{\mathcal{H}} = \partial_{x_\beta}\phi_\alpha(x)$ (Zhou, 2008), we can

write the objective function as:

$$\Big(\sum_\alpha \sum_j \big(u_\alpha^{(j)}\phi_\alpha(x^{(j)}) - \partial_{x_\alpha^{(j)}}\phi_\alpha(x^{(j)})\big)\Big)^2$$
$$= \left(\sum_\alpha \left\langle \sum_j \big(u_\alpha^{(j)} K(x^{(j)},\cdot) - \partial_{x_\alpha^{(j)}} K(x^{(j)},\cdot)\big), \phi_\alpha(\cdot)\right\rangle_{\mathcal{H}}\right)^2$$
$$= \left\langle \sum_j \big(u^{(j)} K(x^{(j)},\cdot) - \nabla_{x^{(j)}} K(x^{(j)},\cdot)\big), \phi(\cdot)\right\rangle_{\mathcal{H}^D}^2.$$

We denote $\zeta := \sum_j \big(u^{(j)} K(x^{(j)},\cdot) - \nabla_{x^{(j)}} K(x^{(j)},\cdot)\big) \in \mathcal{H}^D$. Then the optimal value of the objective after maximizing out ϕ is $\|\zeta\|^2_{\mathcal{H}^D} = \sum_{i,j}\big(u^{(i)}u^{(j)}K(x^{(i)},x^{(j)}) - 2u^{(i)}\nabla_{x^{(j)}}K(x^{(j)},x^{(i)}) + \nabla_{x^{(i)}}\nabla_{x^{(j)}}K(x^{(i)},x^{(j)})\big) = \operatorname{tr}(\hat{u}\hat{K}\hat{u}^\top) - 2\operatorname{tr}(\hat{K}'\hat{u}^\top) + \text{const}$, where $\hat{u}_{:,i} := u^{(i)}$, and $\hat{K}$, $\hat{K}'$ are defined in the main text. To minimize this quadratic function with respect to $\hat{u}$, we further differentiate it with respect to $\hat{u}$ and solve for the stationary point. This finally gives the result $\hat{u}^{\text{GFSF}} = \hat{K}'\hat{K}^{-1}$.

B.2: Derivation with Scalar-Valued Functions

We denote φ as scalar-valued functions on $\mathcal{X}$. For the equality $u(x) = -\nabla \log q(x)$, or $u(x)q(x) + \nabla q(x) = 0$, to hold in the weak sense with scalar-valued test function, we mean:

$$\mathbb{E}_{q(x)}[\varphi(x)u(x) - \nabla\varphi(x)] = 0, \forall \varphi \in C_c^\infty.$$

Let $\{x^{(j)}\}_j$ be a set of samples of $q(x)$. Then the above requirement on $u(x)$ is:

$$\sum_j \Big(\varphi(x^{(j)})u^{(j)} - \nabla\varphi(x^{(j)})\Big) = 0, \forall \varphi \in C_c^\infty, \quad (11)$$

where $u^{(j)} = u(x^{(j)})$. As analyzed above, for a valid vector field, we have to smooth the function φ.

For the above considerations, we restrict φ in Eq. (11) to be in the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ of some kernel $K(\cdot,\cdot)$, and convert the equation as the following optimization problem:

$$\min_{\hat{u}\in\mathbb{R}^{D\times N}} \max_{\substack{\varphi\in\mathcal{H},\\ \|\varphi\|_{\mathcal{H}}=1}} J(\hat{u},\varphi),$$
$$J(\hat{u},\varphi) := \sum_{j,\alpha}\Big(\varphi(x^{(j)})\hat{u}_{\alpha j} - \partial_{x_\alpha^{(j)}}\varphi(x^{(j)})\Big)^2,$$

where $\hat{u}_{\alpha j} := u_\alpha(x^{(j)})$. By using the reproducing properties of RKHS, we can write $J(\hat{u},\varphi)$ as:

$$J(\hat{u},\varphi) = \sum_\alpha \langle \varphi(\cdot), \zeta_\alpha(\cdot)\rangle^2_{\mathcal{H}},$$
$$\zeta_\alpha(\cdot) := \sum_j \Big(\hat{u}_{\alpha j}K(x^{(j)},\cdot) - \partial_{x_\alpha^{(j)}}K(x^{(j)},\cdot)\Big).$$

By linear algebra operations, we have:

$$\max_{\varphi\in\mathcal{H},\|\varphi\|_{\mathcal{H}}=1} J(\hat{u},\varphi) = \lambda_1(A(\hat{u})),$$

where $\lambda_1(A(\hat{u}))$ is the largest eigenvalue of matrix A, and $A(\hat{u})_{\alpha\beta} = \langle \zeta_\alpha(\cdot), \zeta_\beta(\cdot)\rangle_{\mathcal{H}}$, or:

$$A(\hat{u}) = \hat{u}\hat{K}\hat{u}^\top - (\hat{K}'\hat{u}^\top + \hat{u}\hat{K}'^\top) + \hat{K}'',$$

with $\hat{K}''_{\alpha\beta} := \sum_{i,j}\partial_{x_\alpha^{(i)}}\partial_{x_\beta^{(j)}}K(x^{(i)},x^{(j)})$. For distinct samples $\hat{K}$ is positive-definite, so we can conduct Cholesky decomposition: $\hat{K} = GG^\top$ with G non-singular. Note that $A(\hat{u}) = (\hat{u}G - \hat{K}'G^{-1\top})(\hat{u}G - \hat{K}'G^{-1\top})^\top + (\hat{K}'' - \hat{K}'\hat{K}^{-1}\hat{K}'^\top)$. So whenever $\hat{u}G \neq \hat{K}'G^{-1\top}$, the first term will be positive semidefinite with positive largest eigenvalue, which makes $\lambda_1(A(u)) > \lambda_1(\hat{K}'' - \hat{K}'\hat{K}^{-1}\hat{K}'^\top)$, a constant with respect to $\hat{u}$. So to minimize $\lambda_1(A(\hat{u}))$, we require $\hat{u}G = \hat{K}'G^{-1\top}$, *i.e.*, $\hat{u} = \hat{K}'(GG^\top)^{-1} = \hat{K}'\hat{K}^{-1}$. This result coincides with the one for vector-valued functions $\phi \in \mathcal{H}^D$.

In practice, for numerical stability, we add a small diagonal matrix to $\hat{K}$ before conducting inversion. This is a common practice. Particularly, it is adopted in Li & Turner (2017) for the same estimate.

C. Details on Accelerated First-Order Methods on the Wasserstein Space $\mathcal{P}_2(\mathcal{X})$

C.1: Simplification of Riemannian Accelerated Gradient (RAG) with Approximations

We consider the general version of RAG (Alg. 2 of Liu et al. (2017b)). It updates the target variable q_k as:

$$q_k = \operatorname{Exp}_{r_{k-1}}(\varepsilon v_{k-1}),$$

where $v_{k-1} := -\operatorname{grad}\mathrm{KL}(r_{k-1})$. The update rule for the auxiliary variable r_k is given by the solution of the following non-linear equation (see Alg. 2 and Eq. (5) of Liu et al. (2017b)):

$$\Gamma_{r_k}^{r_{k-1}}\left(\frac{k}{\alpha-1}\operatorname{Exp}_{r_k}^{-1}(q_k) + \frac{Dv_k}{\|v_k\|_{r_k}}\right)$$
$$= \frac{k-1}{\alpha-1}\operatorname{Exp}_{r_{k-1}}^{-1}(q_{k-1}) - \frac{k+\alpha-2}{\alpha-1}\varepsilon v_{k-1} + \frac{Dv_{k-1}}{\|v_{k-1}\|_{r_{k-1}}}.$$

Here we focus on simplifying this complicated update rule for r_k with moderate approximations. We note that the original work of RAG (Liu et al., 2017b) actually adopted these approximations in experiments, but the simplification of the general algorithm is not given in the work.

Applying $\left(\Gamma_{r_k}^{r_{k-1}}\right)^{-1}$ to both sides of the equation and noticing that $\left(\Gamma_{r_k}^{r_{k-1}}\right)^{-1} = \Gamma_{r_{k-1}}^{r_k}$, the above equation can be

reformulated as:

$$\begin{aligned}&\frac{k}{\alpha-1}\operatorname{Exp}_{r_k}^{-1}(q_k)+\frac{Dv_k}{\|v_k\|_{r_k}}\\=&\Gamma_{r_{k-1}}^{r_k}\left(\frac{k-1}{\alpha-1}\operatorname{Exp}_{r_{k-1}}^{-1}(q_{k-1})-\frac{k+\alpha-2}{\alpha-1}\varepsilon v_{k-1}\right)\\&+\frac{D\Gamma_{r_{k-1}}^{r_k}(v_{k-1})}{\|v_{k-1}\|_{r_{k-1}}}.\end{aligned}$$

Approximating v_k on the left hand side of the equation by $\Gamma_{r_{k-1}}^{r_k}(v_{k-1})$ and noting that $\left\|\Gamma_{r_{k-1}}^{r_k}(v_{k-1})\right\|_{r_k}=\|v_{k-1}\|_{r_{k-1}}$, we have:

$$\begin{aligned}&\frac{k}{\alpha-1}\operatorname{Exp}_{r_k}^{-1}(q_k)\\=&\Gamma_{r_{k-1}}^{r_k}\left(\frac{k-1}{\alpha-1}\operatorname{Exp}_{r_{k-1}}^{-1}(q_{k-1})-\frac{k+\alpha-2}{\alpha-1}\varepsilon v_{k-1}\right).\end{aligned}$$

Using the fact that $\operatorname{Exp}_{r_k}^{-1}(q_k)=-\Gamma_{q_k}^{r_k}(\operatorname{Exp}_{q_k}^{-1}(r_k))$ and applying $\left(\Gamma_{q_k}^{r_k}\right)^{-1}=\Gamma_{r_k}^{q_k}$ to both sides of the equation, we have:

$$\begin{aligned}&\operatorname{Exp}_{q_k}^{-1}(r_k)\\=&-\Gamma_{r_k}^{q_k}\Gamma_{r_{k-1}}^{r_k}\left(\frac{k-1}{k}\operatorname{Exp}_{r_{k-1}}^{-1}(q_{k-1})-\frac{k+\alpha-2}{k}\varepsilon v_{k-1}\right).\end{aligned}$$

Approximating $\Gamma_{r_k}^{q_k}\Gamma_{r_{k-1}}^{r_k}$ by $\Gamma_{r_{k-1}}^{q_k}$, we finally have $r_k=$

$$\operatorname{Exp}_{q_k}\left[-\Gamma_{r_{k-1}}^{q_k}\left(\frac{k-1}{k}\operatorname{Exp}_{r_{k-1}}^{-1}(q_{k-1})-\frac{k+\alpha-2}{k}\varepsilon v_{k-1}\right)\right].$$

C.2: Reformulation of Riemannian Nesterov's Method (RNes)

We consider the constant step version of RNes (Alg. 2 of (Zhang & Sra, 2018)). It introduces an additional auxiliary variable $s_k\in\mathcal{P}_2(\mathcal{X})$, and update the variables in iteration k as:

$$r_{k-1}=\operatorname{Exp}_{q_{k-1}}\left(c_1\operatorname{Exp}_{q_{k-1}}^{-1}(s_{k-1})\right), \tag{12a}$$

$$q_k=\operatorname{Exp}_{r_{k-1}}(\varepsilon v_{k-1}), \tag{12b}$$

$$s_k=\operatorname{Exp}_{r_{k-1}}\left(\frac{1-\alpha}{1+\beta}\operatorname{Exp}_{r_{k-1}}^{-1}(s_{k-1})+\frac{\alpha}{(1+\beta)\gamma}v_{k-1}\right), \tag{12c}$$

where $v_{k-1}:=-\operatorname{grad}\mathrm{KL}(r_{k-1})$, and the coefficients α,γ,c_1 are set by a step size $\varepsilon>0$, a shrinkage parameter $\beta>0$, and a parameter $\mu>0$ upper bounding the Lipschitz coefficient of the gradient of the objective, in the following way:

$$\begin{aligned}\alpha&=\frac{\sqrt{\beta^2+4(1+\beta)\mu\varepsilon}-\beta}{2},\\\gamma&=\frac{\sqrt{\beta^2+4(1+\beta)\mu\varepsilon}-\beta}{\sqrt{\beta^2+4(1+\beta)\mu\varepsilon}+\beta}\mu,\\c_1&=\frac{\alpha\gamma}{\gamma+\alpha\mu}.\end{aligned} \tag{13}$$

Now we simplify the update rule by collapsing the variable s. Referring to Eq. (12a), the variable s_{k-1} can be expressed by:

$$s_{k-1}=\operatorname{Exp}_{q_{k-1}}\left(\frac{1}{c_1}\operatorname{Exp}_{q_{k-1}}^{-1}(r_{k-1})\right).$$

This result indicates that s_{k-1} lies on the $1/c_1$ portion of the geodesic from q_{k-1} to r_{k-1}, which is the $(1-1/c_1)$ portion of the geodesic from r_{k-1} to q_{k-1}. According to this knowledge, we have:

$$\operatorname{Exp}_{r_{k-1}}^{-1}(s_{k-1})=\left(1-\frac{1}{c_1}\right)\operatorname{Exp}_{r_{k-1}}^{-1}(q_{k-1}).$$

Substitute this result into Eq. (12c), we have:

$$\begin{aligned}s_k=\operatorname{Exp}_{r_{k-1}}\Bigg(&\frac{1-\alpha}{1+\beta}\left(1-\frac{1}{c_1}\right)\operatorname{Exp}_{r_{k-1}}^{-1}(q_{k-1})\\&+\frac{\alpha}{(1+\beta)\gamma\varepsilon}\operatorname{Exp}_{r_{k-1}}^{-1}(q_k)\Bigg),\end{aligned}$$

where we have also substituted v_{k-1} with $\frac{1}{\varepsilon}\operatorname{Exp}_{r_{k-1}}^{-1}(q_k)$ according to Eq. (12b). Leveraging Eq. (13) to simplify the coefficients in the above equation, we get:

$$s_k=\operatorname{Exp}_{r_{k-1}}\left((1-c_2)\operatorname{Exp}_{r_{k-1}}^{-1}(q_{k-1})+c_2\operatorname{Exp}_{r_{k-1}}^{-1}(q_k)\right),$$

where the coefficient $c_2:=1/\alpha$. Replacing $k\to k+1$ in Eq. (12a) and substitute with the above result, we have the update rule for r_k:

$$\begin{aligned}r_k=\operatorname{Exp}_{q_k}\Big\{c_1\operatorname{Exp}_{q_k}^{-1}\Big[\operatorname{Exp}_{r_{k-1}}\Big(&(1-c_2)\operatorname{Exp}_{r_{k-1}}^{-1}(q_{k-1})\\&+c_2\operatorname{Exp}_{r_{k-1}}^{-1}(q_k)\Big)\Big]\Big\},\end{aligned}$$

which builds the update rule of RNes together with Eq. (12a).

In our implementation, the parameters are tackled with ε,β,μ instead of setting c_1,c_2 directly. The shrinkage parameter β is set in the scale of $\sqrt{\mu\varepsilon}$. In our Alg. 1, the coefficient $c_1(c_2-1)$ can be expressed as:

$$1+\beta-\frac{2(1+\beta)(2+\beta)\mu\varepsilon}{\sqrt{\beta^2+4(1+\beta)\mu\varepsilon}-\beta+2(1+\beta)\mu\varepsilon}.$$

C.3: Deduction of Wasserstein Accelerated Gradient (WAG) and Wasserstein Nesterov's Method (WNes) (Alg. 1)

First consider developing WAG based on RAG. We denote the vector field $\zeta_{k-1} := \frac{k-1}{k}\operatorname{Exp}^{-1}_{r_{k-1}}(q_{k-1}) - \frac{k+\alpha-2}{k}\varepsilon v_{k-1}$ for simplicity, so $r_k = \operatorname{Exp}_{q_k}\left[-\Gamma^{q_k}_{r_{k-1}}(\zeta_{k-1})\right]$, due to the update rule of RAG. We assume that $\{x^{(i)}_{k-1}\}_{i=1}^N$ of q_{k-1} and $\{y^{(i)}_{k-1}\}_{i=1}^N$ of r_{k-1} are pairwise close, so from Section 4 we know that $\operatorname{Exp}^{-1}_{r_{k-1}}(q_{k-1})(y^{(i)}_{k-1}) = x^{(i)}_{k-1} - y^{(i)}_{k-1}$, thus $\zeta_{k-1}(y^{(i)}_{k-1}) = \frac{k-1}{k}(x^{(i)}_{k-1} - y^{(i)}_{k-1}) - \frac{k+\alpha-2}{k}\varepsilon v^{(i)}_{k-1}$. Due to the update rule for q_k that we already discovered: $x^{(i)}_k = y^{(i)}_{k-1} + \varepsilon v^{(i)}_{k-1}$, we know that $\{x^{(i)}_k\}_{i=1}^N$ of q_k and $\{y^{(i)}_{k-1}\}_{i=1}^N$ of r_{k-1} are pairwise close, for small enough step size ε. Using the parallel transport estimate developed above with Schild's ladder method, $\big(\Gamma^{q_k}_{r_{k-1}}(\zeta_{k-1})\big)(x^{(i)}_k) \approx \zeta_{k-1}(y^{(i)}_{k-1})$. So finally, we assign $y^{(i)}_k = x^{(i)}_k - \big(\Gamma^{q_k}_{r_{k-1}}(\zeta_{k-1})\big)(x^{(i)}_k) \approx x^{(i)}_k - \zeta_{k-1}(y^{(i)}_{k-1}) = x^{(i)}_k - \frac{k-1}{k}(x^{(i)}_{k-1} - y^{(i)}_{k-1}) + \frac{k+\alpha-2}{k}\varepsilon v^{(i)}_{k-1}$ as a sample of r_k.

We note that initially $x^{(i)}_0 = y^{(i)}_0$. Assume $\{x^{(i)}_{k-1}\}_{i=1}^N$ and $\{y^{(i)}_{k-1}\}_{i=1}^N$ are pairwise close, so for sufficiently small ε, $\zeta_{k-1}(y^{(i)}_{k-1})$ is an infinitesimal vector for all i. This, in turn, indicates that $\{x^{(i)}_k\}_{i=1}^N$ of q_k and $\{y^{(i)}_k\}_{i=1}^N$ of r_k are pairwise close, which provides the assumption for the next iteration. The derivation of WNes based on RNes can be developed similarly, and we omit verbosing the procedure.

D. Details on the HE Method for Bandwidth Selection

We first note that the bandwidth selection problem cannot be solved using theories of heat kernels, which aims to find the evolving density under the Brownian motion with known initial distribution, while in our case the density is unknown and we want to find an update on samples to approximate the effect of Brownian motion.

According to the derivation in the main context, we write the dimensionless final objective explicitly:

$$\begin{aligned}&\frac{1}{h^{D+2}}\sum_k \lambda(x^{(k)})^2\\ =&\frac{1}{h^{D+2}}\sum_k \Big[\Delta\tilde{q}(x^{(k)};\{x^{(i)}\}_i)\\ &+\sum_j \nabla_{x^{(j)}}\tilde{q}(x^{(k)};\{x^{(i)}\}_i)\cdot\nabla\log\tilde{q}(x^{(j)};\{x^{(i)}\}_i)\Big]^2.\end{aligned}$$

For $\tilde{q}(x;\{x^{(j)}\}_j) = (1/Z)\sum_j c(\|x-x^{(j)}\|^2/(2h))$, the above objective becomes:

$$\sum_k\Big(\sum_j\Big[c_j''(x)\|x-x^{(j)}\|^2 + Dhc_j'(x) + \frac{\big(\sum_i c_{ij}'x^{(i)}\big) - \big(\sum_i c_{ij}'\big)x^{(j)}}{\big(\sum_i c_{ij}\big)}\cdot(x-x^{(j)})c_j'(x)\Big]\Big)^2,$$

where $c_j'(x) = c'(\|x-x^{(j)}\|^2/(2h))$, $c_{ij}' = c_j'(x^{(i)})$, $c_{ij} = c(\|x^{(i)}-x^{(j)}\|^2/(2h))$. For Gaussian kernel $c(r) = (2\pi h)^{-\frac{D}{2}}e^{-r}$, denoting $g_k^2(h)$ as the summand for k of the l.h.s. of the above equation, we have:

$$\begin{aligned}&(2\pi)^{\frac{D}{2}}g_k(h)\\ =\;&\big(\sum_j e_{kj}\|d_{kj}\|^2\big) - hD\big(\sum_j e_{kj}\big)\\ &-\sum_j\big(\sum_i e_{ij}\big)^{-1}e_{jk}d_{jk}\cdot\big(\sum_i e_{ij}d_{ij}\big),\\ &(2\pi)^{\frac{D}{2}}g_k'(h)\\ =\;&\frac{1}{2h^2}\big(\sum_j e_{jk}\|d_{jk}\|^4\big) - \frac{D}{h}\big(\sum_j e_{jk}\|d_{jk}\|^2\big)\\ &+\big(\frac{D^2}{2}-D\big)\big(\sum_j e_{jk}\big)\\ &-\frac{1}{2h^2}\sum_j\big(\sum_i e_{ij}\big)^{-1}e_{jk}d_{jk}\cdot\big(\sum_i e_{ij}\|d_{ij}\|^2d_{ij}\big)\\ &-\frac{1}{2h^2}\sum_j\big(\sum_i e_{ij}\big)^{-1}e_{jk}\|d_{jk}\|^2d_{jk}\cdot\big(\sum_i e_{ij}d_{ij}\big)\\ &+\frac{1}{2h^2}\sum_j\big(\sum_i e_{ij}\big)^{-2}\big(\sum_i e_{ij}\|d_{ij}\|^2\big)e_{jk}d_{jk}\cdot\big(\sum_i e_{ij}d_{ij}\big)\\ &+\frac{D}{2h}\sum_j\big(\sum_i e_{ij}\big)^{-1}e_{jk}d_{jk}\cdot\big(\sum_i e_{ij}d_{ij}\big),\end{aligned}$$

where $d_{ij} = x^{(i)} - x^{(j)}$, $e_{ij} = e^{-\|d_{ij}\|^2/(2h)-(D/2)\log h}$.

Although the evaluation of $g_k(h)$ may induce some computation cost, the optimization is with respect to a scalar. In each particle update iteration, before estimating the vector field v, we first update the previous bandwidth by one-step exploration with quadratic interpolation, which only requires one derivative evaluation and two value evaluations.

E. Detailed Settings and Parameters of Experiments

E.1: Detailed Settings and Parameters of the Synthetic Experiment

The bimodal toy target distribution is inspired by the one of (Rezende & Mohamed, 2015). The logarithm of the target

density $p(z)$ for $z = (z_1, z_2) \in \mathbb{R}^2$ is given by:

$$\log p(z) = -2(\|z\|_2^2 - 3)^2 + \log(e^{-2(z_1-3)^2} + e^{-2(z_1+3)^2}) + \text{const.}$$

The region shown in each figure is $[-4, 4] \times [-4, 4]$. The number of particles is 200, and all particles are initialized with standard Gaussian $\mathcal{N}(0, 1)$.

All methods are run for 400 iterations, and all follows the plain WGD method. SVGD uses fixed step size 0.3, while other methods (Blob, GFSD, GFSF) share the fixed step size 0.01. This is because that the updating direction of SVGD is a kernel smoothed one, so it may have a different scale from other methods. Note that the AdaGrad with momentum method in the original SVGD paper (Liu & Wang, 2016) is not used. For GFSF, a small diagonal matrix $0.01I$ is added to $\hat{K}$ before conducting inversion, as discussed at the end of Appendix B.2.

E.2: Detailed Settings and Parameters of the BLR Experiment

We adopt the same settings as (Liu & Wang, 2016), which is also adopted by (Chen et al., 2018a). The Covertype dataset contains 581,012 items with each 54 features. Each run uses a random 80%-20% split of the dataset.

For the model, parameters of the Gamma prior on the precision of the Gaussian prior of the weight are $a_0 = 1.0$, $b_0 = 100$ (b_0 is the scale parameter, not the rate parameter). All methods use 100 particles, randomly initialized by the prior. Batch size for all methods is 50.

Detailed parameters of various methods are provided in Table 2. The WGD column provides the step size. The format of the PO column is "PO parameters(decaying exponent, remember rate, injected noise variance), step size". Both methods use a fixed step size, while the WAG and WNes methods use a decaying step size. The format of WAG column is "WAG parameter α, (step size decaying exponent, step size)" (see Alg. 1), and the format of WNes column is "Wnes parameters (μ, β), (step size decaying exponent, step size)" (see Appendix C.2). One exception is that SVGD-WGD uses the AdaGrad with momentum method to reproduce the results of (Liu & Wang, 2016), which uses remember rate 0.9 and step size 0.03. For GFSF, the small diagonal matrix is (1e-5)I.

E.3: Detailed Settings and Parameters of the BNN Experiment

We follow the same settings as Liu & Wang (2016). For each run, a random 90%-10% train-test split is conducted. The BNN model contains one hidden layer with 50 hidden nodes, and sigmoid activation is used. The parameters of the Gamma prior on the precision parameter of the Gaussian prior of the weights are $a_0 = 1.0$, $b_0 = 0.1$. Batch size is set to 100. Number of particles is fixed as 20 for all methods. Results are collected after 8,000 iterations for every method.

Table 2. Parameters of various methods in the BLR experiment

	WGD	PO
SVGD	3e-2	(1.0, 0.7, 1e-7), 3e-6
Blob	1e-6	(1.0, 0.7, 1e-7), 3e-7
GFSD	1e-6	(1.0, 0.7, 1e-7), 3e-7
GFSF	1e-6	(1.0, 0.7, 1e-7), 3e-7
	WAG	WNes
SVGD	3.9, (0.9, 1e-6)	(300, 0.2), (0.8, 3e-4)
Blob	3.9, (0.9, 1e-6)	(1000, 0.2), (0.9, 1e-5)
GFSD	3.9, (0.9, 1e-6)	(1000, 0.2), (0.9, 1e-5)
GFSF	3.9, (0.9, 1e-6)	(1000, 0.2), (0.9, 1e-5)

Table 3. Parameters of various methods in the BNN experiment

	WGD	PO
SVGD	1e-3	(1.0, 0.6, 1e-7), 1e-4
Blob	(0.5, 3e-5)	(1.0, 0.8, 1e-7), (0.5, 3e-5)
GFSD	(0.5, 3e-5)	(1.0, 0.8, 1e-7), (0.5, 3e-5)
GFSF	(0.5, 3e-5)	(1.0, 0.8, 1e-7), (0.5, 3e-5)
	WAG	WNes
SVGD	3.6, 1e-6	(1000, 0.2), 1e-4
Blob	3.5, (0.5, 1e-5)	(3000, 0.2), (0.6, 1e-4)
GFSD	3.5, (0.5, 1e-5)	(3000, 0.2), (0.6, 1e-4)
GFSF	3.5, (0.5, 1e-5)	(3000, 0.2), (0.6, 1e-4)

Detailed parameters of various methods are provided in Table 3. The format of each column is the same as illustrated in Appendix E.2, except that all SVGD methods uses the AdaGrad with momentum method with remember rate 0.9, so we only provide the step size. WGD and PO methods also adopt the decaying step size, so we provide the decaying exponent. For GFSF, the small diagonal matrix is (1e-2)I.

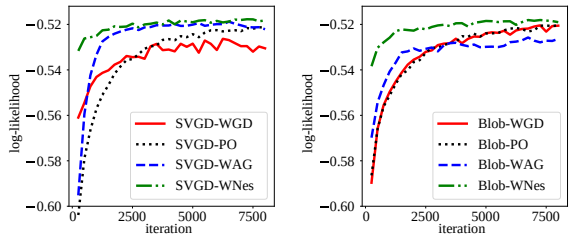
E.4: Detailed Settings and Parameters of the LDA Experiment

We follow the same settings as Ding et al. (2014). The dataset is the ICML dataset[5] that contains 765 documents with vocabulary size 1,918. We also adopt the Expanded-Natural parameterization (Patterson & Teh, 2013), and collapsed Gibbs sampling for stochastic gradient estimation. In each document, 90% of words are used for training the topic proportion of the document, and the left 10% words are used for evaluation. For each run, a random 80%-20%

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

Table 4. Parameters of various methods in the LDA experiment

	WGD	PO
SVGD	3.0	(0.7, 0.7, 1e-4), 10.0
Blob	0.3	(0.7, 0.7, 1e-4), 0.30
GFSD	0.3	(0.7, 0.7, 1e-4), 0.30
GFSF	0.3	(0.7, 0.7, 1e-4), 0.30
	WAG	**WNes**
SVGD	2.5, 3.0	(3.0, 0.2), 10.0
Blob	2.1, 3e-2	(0.3, 0.2), 0.30
GFSD	2.1, 3e-2	(0.3, 0.2), 0.30
GFSF	2.1, 3e-2	(0.3, 0.2), 0.30



(a) SVGD (b) Blob

(c) GFSD (d) GFSF

Figure 8. Acceleration effect of WAG and WNes on BLR on the Covertype dataset, measured in log-likelihood. See Appendix E.2 for detailed experiment settings and parameters.

train-test split of the dataset is done.

For the LDA model, we fix the parameter of the Dirichlet prior on topics as 0.1, and mean and standard deviation of the Gaussian prior on the topic proportion as 0.1 and 1.0, respectively. The number of topics is fixed as 30, and batch size is fixed as 100 for all methods. Collapsed Gibbs sampling is run for 50 iterations for each stochastic gradient estimation. Particle size is fixed as 20 for all methods.

Detailed parameters of all methods are provided in Table 4. The format of each column is the same as illustrated in Appendix E.2, except that all methods uses a decaying step size with decaying exponent 0.55 and initial steps 1,000, so we only provide the step size for all methods. SVGD methods do not use the AdaGrad with momentum method. For GFSF, the small diagonal matrix is (1e-5)I.

For SGNHT, both its sequential and parallel simulations use the fixed step size of 0.03, and its mass and diffusion parameters are set to 1.0 and 22.4.

F. More Experimental Results

F.1: More Results on the BLR Experiment

The results measured in log-likelihood on test dataset corresponding to the results measured in test accuracy in Fig. 4 is provided in Fig. 8. Acceleration effect of our WAG and WNes methods is again clearly demonstrated, making a consistent support. Particularly, the WNes method is more stable than the WAG method.