A proofs

Proof of Lemma 4.1. When $\mathcal{H}$ is the RKHS with kernel k

$$\begin{aligned}
&\mathbb{E}_{q_t}[(D(s_\pi - s_{q_t}))^\top u] - \frac{1}{2}\|Du\|_{\mathcal{H}}^2 \\
&= \mathbb{E}_{q_t}[(D(s_\pi - s_{q_t}))^\top Du] - \frac{1}{2}\|Du\|_{\mathcal{H}}^2 \\
&= \langle R_{D(s_\pi - s_{q_t})}, Du\rangle_{\mathcal{H}} - \frac{1}{2}\|Du\|_{\mathcal{H}}^2
\end{aligned}$$

where R_v is the Riesz representation of the linear function $u \to \mathbb{E}_{q_t} v^\top u$. Therefore a solution to (8) is

$$Du = DR_{D(s_\pi - s_{q_t})}$$

which can be written as

$$\begin{aligned}
Du(x) &= \int D(x)k(x,y)D(y)(s_\pi(y) - s_q(y))q_t(y)dy \\
&= \int k_\perp(x,y)(s_\pi(y) - s_q(y))q_t(y)dy \\
&= \int (k_\perp(x,y)s_\pi(y) + \nabla_y k_\perp(x,y))q_t(y)dy \\
&= \mathbb{E}_{y\sim q_t}(k_\perp(x,y)s_\pi(y) + \nabla_y k_\perp(x,y))
\end{aligned}$$

□

Proof of Theorem 4.2. First, we show the components of r can be written as

$$r_i(x) = \sum_j \partial_j D_{i,j}(x) = -\frac{\sum_j \partial_{i,j}^2 g \partial_j g + \partial_i g \partial_{j,j} g}{\|\nabla g\|^2} + \frac{2(\nabla g^T \nabla^2 g \nabla g)\partial_i g}{\|\nabla g\|^4}.$$

To this end, note that

$$D_{i,j}(x) = 1 - \frac{\partial_i g \partial_j g}{\|\nabla g\|^2}$$

Then by product rule, and the fact that $\partial_j \|\nabla g\|^2 = 2\sum_k \partial_{k,j}^2 g \partial_k g = 2[\nabla^2 g \nabla g]_j$

$$\partial_j D_{i,j}(x) = \frac{-\partial_{i,j}^2 g \partial_j g - \partial_i g \partial_{jj}^2 g}{\|\nabla g\|^2} + \frac{2\partial_i g \partial_j g [\nabla^2 g \nabla g]_j}{\|\nabla g\|^4}.$$

So

$$\sum_j \partial_j D_{i,j}(x) = \frac{-\sum_j \partial_{i,j}^2 g \partial_j g - \sum_j \partial_i g \partial_{jj}^2 g}{\|\nabla g\|^2} + \frac{2\partial_i g (\nabla g^\top \nabla^2 g \nabla g)}{\|\nabla g\|^4}.$$

Comparing with the each component of

$$r = -\frac{\nabla^2 g \nabla g}{\|\nabla g\|^2} - \frac{\mathrm{tr}(\nabla^2 g)}{\|\nabla g\|^2}\nabla g + \frac{2\nabla g^T \nabla^2 g \nabla g}{\|\nabla g\|^4}\nabla g,$$

it is easy to see they are the same.

Next, recall the O-gradient density flow (10)

$$\begin{aligned}
\frac{d}{dt}q &= -\nabla\cdot(\phi(x)q(x)) + \nabla\cdot(D(x)\nabla q(x)) \\
&= -\nabla\cdot(\phi(x)q(x)) + \sum_{i,j}\partial_i(D_{i,j}(x)\partial_j q(x)) \\
&= -\nabla\cdot(\phi(x)q(x)) + \sum_{i,j}\partial_i D_{i,j}(x)\partial_j q(x) + \sum_{i,j} D_{i,j}(x)\partial_{i,j}^2 q(x) \\
&=: -\nabla\cdot(\phi(x)q(x)) + (I).
\end{aligned}$$

Meanwhile, the FPE of the SDE follows

$$\frac{d}{dt}q = -\nabla \cdot (\phi(x)q(x)) - \nabla \cdot (r(x)q(x)) + \sum_{i,j} \partial^2_{i,j}(D_{i,j}(x)q(x)) \tag{18}$$

Then note that

$$\begin{aligned}(II) :&= -\nabla \cdot (r(x)q(x)) = -\sum_i \partial_i \left(\sum_j \partial_j D_{i,j}(x) q(x) \right) \\ &= -\sum_{i,j} \partial^2_{i,j} D_{i,j}(x) q(x) - \sum_{i,j} \partial_j D_{i,j}(x) \partial_i q(x).\end{aligned}$$

Also note that

$$\begin{aligned}\sum_{i,j} \partial^2_{i,j}(D_{i,j}(x)q(x)) &= \sum_{i,j} \partial^2_{i,j} D_{i,j}(x) q(x) + \sum_{i,j} \partial_j D_{i,j}(x) \partial_i q(x) \\ &\quad + \sum_{i,j} \partial_i D_{i,j}(x) \partial_j q(x) + \sum_{i,j} D_{i,j}(x) \partial^2_{i,j} q(x) \\ &= (I) - (II).\end{aligned}$$

So we arrive at our first claim.

For the second claim, using $\nabla g(x)^T D(x) = 0$, the Ito formula indicates that

$$\begin{aligned}dg(X_t) &= \nabla g(X_t)^T [D(X_t) \nabla \log \pi(X_t) - \frac{c g(X_t)}{\|\nabla g(X_t)\|^2} \nabla g(X_t) + r(X_t) + D(X_t)\sqrt{2} dW_t] \\ &\quad + \text{tr}(D(X_t) \nabla^2 g(X_t) D(X_t)) dt \\ &= -c g(X_t) dt + \nabla g(X_t)^T r(X_t) + \text{tr}(D(X_t) \nabla^2 g(X_t) D(X_t)) dt.\end{aligned}$$

So it suffices to show that

$$\nabla g(x)^T r(x) + \text{tr}(D(x) \nabla^2 g(x) D(x)) = 0.$$

To continue, we suppress the expression of x in below to keep formulas short. We first note that $D_{i,j} = 1_{i=j} - \partial_i g \partial_j g / \|\nabla g\|^2$,

$$\partial_j D_{i,j} = \frac{-\partial_{i,j} g \partial_j g - \partial_i g \partial_{j,j} g}{\|\nabla g\|^2} + \frac{2 \sum_k \partial_i g \partial_j g \partial^2_{j,k} g \partial_k g}{\|\nabla g\|^4}$$

We plug this into the computation of $\nabla g^T r = \sum_i \partial_i g \partial_j D_{i,j}$. We note that

$$\sum_{i,j} -\frac{\partial_i g \partial_{i,j} g \partial_j g}{\|\nabla g\|^2} = \frac{-(\nabla g)^T \nabla^2 g \nabla g}{\|\nabla g\|^2}.$$

$$\sum_{i,j} -\frac{(\partial_i g)^2 \partial_{j,j} g \partial_j g}{\|\nabla g\|^2} = -\text{tr}(\nabla^2 g).$$

$$\frac{2 \sum_{i,j} \sum_k \partial_i g \partial_i g \partial_j g \partial^2_{j,k} g \partial_k g}{\|\nabla g\|^4} = \frac{\sum_{j,k} \|\nabla g\|^2 \partial_j g \partial^2_{j,k} g \partial_k g}{\|\nabla g\|^4} = \frac{2 (\nabla g)^T \nabla^2 g \nabla g}{\|\nabla g\|^2}$$

Therefore

$$\begin{aligned}\sum_i \partial_i g \partial_j D_{i,j} &= \frac{(\nabla g)^T \nabla^2 g \nabla g}{\|\nabla g\|^2} - \text{tr}(\nabla^2 g) \\ &= \text{tr}\left(-\nabla^2 g + \nabla^2 g \frac{\nabla g (\nabla g)^T}{\|\nabla g\|^2} \right) \\ &= \text{tr}\left(-\nabla^2 g D\right) = \text{tr}\left(-D \nabla^2 g D\right), \quad \text{since } D^2 = D.\end{aligned}$$

This completes our proof.

For the last claim, we note that

$$\begin{aligned}\mathcal{L}f &= \nabla f^T[D\nabla \log \pi - \frac{\psi(g)}{\|\nabla g\|^2}\nabla g + r] + \text{tr}(D\nabla^2 f D)\\ &= \nabla f^T[\nabla \log \pi - \frac{\nabla^2 g}{\|\nabla g\|^2}\nabla g] + \text{tr}(D\nabla^2 f)\\ &= \nabla f^T \nabla \log \pi + \text{tr}(\nabla^2 f) - \frac{\nabla f^T \nabla^2 g \nabla g}{\|\nabla g\|^2} + \frac{\nabla g^T \nabla^2 f \nabla g}{\|\nabla g\|^2}\end{aligned}$$

Finally, we note that $\nabla f^T \nabla g = 0$. Take derivative of this identity we find

$$\nabla^2 f \nabla g + \nabla^2 g \nabla f = 0$$

Therefore

$$\nabla f^T \nabla^2 g \nabla g = -\nabla g^T \nabla^2 f \nabla g$$

and

$$\mathcal{L}f = \nabla f^T \nabla \log \pi + \text{tr}(\nabla^2 f)$$

which is the same as the generator of the Langevin diffusion. □

Remark If we drop r and implementing a naive SDE:

$$dx_t = \phi(x) + \sqrt{2}D(x)dw_t,$$

the FPE of this SDE will be

$$\frac{d}{dt}q = -\nabla \cdot (\phi(x)q(x)) + \sum_{i,j} \partial^2_{i,j}(D_{i,j}(x)q(x)).$$

It is identical to (18) but without the term $-\nabla \cdot (r(x)q(x)) = (II)$. In other words, it will not match (10) unless $(II) = 0$, which happens when $\sum_{i,j} \partial^2_{i,j} D_{i,j} = \sum_j \partial_j D_{i,j} \equiv 0$. But it should be pointed out that if g is an affine function, $D(x)$ is a constant matrix, then $(II) \equiv 0$, and r is safe to be dropped out.

Proof of Proposition 5.2. We will first show that for any function f, the following holds

$$\int \pi^g(z) f(x) \pi_{\eta,z}(x) dx dz \to \int f(x)\pi(x)dx,$$

when $\eta \to 0$. So (13) holds for Π_z as the weak limit of $\pi_{\eta,z}$.

We note that

$$\pi_{\eta,z}(x) = \frac{\pi(x)\exp(-\frac{1}{2\eta}(g(x)-z)^2)}{\sqrt{2\eta\pi}Z_{\eta,z}}$$

where the normalizing constant is given by

$$Z_{\eta,z} = \frac{1}{\sqrt{2\pi\eta}}\int \pi(x)\exp(-\frac{1}{2\eta}(g(x)-z)^2)dx.$$

Because π^g is the density of $g(X)$, so for any function h:

$$\int h(g(x))\pi(x)dx = \mathbb{E}_{X\sim\pi}[h(g(X))] = \int h(y)\pi^g(y)dy.$$

We pick $h(y) = \exp(-\frac{1}{2\eta}(y-z)^2)$, we obtain that

$$\begin{aligned}Z_{\eta,z} &= \int \frac{1}{\sqrt{2\pi\eta}}\exp(-\frac{1}{2\eta}(y-z)^2)\pi^g(y)dy\\ &= \mathbb{E}\pi^g(z+\sqrt{\eta}\xi) = \pi^g(z)(1+R(z))^{-1}.\end{aligned}$$

where the $|R| \le 2L\sqrt{\eta}$ and L is the regularity constant of π^g. Therefore

$$\begin{aligned}
&\int \pi^g(z) f(x) \pi_{\eta,z}(x) dx dz \\
&= \int \pi(x) f(x) \left(\int \frac{1}{\sqrt{2\pi\eta}} \exp(-\frac{1}{2\eta}(g(x)-z)^2) \frac{\pi^g(z)}{Z_{\eta,z}} dz \right) dx \\
&= \int \pi(x) f(x) \left(\int \frac{1}{\sqrt{2\pi\eta}} \exp(-\frac{1}{2\eta}(g(x)-z)^2) dz \right) dx + \mathbb{E}R(g(X)) \\
&= \int \pi(x) f(x) dx + \mathbb{E}R(g(X)).
\end{aligned}$$

Since $\mathbb{E}R(g(X)) \le 2L\sqrt{\eta}$, we find our first claim when $\eta \to 0$.

Next note that if we pick $f(x) = 1_{|g(x)-z| \ge \epsilon}$

$$\begin{aligned}
&\int f(x) \pi_{\eta,z}(x) dx \\
&= \int 1_{|g(x)-z|\ge\epsilon} \frac{1}{\sqrt{2\pi\eta}} \exp(-\frac{1}{2\eta}(g(x)-z)^2) \frac{\pi(x)}{Z_{\eta,z}} dx \\
&\le \int \frac{1}{\sqrt{2\pi\eta}} \exp(-\frac{\epsilon^2}{2\eta}) \frac{\pi(x)}{Z_{\eta,z}} dx = \frac{1}{\sqrt{2\pi\eta} Z_{\eta,z}} \exp(-\frac{\epsilon^2}{2\eta}).
\end{aligned}$$

When $\eta \to 0$, since $Z_{\eta,z} \to \pi^g(z)$, and $\frac{1}{\sqrt{2\eta}} \exp(-\frac{\epsilon^2}{2\eta}) \to 0$, we find that

$$\Pi_z(|g(X)-z| \ge \epsilon) = \lim_{\eta \to 0} \int f(x) \pi_{\eta,z}(x) dx \to 0$$

For the Stein equation part, note that for each $\pi_{\eta,z}$, we have the following by Stein's identity:

$$\mathbb{E}_{\pi_{\eta,z}}[(\nabla \log \pi(x) - \frac{1}{2\eta}(g(x)-z)\nabla g(x))^\top \phi(x) + \nabla^\top \phi(x)] = 0.$$

But $\nabla g(x)^\top \phi(x) = 0$ for $\phi \in \mathcal{H}_\perp$. This gives

$$\mathbb{E}_{\pi_{\eta,z}}[\mathcal{A}_\pi \phi] = 0, \quad \forall \eta.$$

Taking $\eta \to 0$ yields that $\mathbb{E}_{\pi_z}[\mathcal{A}_\pi \phi] = 0$. □

Lemma A.1. *Suppose $q(x), q^g(z)$ and $\pi(x), \pi^g(z), \pi_z(x)$ are all C^1 functions, then the Radon–Nikodym derivative between q_z and π_z can be written as*

$$\frac{dq_z}{d\pi_z}(x) = \frac{\pi^g(z) q(x)}{q^g(z)\pi(x)}, \quad z = g(x).$$

In particular,

$$D(s_{q_z}(x) - s_{\pi_z}(x)) = D(s_q(x) - s_\pi(x)).$$

Proof of Lemma A.1. We will show (13) holds, where $dq_z = \frac{dq_z}{d\pi_z} d\pi_z$ with our choice of RN derivative. This can be done using

$$\begin{aligned}
\mathbb{E}_{z\sim q^g} \mathbb{E}_{x\sim q_z}[f(x)] &= \int_R dz q^g(z) \int_{\mathcal{G}_z} f dq_z \\
&= \int_R dz q^g(z) \int_{\mathcal{G}_z} \frac{\pi^g(z) q}{q^g(z)\pi} f d\pi_z \\
&= \int_R \pi^g(z) dz \int_{\mathcal{G}_z} \frac{q}{\pi} f d\pi_z \\
&= \mathbb{E}_{z\sim\pi^g} \mathbb{E}_{y\sim\pi_z}[f(x)q(x)/\pi(x)] \\
&= \mathbb{E}_{x\sim\pi}[f(x)q(x)/\pi(x)] = \mathbb{E}_{x\sim q}[f(x)].
\end{aligned}$$

Moreover

$$D(s_{q_z}(x) - s_{\pi_z}(x)) = D(s_q(x) - s_\pi(x)) + D(\nabla g(x) s_{q^g}(g(x)) - \nabla g(x) s_{q^g}(g(x))) = D(s_q(x) - s_\pi(x)).$$

□

Proof of Proposition 5.4. Note that

$$\left|\mathbb{E}_{\Pi_0}[f] - \int_{-\delta}^{\delta} q_g(z)\mathbb{E}_{\Pi_z}[f]dz\right| \leq \int_{-\delta}^{\delta} q_g(z)\left|\mathbb{E}_{\Pi_0}[f] - \mathbb{E}_{\Pi_z}[f]\right| dz$$
$$\leq \max_{|z|\leq\delta} |\mathbb{E}_{\Pi_z}[f] - \mathbb{E}_{\Pi_0}[f]|.$$

Meanwhile

$$\mathbb{E}_q[f] - \int_{-\delta}^{\delta} q_g(z)\mathbb{E}_{\Pi_z}[f]dz = \int_{-\delta}^{\delta} q_g(z)\left(\mathbb{E}_{q_z}[f] - \mathbb{E}_{\Pi_z}[f]\right) dz.$$

Then we note the following holds when we restrict f on $\mathcal{G}_z$. We use a paramaterization of $\mathcal{G}_z$ with dummy variable y and the inherited metric form $\mathbb{R}^d$,

$$\begin{aligned}
(\mathbb{E}_{q_z}[f] - \mathbb{E}_{\Pi_z}[f])^2 &= \left(\int_{\mathcal{G}_z} \Pi_z(y)(\frac{q_z(y)}{\Pi_z(y)} - 1)f(y)dy\right)^2 \\
&\leq \int_{\mathcal{G}_z} \Pi_z(y)(\sqrt{\frac{q_z(y)}{\Pi_z(y)}} - 1)^2 dy \cdot \int_{\mathcal{G}_z} \Pi_z(y)(\sqrt{\frac{q_z(y)}{\Pi_z(y)}} + 1)^2 f(y)^2 dy \\
&\leq 2\int_{\mathcal{G}_z} \Pi_z(y)(\sqrt{\frac{q_z(y)}{\Pi_z(y)}} - 1)^2 dy \cdot \int_{\mathcal{G}_z} \Pi_z(y)(\frac{q_z(y)}{\Pi_z(y)} + 1) dy \\
&= 4\int_{\mathcal{G}_z} \Pi_z(y)(\sqrt{\frac{q_z(y)}{\Pi_z(y)}} - 1)^2 dy \\
&= 8(1 - \mathbb{E}_{\Pi_z}\sqrt{\frac{q_z(y)}{\Pi_z(y)}}) = 8(1-b) \text{ with } b = \mathbb{E}_{\Pi_z}\sqrt{\frac{q_z(y)}{\Pi_z(y)}} \leq 1.
\end{aligned}$$

Then note that

$$\text{var}_{\Pi_z}\sqrt{\frac{q_z(y)}{\Pi_z(y)}} = \mathbb{E}_{\Pi_z}[\frac{q_z(y)}{\Pi_z(y)}] - b^2 = 1 - b^2.$$

Therefore by κ-PI, we have

$$\begin{aligned}
1 - b \leq (1 - b^2) &\leq \kappa \int_{\mathcal{G}_z} \Pi_z(y) \left\|\nabla\sqrt{\frac{q_z(y)}{\Pi_z(y)}}\right\|_{\mathcal{G}_z}^2 dy \\
&= \kappa \int_{\mathcal{G}_z} q_z(y)\|(s_{q_z}(y) - s_{\Pi_z}(y))\|_{\mathcal{G}_z}^2 dy \\
&= \kappa \int_{\mathcal{G}_z} q_z(y)\|D(y)(s_{q_z}(y) - s_{\Pi_z}(y))\|^2 dy \\
&= \kappa\mathbb{E}_{q_z}\|D(s_{q_z} - s_{\Pi_z})\|^2.
\end{aligned}$$

So in combination

$$(\mathbb{E}_{q_z}[f] - \mathbb{E}_{\Pi_z}[f])^2 \leq 8\kappa\mathbb{E}_{q_z}\|D(s_{q_z} - s_{\Pi_z})\|^2.$$

Then by Lemma A.1 we have

$$D\nabla(\log q_z - \log \Pi_z) = D(s_q - s_\pi).$$

So we find that

$$(\mathbb{E}_{q_z}[f] - \mathbb{E}_{\Pi_z}[f])^2 \leq 4\kappa\mathbb{E}_{q_z}\|D(s_q - s_\pi)\|^2.$$

Integrating both sides with $q_g(z)$ we find our final claim.

□

Proof of Proposition 5.5. Fix a particle z_t in the density of q_t, we track its g-value trajectory:

$$\frac{d}{dt}g(z_t) = \nabla g(z_t)^\top v_t(z_t) = -\psi(g(z_t)),$$

we will show that $g(z_t) \leq M_t$ for all t (The proof for $g(z_t) \geq -M_t$ is identifical is omitted). Suppose $g(z_t) > M_t + \epsilon$ for some t and $\epsilon > 0$. Let $t_0 = \inf\{t > 0, g(z_t) > M_t + \epsilon\}$. Then $\frac{d}{dt}g(z_{t_0}) = -\psi(M_{t_0} + \epsilon) < -\psi(M_{t_0})$, so for a sufficiently small $\delta > 0$, $g(z_{t_0-\delta}) > M_{t_0-\delta}$, this contradicts the definition of t_0.

For the second claim, note that

$$\begin{aligned}
&\frac{d}{dt}\mathrm{KL}(q_t\|\pi) = -\int q_t(x)v_t(x)^\top(s_\pi(x) - s_{q_t}(x))dx \\
&= -\int q_t(x)(v_\perp(x) + v_\sharp(x))^\top(s_\pi(x) - s_{q_t}(x))dx \\
&= -F_\perp(q_t, \pi) + \int q_t(x)\frac{\psi(g(x))\nabla g(x)^\top(s_\pi(x) - s_{q_t}(x))}{\|\nabla g(x)\|^2}dx \\
&= -F_\perp(\pi, q_t) + \int \frac{\psi(g(x))\nabla g(x)^\top s_\pi(x)}{\|\nabla g(x)\|^2}dx \\
&\quad + \int \frac{\|\nabla g(x)\|^2\dot{\psi}(g(x)) + \psi(g(x))\Delta g(x)}{\|\nabla g(x)\|^2}q_t(x)dx \\
&\quad - \int \frac{2\psi(g(x))\nabla g(x)^\top\nabla^2 g(x)\nabla g(x)}{\|\nabla g(x)\|^4}q_t(x)dx \\
&\leq -F_\perp(\pi, q_t) + \mathbb{E}_{q_t}[|\dot{\psi}(g)|] + C_0\mathbb{E}_{q_t}[|\psi(g)|\|\nabla g\|^2].
\end{aligned}$$

□

Proof of Theorem 5.6. We use the notation M_t from Proposition 5.5. When we take $\psi(z) = -\alpha\mathrm{sgn}(z)z^{1+\beta}$, we find that for some c_0 that depends on M_0

$$\frac{d}{dt}M_t = -\alpha|M_t|^{1+\beta} \Rightarrow M_t = (\alpha\beta)^{-\frac{1}{\beta}}(t + c_0)^{-\frac{1}{\beta}}.$$

Moreover, we have that for some constant C_1

$$\begin{aligned}
&\frac{d}{dt}\mathrm{KL}(q_t\|\pi) \\
&\leq -F_\perp(q_t, \pi) + \alpha\mathbb{E}_{q_t}(1 + \beta)[|g|^\beta] + \alpha C_0\mathbb{E}_{q_t}[|g|^{1+\beta}\|\nabla g\|^2] \\
&\leq -F_\perp(q_t, \pi) + \frac{C_1}{t + c_0} + \frac{C_1}{(t + c_0)^{1+1/\beta}}.
\end{aligned}$$

Integrating both sides yields the following for some constant M_α

$$\begin{aligned}
\int_{T/2}^{T} F_\perp(q_t, \pi)dt &\leq \int_0^T F_\perp(q_t, \pi)dt \\
&\leq \mathrm{KL}(q_0, \pi) + C_1\log\frac{T + c_0}{c_0} + C_1M_\alpha.
\end{aligned}$$

So

$$\min_{T/2\leq t_0\leq T} F_\perp(q_t, \pi) \leq \frac{2}{T}\mathrm{KL}(q_0, \pi) + \frac{2C_1}{T}\log\frac{T + c_0}{c_0} + \frac{2}{T}C_1M_\alpha.$$

Finally, we note that if $r^D_{q,\pi}$ is the Riez representation of $D(s_\pi - s_q)^T$,

$$\mathbb{E}_q\mathcal{A}_\pi\phi = \mathbb{E}_q(D(s_\pi - s_q))^\top\phi = \langle r^D_{q,\pi}, \phi\rangle_\mathcal{H} \leq \frac{1}{2}\|r^D_{q,\pi}\|_\mathcal{H} = \frac{1}{2}\sqrt{F_\perp(q, \pi)}.$$

□

B Additional Experiments Results and Setting Details

We use NVIDIA GeForce RTX 2080 Ti for neural network experiments.

B.1 Synthetic Distribution

For both O-Langevin and O-SVGD, we use $\alpha = 100$ and $\beta = 0$. We set $\eta = 0.01$ and 0.5 for O-Langevin and O-SVGD respectively. For CLangevin and CHMC, we use a python implementation [1] and tune the step size and the number of leapfrog steps. We report the best results which are achieved at step size $= 0.3$ for CLangevin, and step size $= 1$ and number of leapfrog steps $= 2$ for CHMC.

We use energy distance to measure the difference between the approximated distributions by sampling methods and the target distribution. Energy distance is a statistical distance between probability distributions and has been used in the literature, e.g. [29, 30, 11]. Formally speaking, the energy distance between probability distributions P and Q is defined by

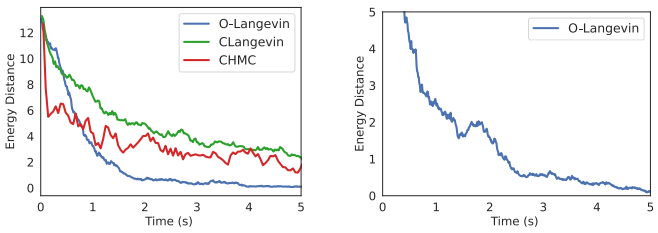
$$D(P, Q) = 2E_{Z,W} \|Z - W\|_2 - E_{Z,Z'} \|Z - Z'\|_2 - E_{W,W'} \|W - W'\|_2$$

where $Z, Z' \sim P$ and $W, W' \sim Q$.

Runtime Comparison We report runtime comparison in Figure 5. We did not include O-SVGD since one iteration of it already $>$ 5s (SVGD is known to take more time per iteration than MCMC due to computing particle interaction). When starting on the manifold, we observe that O-Langevin converges much faster than previous manifold sampling methods. It takes about 4s to fully converge whereas previous methods have not fully converged after 5s. Previous methods cannot work with initializations outside the manifold, thus we only report the runtime of O-Langevin in Figure 5b.

Effect of Hyperparameters Besides the hyperparameter step size η as in standard Langevin and SVGD, our methods have hyperparameters $\alpha > 0$ and $\beta \in (0, 1]$ in $\psi(x) = \alpha \text{sign}(x)|x|^{1+\beta}$ to control the speed of the sampler to approach the manifold and the closeness of the sampler to the manifold after converging. In theory, as α increases, the sampler approaches the manifold faster and stays closer. As β increases, the sampler first converges faster to the manifold but stays relatively far away after converging. We report the results of O-Langevin with varying α (with fixed $\beta = 0$) and β (with fixed $\alpha = 1$) when starting outside the manifold in Figure 6. We can see that the results of MAE, which measures the closeness to the manifold, align with the theoretical analysis. The energy distance with varying α and β are similar while $\alpha = 10$ and $\beta = 0$ perform slightly better. The theoretical analysis could not tell which values of hyperparameters give the fastest convergence to the target distribution thus we still need to tune α and β to achieve the optimal performance. In practice, we recommend to set $\beta = 0$ (though our theoretical results apply to $\beta \in (0, 1]$, we find $\beta = 0$ generally works well in practice.) and tune α to achieve a desirable MAE and energy distance.

Density Estimation To compare the estimated density with the ground truth, we plot the collected samples after 5000 epochs when starting on the manifold and 8000 epochs when starting outside the manifold in Figure 7. The density estimation from our methods is closer to the ground truth than previous methods, aligning with the results of the energy distance in Figure 2.



(a) (b)

Figure 5: Runtime comparison when (a) starting on the manifold and (b) starting outside the manifold.

[1] https://matt-graham.github.io/mici/

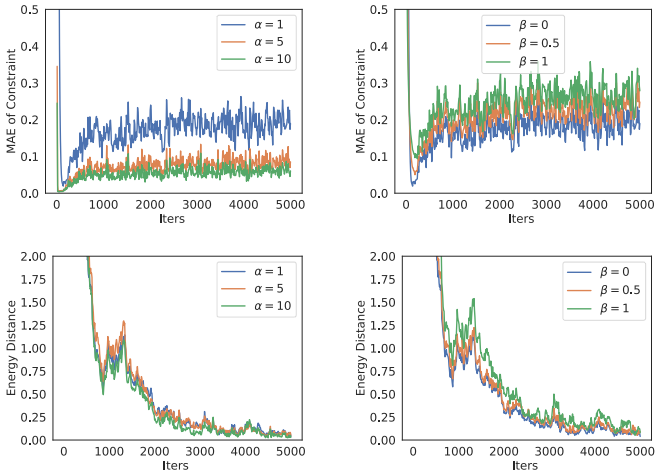
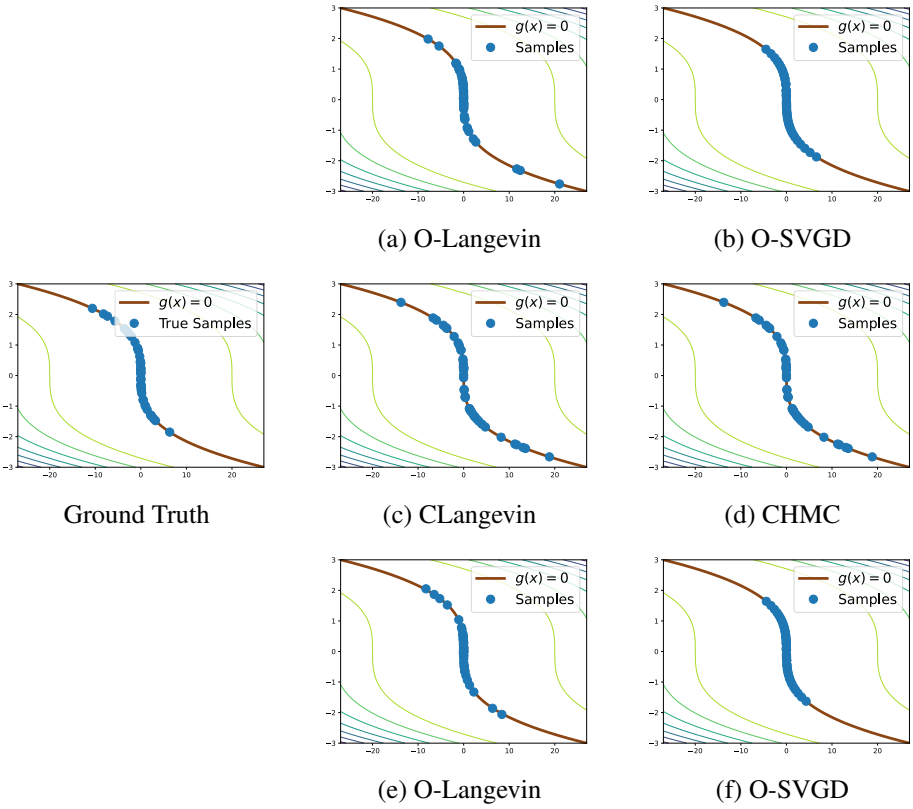


Figure 6: Effect of hyperparameters α and β



(a) O-Langevin (b) O-SVGD

Ground Truth (c) CLangevin (d) CHMC

(e) O-Langevin (f) O-SVGD

Figure 7: Density estimation when (a)-(d) starting on the manifold and (e)-(f) starting outside the manifold.

B.2 Income Classification with Fairness Constraint

The Adult Income dataset contains 30,162 training samples and 15,060 test samples. The feature dimension is 86. Following previous work [26, 24], we obtain the training set by randomly subsampling 20,000 data points from the training samples. The model is a two-layer multilayer perceptron

(MLP), which has 50 hidden units and RELU nonlinearities. The metric values are the mean over all particles. For both methods, we use $n = 10$ particles and $\beta = 0$. For O-Langevin, $\alpha = 100$ and $\eta = 10^{-5}$. For O-SVGD, $\alpha = 130$ and $\eta = 10^{-4}$. The results are averaged over 3 runs with the standard error as the error bar.

B.3 Loan Classification with Logic Rules

The dataset[2] contains loans issued through 2007-2015 of several banks. Each data point contains 28 features such as the current loan status and latest payment information. We define the logic loss to be the binary cross-entropy loss. The metric values are the mean over all particles. For both methods, we use $n = 10$ and $\beta = 0$. For O-Langevin, $\alpha = 80$ and $\eta = 10^{-4}$. For O-SVGD, $\alpha = 100$ and $\eta = 10^{-3}$. The results are averaged over 3 runs with the standard error as the error bar.

B.4 Prior-Agnostic Bayesian Neural Networks

For large models, such as ResNet-18 on this task, computing second-order derivatives is slow. To speed up our methods on large models, we ignore the second-order terms in O-Langevin and O-SVGD and empirically find that they still perform well. We leave the theoretical analysis for future work.

Specifically, we ignore the r term in the update of O-Langevin and obtain

$$x_{t+1} = x_t + \eta \cdot v_\sharp(x_t) + \text{Langevin}_\perp(x_t),$$
$$\text{where } \text{Langevin}_\perp(x_t) = \eta D(x_t)\nabla \log \pi(x_t) + \sqrt{2\eta} D(x_t)\xi_t, \;\; \xi_t \sim \mathcal{N}(0, I).$$

For O-SVGD, the update becomes

$$x_{i,t+1} = x_{i,t} + \eta \cdot \left(v_\sharp(x_{i,t}) + \text{SVGD}_{K_\perp}(x_{i,t})\right),$$
$$\text{where } \text{SVGD}_{K_\perp}(x_{i,t}) = \frac{1}{n}\sum_{j=1}^{n} k_\perp(x_{i,t}, x_{j,t})\nabla_{x_{j,t}} \log \pi(x_{j,t}) + \tilde{\nabla}_{x_{j,t}} k_\perp(x_{i,t}, x_{j,t})$$
$$\text{and } \tilde{\nabla}_{x_{j,t}} k_\perp(x_{i,t}, x_{j,t}) = D(x_{i,t})(D(x_{j,t})\nabla_{x_{j,t}} k(x_{i,t}, x_{j,t}))$$

For all results, we use 200 epochs, 64 batchsize, $n = 4$, $\beta = 0$, $\alpha = 1000$ and $\eta = 10^{-4}$. During testing, we do Bayesian model averaging to obtain test error, ECE and AUROC.

B.5 Computational Cost Comparison

Our method is the first constraint sampling without the requirement of initialization on the manifold, so there is essentially no baseline that can achieve the same effect. Compared to the unconstrained Langevin and SVGD, our method additionally computes the gradient and the Hessian of the constraint function. Compared to previous manifold sampling methods which require expensive projection subroutines, our method has a much cheaper and faster update. For example, in the synthetic distribution experiment, one update of O-Langevin (ours) takes 0.023s whereas the previous method CLangevin takes 0.08s. From Figure 2a, we can see that O-Langevin also converges faster than CLangevin in terms of the number of iterations.

B.6 Further Comparison to Previous Methods

Manifold Sampling Methods Previous manifold sampling methods assume that the initialization is on the manifold. One may wonder if we can obtain such an initialization by optimization algorithms so that we can still use previous methods when the sampler starts outside the manifold. This will not work because the initialization must be exactly on the manifold whereas the solutions found by optimization always have some intolerable error. Finding a point that is exactly on the manifold without any prior knowledge is by itself a hard problem. Therefore, we are not able to compare our methods with previous methods when there are no known in-domain points, such as the income, loan and image classification tasks in Section 6.

[2] https://www.kaggle.com/wendykan/lending-club-loan-data

Moment Constraints As mentioned in the related work, sampling with moment constraints $\mathbb{E}_q[g]$ cannot guarantee every sample to satisfy the constraint. To empirically show the difference between our methods and this type of methods, we compare O-Langevin to Control+ Langevin, which is a recently proposed moment constraint sampling method [25], on the income classification task. We report the mean and the maximum value of the fairness loss in Figure 8. While both methods have small mean fairness loss, the maximum value of O-Langevin is always much smaller than that of Control+Langevin. This suggests that every sample of our method satisfies the constraint well whereas some samples of Control+Langevin violate the constraint significantly, since the moment constraint can only guarantee the mean value instead of the value of each sample.

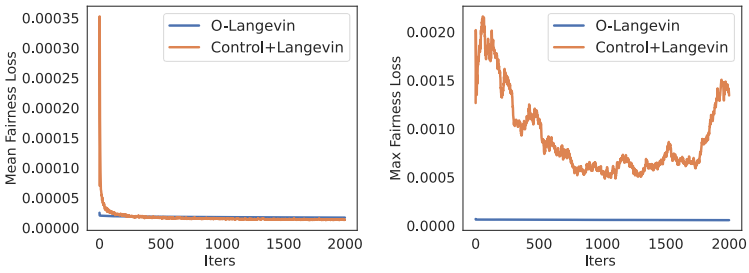


Figure 8: Every sample of O-Langevin satisfies the constraint well whereas some samples of Control+ Langevin violate the constraint significantly.