A Density Function of the Multi-Mode Distribution in Section 3

The negative log-density function of the multi-mode distribution in Section 3 is defined as:

$$U(\boldsymbol{\theta}) \triangleq e^{\frac{3}{4}\boldsymbol{\theta}^2 - \frac{3}{2}\sum_{i=1}^{10} c_i \sin\left(\frac{1}{4}\pi i(\boldsymbol{\theta}+4)\right)} ,$$

where $c = (-0.47, -0.83, -0.71, -0.02, 0.24, 0.01, 0.27, -0.37, 0.87, -0.37)$ is a vector, c_i is the i-th element of c.

B Gronwall Lemma

The Gronwall Lemma plays an important role in parts of our proofs, which is stated in Lemma 11.

Lemma 11 (Gronwall Lemma) *Let $\mathcal{I}$ denotes an interval of the form $[a, +\infty)$ for some $a \in \mathbb{R}$. If $v(\tau)$, defined on $\mathcal{I}$, is differentiable in $\mathcal{I}$ and satisfies the following inequality:*

$$v'(\tau) \leq \beta(\tau)v(\tau) ,$$

where $\beta(\tau)$ is a real-value continuous function defined on $\mathcal{I}$. Then $v(\tau)$ can be bounded as:

$$v(\tau) \leq v(a) \exp\left(\int_a^\tau \beta(s)\mathrm{d}s\right)$$

C Proof of Theorem 3

Proofs of Theorem 2 and 4 rely on techniques in the proofs for Section 4. As a result, we defer the proofs of Theorem 2 and 4 to the later part.

To prove Theorem 3, we rely on the definition of generalized derivative in Definition 1.

Definition 1 (Generalized Derivative) *Let g and ϕ be locally integrable functions on an open set $\Omega \subset \mathbb{R}^d$, that is, Lebesgue integrable on any closed bounded set $\mathcal{F} \subset \omega$. Then ϕ is the generalized derivative of g with respect to $\boldsymbol{\theta}_j$ on Ω, written as $\phi = \partial_{\boldsymbol{\theta}_j} g$, if for any infinitely-differentiable function u with compact support in Ω, we have*

$$\int_\Omega g(\boldsymbol{\theta})\partial_{\boldsymbol{\theta}_j} u(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = -\int_\Omega \phi(\boldsymbol{\theta})u(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} .$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_d)$ on $\Omega \subset \mathbb{R}^d$.

Proof The proof relies on further expansions on the definition of generalized derivative on specific functions. Specifically, let the function g in Definition 1 be in a form of $g \triangleq Gf$ for the product of two functions G and f (specified below). The generalized derivative of (Gf) with respect to $\boldsymbol{\theta}_j$, written as $\partial_{\boldsymbol{\theta}_j}(Gf)$, satisfies

$$\int \partial_{\boldsymbol{\theta}_j}(Gf)\ u(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = -\int Gf\ \partial_{\boldsymbol{\theta}_j} u(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{17}$$

for all differentiable function $u(\cdot)$.

In Theorem 3, we want to prove a particle representation of the following PDE:

$$\partial_\tau \nu_\tau = F_1 = \nabla_{\boldsymbol{\theta}} \cdot (\nu_\tau F(\boldsymbol{\theta}) + (\mathcal{K} * \nu_\tau)\nu_\tau) \triangleq -\sum_j^d \partial_{\boldsymbol{\theta}_j}(Gf) ,$$

where we set $f(\boldsymbol{\theta}) = \nu_\tau(\boldsymbol{\theta})$ and $G(\boldsymbol{\theta}) \triangleq -F(\boldsymbol{\theta}) - (\mathcal{K} * \nu_\tau)(\boldsymbol{\theta})$. Taking integration on both sides for any continuous function $u(\boldsymbol{\theta})$, we have

$$\int \partial_\tau \nu_\tau u(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = -\int \sum_j^d \partial_{\boldsymbol{\theta}_j}(Gf)\ u(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$

$$\Rightarrow \int \partial_\tau f\ u(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = -\sum_j^d \int \partial_{\boldsymbol{\theta}_j}(Gf)\ u(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{18}$$

By applying (17) in (18), we have

$$\begin{aligned}\int \partial_\tau f \; u(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} &= -\sum_j^d \int \partial_{\boldsymbol{\theta}_j}(Gf) \; u(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \sum_j^d \int Gf\partial_{\boldsymbol{\theta}_j} u(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta} \; .\end{aligned}$$

Since $f = \nu_\tau(\boldsymbol{\theta})$ and we can set $u(\boldsymbol{\theta}) = \boldsymbol{\theta}$, we will derive

$$\begin{aligned}&\int \partial_\tau \nu_\tau(\boldsymbol{\theta}) u(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \sum_j^d \int G(\boldsymbol{\theta})\nu_\tau(\boldsymbol{\theta}) \; \partial_{\boldsymbol{\theta}_j} u(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ \Rightarrow &\frac{\mathrm{d}}{\mathrm{d}\tau}\int \nu_\tau(\boldsymbol{\theta})\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta} = \int G(\boldsymbol{\theta}) \; \nu_\tau(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ \Rightarrow &\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbb{E}_{\nu_\tau}[\boldsymbol{\theta}] = \mathbb{E}_{\nu_\tau}[G(\boldsymbol{\theta})] \; . \end{aligned} \tag{19}$$

In particle approximation, we have $\nu_\tau(\boldsymbol{\theta}) \approx \frac{1}{M}\sum_{i=1}^M \delta_{(\boldsymbol{\theta}_\tau^{(i)})}(\boldsymbol{\theta})$. For each particle,according to the definition of $\mathcal{K} * \nu_\tau$ in Sec 2.3, (19) reduces to the following equation:

$$\begin{aligned}\mathrm{d}\boldsymbol{\theta}_\tau^{(i)} &= G(\boldsymbol{\theta}_\tau^{(i)})\mathrm{d}\tau \\ &= -\beta^{-1}F(\boldsymbol{\theta}_\tau^{(i)})\mathrm{d}\tau - \frac{1}{M}\sum_{j=1}^M K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)})F(\boldsymbol{\theta}_\tau^{(j)})\mathrm{d}\tau + \frac{1}{M}\sum_{j=1}^M \nabla K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)})\mathrm{d}\tau,\end{aligned}$$

which is the update equation in Theorem 3. This completes the proof. ■

D Proof of Theorem 5

Note that one challenge in our analysis compared with the analysis for diffusion-based methods, such as those for SG-MCMC [Vollmer et al., 2016, Chen et al., 2015], is how to bound the gap between the original nonlinear PDE (4) and the reduced SDE (8). Following the analysis of granular media equations such as [Malrieu, 2003, Cattiaux et al., 2008, Durmus et al., 2018], we introduce a intermediate SDE in-between (6) and (8), defined as:

$$\begin{cases}\mathrm{d}\bar{\boldsymbol{\theta}}_\tau = & -\beta^{-1}F(\bar{\boldsymbol{\theta}}_\tau)\mathrm{d}\tau - \mathbb{E}_{Y\sim\nu_\tau}K(\bar{\boldsymbol{\theta}}_\tau - Y)F(Y)\mathrm{d}\tau + \nabla K * \nu_\tau(\bar{\boldsymbol{\theta}}_\tau)\mathrm{d}\tau + \sqrt{2\beta^{-1}}\mathrm{d}\bar{\mathcal{W}}_\tau \\ \mathcal{L}(\bar{\boldsymbol{\theta}}_\tau) = & \nu_\tau \mathrm{d}\boldsymbol{\theta}\end{cases} \tag{20}$$

where $\mathcal{L}(\bar{\boldsymbol{\theta}}_\tau)$ denotes the probability law of $\bar{\boldsymbol{\theta}}_\tau$, $\bar{\mathcal{W}}_\tau \in \mathbb{R}^d$ is a d-dimensional Brownian motion independent of $\bar{\boldsymbol{\theta}}_\tau$ and Y is a random variable independent of $\bar{\boldsymbol{\theta}}_\tau$, which is integrated out. In order to match $\bar{\boldsymbol{\theta}}_\tau$ with the particles $\{\boldsymbol{\theta}_\tau^{(i)}\}_{i=1}^M$ in the SDE system (8), we duplicate (20) M times, each endowing an exact solution $\bar{\boldsymbol{\theta}}_\tau^{(i)}$ indexed by i. The distribution of each particles $\{\bar{\boldsymbol{\theta}}_\tau^{(i)}\}_{i=1}^M$ is denoted as ν_τ. Note since (20) is introduced for the purpose of proof convenience without any restrictions, we construct it in a way such that all the $\bar{\mathcal{W}}_\tau^{(i)}$ are *exactly the same*, but independent of each $\mathcal{W}_\tau^{(i)}$, *i.e.*,

$$\begin{cases}\mathrm{d}\bar{\boldsymbol{\theta}}_\tau^{(i)} = & -\beta^{-1}F(\bar{\boldsymbol{\theta}}_\tau^{(i)})\mathrm{d}\tau - \mathbb{E}_{Y_i\sim\nu_\tau}K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - Y_i)F(Y_i)\mathrm{d}\tau + \nabla K * \nu_\tau(\bar{\boldsymbol{\theta}}_\tau^{(i)})\mathrm{d}\tau + \sqrt{2\beta^{-1}}\mathrm{d}\bar{\mathcal{W}}_\tau^{(i)} \\ \mathcal{L}(\bar{\boldsymbol{\theta}}_\tau^{(i)}) = & \nu_\tau \mathrm{d}\boldsymbol{\theta}\end{cases} \tag{21}$$

where, similarly, Y_i is a random variable independent of $\bar{\boldsymbol{\theta}}_\tau^{(i)}$, introduced for the convenience of the proof. Furthermore, we set all the $\bar{\boldsymbol{\theta}}_0^{(i)}$ exact the same but independent of each $\boldsymbol{\theta}_0^{(i)}$. Consequently, all the $\bar{\boldsymbol{\theta}}_\tau^{(i)}$ are also exactly the same but independent of each other. Please note these settings do not affect our algorithm, as (21) are only introduced for the purpose of proof. Now it is ready to prove Theorem 5.

Proof [Proof of Theorem 5] First, from the definitions, we have

$$
\begin{aligned}
\mathrm{d}\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right) &= -\beta^{-1}\left(F(\boldsymbol{\theta}_\tau^{(i)}) - F(\bar{\boldsymbol{\theta}}_\tau^{(i)})\right)\mathrm{d}\tau \\
&+ \frac{1}{M}\sum_j^M \left[\nabla K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}) - \nabla K * \nu_\tau(\bar{\boldsymbol{\theta}}_\tau^{(i)})\right]\mathrm{d}\tau \\
&- \frac{1}{M}\sum_j^M \left(F(\boldsymbol{\theta}_\tau^{(j)})K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}) - \mathbb{E}_{Y_j\sim\nu_\tau}F(Y_j)K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - Y_j)\right)\mathrm{d}\tau \\
&+ (\sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_\tau^{(i)} - \sqrt{2\beta^{-1}}\mathrm{d}\bar{\mathcal{W}}_\tau^{(i)})\mathrm{d}\tau
\end{aligned}
$$

Hence,

$$
\Rightarrow \mathrm{d}\left(\sum_i^M \left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2\right) \tag{22}
$$

$$
= \frac{2}{M}\sum_{i,j}^M (A_{ij}(\tau) + B_{ij}(\tau) + C_{ij}(\tau) + F_{ij}(\tau) + G_{ij}(\tau) + H_{ij}(\tau) + I_{i,j}(\tau))\mathrm{d}\tau\ , \tag{23}
$$

where

$$
\begin{aligned}
A_{ij}(\tau) &= -\beta^{-1}\left(F(\boldsymbol{\theta}_\tau^{(i)}) - F(\bar{\boldsymbol{\theta}}_\tau^{(i)})\right)\cdot\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right) \\
B_{ij}(\tau) &= \left(\nabla K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}) - \nabla K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right)\cdot\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right) \\
C_{ij}(\tau) &= \left(\nabla K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(j)}) - \nabla K * \nu_\tau(\bar{\boldsymbol{\theta}}_\tau^{(i)})\right)\cdot\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right) \\
F_{ij}(\tau) &= -\left(F(\boldsymbol{\theta}_\tau^{(j)})K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}) - F(\bar{\boldsymbol{\theta}}_\tau^{(j)})K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)})\right)\cdot\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right) \\
G_{ij}(\tau) &= -\left(F(\bar{\boldsymbol{\theta}}_\tau^{(j)})K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}) - F(\bar{\boldsymbol{\theta}}_\tau^{(j)})K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right)\cdot\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right) \\
H_{ij}(\tau) &= -\left(F(\bar{\boldsymbol{\theta}}_\tau^{(j)})K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(j)}) - \mathbb{E}_{Y_j\sim\nu_\tau}F(Y_j)K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - Y_j)\right)\cdot\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right) \\
I_{i,j}(\tau) &= (\sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_\tau^{(i)} - \sqrt{2\beta^{-1}}\mathrm{d}\bar{\mathcal{W}}_\tau^{(i)})\cdot\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right)
\end{aligned}
$$

Next, we bound these terms in the following. For the $A_{ij}(\tau)$ term, according to bullet *i)* in Assumption 3 for F, we have

$$
\begin{aligned}
\mathbb{E}\sum_{ij} A_{ij}(\tau) &= -\mathbb{E}\sum_{ij}\beta^{-1}\left(F(\boldsymbol{\theta}_\tau^{(i)}) - F(\bar{\boldsymbol{\theta}}_\tau^{(i)})\right)\cdot\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right) \\
&\leq -\beta^{-1}m_F M\sum_i \mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2
\end{aligned}
$$

For the $B_{ij}(\tau)$ term, applying the oddness of ∇K in Assumption 1, we have

$$
\begin{aligned}
&\mathbb{E}\sum_{ij} B_{ij}(\tau) \\
=&\mathbb{E}\sum_{ij}\left(\nabla K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}) - \nabla K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right)\cdot\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right) \\
=&\mathbb{E}\frac{1}{2}\sum_{ij}\left(\nabla K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}) - \nabla K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right)\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)} - (\boldsymbol{\theta}_\tau^{(j)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right) \\
\leq&\mathbb{E}\frac{1}{2}L_{\nabla K}\sum_{ij}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)} - (\boldsymbol{\theta}_\tau^{(j)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right\|^2 \leq 2L_{\nabla K}M\mathbb{E}\sum_i\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2
\end{aligned}
$$

For the $C_{ij}(\tau)$ term, we have

$$\begin{aligned}
\mathbb{E}\sum_j C_{ij}(\tau) &\overset{(1)}{\leq} \left(\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2\right)^{1/2} \left(\mathbb{E}\left\|\sum_j \left(\nabla K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(j)}) - \nabla K * \nu_\tau(\bar{\boldsymbol{\theta}}_\tau^{(i)})\right)\right\|^2\right)^{1/2} \\
&\overset{(2)}{=} \left(\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2\right)^{1/2} \left(\sum_j \mathbb{E}\left(\nabla K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(j)}) - \nabla K * \nu_\tau(\bar{\boldsymbol{\theta}}_\tau^{(i)})\right)^2\right)^{1/2} \\
&\overset{(3)}{\leq} 2H_{\nabla K}\sqrt{M}\left(\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2\right)^{1/2}
\end{aligned}$$

where (1) is obtained by applying the Cauchy-Schwarz inequality, and (2) by the fact that $\mathbb{E}\left(K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(j)}) - K * \nu_\tau(\bar{\boldsymbol{\theta}}_\tau^{(i)})\right) = 0$. Furthermore, due to the fact that $\nabla K = \exp(-\frac{\|\boldsymbol{\theta}\|^2}{\eta^2})\frac{2}{\eta^2}\boldsymbol{\theta}$, we can bound the $\nabla K(\boldsymbol{\theta})$ with $\|\nabla K\| \leq \exp(-\frac{\|\boldsymbol{\theta}\|^2}{\eta^2})\frac{2}{\eta^2}\|\boldsymbol{\theta}\|$. Hence there exists some positive constant $H_{\nabla K}$ such that $\|\nabla K(\boldsymbol{\theta})\| \leq H_{\nabla K}$.

Similarly, since $K \leq 1$, we have the following result for the $H_{ij}(\tau)$ term,

$$\begin{aligned}
&\mathbb{E}\sum_j H_{ij}(\tau) \\
\leq &\left(\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2\right)^{1/2} \left(\mathbb{E}\left\|\sum_j \left(F(\bar{\boldsymbol{\theta}}_\tau^{(j)})K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(j)}) - \mathbb{E}_{Y_j\sim\nu_\tau}F(Y_j)K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - Y_j)\right)\right\|^2\right)^{1/2} \\
= &\left(\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2\right)^{1/2} \left(\sum_j \mathbb{E}\left(F(\bar{\boldsymbol{\theta}}_\tau^{(j)})K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(j)}) - \mathbb{E}_{Y_j\sim\nu_\tau}F(Y_j)K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - Y_j)\right)^2\right)^{1/2} \\
\leq &2H_\theta\sqrt{M}\left(\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2\right)^{1/2}
\end{aligned}$$

The last inequality follows from the fact that $\sqrt{E[F(\bar{\boldsymbol{\theta}}_\tau^{(j)})^2} \leq L_F\sqrt{E\|\bar{\boldsymbol{\theta}}_\tau^{(j)}\|^2} \leq L_F\sqrt{\gamma_0 + \frac{d}{m'\beta}}$, which is derived in Theorem 18. We denote $L_F\sqrt{\gamma_0 + \frac{d}{m'\beta}}$ as H_θ.

For the $F_{ij}(\tau)$ and $G_{ij}(\tau)$ terms, we have:

$$\begin{aligned}
\mathbb{E}\sum_{ij} F_{ij}(\tau) &= -\mathbb{E}\sum_{ij}\left(F(\boldsymbol{\theta}_\tau^{(j)})K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}) - F(\bar{\boldsymbol{\theta}}_\tau^{(j)})K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)})\right)\cdot\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right) \\
&\leq \mathbb{E}\sum_{ij} L_F\left\|\boldsymbol{\theta}_\tau^{(j)} - \bar{\boldsymbol{\theta}}_\tau^{(j)}\right\|\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\| \\
&\leq L_F M\mathbb{E}\sum_i\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2 .
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\sum_{ij} G_{ij}(\tau) =&\mathbb{E}\sum_{ij}\left(F(\bar{\boldsymbol{\theta}}_\tau^{(j)})K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}) - F(\bar{\boldsymbol{\theta}}_\tau^{(j)})K(\bar{\boldsymbol{\theta}}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right)\cdot\left(\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right) \\
&\leq L_F L_K\sum_{i,j}\mathbb{E}\left(\left\|\bar{\boldsymbol{\theta}}_\tau^{(j)}\right\|\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)} - (\boldsymbol{\theta}_\tau^{(j)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right\|\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|\right) \\
&\leq L_F L_K\sum_{i,j}\sqrt{\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2}\sqrt{\mathbb{E}\left(\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)} - (\boldsymbol{\theta}_\tau^{(j)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right\|^2\left\|\bar{\boldsymbol{\theta}}_\tau^{(j)}\right\|^2\right)}
\end{aligned}$$

$$
\begin{aligned}
&\leq L_F L_K \sum_{i,j} \sqrt{\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2} \sqrt{\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)} - (\boldsymbol{\theta}_\tau^{(j)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right\|^2 \mathbb{E}\left\|\bar{\boldsymbol{\theta}}_\tau^{(j)}\right\|^2} \\
&\leq H_\theta L_K \sum_{i,j} \sqrt{\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)} - (\boldsymbol{\theta}_\tau^{(j)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right\|^2 \mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2} \\
&\leq H_\theta L_K \sum_{i,j} \sqrt{\left(2\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2 + 2\mathbb{E}\left\|(\boldsymbol{\theta}_\tau^{(j)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right\|^2\right) \mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2} \\
&\leq \frac{1}{2} H_\theta L_K \sum_{i,j} \left(2\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2 + 3\mathbb{E}\left\|(\boldsymbol{\theta}_\tau^{(j)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})\right\|^2\right) \\
&\leq \frac{5}{2} H_\theta L_K M \mathbb{E} \sum_{i} \left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2
\end{aligned}
$$

The above result is derived with Cauchy-Schwarz inequality and the independence between $\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)} - (\boldsymbol{\theta}_\tau^{(j)} - \bar{\boldsymbol{\theta}}_\tau^{(j)})$ and $\bar{\boldsymbol{\theta}}_\tau^{(j)}$. The independency come from the following argument: According to our constructions of all the $\bar{\boldsymbol{\theta}}_\tau^{(i)}$ in (21), we conclude that all the $\bar{\boldsymbol{\theta}}_\tau^{(i)}$ are identical. Hence, we have $\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)} - (\boldsymbol{\theta}_\tau^{(j)} - \bar{\boldsymbol{\theta}}_\tau^{(j)}) = \boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}$, which is obvious independent of $\bar{\boldsymbol{\theta}}_\tau^{(j)}$.

For the $I_{i,j}(\tau)$ term, following the analysis in [Malrieu, 2003, Cattiaux et al., 2008, Durmus et al., 2018] and applying the independency between $\mathcal{W}_\tau^{(i)} - \bar{\mathcal{W}}_\tau^{(i)}$ and $\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}$, we have

$$\mathbb{E}\sum_{ij} I_{ij}(\tau) = 0.$$

Denote $\gamma_i(\tau) \triangleq \mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2$. Due to the exchangeability of the particles, $\gamma_i(\tau)$ have the same value for all the particles, denoted as $\gamma(\tau)$. According to (23) and the bounds derived above, we have

$$\gamma'(\tau) \leq -2\lambda_1\gamma(\tau) + \frac{2H_{\nabla K} + 2H_\theta}{\sqrt{M}}\sqrt{\gamma(\tau)}\ .$$

where $\lambda_1 = \beta^{-1}m_F - \frac{5}{2}H_\theta L_K - L_F - 2L_{\nabla K}$. After some algebra, the above inequality can be transformed to

$$
\begin{aligned}
&(\sqrt{\gamma(\tau)} - \frac{2(H_{\nabla K} + H_\theta)}{\sqrt{M}(\beta^{-1}m_F - \frac{5}{2}H_\theta L_K - L_F - 2L_{\nabla K})})' \\
\leq& -\lambda_1(\sqrt{\gamma(\tau)} - \frac{2(H_{\nabla K} + H_\theta)}{\sqrt{M}(\beta^{-1}m_F - \frac{5}{2}H_\theta L_K - L_F - 2L_{\nabla K})})
\end{aligned}
$$

Note that $\boldsymbol{\theta}_\tau^{(i)}$ and $\bar{\boldsymbol{\theta}}_\tau^{(i)}$ are initialized with the same initial distribution $\mu_0 = \nu_0$ but independent of each other. From the proof of Theorems 16 and 18, we can have $\gamma(0) \leq 4\gamma_0$ for some constant γ_0. When we set γ_0 small enough, we can have the following results according to the Gronwall Lemma.

$$\sqrt{\gamma(\tau)} \leq \frac{2(H_{\nabla K} + H_\theta)}{\sqrt{M}(\beta^{-1}m_F - \frac{5}{2}H_\theta L_K - L_F - 2L_{\nabla K})}$$

Hence, there exist some positive constant (c_1, c_2) such that:

$$
\begin{aligned}
W_1(\rho_\tau, \nu_\tau) &\overset{(1)}{\leq} W_2(\rho_\tau, \nu_\tau) \\
&\overset{(2)}{\leq} \sqrt{\mathbb{E}\left\|\boldsymbol{\theta}_\tau^{(i)} - \bar{\boldsymbol{\theta}}_\tau^{(i)}\right\|^2} \overset{(3)}{\leq} \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)},
\end{aligned}
\tag{24}
$$

where (1) holds due to the relationship between W_1 and W_2 metric [Givens and Shortt, 1984], (2) due to the definition of W_2, and (3) due to the result from the previous proof.

■

E Proof of Theorem 6

Proof [Proof of Theorem 6] First, note our goal is to bound $W_1(\nu_\tau, \nu_\infty) \leq c_3 \exp(-2\lambda_1 \tau)$. According to the relationship between W_1 and W_2 metric that $W_1 \leq W_2$ [Givens and Shortt, 1984], once we bound $W_2(\nu_\tau, \nu_\infty)$ as $W_2(\nu_\tau, \nu_\infty) \leq c_3 \exp(-2\lambda_1 \tau)$, the bound for W_1 will automatically hold.

In the following, we will bound W_2. We first note the following cases based on equation (8):

- We set the initial distribution of each particle to be ν_0, which means $\rho_0 = \mathcal{L}(\boldsymbol{\theta}_0^{(i)}) = \nu_0$. In this case, the M evolved particles are denoted as $\{\boldsymbol{\theta}_{\tau,1}^{(i)}\}_{i=1}^M$. We denote the distribution of each $\boldsymbol{\theta}_{\tau,1}^{(i)}$ at τ as $\rho_{\tau,1}$.
- We set the initial distribution of each particle to be ν_∞, which means $\rho_0 = \mathcal{L}(\boldsymbol{\theta}_0^{(i)}) = \nu_\infty$. In this case, the M evolved particles are denoted as $\{\boldsymbol{\theta}_{\tau,2}^{(i)}\}_{i=1}^M$. We denote the distribution of each $\boldsymbol{\theta}_{\tau,2}^{(i)}$ at τ as $\rho_{\tau,2}$.

To bound $W_2(\nu_\tau, \nu_\infty)$, we decompose it as:

$$W_2(\nu_\tau, \nu_\infty) \leq W_2(\nu_\tau, \rho_{\tau,1}) + W_2(\rho_{\tau,1}, \rho_{\tau,2}) + W_2(\rho_{\tau,2}, \nu_\infty) \; . \tag{25}$$

Note that $\rho_{0,1} = \nu_0$ and $\rho_{0,2} = \nu_\infty$. According to (24), we have

$$\begin{aligned} W_2(\nu_\tau, \rho_{\tau,1}) &\leq \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)} \\ W_2(\rho_{\tau,2}, \nu_\infty) &\leq \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)} \end{aligned}$$

It remains to bound the term $W_2(\rho_{\tau,1}, \rho_{\tau,2})$. It is worth mentioning that the reason of introducing $\{\boldsymbol{\theta}_{\tau,1}^{(i)}\}_{i=1}^M$ and $\{\boldsymbol{\theta}_{\tau,2}^{(i)}\}_{i=1}^M$ is to bound the term $W_2(\nu_\tau, \nu_\infty)$, which consequently is to bound $W_2(\rho_{\tau,1}, \rho_{\tau,2})$. For some special settings of $\{\boldsymbol{\theta}_{\tau,1}^{(i)}\}_{i=1}^M$ and $\{\boldsymbol{\theta}_{\tau,2}^{(i)}\}_{i=1}^M$, it will allow us to bound $W_2(\rho_{\tau,1}, \rho_{\tau,2})$ easier. To this end, we set all the $\{\boldsymbol{\theta}_{0,1}^{(i)}\}_{i=1}^M$ and the corresponding $\mathcal{W}_{\tau,1}^{(i)}$ to be exactly the same. Consequently, all the $\{\boldsymbol{\theta}_{\tau,1}^{(i)}\}_{i=1}^M$ will be identical. In this setting, the bound proved above for $W_2(\nu_\tau, \rho_{\tau,1})$ still holds since this is just a specific case for Theorem 5. The same argument goes for $\{\boldsymbol{\theta}_{\tau,2}^{(i)}\}_{i=1}^M$. And we are left to prove the bound for $W_2(\rho_{\tau,1}, \rho_{\tau,2})$.

Since $W_2(\rho_{\tau,1}, \rho_{\tau,2}) \leq \mathbb{E}\left(\left\|\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right\|^2\right) \triangleq r(\tau)$, we will derive a bound for $\mathbb{E}\left(\left\|\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right\|^2\right)$ in the following:

$$\begin{aligned} &\mathrm{d}\left(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right) = -\beta^{-1}\left(F(\boldsymbol{\theta}_{\tau,1}^{(i)}) - F(\boldsymbol{\theta}_{\tau,2}^{(i)})\right)\mathrm{d}\tau \\ &+ \frac{1}{M}\sum_j^M \left[\nabla K(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,1}^{(j)}) - \nabla K(\boldsymbol{\theta}_{\tau,2}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(j)})\right]\mathrm{d}\tau \\ &- \frac{1}{M}\sum_j^M \left(F(\boldsymbol{\theta}_{\tau,1}^{(j)})K(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,1}^{(j)}) - F(\boldsymbol{\theta}_{\tau,2}^{(j)})K(\boldsymbol{\theta}_{\tau,2}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(j)})\right)\mathrm{d}\tau \\ &+ (\sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_{\tau,1}^{(i)} - \sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_{\tau,2}^{(i)})\mathrm{d}\tau \end{aligned}$$

As a result, we have

$$\mathrm{d}\left(\sum_i^M \left\|\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right\|^2\right) = \frac{2}{M}\sum_{i,j}^M (\xi_{ij}^1(\tau) + \xi_{ij}^2(\tau) + \xi_{ij}^3(\tau) + \xi_{ij}^4(\tau) + \xi_{ij}^5(\tau))\mathrm{d}\tau$$

where

$$\begin{aligned}
\xi_{ij}^1(\tau) &= -\beta^{-1}\left(F(\boldsymbol{\theta}_{\tau,1}^{(i)}) - F(\boldsymbol{\theta}_{\tau,2}^{(i)})\right)\cdot\left(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right)\\
\xi_{ij}^2(\tau) &= \left(\nabla K(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,1}^{(j)}) - \nabla K(\boldsymbol{\theta}_{\tau,2}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(j)})\right)\cdot\left(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right)\\
\xi_{ij}^3(\tau) &= -\left(F(\boldsymbol{\theta}_{\tau,1}^{(j)})K(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,1}^{(j)}) - F(\boldsymbol{\theta}_{\tau,2}^{(j)})K(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,1}^{(j)})\right)\cdot\left(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right)\\
\xi_{ij}^4(\tau) &= -\left(F(\boldsymbol{\theta}_{\tau,2}^{(j)})K(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,1}^{(j)}) - F(\boldsymbol{\theta}_{\tau,2}^{(j)})K(\boldsymbol{\theta}_{\tau,2}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(j)})\right)\cdot\left(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right)\\
\xi_{ij}^5(\tau) &= (\sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_{\tau,1}^{(i)} - \sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_{\tau,2}^{(i)}))\cdot\left(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right)
\end{aligned}$$

For the $\xi_{ij}^1(\tau)$ terms, according to Assumption 3 for F, we have

$$\begin{aligned}
\mathbb{E}\sum_{ij}\xi_{ij}^1(\tau) &= -\mathbb{E}\sum_{ij}\beta^{-1}\left(F(\boldsymbol{\theta}_{\tau,1}^{(i)}) - F(\boldsymbol{\theta}_{\tau,2}^{(i)})\right)\cdot\left(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right)\\
&\le -\beta^{-1}m_F M\mathbb{E}\sum_i\left\|\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right\|^2 .
\end{aligned}$$

For the $\xi_{ij}^2(\tau)$ term, applying the concave condition for K and the oddness of ∇K in Assumption 1, we have

$$\begin{aligned}
&\mathbb{E}\sum_{ij}\xi_{ij}^2(\tau) = \mathbb{E}\sum_{ij}^{M}\left(\nabla K(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,1}^{(j)}) - \nabla K(\boldsymbol{\theta}_{\tau,2}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(j)})\right)\cdot\left(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right)\\
&= \frac{1}{2}\sum_{ij}^{M}\mathbb{E}\left(\nabla K(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,1}^{(j)}) - \nabla K(\boldsymbol{\theta}_{\tau,2}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(j)})\right)\cdot\left(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)} - (\boldsymbol{\theta}_{\tau,1}^{(j)} - \boldsymbol{\theta}_{\tau,2}^{(j)})\right)\\
&\le \frac{1}{2}L_K\mathbb{E}\sum_{ij}^{M}\left\|\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)} - (\boldsymbol{\theta}_{\tau,1}^{(j)} - \boldsymbol{\theta}_{\tau,2}^{(j)})\right\|^2 \le 2L_K M\mathbb{E}\sum_i\left\|\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right\|^2 .
\end{aligned}$$

For the $\xi_{ij}^3(\tau)$ terms, applying the L_F-Lipschitz property for F and using $K \le 1$, we have

$$\begin{aligned}
&\mathbb{E}\sum_{ij}\xi_{ij}^3(\tau)\\
=&\mathbb{E}\sum_{ij}-\left(F(\boldsymbol{\theta}_{\tau,1}^{(j)})K(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,1}^{(j)}) - F(\boldsymbol{\theta}_{\tau,2}^{(j)})K(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,1}^{(j)})\right)\cdot\left(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right)\\
\le&\mathbb{E}\sum_{ij}L_F\left\|\boldsymbol{\theta}_{\tau,1}^{(j)} - \boldsymbol{\theta}_{\tau,2}^{(j)}\right\|\left\|\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right\|\\
\le&L_F M\mathbb{E}\sum_i\left\|\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right\|^2 .
\end{aligned}$$

For the $\xi_{ij}^4(\tau)$ terms, recall that all the $\boldsymbol{\theta}_{\tau,1}^{(i)}$ are identical (and all the $\boldsymbol{\theta}_{\tau,2}^{(i)}$ are identical), we have

$$\begin{aligned}
&\mathbb{E}\sum_{ij}\xi_{ij}^4(\tau)\\
&= -\mathbb{E}\sum_{ij}\left(F(\boldsymbol{\theta}_{\tau,2}^{(j)})K(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,1}^{(j)}) - F(\boldsymbol{\theta}_{\tau,2}^{(j)})K(\boldsymbol{\theta}_{\tau,2}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(j)})\right)\cdot\left(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right) = 0 .
\end{aligned}$$

Similar to the proof of the Theorem 5, we have

$$\begin{aligned}
\mathbb{E}\sum_{i,j}\xi_{ij}^5(\tau) &= \mathbb{E}\sum_{i,j}(\sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_{\tau,1}^{(i)} - \sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_{\tau,2}^{(i)}))\cdot\left(\boldsymbol{\theta}_{\tau,1}^{(i)} - \boldsymbol{\theta}_{\tau,2}^{(i)}\right)\\
&= 0
\end{aligned}$$

Combining these bounds, we have

$$r'(\tau) \leq -2(\beta^{-1}m_F - L_F - 2L_K)r(\tau) \ .$$

According to the Gronwall lemma, we have

$$r(\tau) \leq r(0)e^{-2\lambda_1\tau},$$

where $\lambda_1 = \beta^{-1}m_F - L_F - 2L_K$.

Consequently, there exists some positive constant c_3 such that

$$W_2(\rho_{\tau,1}, \rho_{\tau,2}) \leq c_3 e^{-2\lambda_1\tau}$$

Combing all bounds for (25), we have

$$W_2(\nu_\tau, \nu_\infty) \leq c_3 e^{-2\lambda_1\tau} + \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)} + \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)}$$

We can further tighten the above bound by noting that ν_τ is the solution of (6), which has nothing to do with the number of particles M. As a result, we can set $M \to \infty$, resulting in

$$W_2(\nu_\tau, \nu_\infty) \leq c_3 e^{-2\lambda_1\tau} \ ,$$

which completes the proof.

■

F Proof of Theorem 7

To bound the $W_1(\mu_T, \rho_{\sum_{k=0}^{T-1} h_k})$ term, note the original SDE driving the particles $\{\boldsymbol{\theta}_\tau^{(i)}\}$ in (8) corresponds to is a nonlinear PDE, which is hard to deal with. Fortunately, (8) can be turned into a diffusion-based SDE by concatenating the particles at each time into a single vector representation, *i.e.*, by defining the new parameter at time τ as $\boldsymbol{\Theta}_\tau \triangleq [\boldsymbol{\theta}_\tau^{(1)}, \cdots, \boldsymbol{\theta}_\tau^{(M)}] \in \mathbb{R}^{Md}$. Consequently, $\boldsymbol{\Theta}_\tau$ is driven by the following SDE:

$$\mathrm{d}\boldsymbol{\Theta}_\tau = -F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau)\mathrm{d}\tau + \sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_\tau^{(Md)} \ , \tag{26}$$

where

$$\begin{aligned} F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau) \triangleq [&\beta^{-1}F(\boldsymbol{\theta}_\tau^{(1)}) - \frac{1}{M}\sum_{j=1}^M \nabla K(\boldsymbol{\theta}_\tau^{(1)} - \boldsymbol{\theta}_\tau^{(j)}) + \frac{1}{M}\sum_{j=1}^M K(\boldsymbol{\theta}_\tau^{(1)} - \boldsymbol{\theta}_\tau^{(j)})F(\boldsymbol{\theta}_\tau^{(j)}), \cdots, \\ &\beta^{-1}F(\boldsymbol{\theta}_\tau^{(M)}) - \frac{1}{M}\sum_{j=1}^M \nabla K(\boldsymbol{\theta}_\tau^{(M)} - \boldsymbol{\theta}_\tau^{(j)}) + \frac{1}{M}\sum_{j=1}^M K(\boldsymbol{\theta}_\tau^{(M)} - \boldsymbol{\theta}_\tau^{(j)})F(\boldsymbol{\theta}_\tau^{(j)})] \end{aligned}$$

is a vector function $\mathbb{R}^{Md} \to \mathbb{R}^{Md}$, and $\mathcal{W}\tau^{(Md)}$ is Brownian motion of dimension Md.

Now we define $F_{(q)\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau) \triangleq [\beta^{-1}F_q(\boldsymbol{\theta}_\tau^{(1)}) - \frac{1}{MN}\sum_{j=1}^M \nabla K(\boldsymbol{\theta}_\tau^{(1)} - \boldsymbol{\theta}_\tau^{(j)}) + \frac{1}{M}\sum_{j=1}^M K(\boldsymbol{\theta}_\tau^{(1)} - \boldsymbol{\theta}_\tau^{(j)})F_q(\boldsymbol{\theta}_\tau^{(j)}), \cdots, \beta^{-1}F_q(\boldsymbol{\theta}_\tau^{(M)}) - \frac{1}{MN}\sum_{j=1}^M \nabla K(\boldsymbol{\theta}_\tau^{(M)} - \boldsymbol{\theta}_\tau^{(j)}) + \frac{1}{M}\sum_{j=1}^M K(\boldsymbol{\theta}_\tau^{(M)} - \boldsymbol{\theta}_\tau^{(j)})F_q(\boldsymbol{\theta}_\tau^{(j)})]$. We can verify that $F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau) = \sum_{q=1}^N F_{(q)\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau)$.

Define $\Theta_k \triangleq [\theta_k^{(1)}, \cdots, \theta_k^{(M)}]$ and $G_{\mathcal{I}_k}^{\Theta} \triangleq \frac{N}{B_k}\sum_{q \in \mathcal{I}_k} F_{(q)\boldsymbol{\Theta}}(\Theta_k)$. It is seen that the following result holds:

$$\Theta_{k+1} = \Theta_k - G_{\mathcal{I}_k}^{\Theta} h_k + \sqrt{2\beta^{-1}h_k}\Xi_k \ , \tag{27}$$

where $\Xi_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{Md\times Md})$. As a result, we have that Θ_k of (27) is accutually the numerical solution of the SDE (26) via stochastic gradients.

Denote the distribution of Θ_k as μ_k^{Θ}, and the distribution of $\boldsymbol{\Theta}_\tau$ as $\rho_\tau^{\boldsymbol{\Theta}}$. Before proceeding to our theoretical results, we first present the following Lemmas, which is very important in our proof.

Lemma 12 $W_1(\mu_k, \rho_\tau) \leq \frac{1}{\sqrt{M}} W_1(\mu_k^{\Theta}, \rho_\tau^{\boldsymbol{\Theta}})$

Proof [Proof of Lemma 12] Let us recall the definition of W_1 metric and its Kantorovich-Rubinstein duality [Villani, 2008], *i.e.* $W_1(\mu,\nu) \triangleq \sup_{\|g\|_{lip}\leq 1} |\mathbb{E}_{\boldsymbol{\theta}\sim\mu}[g(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim\nu}[g(\boldsymbol{\theta})]|$. We can prove the fact that if $g(\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}$ is a L_g-Lipschitz function in $\mathbb{R}^d$, the $g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})$, defined as $g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) = \frac{1}{\sqrt{M}}\sum_i^M g(\boldsymbol{\theta}^{(i)})$, is a L_g-Lipschitz function in $\mathbb{R}^{Md}$, where $\boldsymbol{\Theta} \triangleq [\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(M)}]$. The proof is as follows:

$$
\begin{aligned}
&\|g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_1) - g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_2)\| \leq \frac{1}{\sqrt{M}}\sum_{i=1}^M \|g(\boldsymbol{\theta}_1^{(i)}) - g(\boldsymbol{\theta}_2^{(i)})\| \\
&\leq \frac{L_g}{\sqrt{M}}\sum_{i=1}^M \|\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_2^{(i)}\| \\
&\leq \frac{L_g}{\sqrt{M}}\sqrt{M}\sqrt{\sum_{i=1}^M \|\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_2^{(i)}\|^2} = L_g\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|
\end{aligned}
$$

As a result, we have:

$$
\begin{aligned}
&\frac{1}{M}\sum_{i=1}^M \left|\mathbb{E}_{\theta_k^{(i)}\sim\mu_k}[g(\theta_k^{(i)})] - \mathbb{E}_{\boldsymbol{\theta}_\tau^{(i)}\sim\rho_\tau}[g(\boldsymbol{\theta}_\tau^{(i)})]\right| \\
\overset{(1)}{=}&\frac{1}{\sqrt{M}}\left|\frac{1}{\sqrt{M}}\sum_{i=1}^M (\mathbb{E}_{\theta_k^{(i)}\sim\mu_k}[g(\theta_k^{(i)})] - \mathbb{E}_{\boldsymbol{\theta}_\tau^{(i)}\sim\rho_\tau}[g(\boldsymbol{\theta}_\tau^{(i)})])\right| \\
=&\frac{1}{\sqrt{M}}\left|\mathbb{E}_{\Theta_k\sim\mu_k}[g_{\boldsymbol{\Theta}}(\Theta_k)] - \mathbb{E}_{\boldsymbol{\Theta}_\tau\sim\rho_\tau}[g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau)]\right| \ ,
\end{aligned}
$$

where (1) holds because $\mathbb{E}_{\theta_k^{(1)}\sim\mu_k}[g(\theta_k^{(1)})] = \cdots = \mathbb{E}_{\theta_k^{(M)}\sim\mu_k}[g(\theta_k^{(M)})]$ for all the particles $\theta_k^{(i)}$, and $\mathbb{E}_{\boldsymbol{\theta}_\tau^{(1)}\sim\rho_\tau}[g(\boldsymbol{\theta}_\tau^{(1)})] = \cdots = \mathbb{E}_{\boldsymbol{\theta}_\tau^{(M)}\sim\rho_\tau}[g(\boldsymbol{\theta}_\tau^{(M)})]$ for all the particles $\boldsymbol{\theta}_\tau^{(i)}$. According to the definition of W_1 metric, we derive that

$$
\begin{aligned}
&W_1(\mu_k, \rho_\tau) \\
=&\sup_{\|g\|_{lip}\leq 1}\frac{1}{M}\sum_{i=1}^M \left|\mathbb{E}_{\theta_k^{(i)}\sim\mu_k}[g(\theta_k^{(i)})] - \mathbb{E}_{\boldsymbol{\theta}_\tau^{(i)}\sim\rho_\tau}[g(\boldsymbol{\theta}_\tau^{(i)})]\right| \\
=&\frac{1}{\sqrt{M}}\sup_{\|g\|_{lip}\leq 1}|\mathbb{E}_{\Theta_k\sim\mu_k}[g_{\boldsymbol{\Theta}}(\Theta_k)] - \mathbb{E}_{\boldsymbol{\Theta}_\tau\sim\rho_\tau}[g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau)]| \\
=&\frac{1}{\sqrt{M}}\sup_{\|g_{\boldsymbol{\Theta}}\|_{lip}\leq 1}|\mathbb{E}_{\Theta_k\sim\mu_k}[g_{\boldsymbol{\Theta}}(\Theta_k)] - \mathbb{E}_{\boldsymbol{\Theta}_\tau\sim\rho_\tau}[g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau)]| \\
&\leq \frac{1}{\sqrt{M}}W_1(\mu_k^{\Theta}, \rho_\tau^{\boldsymbol{\Theta}}) \ ,
\end{aligned}
$$

which completes the proof. ■

Lemma 13 *Assuming $F(\mathbf{0}) = \mathbf{0}$. If F in (9) is Lipschitz with constant L_F, and satisfies the dissipative property that $\langle F(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle \geq m_F \|\boldsymbol{\theta}\|^2 - b$. Then $F_{\boldsymbol{\Theta}}$ in (26) satisfies $\langle F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}), \boldsymbol{\Theta}\rangle \geq (\beta^{-1}m_F - m')\|\boldsymbol{\Theta}\|^2 - \beta^{-1}Mb$, where l' and m' are some positive constants. Besides we have $\mathbb{E}\|F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_1) - F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_2)\|^2 \leq (\sqrt{2}\beta^{-1}L_F + l')\mathbb{E}\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|^2$ if the same settings as in the proof of Theorem 6 is adopted.*

Proof [Proof of Lemma 13]

We will bound $\langle F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}), \boldsymbol{\Theta}\rangle$ by noting that:

$$\begin{aligned}&\langle F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}), \boldsymbol{\Theta}\rangle\\ =&\sum_i^M \left(\beta^{-1}F(\boldsymbol{\theta}^{(i)})\boldsymbol{\theta}^{(i)} + \frac{1}{M}\sum_j^M K(\boldsymbol{\theta}^{(i)}-\boldsymbol{\theta}^{(j)})F(\boldsymbol{\theta}^{(j)})\boldsymbol{\theta}^{(i)} - \frac{1}{M}\sum_j^M \nabla K(\boldsymbol{\theta}^{(i)}-\theta^{(j)})\boldsymbol{\theta}^{(i)}\right)\end{aligned}$$

Notice that:

$$\begin{aligned}\sum_i^M \beta^{-1}F(\boldsymbol{\theta}^{(i)})\boldsymbol{\theta}^{(i)} &\geq \beta^{-1}m_F\sum_i^M \|\boldsymbol{\theta}^{(i)}\|^2 - \beta^{-1}Mb\\ &= \beta^{-1}m_F\|\boldsymbol{\Theta}\|^2 - \beta^{-1}Mb\end{aligned}$$

Furthermore, since it is assumed that $F(0) = \mathbf{0}$, we have:

$$\begin{aligned}&\sum_i^M \frac{1}{M}\sum_j^M K(\boldsymbol{\theta}^{(i)}-\boldsymbol{\theta}^{(j)})F(\boldsymbol{\theta}^{(j)})\boldsymbol{\theta}^{(i)}\\ \geq& -\frac{1}{M}\sum_i^M\sum_j^M L_F\|\boldsymbol{\theta}^{(i)}\|\|\boldsymbol{\theta}^{(j)}\|\\ \geq& -L_F\sum_{i=1}^M \|\boldsymbol{\theta}^{(i)}\|^2 = -L_F\|\boldsymbol{\Theta}\|^2\end{aligned}$$

In addition, since ∇K is an odd function, we have:

$$\begin{aligned}&\sum_i^M \frac{1}{M}\sum_j^M \nabla K(\boldsymbol{\theta}^{(i)}-\boldsymbol{\theta}^{(j)})\boldsymbol{\theta}^{(i)}\\ \geq& -\sum_i^M \frac{1}{M}\sum_j^M \frac{2}{\eta^2}\|\boldsymbol{\theta}^{(i)}-\boldsymbol{\theta}^{(j)}\|\|\boldsymbol{\theta}^{(i)}\|\\ \geq& -\frac{4}{\eta^2}\sum_i^M \|\boldsymbol{\theta}^{(i)}\|^2 = -\frac{4}{\eta^2}\|\boldsymbol{\Theta}\|^2\end{aligned}$$

As a result, we arrive at the following result:

$$\langle F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}), \boldsymbol{\Theta}\rangle \geq (\beta^{-1}m - L_F - \frac{4}{\eta^2})\|\boldsymbol{\Theta}\|^2 - \beta^{-1}Mb\ .$$

Furthermore, for the other conclusion, we have:

$$\begin{aligned}&\mathbb{E}\|F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_1) - F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_2)\|^2 = \mathbb{E}\sum_i^M \left\|\omega_i^1+\omega_i^2+\omega_i^3+\omega_i^4\right\|^2\\ \leq& \ \mathbb{E}\sum_i^M \left(\|\omega_i^1\|+\|\omega_i^2\|+\|\omega_i^3\|+\|\omega_i^4\|\right)^2\\ \leq& \ 4\mathbb{E}\sum_i^M \left(\|\omega_i^1\|^2+\|\omega_i^2\|^2+\|\omega_i^3\|^2+\|\omega_i^4\|^2\right)\end{aligned}$$

where

$$
\begin{aligned}
\sum_i^M \mathbb{E}\|\omega_i^1\|^2 &= \sum_i^M \mathbb{E}\|\beta^{-1}F(\boldsymbol{\theta}_1^{(i)}) - \beta^{-1}F(\boldsymbol{\theta}_2^{(i)})\|^2 \le \sum_i^M \beta^{-2}L_F^2\mathbb{E}\|\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_2^{(i)}\|^2 \\
\sum_i^M \mathbb{E}\|\omega_i^2\|^2 &= \frac{1}{M^2}\mathbb{E}\|\sum_j^M K(\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_1^{(j)})F(\boldsymbol{\theta}_1^{(j)}) - \sum_j^M K(\boldsymbol{\theta}_2^{(i)} - \boldsymbol{\theta}_2^{(j)})F(\boldsymbol{\theta}_1^{(j)})\|^2 \\
&= \frac{1}{M}\sum_j^M \mathbb{E}\|F(\boldsymbol{\theta}_1^{(j)})(K(\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_1^{(j)}) - K(\boldsymbol{\theta}_2^{(i)} - \boldsymbol{\theta}_2^{(j)}))\|^2 = 0 \\
\sum_i^M \mathbb{E}\|\omega_i^3\|^2 &= \sum_i^M \mathbb{E}\|\frac{1}{M}(\sum_j^M K(\boldsymbol{\theta}_2^{(i)} - \boldsymbol{\theta}_2^{(j)})F(\boldsymbol{\theta}_1^{(j)}) - \sum_j^M K(\boldsymbol{\theta}_2^{(i)} - \boldsymbol{\theta}_2^{(j)})F(\boldsymbol{\theta}_2^{(j)}))\|^2 \\
&\le \sum_i^M L_F^2\mathbb{E}\|\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_2^{(i)}\|^2 \\
\sum_i^M \mathbb{E}\|\omega_i^4\|^2 &= \sum_i^M \mathbb{E}\| -\frac{1}{M}(\sum_j^M \nabla K(\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_1^{(j)}) - \sum_j^M \nabla K(\boldsymbol{\theta}_2^{(i)} - \boldsymbol{\theta}_2^{(j)}))\|^2 \\
&\le \sum_i^M \mathbb{E}\frac{L_{\nabla K}^2}{M}(\sum_j^M \|\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_2^{(i)} - (\boldsymbol{\theta}_1^{(j)} - \boldsymbol{\theta}_2^{(j)})\|^2) \\
&\le \sum_i^M \mathbb{E}L_{\nabla K}^2(2\|\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_2^{(i)}\|^2 + \frac{2}{M}\sum_j^M \|\boldsymbol{\theta}_1^{(j)} - \boldsymbol{\theta}_2^{(j)}\|^2) \\
&\le 4L_{\nabla K}^2\sum_i^M \mathbb{E}\|\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_2^{(i)}\|^2
\end{aligned}
$$

Hence, we have

$$\mathbb{E}\|F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_1) - F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_2)\|^2 \le (\beta^{-2}L_F^2 + L_F^2 + 4L_{\nabla K}^2)\mathbb{E}\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|^2$$

■

Now it is ready to prove Theorem 7. It is worth noting that with the assumption of $F(0) = \mathbf{0}$, the first bullet in Assumption 1 recovers the dissipative assumption as $\langle F(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle \ge m_F\|\boldsymbol{\theta}\|^2$.

Proof We use Lemma C.5 in [Xu et al., 2018] to verify that $F_{\boldsymbol{\Theta}}$ satisfies the assumptions in [Raginsky et al., 2017] by setting $\delta = \frac{a'}{B}$ with a' a positive constant and B the size of the random set $\mathcal{I}$.

Let $\mu_k^{\Theta} := \mathcal{L}(\Theta_k)$ and $\rho_\tau^{\boldsymbol{\Theta}} := \mathcal{L}(\boldsymbol{\Theta}_\tau)$. We make some modifications to the proof of Lemma 3.6 in [Raginsky et al., 2017] and derive the following results. The relative entropy $D_{KL}(\mu_k^{\Theta}\|\rho_{kh}^{\boldsymbol{\Theta}})$ satisfies:

$$D_{KL}(\mu_k^{\Theta}\|\rho_{kh}^{\boldsymbol{\Theta}}) \le (A_0\beta\frac{a\prime}{B} + A_1 h)kh$$

with

$$
\begin{aligned}
A_0 &= \left(2(\beta^{-2}L_F^2 + L_F^2 + 4L_{\nabla K}^2)\left(a_2 + 2(1 \vee \frac{1}{\beta^{-1}m_F - m'}) \cdot (2a_1^2 + \frac{Md}{\beta})\right) + a_1^2\right) \\
A_1 &= 6(\beta^{-2}L_F^2 + L_F^2 + 4L_{\nabla K}^2)(\beta A_0 + Md)
\end{aligned}
$$

and a_1, a_2 are some positive constants. When β is small enough such that the subtraction terms in the above bounds are positive, there exist some positive constants a_3, a_4 such that

$$A_0 \le a_3\frac{Md}{\beta^3}, \text{ and } A_1 \le a_4\frac{Md}{\beta^4}$$

Similar to the proof of Lemma 13, it is easy to verify that there exists some positive constant a_5 such that $\langle F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_1) - F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_2), \boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2 \rangle \geq (\beta^{-1} m_F - a_5) \|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|^2$. Note that when β is small enough, (26) satisfies the conditions of Proposition 4.2 in [Cattiaux et al., 2008]. Hence, there exits some positive constant $\mathcal{C}$ such that $W_1(\mu_k^{\Theta}, \rho_{kh}^{\boldsymbol{\Theta}}) \leq \mathcal{C}\sqrt{D_{KL}(\mu_k^{\Theta} \| \rho_{kh}^{\boldsymbol{\Theta}})}$.

According to Corollary 4 and Lemma 8 in [Bolley and Villani, 2005], we can derive an explicit expression for $\mathcal{C}$:

$$\mathcal{C} \leq a_6 \beta^{-1} M d \ ,$$

when β is a small enough constant and a_6 is some positive constant.

Applying Lemma 12, we have

$$\begin{aligned}
W_1(\mu_k, \rho_{kh}) \leq & \frac{1}{\sqrt{M}} W_1(\mu_k^{\Theta}, \rho_{kh}^{\boldsymbol{\Theta}}) \\
\leq & a_6 M d^{\frac{3}{2}} \beta^{-3} (a_3 a' \beta^2 B^{-1} + a_4 h)^{\frac{1}{2}} k^{\frac{1}{2}} h^{\frac{1}{2}}
\end{aligned}$$

Setting $k = T$ completes the poof. ∎

G Proof of Theorem 8

Proof Our proof is based on the techniques in the proof of Lemma 3.6 in [Raginsky et al., 2017]. Firstly, adopting the same notation as in Section F, we have the following update:

$$\Theta_{k+1} = \Theta_k - \beta^{-1} G_{\mathcal{I}_k}^{\Theta} h_k + \sqrt{2\beta^{-1} h_k} \Xi_k \ , \tag{28}$$

where $\Xi_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{Md \times Md})$ and $h_k = \frac{h_0}{k+1}$. We note the unbiasness of $G_{\mathcal{I}_k}^{\Theta}$, *i.e.*, $\mathbb{E}(G_{\mathcal{I}_k}^{\Theta}) = F_{\boldsymbol{\Theta}}(\Theta_k), \ \ \forall \boldsymbol{\Theta} \in \mathbb{R}^{Md}$, due to the way we choose the minibatch $\mathcal{I}_k$. We need to define $q(\tau)$, which will be used in the following proof:

$$q(\tau) = \{k \in \mathbb{R} | \sum_{i=0}^{k-1} h_i \leq \tau < \sum_{i=0}^{k} h_i\} \ .$$

Furthermore, define $\sum_{i=0}^{-1} h_i \triangleq 0$ and $\sum_{i=0}^{0} h_i \triangleq h_0$ for the convenience of statement in the following.

Now we focus on the following continuous-time interpolation of Θ_k:

$$\underline{\boldsymbol{\Theta}}(\tau) = \boldsymbol{\Theta}_0 - \int_0^{\tau} \tilde{G}_{\mathcal{I}(s)}^{\boldsymbol{\Theta}} \left(\underline{\boldsymbol{\Theta}}(\sum_{i=0}^{q(s)-1} h_i) \right) \mathrm{d}s + \sqrt{\frac{2}{\beta}} \int_0^{\tau} \mathcal{W}_s^{(Md)},$$

where $\mathcal{I}(s) \equiv \mathcal{I}_k$ for $\tau \in \left[\sum_{i=0}^{k-1} h_i, \sum_{i=0}^{k} h_i\right)$, $\tilde{G}_{\mathcal{I}(s)}^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) \triangleq \frac{N}{B(s)} \sum_{q \in \mathcal{I}(s)} F_{(q)\boldsymbol{\Theta}}(\boldsymbol{\Theta})$ and $B(s)$ is the size of the minibatch $\mathcal{I}(s)$. It is easily seen that for each k, $\underline{\boldsymbol{\Theta}}(\sum_{i=0}^{k-1} h_i)$ and Θ_k have the same probability law ρ_k^{Θ}. Besides we need some similar settings in the proof of Theorem 6 for $\mathcal{W}_s^{(Md)}$. Since $\underline{\boldsymbol{\Theta}}(\tau)$ is not a Markov process, we define the following Itô process which has the same one-time marginals as $\underline{\boldsymbol{\Theta}}(\tau)$

$$\Lambda(\tau) = \Theta_0 - \int_0^{\tau} \underline{G}_s \left(\Lambda(s)\right) \mathrm{d}s + \sqrt{\frac{2}{\beta}} \int_0^{\tau} \mathcal{W}_s^{(Md)}$$

$$\text{where } \underline{G}_{\tau}(x) := \mathbb{E}\left[\tilde{G}_{\mathcal{I}(\tau)}^{\boldsymbol{\Theta}} \left(\underline{\boldsymbol{\Theta}}(\sum_{i=0}^{q(\tau)-1} h_i) \right) | \underline{\boldsymbol{\Theta}}(\tau) = x \right] \ .$$

Let the probability laws $\mathbf{P}_{\Lambda}^{\tau} := \mathcal{L}\left(\Lambda(s) : 0 \leq s \leq \tau\right)$ and $\mathbf{P}_{\boldsymbol{\Theta}}^{\tau} := \mathcal{L}\left(\boldsymbol{\Theta}(s) : 0 \leq s \leq \tau\right)$. According to the proof of lemma 3.6 in [Raginsky et al., 2017], we can derive a similar result for the relative entropy of $\mathbf{P}_{\Lambda}^{\tau}$ and $\mathbf{P}_{\boldsymbol{\Theta}}^{\tau}$:

$$\begin{aligned}
D_{KL}(\mathbf{P}_{\Lambda}^{\tau} \,\|\, \mathbf{P}_{\boldsymbol{\Theta}}^{\tau}) &= -\int \mathrm{d}\,\mathbf{P}_{\Lambda}^{\tau} \log \frac{\mathrm{d}\,\mathbf{P}_{\Lambda}^{\tau}}{\mathrm{d}\,\mathbf{P}_{\boldsymbol{\Theta}}^{\tau}} \\
=& \frac{\beta}{4} \int_{0}^{\tau} \mathbb{E}\|F_{\boldsymbol{\Theta}}(\Lambda(s)) - \underline{G}_{s}\left(\Lambda(s)\right)\|^{2} \mathrm{d}s \\
=& \frac{\beta}{4} \int_{0}^{\tau} \mathbb{E}\|F_{\boldsymbol{\Theta}}(\underline{\boldsymbol{\Theta}}(s)) - \underline{G}_{s}\left(\underline{\boldsymbol{\Theta}}(s)\right)\|^{2} \mathrm{d}s \ ,
\end{aligned}$$

where the last line follows because $\mathcal{L}(\underline{\boldsymbol{\Theta}}(s)) = \mathcal{L}(\Lambda(s)),\ \forall s$.

In the following proof, we let $\tau = \sum_{i=0}^{k-1} h_i$ for some $k \in \mathbb{R}$. Now we can use the martingale property (conditional independence) of Itô integral to derive:

$$\begin{aligned}
& D_{KL}(\mathbf{P}_{\Lambda}^{\sum_{i=0}^{k-1} h_i} \,\|\, \mathbf{P}_{\boldsymbol{\Theta}}^{\sum_{i=0}^{k-1} h_i}) \\
&= \frac{\beta}{4} \sum_{j=0}^{k-1} \int_{\sum_{i=0}^{j-1} h_i}^{\sum_{i=0}^{j} h_i} \mathbb{E}\|F_{\boldsymbol{\Theta}}(\underline{\boldsymbol{\Theta}}(s)) - \underline{G}_{s}\left(\underline{\boldsymbol{\Theta}}(s)\right)\|^{2} \mathrm{d}s \\
&\leq \frac{\beta}{2} \sum_{j=0}^{k-1} \int_{\sum_{i=0}^{j-1} h_i}^{\sum_{i=0}^{j} h_i} \mathbb{E}\|F_{\boldsymbol{\Theta}}(\underline{\boldsymbol{\Theta}}(s)) - F_{\boldsymbol{\Theta}}(\underline{\boldsymbol{\Theta}}(\sum_{i=0}^{q(s)-1} h_i))\|^{2} \mathrm{d}s \\
&\quad + \frac{\beta}{2} \sum_{j=0}^{k-1} \int_{\sum_{i=0}^{j-1} h_i}^{\sum_{i=0}^{j} h_i} \mathbb{E}\left\| F_{\boldsymbol{\Theta}}(\underline{\boldsymbol{\Theta}}(\sum_{i=0}^{q(s)-1} h_i)) - \tilde{G}_{\mathcal{I}(s)}^{\boldsymbol{\Theta}}\left(\underline{\boldsymbol{\Theta}}(\sum_{i=0}^{q(s)-1} h_i)\right) \right\|^{2} \mathrm{d}s \\
&\leq \frac{\beta L_{F_{\boldsymbol{\Theta}}}^{2}}{2} \sum_{j=0}^{k-1} \int_{\sum_{i=0}^{j-1} h_i}^{\sum_{i=0}^{j} h_i} \mathbb{E}\|\underline{\boldsymbol{\Theta}}(s) - \underline{\boldsymbol{\Theta}}(\sum_{i=0}^{q(s)-1} h_i)\|^{2} \mathrm{d}s \qquad (29) \\
&+ \frac{\beta}{2} \sum_{j=0}^{k-1} \int_{\sum_{i=0}^{j-1} h_i}^{\sum_{i=0}^{j} h_i} \mathbb{E}\left\| F_{\boldsymbol{\Theta}}(\underline{\boldsymbol{\Theta}}(\sum_{i=0}^{q(s)-1} h_i)) - \tilde{G}_{\mathcal{I}(s)}^{\boldsymbol{\Theta}}\left(\underline{\boldsymbol{\Theta}}(\sum_{i=0}^{q(s)-1} h_i)\right) \right\|^{2} \mathrm{d}s \ , \qquad (30)
\end{aligned}$$

where $L_{F_{\boldsymbol{\Theta}}}^{2} \triangleq \beta^{-2} L_{F}^{2} + L_{F}^{2} + 4L_{\nabla K}^{2}$.

For the first part of (29), consider some $s \in [\sum_{i=0}^{j-1} h_i, \sum_{i=0}^{j} h_i)$. From the definitions, the following equation holds:

$$\begin{aligned}
& \underline{\boldsymbol{\Theta}}(s) - \underline{\boldsymbol{\Theta}}(\sum_{i=0}^{j-1} h_i) \\
=& -(s - \sum_{i=0}^{j-1} h_i) G_{\mathcal{I}_j}^{\Theta} + \sqrt{2/\beta}(\mathcal{W}_{s}^{(Md)} - \mathcal{W}_{\sum_{i=0}^{j-1} h_i}^{(Md)}) \\
=& -(s - \sum_{i=0}^{j-1} h_i) G_{\mathcal{I}_j}^{\Theta} + (s - \sum_{i=0}^{j-1} h_i)(F_{\boldsymbol{\Theta}}(\Theta_j) - G_{\mathcal{I}_j}^{\Theta}) + \sqrt{2/\beta}(\mathcal{W}_{s}^{(Md)} - \mathcal{W}_{\sum_{i=0}^{j-1} h_i}^{(Md)})
\end{aligned}$$

Applying results from Lemma 3.1 and 3.2 in [Raginsky et al., 2017], and Lemma C.5 in [Xu et al., 2018], we have:

$$\begin{aligned}
& \mathbb{E}\|\underline{\boldsymbol{\Theta}}(s) - \underline{\boldsymbol{\Theta}}(\sum_{i=0}^{j-1} h_i)\|^{2} \\
\leq& 3\frac{{h_0}^{2}}{(j+1)^{2}} \mathbb{E}\|G_{\mathcal{I}_j}^{\Theta}\|^{2} + 3\frac{{h_0}^{2}}{(j+1)^{2}} \mathbb{E}\|F_{\boldsymbol{\Theta}}(\Theta_j) - G_{\mathcal{I}_j}^{\Theta}\|^{2} + \frac{6h_0 Md}{\beta(j+1)} \\
\leq& 12\frac{{h_0}^{2}}{(j+1)^{2}} \max_{0 \leq j \leq k-1} (L_{F_{\boldsymbol{\Theta}}}^{2} \mathbb{E}\|\Theta_j\|^{2} + b_1) + \frac{6h_0 Md}{\beta(j+1)}
\end{aligned}$$

where b_1 is some positive constant.

Consequently, the first part of (29) can be bounded as:

$$\begin{aligned}
&\frac{\beta L_{F_{\boldsymbol{\Theta}}}^2}{2}\sum_{j=0}^{k-1}\int_{\sum_{i=0}^{j-1}h_i}^{\sum_{i=0}^{j}h_i}\mathbb{E}\|\underline{\boldsymbol{\Theta}}(s)-\underline{\boldsymbol{\Theta}}(\sum_{i=0}^{q(s)-1}h_i)\|^2\mathrm{d}s\\
\leq&\frac{\beta L_{F_{\boldsymbol{\Theta}}}^2}{2}\sum_{j=0}^{k-1}\left[12\frac{{h_0}^3}{(j+1)^3}\max_{0\leq j\leq K-1}(L_{F_{\boldsymbol{\Theta}}}^2\mathbb{E}\|\Theta_j\|^2+b_1)+\frac{6{h_0}^2Md}{\beta(j+1)^2}\right]\\
\leq&\pi^2\beta L_{F_{\boldsymbol{\Theta}}}^2{h_0}^3\max_{0\leq j\leq K-1}(L_{F_{\boldsymbol{\Theta}}}^2\mathbb{E}\|\Theta_j\|^2+b_1)+\frac{\pi^2L_{F_{\boldsymbol{\Theta}}}^2{h_0}^2Md}{2}\ ,
\end{aligned}$$

where the last inequality follows from the fact that

$$\sum_{j=0}^{k-1}\frac{1}{(j+1)^3}\leq\sum_{j=0}^{k-1}\frac{1}{(j+1)^2}\leq\sum_{j=0}^{\infty}\frac{1}{(j+1)^2}=\frac{\pi^2}{6}\ .$$

Now we bound the second part (30). According to Lemma C.5 in [Xu et al., 2018], we have:

$$\begin{aligned}
&\frac{\beta}{2}\sum_{j=0}^{k-1}\int_{\sum_{i=0}^{j-1}h_i}^{\sum_{i=0}^{j}h_i}\mathbb{E}\left\|F_{\boldsymbol{\Theta}}(\underline{\boldsymbol{\Theta}}(\sum_{i=0}^{q(s)-1}h_i))-\tilde{G}_{\mathcal{I}(s)}^{\boldsymbol{\Theta}}\left(\underline{\boldsymbol{\Theta}}(\sum_{i=0}^{q(s)-1}h_i)\right)\right\|^2\mathrm{d}s\\
=&\sum_{j=0}^{k-1}\frac{\beta h_0}{2(j+1)}\mathbb{E}\|F_{\boldsymbol{\Theta}}(\Theta_j)-G_{\mathcal{I}_j}^{\Theta}\|^2\\
\leq&\beta h_0\max_{0\leq j\leq k-1}(L_{F_{\boldsymbol{\Theta}}}^2\mathbb{E}\|\Theta_j\|^2+b_1)\cdot\left(\frac{4}{B_0}+\sum_{j=1}^{k-1}\frac{4}{(j+1)(B_0+\log^{\frac{100}{99}}(j+1))}\right)\\
\leq&\beta h_0\max_{0\leq j\leq k-1}(L_{F_{\boldsymbol{\Theta}}}^2\mathbb{E}\|\Theta_j\|^2+b_1)\cdot\left(\frac{4}{B_0}+\sum_{j=1}^{k-1}\frac{4}{(j+1)\log^{\frac{100}{99}}(j+1)}\right)\\
\leq&(b_2+\frac{4}{B_0})\beta h_0\max_{0\leq j\leq k-1}(L_{F_{\boldsymbol{\Theta}}}^2\mathbb{E}\|\Theta_j\|^2+b_1)\ ,
\end{aligned}$$

where the last inequality follows from the fact that when $r>1$,

$$\begin{aligned}
&\sum_{j=1}^{k-1}\frac{4}{(j+1)\log^r(j+1)}\\
\leq&\sum_{j=1}^{\infty}\frac{4}{(j+1)\log^r(j+1)}\leq\frac{4\log^{1-r}2}{r-1}\ .
\end{aligned}$$

Denote $\mu_k^{\Theta}:=\mathcal{L}(\Theta_k)$ and $\rho_\tau^{\boldsymbol{\Theta}}:=\mathcal{L}(\boldsymbol{\Theta}_\tau)$. Due to the data-processing inequality for the relative entropy, we have

$$\begin{aligned}
&D_{KL}(\mu_k^{\Theta}\|\rho_{\sum_{i=0}^{k-1}h_i}^{\boldsymbol{\Theta}})\leq D_{KL}(\mathbf{P}_\Lambda^{\sum_{i=0}^{k-1}h_i}\parallel\mathbf{P}_{\boldsymbol{\Theta}}^{\sum_{i=0}^{k-1}h_i})\\
&\leq\pi^2\beta L_{F_{\boldsymbol{\Theta}}}^2{h_0}^3\max_{0\leq j\leq k-1}(L_{F_{\boldsymbol{\Theta}}}^2\mathbb{E}\|\Theta_j\|^2+b_1)+\frac{\pi^2L_{F_{\boldsymbol{\Theta}}}^2{h_0}^2Md}{2}+(b_2+\frac{4}{B_0})\beta h_0\max_{0\leq j\leq k-1}(L_{F_{\boldsymbol{\Theta}}}^2\mathbb{E}\|\tilde{\boldsymbol{\Theta}}_j\|^2+b_1)\\
&\leq(\pi^2\beta L_{F_{\boldsymbol{\Theta}}}^2{h_0}^3+b_2\beta h_0+\frac{4}{B_0}\beta h_0)\max_{0\leq j\leq k-1}(L_{F_{\boldsymbol{\Theta}}}^2\mathbb{E}\|\Theta_j\|^2+b_1)+\frac{\pi^2L_{F_{\boldsymbol{\Theta}}}^2{h_0}^2Md}{2}\ .
\end{aligned}$$

Theorem 14 has provided a uniform bound to $\max_{0\leq j\leq k-1}(L_{F_{\boldsymbol{\Theta}}}^2\mathbb{E}\|\Theta_j\|^2+b_1)$. Hence it can be concluded that $D_{KL}(\mathbf{P}_\Lambda^{\sum_{i=0}^{k-1}h_i}\parallel\mathbf{P}_{\boldsymbol{\Theta}}^{\sum_{i=0}^{k-1}h_i})$ would not increase w.r.t. k. This is a nice property that the fixed-step-size SPOS does not endow. Since $L^2_{F_{\boldsymbol{\Theta}}} \triangleq \beta^{-2}L_F^2 + L_F^2 + 4L^2_{\nabla K}$, it is easy to verify that when β is small enough, there exists some positive constants b_3, b_4, b_5 and b_6 such that:

$$\begin{aligned}
&D_{KL}(\mu_k^{\Theta}\|\rho^{\boldsymbol{\Theta}}_{\sum_{i=0}^{k-1}h_i}) \\
\leq&(\pi^2\beta L^2_{F_{\boldsymbol{\Theta}}}{h_0}^3 + b_2\beta h_0 + \frac{4\beta h_0}{B_0}) \times \max_{0\leq j\leq K-1}(L^2_{F_{\boldsymbol{\Theta}}}\mathbb{E}\|\tilde{\boldsymbol{\Theta}}_j\|^2 + b_1) + \frac{\pi^2 L^2_{F_{\boldsymbol{\Theta}}}{h_0}^2 Md}{2} \\
\leq&(b_3h_0^3 + \frac{b_4\beta^3 h_0}{B_0} + b_5h_0^2\beta^2)\frac{Md}{\beta^4}\ .
\end{aligned}$$

Similar to the proof of Theorem 7, we can bound the $W_1(\mu_k^{\Theta}\|\rho^{\boldsymbol{\Theta}}_{\sum_{i=0}^{k-1}h_i})$ term with Corollary 4, Lemma 8 in [Bolley and Villani, 2005] and Proposition 4.2 in [Cattiaux et al., 2008]. Specifically, when β is small enough, there exist some positive constant a_6 such that:

$$\begin{aligned}
&W_1(\mu_k^{\Theta}\|\rho^{\boldsymbol{\Theta}}_{\sum_{i=0}^{k-1}h_i}) \leq a_6(\frac{Md}{\beta})\sqrt{D_{KL}(\mu_k^{\Theta}\|\rho^{\boldsymbol{\Theta}}_{\sum_{i=0}^{k-1}h_i})} \\
\leq&a_6\beta^{-3}M^{\frac{3}{2}}d^{\frac{3}{2}}(b_3h_0^3 + \frac{b_4\beta^3 h_0}{B_0} + b_5h_0^2\beta^2)^{\frac{1}{2}}\ .
\end{aligned}$$

According to Lemma 12, we have

$$\begin{aligned}
&W_1(\mu_k, \rho_{kh}) \leq \frac{1}{\sqrt{M}}W_1(\mu_k^{\Theta}\|\rho^{\boldsymbol{\Theta}}_{\sum_{i=0}^{k-1}h_i}) \\
=&a_6\beta^{-3}Md^{\frac{3}{2}}(b_3h_0^3 + \frac{b_4\beta^3 h_0}{B_0} + b_5h_0^2\beta^2)^{\frac{1}{2}}
\end{aligned}$$

Setting $k = T$ finishes the proof. ■

H Proof of Theorems 2 and 4

Proof [Proof for Theorem 2] The proof is by direct calculation:

$$\begin{aligned}
&\mathrm{d}\left(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right) = \\
&\frac{1}{M}\sum_q^M\left[\nabla K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(q)}) - \nabla K(\boldsymbol{\theta}_\tau^{(j)} - \boldsymbol{\theta}_\tau^{(q)})\right]\mathrm{d}\tau \\
&-\frac{1}{M}\sum_q^M\left(F(\boldsymbol{\theta}_\tau^{(q)})K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(q)}) - F(\boldsymbol{\theta}_\tau^{(q)})K(\boldsymbol{\theta}_\tau^{(j)} - \boldsymbol{\theta}_\tau^{(q)}))\right)\mathrm{d}\tau
\end{aligned}$$

$$\begin{aligned}
&\Rightarrow\mathrm{d}\left(\mathbb{E}\sum_{ij}^M\left\|\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right\|^2\right) = \\
&\mathbb{E}\sum_{ij}^M\frac{2}{M}\sum_q^M\left[\nabla K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(q)}) - \nabla K(\boldsymbol{\theta}_\tau^{(j)} - \boldsymbol{\theta}_\tau^{(q)})\right]\times\left(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right)\mathrm{d}\tau \\
&-\mathbb{E}\sum_{ij}^M\frac{2}{M}\sum_q^M\left(F(\boldsymbol{\theta}_\tau^{(q)})K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(q)}) - F(\boldsymbol{\theta}_\tau^{(q)})K(\boldsymbol{\theta}_\tau^{(j)} - \boldsymbol{\theta}_\tau^{(q)}))\right)\left(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right)\mathrm{d}\tau \\
&\leq -2m_K\mathbb{E}\sum_{ij}^M\left\|\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right\|^2\mathrm{d}\tau + 2H_FL_K\mathbb{E}\sum_{ij}^M\left\|\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right\|^2\mathrm{d}\tau\ ,
\end{aligned}$$

where H_F is the maximum value of $\|F(\theta)\|$ on the bounded space. Denote $z(\tau) = \mathbb{E}\sum_{ij}^M \left\|\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right\|^2$. We have

$$z(\tau)' \leq -(2m_K - 2H_F L_K)z(\tau) \tag{31}$$

Applying Gronwall Lemma on (31) finishes the proof. ■

Proof [Proof of Theorem 4]

For the SPOS, we have

$$\begin{aligned}
&\mathrm{d}\left(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right) = -\beta^{-1}\left(F(\boldsymbol{\theta}_\tau^{(i)}) - F(\boldsymbol{\theta}_\tau^{(j)})\right)\mathrm{d}\tau \\
&+ \frac{1}{M}\sum_q^M \left[\nabla K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(q)}) - \nabla K(\boldsymbol{\theta}_\tau^{(j)} - \boldsymbol{\theta}_\tau^{(q)})\right]\mathrm{d}\tau \\
&- \frac{1}{M}\sum_q^M \left(F(\boldsymbol{\theta}_\tau^{(q)})K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(q)}) - F(\boldsymbol{\theta}_\tau^{(q)})K(\boldsymbol{\theta}_\tau^{(j)} - \boldsymbol{\theta}_\tau^{(q)}))\right)\mathrm{d}\tau \\
&+ \sqrt{\frac{2}{\beta}}(\mathrm{d}\mathcal{W}_\tau^{(i)} - \mathrm{d}\mathcal{W}_\tau^{(j)})
\end{aligned}$$

Hence we have

$$\begin{aligned}
&\Rightarrow \mathrm{d}\left(\mathbb{E}\sum_{ij}^M \left\|\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right\|^2\right) \qquad (32)\\
=&-\mathbb{E}2\sum_{ij}^M \beta^{-1}\left(F(\boldsymbol{\theta}_\tau^{(i)}) - F(\boldsymbol{\theta}_\tau^{(j)})\right)\left(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right)\mathrm{d}\tau \\
&+\mathbb{E}\sum_{ij}^M \frac{2}{M}\sum_q^M \left[\nabla K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(q)}) - \nabla K(\boldsymbol{\theta}_\tau^{(j)} - \boldsymbol{\theta}_\tau^{(q)})\right]\left(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right)\mathrm{d}\tau \\
&-\mathbb{E}\sum_{ij}^M \frac{2}{M}\sum_q^M \left(F(\boldsymbol{\theta}_\tau^{(q)})K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(q)}) - F(\boldsymbol{\theta}_\tau^{(q)})K(\boldsymbol{\theta}_\tau^{(j)} - \boldsymbol{\theta}_\tau^{(q)}))\right)\left(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right)\mathrm{d}\tau \\
&+\mathbb{E}2\sum_{ij}^M \sqrt{\frac{2}{\beta}}(\mathrm{d}\mathcal{W}_\tau^{(i)} - \mathrm{d}\mathcal{W}_\tau^{(j)})\left(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right) \\
&\leq -2\beta^{-1}m_F\mathbb{E}\sum_{ij}^M \left\|\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right\|^2\mathrm{d}\tau \\
&-2m_K\mathbb{E}\sum_{ij}^M \left\|\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right\|^2\mathrm{d}\tau + 2H_F L_K\mathbb{E}\sum_{ij}^M \left\|\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right\|^2\mathrm{d}\tau \\
&+2\sqrt{\frac{2}{\beta}}\left(\mathbb{E}\sum_{ij}^M(\mathrm{d}\mathcal{W}_\tau^{(i)} - \mathrm{d}\mathcal{W}_\tau^{(j)})^2\right)^{1/2}\left(\mathbb{E}\sum_{ij}^M \left\|\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right\|^2\right)^{1/2}.
\end{aligned}$$

Denote $z(\tau) = \mathbb{E}\sum_{ij}^M \left\|\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\right\|^2$. We have

$$z(\tau)' \leq -(2\beta^{-1}m_F + 2m_K - 2H_F L_K)z(\tau) + 4M\sqrt{\frac{d}{\beta}z(\tau)} \tag{33}$$

Applying Gronwall Lemma on (33) finished the proof. ■

Based on the bound, we can see that the particles in SPOS will not converge to one point, overcoming the pitfall of SVGD.

I The uniform bounds on SPOS

Following the formulations in Section F, we will derive the following theorem about the uniform bound on each particle in SPOS.

Theorem 14 *For the Θ_k in (27), we have*

$$E\|\Theta_k\|^2 \leq M\gamma_0 + (1 \vee \frac{1}{m'})\frac{2Md}{\beta}$$

First we need to prove the following lemma.

Lemma 15 *For the $F_{\boldsymbol{\Theta}}$ in the (26) and Θ_k in the (27), we have the following result*

$$\|F_{\boldsymbol{\Theta}}(\Theta_k)\|^2 \leq (3\beta^{-2}L_F^2 + 3L_F^2 + 48/\eta^4)\|\Theta_k\|^2 \tag{34}$$

Proof [Proof of Lemma 15]

$$\begin{aligned}\|F_{\boldsymbol{\Theta}}(\Theta_k)\|^2 =& \sum_i^M \left\|\omega_i^1 + \omega_i^2 + \omega_i^3\right\|^2 \\ &\leq \sum_i^M \left(\|\omega_i^1\| + \|\omega_i^2\| + \|\omega_i^3\|\right)^2 \leq 3\sum_i^M \left(\|\omega_i^1\|^2 + \|\omega_i^2\|^2 + \|\omega_i^3\|^2\right)\end{aligned}$$

where

$$\begin{aligned}\|\omega_i^1\| &= \|\beta^{-1}F(\boldsymbol{\theta}_k^{(i)})\| \leq \beta^{-1}L_F\|\boldsymbol{\theta}_k^{(i)}\| \\ \|\omega_i^2\| &= \|\frac{1}{M}\sum_j^M K(\boldsymbol{\theta}_k^{(i)} - \boldsymbol{\theta}_k^{(j)})F(\boldsymbol{\theta}_k^{(j)})\| \\ &\leq L_F\frac{1}{M}\sum_j^M \|\boldsymbol{\theta}_k^{(j)}\| \\ \|\omega_i^3\| &= \| - \frac{1}{M}\sum_j^M \nabla K(\boldsymbol{\theta}_k^{(i)} - \boldsymbol{\theta}_k^{(j)})\| \\ &\leq \frac{2/\eta^2}{M}\sum_j^M \|\boldsymbol{\theta}_k^{(i)} - \boldsymbol{\theta}_k^{(j)}\| \\ &\leq \frac{2}{\eta^2}(\|\boldsymbol{\theta}_k^{(i)}\| + \frac{1}{M}\sum_j^M \|\boldsymbol{\theta}_k^{(j)}\|)\end{aligned}$$

Substituting the above bounds into $F_{\boldsymbol{\Theta}}(\Theta_k)$, it is easy to verify that

$$\begin{aligned}\|F_{\boldsymbol{\Theta}}(\Theta_k)\|^2 \leq& 3\sum_i^M \left(\beta^{-2}L_F^2\|\boldsymbol{\theta}_k^{(i)}\|^2 + \frac{L_F^2}{M}\sum_j^M \|\boldsymbol{\theta}_k^{(j)}\|^2 + 2(2/\eta^2)^2\|\boldsymbol{\theta}_k^{(i)}\|^2 + \frac{2(2/\eta^2)^2}{M}\sum_j^M \|\boldsymbol{\theta}_k^{(j)}\|^2\right) \\ &\leq (3\beta^{-2}L_F^2 + 3L_F^2 + 48/\eta^4)\|\Theta_k\|^2\end{aligned}$$

With the Lemma 13 and 15, we can now derive the the uniform bound on each particle in SPOS. Our proof is based on the proof of Lemma 3.2 in [Raginsky et al., 2017]

Proof [Proof of Theorem 14] From (27), it follows that

$$\begin{aligned} E\|\Theta_{k+1}\|^2 =& E\|\Theta_k - G^{\Theta}_{\mathcal{I}_k} h_k\|^2 + \sqrt{\frac{8h_k}{\beta}} E\langle \Theta_k - G^{\Theta}_{\mathcal{I}_k} h_k, \Xi_k\rangle + \frac{2h_k}{\beta} E\|\Xi_k\|^2 \\ =& E\|\Theta_k - G^{\Theta}_{\mathcal{I}_k} h_k\|^2 + \frac{2h_k M d}{\beta} \end{aligned}$$

where the second step uses independence of $\Theta_k - G^{\Theta}_{\mathcal{I}_k} h_k$ and Ξ_k,the unbiasedness property that $E[G^{\Theta}_{\mathcal{I}_k}] = F_{\boldsymbol{\Theta}}(\Theta_k)$ and $E[\Xi_k] = 0$

$$\begin{aligned} E\|\Theta_k - G^{\Theta}_{\mathcal{I}_k} h_k\|^2 &= E\|\Theta_k - F_{\boldsymbol{\Theta}}(\Theta_k) h_k\|^2 + 2h_k E\langle \Theta_k - F_{\boldsymbol{\Theta}}(\Theta_k) h_k, F_{\boldsymbol{\Theta}}(\Theta_k) - G^{\Theta}_{\mathcal{I}_k}\rangle + h_k^2 E\|F_{\boldsymbol{\Theta}}(\Theta_k) - G^{\Theta}_{\mathcal{I}_k}\|^2 \\ &= E\|\Theta_k - F_{\boldsymbol{\Theta}}(\Theta_k) h_k\|^2 + h_k^2 E\|F_{\boldsymbol{\Theta}}(\Theta_k) - G^{\Theta}_{\mathcal{I}_k}\|^2 \end{aligned} \tag{35}$$

The first term in (35) can estimated as

$$\begin{aligned} E\|\Theta_k - F_{\boldsymbol{\Theta}}(\Theta_k) h_k\|^2 &= E\|\Theta_k\|^2 - 2h_k E\langle \Theta_k, F_{\boldsymbol{\Theta}}(\Theta_k)\rangle + h_k^2 E\|F_{\boldsymbol{\Theta}}(\Theta_k)\|^2 \\ &\le E\|\Theta_k\|^2 + 2h_k(-(\beta^{-1}m - L_F - \frac{4}{\eta^2}) E\|\Theta_k\|^2) + h_k^2(3\beta^{-2}L_F^2 + 3L_F^2 + 48/\eta^4) E\|\Theta_k\|^2 \\ &\le (1 - 2h_k m'_F + h_k^2 L') E\|\Theta_k\|^2 \end{aligned}$$

where $m' \triangleq \beta^{-1} m_F - L_F - \frac{4}{\eta^2}$ and $L' \triangleq 3\beta^{-2}L_F^2 + 3L_F^2 + 48/\eta^4$.

Following the Lemma C.5 from [Xu et al., 2018] and some modifications (the settings are a bit different,but the results are the same), we could estimate the the second term in (35) as

$$E\|F_{\boldsymbol{\Theta}}(\Theta_k) - G^{\Theta}_{\mathcal{I}_k}\|^2 \le \frac{2(N-B)}{B(N-1)} L' E\|\Theta_k\|^2 \le 2L' E\|\Theta_k\|^2 \tag{36}$$

Now we can derive that

$$E\|\Theta_{k+1}\|^2 \le (1 - 2h_k m' + 3h_k^2 L') E\|\Theta_k\|^2 + \frac{2h_k M d}{\beta}$$

Fix some $0 < h_0 \le 1 \wedge \frac{m'}{3L'}$, we will show that $\forall k$

$$E\|\Theta_k\|^2 \le E\|\Theta_0\|^2 + (1 \vee \frac{1}{m'})\frac{2Md}{\beta} = M\gamma_0 + (1 \vee \frac{1}{m'})\frac{2Md}{\beta} \tag{37}$$

First, it is easy to see that $(1 - 2h_k m' + 3h_k^2 L')$ increases with the decrease of h_k. Suppose $k^\star$ is the last k that satisfies $(1 - 2h_k m' + 3h_k^2 L') \le 0$, and $\forall k \le k^\star$, $E\|\Theta_k\|^2$ satisfies (37).

Then we will see that if $E\|\Theta_{k-1}\|^2 \le S (k > k^\star)$ and $S > \frac{2Md}{\beta}$, then $E\|\Theta_k\|^2 \le S$.

$$E\|\Theta_k\|^2 \le (1 - 2h_k m' + 3h_k^2 L') S + \frac{2h_k M d}{\beta} \le S - S(2h_k m' - 3h_k^2 L') + \frac{2Md}{\beta} < S$$

Since $M\gamma_0 + (1 \vee \frac{1}{m'})\frac{2Md}{\beta} > \frac{2Md}{\beta}$, it is easy to verify that (37) holds. ■

We next prove the following theorem.

Theorem 16 *For the $\boldsymbol{\Theta}_\tau$ in (26), we have*

$$E\|\boldsymbol{\Theta}_\tau\|^2 \le M\gamma_0 + \frac{Md}{m'\beta} \tag{38}$$

Proof Let $\mathcal{Y}(\tau) \triangleq \|\boldsymbol{\Theta}_\tau\|^2$. The Itô lemma gives

$$d\mathcal{Y}(\tau) = -2\langle \boldsymbol{\Theta}_\tau, F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau)\rangle d\tau + \frac{2Md}{\beta}d\tau + \sqrt{\frac{8}{\beta}}\boldsymbol{\Theta}_\tau^\star d\mathcal{W}_\tau,$$

where $\boldsymbol{\Theta}_\tau^\star d\mathcal{W}_\tau \triangleq \sum_{i=1}^{Md}\boldsymbol{\Theta}_{i,\tau}d\mathcal{W}_{i,\tau}$ and the $\boldsymbol{\Theta}_{i,\tau}, d\mathcal{W}_{i,\tau}$ are the i-th components of $\boldsymbol{\Theta}_\tau$ and $\mathcal{W}_\tau$. Now this can be rewritten as

$$\begin{aligned}&2m'e^{2m'\tau}\mathcal{Y}(\tau)d\tau + e^{2m'\tau}d\mathcal{Y}(\tau) = \\ &= -2e^{2m'\tau}\langle \boldsymbol{\Theta}_\tau, F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau)\rangle d\tau + 2m'e^{2m'\tau}\mathcal{Y}(\tau)d\tau + \frac{2Md}{\beta}e^{2m'\tau}d\tau + \sqrt{\frac{8}{\beta}}e^{2m'\tau}\boldsymbol{\Theta}_\tau^\star d\mathcal{W}_\tau\end{aligned} \tag{39}$$

Since $2m'e^{2m'\tau}\mathcal{Y}(\tau)d\tau + e^{2m'\tau}d\mathcal{Y}(\tau)$ is the total Itô derivative of $e^{2m'\tau}\mathcal{Y}(\tau)$, we arrive at

$$d\left(e^{2m'\tau}\mathcal{Y}(\tau)\right) = -2e^{2m'\tau}\langle \boldsymbol{\Theta}_\tau, F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau)\rangle d\tau + 2m'e^{2m'\tau}\mathcal{Y}(\tau)d\tau + \frac{2Md}{\beta}e^{2m'\tau}d\tau + \sqrt{\frac{8}{\beta}}e^{2m'\tau}\boldsymbol{\Theta}_\tau^\star d\mathcal{W}_\tau \tag{40}$$

With integrating and rearranging, the above equation turns into

$$\begin{aligned}\mathcal{Y}(\tau) =& e^{-2m'\tau}\mathcal{Y}(0) - 2\int_0^\tau e^{2m'(s-\tau)}\langle \boldsymbol{\Theta}_\tau, F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau)\rangle d\tau \\ &+ 2m'\int_0^\tau e^{2m'(s-\tau)}\mathcal{Y}(s)ds + \frac{Md}{m'\beta}(1-e^{-2m'\tau}) + \sqrt{\frac{8}{\beta}}\int_0^\tau e^{2m'(s-\tau)}\boldsymbol{\Theta}_s^\star d\mathcal{W}_s ds\end{aligned} \tag{41}$$

Now with lemma 13, we can write

$$-2\int_0^\tau e^{2m'(s-\tau)}\langle \boldsymbol{\Theta}_\tau, F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau)\rangle d\tau \le -2\int_0^\tau e^{2m'(s-\tau)}(m'\mathcal{Y}(s))d\tau$$

Then, with (46) we arrive at

$$\|\boldsymbol{\Theta}_\tau\|^2 \le e^{-2m'\tau}\|\boldsymbol{\Theta}_0\|^2 + \frac{Md}{\beta m'}(1-e^{-2m'\tau}) + \sqrt{\frac{8}{\beta}}\int_0^\tau e^{2m'(s-\tau)}\boldsymbol{\Theta}_s^\star d\mathcal{W}_s ds$$

Taking expectations and using the martingale property of the Itô integral, we can derive the following result according to the independence of the the initial particles $\boldsymbol{\theta}_0^{(i)}$:

$$\begin{aligned}E\|\boldsymbol{\Theta}_\tau\|^2 &\le e^{-2m'\tau}E\|\boldsymbol{\Theta}_0\|^2 + \frac{Md}{\beta m'}(1-e^{-2m'\tau}) \\ &\le e^{-2m'\tau}M\gamma_0 + \frac{Md}{\beta m'}(1-e^{-2m'\tau})\end{aligned}$$

This finishes the proof. ■

It is easy to get the following corollary with the exchangeability of the particles

Corollary 17 *For the particles $\boldsymbol{\theta}_\tau^{(i)}$ in (3), we have*

$$E\|\boldsymbol{\theta}_\tau\|^2 \le \gamma_0 + \frac{d}{m'\beta} \tag{42}$$

Similarly, now we can provide a uniform bound for the $\bar{\boldsymbol{\theta}}_\tau$ in (20).

Theorem 18 *For the $\bar{\boldsymbol{\theta}}_\tau^{(i)}$ in (21), we have*

$$E\|\bar{\boldsymbol{\theta}}_\tau^{(i)}\|^2 \le \gamma_0 + \frac{d}{m'\beta} \tag{43}$$

Proof Let $\mathcal{Y}(\tau) \triangleq \|\bar{\boldsymbol{\theta}}_\tau\|^2$, where the $\bar{\boldsymbol{\theta}}_\tau$ is from (20). The Itô lemma gives

$$d\mathcal{Y}(\tau) = 2\langle \bar{\boldsymbol{\theta}}_\tau, -\beta^{-1}F(\bar{\boldsymbol{\theta}}_\tau) - E_{Y\sim\nu_\tau}K(\bar{\boldsymbol{\theta}}_\tau - Y)F(Y) + \nabla K * \nu_\tau(\bar{\boldsymbol{\theta}}_\tau)\rangle d\tau + \frac{2d}{\beta}d\tau + \sqrt{\frac{8}{\beta}}\bar{\boldsymbol{\theta}}^\star_\tau d\mathcal{W}_\tau,$$

where $\bar{\boldsymbol{\theta}}^\star_\tau d\mathcal{W}_\tau \triangleq \sum_{i=1}^d \bar{\boldsymbol{\theta}}_{i,\tau} d\mathcal{W}_{i,\tau}$ and the $\bar{\boldsymbol{\theta}}_{i,\tau}, d\mathcal{W}_{i,\tau}$ are the i-th components of $\bar{\boldsymbol{\theta}}_{i,\tau}$ and $\mathcal{W}_\tau$. This can be rewritten as

$$\begin{aligned} 2m'e^{2m'\tau}\mathcal{Y}(\tau)d\tau + e^{2m'\tau}d\mathcal{Y}(\tau) =& 2e^{2m'\tau}\langle \bar{\boldsymbol{\theta}}_\tau, -\beta^{-1}F(\bar{\boldsymbol{\theta}}_\tau) - E_{Y\sim\nu_\tau}K(\bar{\boldsymbol{\theta}}_\tau - Y)F(Y) + \nabla K * \nu_\tau(\bar{\boldsymbol{\theta}}_\tau)\rangle d\tau \\ &+ 2m'e^{2m'\tau}\mathcal{Y}(\tau)d\tau + \frac{2d}{\beta}e^{2m'\tau}d\tau + e^{2m'\tau}\sqrt{\frac{8}{\beta}}\bar{\boldsymbol{\theta}}^\star_\tau d\mathcal{W}_\tau \end{aligned} \tag{44}$$

Since $2m'e^{2m'\tau}\mathcal{Y}(\tau)d\tau + e^{2m'\tau}d\mathcal{Y}(\tau)$ is the total Itô derivative of $e^{2m'\tau}\mathcal{Y}(\tau)$, we arrive at

$$\begin{aligned} d\left(e^{2m'\tau}\mathcal{Y}(\tau)\right) =& 2e^{2m'\tau}\langle \bar{\boldsymbol{\theta}}_\tau, -\beta^{-1}F(\bar{\boldsymbol{\theta}}_\tau) - E_{Y\sim\nu_\tau}K(\bar{\boldsymbol{\theta}}_\tau - Y)F(Y) + \nabla K * \nu_\tau(\bar{\boldsymbol{\theta}}_\tau)\rangle d\tau \\ &+ 2m'e^{2m'\tau}\mathcal{Y}(\tau)d\tau + \frac{2d}{\beta}e^{2m'\tau}d\tau + e^{2m'\tau}\sqrt{\frac{8}{\beta}}\bar{\boldsymbol{\theta}}^\star_\tau d\mathcal{W}_\tau \end{aligned} \tag{45}$$

With integrating and rearranging, the above equation turns into

$$\begin{aligned} \mathcal{Y}(\tau) =& e^{-2m'\tau}\mathcal{Y}(0) + 2\int_0^\tau e^{2m'(s-\tau)}\langle \bar{\boldsymbol{\theta}}_s, -\beta^{-1}F(\bar{\boldsymbol{\theta}}_s) - E_{Y\sim\nu_\tau}K(\bar{\boldsymbol{\theta}}_s - Y)F(Y) + \nabla K * \nu_s(\bar{\boldsymbol{\theta}}_s)\rangle ds \\ &+ 2m'\int_0^\tau e^{2m'(s-\tau)}\mathcal{Y}(s)ds + \frac{d}{m'\beta}(1 - e^{-2m'\tau}) + \int_0^\tau e^{2m'(s-\tau)}\sqrt{\frac{8}{\beta}}\bar{\boldsymbol{\theta}}^\star_s d\mathcal{W}_s ds \end{aligned} \tag{46}$$

With lemma 13, we can write

$$2\int_0^\tau e^{2m'(s-\tau)}\langle \bar{\boldsymbol{\theta}}_s, -\beta^{-1}F(\bar{\boldsymbol{\theta}}_s)\rangle ds \leq -2\int_0^\tau e^{2m'(s-\tau)}(\beta^{-1}m_F\mathcal{Y}(s))ds$$

$$2\int_0^\tau e^{2m'(s-\tau)}\langle \bar{\boldsymbol{\theta}}_s, E_{Y\sim\nu_s}K(\bar{\boldsymbol{\theta}}_s - Y)F(Y)\rangle ds \leq 2\int_0^\tau e^{2m'(s-\tau)}\left(L_F(E_{Y\sim\nu_s}\|Y\|)\|\bar{\boldsymbol{\theta}}_s\|\right)ds$$

$$2\int_0^\tau e^{2m'(s-\tau)}\langle \bar{\boldsymbol{\theta}}_s, \nabla K * \nu_s(\bar{\boldsymbol{\theta}}_s)\rangle ds \leq 2\int_0^\tau e^{2m'(s-\tau)}\left(\frac{2}{\eta^2}(E_{Y\sim\nu_s}\|Y\|)\|\bar{\boldsymbol{\theta}}_s\| + \frac{2}{\eta^2}\mathcal{Y}(s)\right)ds$$

Then, with (46) we arrive at

$$\begin{aligned} \mathcal{Y}(\tau) \leq& e^{-2m't}\mathcal{Y}(0) + \frac{d}{\beta m'}(1 - e^{-2m'\tau}) - 2\int_0^\tau e^{2m'(s-\tau)}(\beta^{-1}m_F\mathcal{Y}(s))ds \\ &+ 2\int_0^\tau e^{2m'(s-\tau)}\left(L_F(E_{Y\sim\nu_s}\|Y\|)\|\bar{\boldsymbol{\theta}}_s\|\right)ds + 2\int_0^\tau e^{2m'(s-\tau)}\left(\frac{2}{\eta^2}(E_{Y\sim\nu_s}\|Y\|)\|\bar{\boldsymbol{\theta}}_s\| + \frac{2}{\eta^2}\mathcal{Y}(s)\right)ds \\ &+ 2m'\int_0^\tau e^{2m'(s-\tau)}\mathcal{Y}(s)ds + \int_0^\tau e^{2m'(s-\tau)}\sqrt{\frac{8}{\beta}}\bar{\boldsymbol{\theta}}^\star_s d\mathcal{W}_s ds \end{aligned}$$

Taking expectations and using the martingale property of the Itô integral, we can derive the following result:

$$\begin{aligned} E\|\bar{\boldsymbol{\theta}}_\tau\|^2 \leq& e^{-2m't}E\|\bar{\boldsymbol{\theta}}_0\|^2 + \frac{d}{\beta m'}(1 - e^{-2m'\tau}) - 2\int_0^\tau e^{2m'(s-\tau)}(\beta^{-1}m_F E\|\bar{\boldsymbol{\theta}}_s\|^2)ds \\ &+ 2\int_0^\tau e^{2m'(s-\tau)}\left(L_F(E_{Y\sim\nu_s}\|Y\|)E\|\bar{\boldsymbol{\theta}}_s\|\right)ds + 2\int_0^\tau e^{2m'(s-\tau)}\left(\frac{2}{\eta^2}(E_{Y\sim\nu_s}\|Y\|)E\|\bar{\boldsymbol{\theta}}_s\| + \frac{2}{\eta^2}E\|\bar{\boldsymbol{\theta}}_s\|^2\right)ds \\ &+ 2m'\int_0^\tau e^{2m'(s-\tau)}E\|\bar{\boldsymbol{\theta}}_s\|^2 ds + \int_0^\tau e^{2m'(s-\tau)}\sqrt{\frac{8}{\beta}}\bar{\boldsymbol{\theta}}^\star_s d\mathcal{W}_s ds \end{aligned}$$

With $\mathcal{L}(\bar{\boldsymbol{\theta}}_\tau) = \nu_\tau \mathrm{d}\boldsymbol{\theta}$ and $m' \triangleq \beta^{-1}m_F - L_F - \frac{4}{\eta^2}$, we can derive the following result with Cauchy–Schwarz. inequality,

$$
\begin{aligned}
E\|\bar{\boldsymbol{\theta}}_\tau\|^2 \le\ & e^{-2m't}E\|\bar{\boldsymbol{\theta}}_0\|^2 + \frac{d}{\beta m'}(1-e^{-2m'\tau}) - 2\int_0^\tau e^{2m'(s-\tau)}(\beta^{-1}m_F E\|\bar{\boldsymbol{\theta}}_s\|^2)ds \\
& + 2\int_0^\tau e^{2m'(s-\tau)}\left(L_F E\|\bar{\boldsymbol{\theta}}_s\|^2\right)ds + 2\int_0^\tau e^{2m'(s-\tau)}\left(\frac{2}{\eta^2}E\|\bar{\boldsymbol{\theta}}_s\|^2 + \frac{2}{\eta^2}E\|\bar{\boldsymbol{\theta}}_s\|^2\right)ds \\
& + 2m'\int_0^\tau e^{2m'(s-\tau)}E\|\bar{\boldsymbol{\theta}}_s\|^2 ds \\
& \le e^{-2m't}E\|\bar{\boldsymbol{\theta}}_0\|^2 + \frac{d}{\beta m'}(1-e^{-2m'\tau}) \\
& \le \gamma_0 + \frac{d}{\beta m'}
\end{aligned}
$$

This completes the proof. ■

J Non-Asymptotic Convergence Analysis: the Nonconvex Case

Since the non-convex case is much more complicated than the convex case, we reply on different assumptions and adopt another distance metric, denoted as $\tilde{\mathcal{B}}$, to characterize the convergence behavior of SPOS under the non-convex case. Note in this section, we give the preliminary convergence results of SPOS under the non-convex setting. A more complete version will be interesting future work.

Specifically, define $\tilde{\mathcal{B}}(\mu,\nu)$ as $\tilde{\mathcal{B}}(\mu,\nu) \triangleq |\mathbb{E}_{\boldsymbol{\theta}\sim\mu}[f(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim\nu}[f(\boldsymbol{\theta})]|$ for a known L_f-continuous function f satisfying Assumption 4 below. Note such metric has also been adopted in [Vollmer et al., 2016, Chen et al., 2015]. Our analysis considers (T, M, h_k) as variables in $\tilde{\mathcal{B}}$. In addition, we use $\{\hat{\theta}_k^{(i)}\}_{i=1}^M$ to denote the particles when full gradients are adopted in (9). The distribution of the particles is denoted as $\hat{\mu}_k$.

Our high-level idea of bounding $\tilde{\mathcal{B}}(\mu_T,\nu_\infty)$ is to decompose it as follows:

$$
\tilde{\mathcal{B}}(\mu_T,\nu_\infty) \le \tilde{\mathcal{B}}(\mu_T,\hat{\mu}_T) + \tilde{\mathcal{B}}(\hat{\mu}_T,\hat{\mu}_\infty) + \tilde{\mathcal{B}}(\hat{\mu}_\infty,\rho_\infty) + \tilde{\mathcal{B}}(\rho_\infty,\nu_\infty) \tag{47}
$$

Similarly, our idea is to concatenate the particles at each time into a single vector representation, *i.e.* defining the new parameter at time τ as $\boldsymbol{\Theta}_\tau \triangleq [\boldsymbol{\theta}_\tau^{(1)},\cdots,\boldsymbol{\theta}_\tau^{(M)}] \in \mathbb{R}^{Md}$. Consequently, the nonlinear PDE system (8) can be turned into an SDE ,which means $\boldsymbol{\Theta}_\tau$ is driven by the following SDE:

$$
\mathrm{d}\boldsymbol{\Theta}_\tau = -F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau)\mathrm{d}\tau + \sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_\tau^{(Md)}\ , \tag{48}
$$

where $F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau) \triangleq [\beta^{-1}F(\boldsymbol{\theta}_\tau^{(1)}) - \frac{1}{M}\sum_{j=1}^M \nabla K(\boldsymbol{\theta}_\tau^{(1)} - \boldsymbol{\theta}_\tau^{(j)}) + \frac{1}{M}\sum_{j=1}^M K(\boldsymbol{\theta}_\tau^{(1)} - \boldsymbol{\theta}_\tau^{(j)})F(\boldsymbol{\theta}_\tau^{(j)}),\cdots,\beta^{-1}F(\boldsymbol{\theta}_\tau^{(M)}) - \frac{1}{M}\sum_{j=1}^M \nabla K(\boldsymbol{\theta}_\tau^{(M)} - \boldsymbol{\theta}_\tau^{(j)}) + \frac{1}{M}\sum_{j=1}^M K(\boldsymbol{\theta}_\tau^{(M)} - \boldsymbol{\theta}_\tau^{(j)})F(\boldsymbol{\theta}_\tau^{(j)})]$ is a vector function $\mathbb{R}^{Md}\to\mathbb{R}^{Md}$, and $\mathcal{W}_\tau^{(Md)}$ is Brownian motion of dimension $M\times d$. Similarly, we can define $\hat{\Theta}_k \triangleq [\hat{\theta}_k^{(1)},\cdots,\hat{\theta}_k^{(M)}] \in \mathbb{R}^{Md}$ for the full-gradient case. Hence, it can be seen that through such a decomposition in (47), the bound related to a nonlinear PDE system (8) reduces to that of an SDE. The second term $\tilde{\mathcal{B}}(\hat{\mu}_T,\hat{\mu}_\infty)$ reflexes the geometric ergodicity of a dynamic system with a numerical method. It is known that even if a dynamic system has an exponential convergence rate to its equilibrium, its corresponding numerical method might not. Our bound for $\tilde{\mathcal{B}}(\hat{\mu}_T,\hat{\mu}_\infty)$ is essentially a specification of the result of [Mattingly et al., 2002], which has also been applied by [Xu et al., 2018]. The third term $\tilde{\mathcal{B}}(\hat{\mu}_\infty,\rho_\infty)$ reflects the numerical error of an SDE, which has been studied in related literature such as [Chen et al., 2015]. To this end, we adopt standard assumptions used in the analysis of SDEs [Vollmer et al., 2016, Chen et al., 2015], rephrased in Assumption 4.

Assumption 4 *For the SDE (48) and a Lipschitz function f, let ψ be the solution functional of the Poisson equation: $\mathcal{G}\psi(\hat{\Theta}_k) = \frac{1}{M}\sum_{i=1}^M f(\hat{\theta}_k^{(i)}) - \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}[f(\boldsymbol{\theta})]$, where $\mathcal{G}$ denotes the infinite generator of the SDE (48). Assume ψ and its up to 4th-order derivatives, $\mathcal{D}^k\psi$, are bounded by a function $\mathcal{V}$, i.e., $\|\mathcal{D}^k\psi\| \le H_k\mathcal{V}^{p_k}$ for $k = (0,1,2,3,4)$, $H_k, p_k > 0$. Furthermore, the expectation of $\mathcal{V}$ on $\{\boldsymbol{\Theta}_\tau\}$ is bounded: $\sup_l \mathbb{E}\mathcal{V}^p(\boldsymbol{\Theta}_\tau) < \infty$, and $\mathcal{V}$ is smooth such that $\sup_{s\in(0,1)}\mathcal{V}^p\left(s\boldsymbol{\Theta} + (1-s)\boldsymbol{\Theta}'\right) \le H\left(\mathcal{V}^p(\boldsymbol{\Theta}) + \mathcal{V}^p(\boldsymbol{\Theta}')\right)$, $\forall\boldsymbol{\Theta},\boldsymbol{\Theta}', p \le \max\{2p_k\}$ for $H > 0$.*

Assumption 5 *i) F, K and ∇K are L_F, L_K and $L_{\nabla k}$ Lipschitz; ii) F satisfies the dissipative property, i.e., $\langle F(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle \geq m\left\|\boldsymbol{\theta}\right\|^2 - b$ for some $m, b > 0$; iii) Remark 3 applies to the nonconvex setting, i.e. $\sup_{\|f\|_{Lip}\leq 1}|\mathbb{E}_{\boldsymbol{\theta}\sim\mu_\infty}[f(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim\nu_\infty}[f(\boldsymbol{\theta})]| = \mathcal{W}_1(\rho_\infty, \nu_\infty) = O(M^{-1/2})$.*

Remark 5 *Assumption 4 is necessary to control the gap between a numerical solution and the exact solution of an SDE. Specifically, it is used to bound the $\tilde{\mathcal{B}}(\hat{\mu}_\infty, \rho_\infty)$ term and the $\tilde{\mathcal{B}}(\mu_T, \hat{\mu}_T)$ term above. Purely relying on the dissipative assumption in Assumption 5 as in non-convex optimization with SG-MCMC [Raginsky et al., 2017, Xu et al., 2018] would induce a bound increasing linearly w.r.t. time τ. Thus it is not suitable for our goal. Finally, iii) in Assumption 5 is a mild condition and reasonable because we expect particles to be able to approximate all distributions equally well in the asymptotic limit of $t \to \infty$ by ergodicity due to the injected noise. How to remove/replace this assumption is an interesting future work.*

Based on the assumptions above, the bounds for $\tilde{\mathcal{B}}(\hat{\mu}_T, \hat{\mu}_\infty)$ and $\tilde{\mathcal{B}}(\hat{\mu}_\infty, \rho_\infty)$ are summarized below.

Theorem 19 *Under Assumption 4–5, if we set the stepsize $h_k = h$, we can have the following results:*

$$\begin{aligned} &\tilde{\mathcal{B}}(\hat{\mu}_T, \hat{\mu}_\infty) \leq C_2\varsigma\sigma^{-Md/2}(1+\varsigma e^{m_{\boldsymbol{\Theta}}h})\exp\left(-2m_{\boldsymbol{\Theta}}Th\sigma^{Md}/\log(\varsigma)\right), \\ &\textit{and } \tilde{\mathcal{B}}(\hat{\mu}_\infty, \rho_\infty) \leq C_3 h/\beta, \end{aligned} \tag{49}$$

where $\varsigma = 2L_{\boldsymbol{\Theta}}(Mb\beta + m_{\boldsymbol{\Theta}}\beta + Md)/m_{\boldsymbol{\Theta}}$, $L_{\boldsymbol{\Theta}} = \sqrt{2}\beta^{-1}L_F + l'$, $m_{\boldsymbol{\Theta}} = \beta^{-1}m - m'$, and $(\sigma, C_2, C_3, l', m')$ are some positive constants independent of (T, M, h) and $\sigma \in (0, 1)$

Remark 6 *In order to make the $\tilde{\mathcal{B}}(\hat{\mu}_T, \hat{\mu}_\infty)$ term asymptotically decrease to zero, the number of running iteration T should increase at a rate faster enough to compensate the effect of increasing M. We believe there is room for improving this bound, which is an interesting future work.*

Next we bound the $\tilde{\mathcal{B}}(\mu_T, \hat{\mu}_T)$ term related to stochastic gradients. By adapting results from analysis of diffusion processes [Xu et al., 2018], $\tilde{\mathcal{B}}(\mu_T, \hat{\mu}_T)$ can be bounded with Theorem 20.

Theorem 20 *Under Assumptions 4–5, if we set $B_k = B$ and $h_k = h$, $\tilde{\mathcal{B}}(\mu_T, \hat{\mu}_T)$ is bounded as*

$$\tilde{\mathcal{B}}(\mu_T, \hat{\mu}_T) \leq C_5 Th(L_{\boldsymbol{\Theta}}\Gamma' + MC_4)\sqrt{(6+2\Gamma')\beta/(BM)},$$

where $\Gamma' = 2(1+1/m_\Theta)(Mb + 2M^2C_4^2 + Md/\beta)$ and , (C_4, C_5) is some positive constant independent of (T, M, h)

Finally, by combining the results from Theorem 19, 20 and *iii*) in Assumption 5, we arrive at a bound for our target $\tilde{\mathcal{B}}(\mu_T, \nu_\infty)$, summarized in Theorem 21.

Theorem 21 *Under Assumptions 4–5, there exist some positive constants $(C_2, C_3, C_4, C_5, C_6)$ such that:*

$$\begin{aligned} \tilde{\mathcal{B}}(\mu_T, \nu_\infty) \leq & C_2\varsigma\sigma^{-Md/2}(1+\varsigma e^{m_{\boldsymbol{\Theta}}h}) \times \exp\left(-2m_{\boldsymbol{\Theta}}Th\sigma^{Md}/\log(\varsigma)\right) + C_3 h/\beta \\ & + C_5 Th(L_{\boldsymbol{\Theta}}\Gamma' + MC_4)\left((6+2\Gamma')\beta/(BM)\right)^{1/2} + C_6/\sqrt{M}, \end{aligned}$$

where σ, ς and Γ' are the same as those in Theorem 19–20.

K Proof of Theorem 19

Proof [Proof of Theorem 19] Our conclusion for $\tilde{\mathcal{B}}(\hat{\mu}_T, \hat{\mu}_\infty)$ is essentially a specification of the result in [Mattingly et al., 2002], which has also been applied in [Xu et al., 2018].

Specifically, we rely on the following lemma, which is essentially Theorem 7.3 in [Mattingly et al., 2002] and Lemma C.3 in [Xu et al., 2018]. Consider the following SDE (eq.48):

$$\mathrm{d}\boldsymbol{\Theta}_\tau = -F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\tau)\mathrm{d}\tau + \sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_\tau^{(Md)}$$

As mentioned in Section 5, we denote the distribution of $\boldsymbol{\Theta}_\tau$ as $\rho_\tau^{\boldsymbol{\Theta}}$, and define $\hat{\Theta}_k \triangleq [\hat{\theta}_k^{(1)}, \cdots, \hat{\theta}_k^{(M)}] \in \mathbb{R}^{Md}$, which is actually the numerical solution of (48) using full gradient with Euler method. Denote the distribution of $\hat{\Theta}_k$ as $\hat{\mu}_k^{\boldsymbol{\Theta}}$.

Lemma 22 *Let $F_{\boldsymbol{\Theta}}$ be Lipschitz-continuous with constant $L_{\boldsymbol{\Theta}}$, and satisfy the dissipative property that $\langle F_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}), \boldsymbol{\Theta}\rangle \geq m_{\boldsymbol{\Theta}} \left\|\boldsymbol{\Theta}\right\|^2 - b_{\boldsymbol{\Theta}}$. Define $V_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) = C_0 + L_{\boldsymbol{\Theta}}/2\|\boldsymbol{\Theta}\|^2$. The Euler method for (48) has a unique invariant measure $\hat{\mu}^{\Theta}_{\infty}$, and for all test function $f_{\boldsymbol{\Theta}}$ such that $|f_{\boldsymbol{\Theta}}| \leq V_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})$, we have*

$$\begin{aligned}&\left|\mathbb{E}[f_{\boldsymbol{\Theta}}(\hat{\Theta}_k))] - \mathbb{E}_{\hat{\Theta}_\infty\sim\hat{\mu}^{\Theta}_\infty}[f(\hat{\Theta}_\infty)]\right|\\ \leq & C\kappa\rho^{-Md/2}(1+\kappa e^{m_{\boldsymbol{\Theta}}h})\exp\left(-\frac{2m_{\boldsymbol{\Theta}}kh\rho^{Md}}{\log(\kappa)}\right) ,\end{aligned}$$

where $\rho \in (0,1)$, $C > 0$ are positive constants, and $\kappa = 2L_{\boldsymbol{\Theta}}(b_{\boldsymbol{\Theta}}\beta + m_{\boldsymbol{\Theta}}\beta + Md)/m_{\boldsymbol{\Theta}}$.

Now we define $f_{\boldsymbol{\Theta}} : \mathbb{R}^{Md} \to \mathbb{R}$ as $f_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) = \frac{1}{M}\sum_i^M f(\boldsymbol{\theta}^{(i)})$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a L_f-Lipschitz function satisfying our Assumption 4, and $\boldsymbol{\Theta} \triangleq [\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(M)}]$. Similar to the proof of Lemma 11, we can find that $f_{\boldsymbol{\Theta}} : \mathbb{R}^{Md}$ is a $L_f/\sqrt{M}$-Lipschitz function. Furthermore, according to Lemma 13, it is easily check that $F_{\boldsymbol{\Theta}}$ is $L_{\boldsymbol{\Theta}}$-Lipschitz where $L_{\boldsymbol{\Theta}} = \sqrt{2}\beta^{-1}L_F + l'$. Hence, when β is small enough, we have $L_f/\sqrt{M} \leq \sqrt{2}\beta^{-1}L_F + l'$. As a result, we can set the C_0 large enough to force $f_{\boldsymbol{\Theta}}$ to satisfy the condition in Lemma 22 that $|f_{\boldsymbol{\Theta}}| \leq V_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})$. According to the exchangeability of the particle system $\{\hat{\theta}_k^{(i)}\}$ and Lemma 13, we can bound $\tilde{\mathcal{B}}(\hat{\mu}_T, \hat{\mu}_\infty)$ as

$$\begin{aligned}\tilde{\mathcal{B}}(\hat{\mu}_T, \hat{\mu}_\infty) \leq & \left|\mathbb{E}[f_{\boldsymbol{\Theta}}(\hat{\Theta}_T))] - \mathbb{E}_{\hat{\Theta}_\infty\sim\hat{\mu}^{\Theta}_\infty}[f(\hat{\Theta}_\infty)]\right|\\ \leq & C_2\varsigma\sigma^{-Md/2}(1+\varsigma e^{m_{\boldsymbol{\Theta}}h})\exp\left(-2m_{\boldsymbol{\Theta}}Th\sigma^{Md}/\log(\varsigma)\right)\end{aligned}$$

where $\varsigma = 2L_{\boldsymbol{\Theta}}(Mb\beta + m_{\boldsymbol{\Theta}}\beta + Md)/m_{\boldsymbol{\Theta}}$, $L_{\boldsymbol{\Theta}} = \sqrt{2}\beta^{-1}L_F + l'$, $m_{\boldsymbol{\Theta}} = \beta^{-1}m - m'$, and (σ, C_2, l', m') are some positive constants independent of (T, M, h) and $\sigma \in (0,1)$.

To prove the bound for $\tilde{\mathcal{B}}(\hat{\mu}_\infty, \rho_\infty)$, since $\hat{\Theta}_k = (\hat{\theta}_k^{(1)}, \cdots, \hat{\theta}_k^{(M)})$ can be considered as a solution to the SDE (48), standard results from linear FP equation can be applied. Specifically, for the $\tilde{\mathcal{B}}(\hat{\mu}_\infty, \rho_\infty)$ term, we rely on the following lemma adapted from Lemma C.4 in [Xu et al., 2018, Chen et al., 2015], which is essentially the result of [Chen et al., 2015] when taking $T \to \infty$.

Lemma 23 *Under the same assumption as in Lemma 22, for the Lipschitz-continuous function $f_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) = \frac{1}{M}\sum_i^M f(\boldsymbol{\theta}^{(i)})$ mentioned above, the following bound is satisfied for some positive constant C:*

$$\left|\frac{1}{T}\sum_{k=1}^{T-1}\mathbb{E}[f_{\boldsymbol{\Theta}}(\hat{\Theta}_k)] - \mathbb{E}_{\boldsymbol{\Theta}_\infty\sim\rho^{\Theta}_\infty}[f(\boldsymbol{\Theta}_\infty)]\right| \leq C(\frac{h}{\beta} + \frac{\beta}{Th}) .$$

The uniqueness of invariant measure of the Euler method from Lemma 22 implies the numerical solution $\hat{\Theta}_k$ to be ergodic. Then similar to the proof of Lemma 4.2 in [Xu et al., 2018], we consider the case where $T \to \infty$. Taking average over the $\{\hat{\Theta}_k\}_{k=0}^{T-1}$, we have

$$\mathbb{E}_{\hat{\Theta}_\infty\sim\hat{\mu}^{\Theta}_\infty}[f_{\boldsymbol{\Theta}}(\hat{\Theta}_\infty)] = \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}[f_{\boldsymbol{\Theta}}(\hat{\Theta}_k)]$$

Now according to the exchangeability of the particle system $\{\hat{\theta}_k^{(i)}\}$ and $\{\boldsymbol{\theta}_\tau^{(i)}\}$, we can bound the $\tilde{\mathcal{B}}(\hat{\mu}_\infty, \rho_\infty)$ as :

$$\begin{aligned}\tilde{\mathcal{B}}(\hat{\mu}_\infty, \rho_\infty) \leq & \left|\mathbb{E}_{\hat{\Theta}_\infty\sim\hat{\mu}^{\Theta}_\infty}[f_{\boldsymbol{\Theta}}(\hat{\Theta}_\infty)] - \mathbb{E}_{\boldsymbol{\Theta}_\infty\sim\rho^{\Theta}_\infty}[f_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_\infty)]\right|\\ \leq & C_3h/\beta ,\end{aligned}$$

where C_3 are some positive constant. ■

L Proof of Theorem 20

Proof [Proof of Theorem 20] Adopting the same notation used in the proof of Theorem 7, we define $\Theta_k \triangleq [\theta_k^{(1)}, \cdots, \theta_k^{(M)}]$ and $G^{\Theta}_{\mathcal{I}_k} \triangleq \frac{N}{B_k}\sum_{q\in\mathcal{I}_k} F_{(q)\boldsymbol{\Theta}}(\Theta_k)$. We denote the distribution of Θ_k as μ^{Θ}_k, where

$$\Theta_{k+1} = \Theta_k - \beta^{-1}G^{\Theta}_{\mathcal{I}_k}h_k + \sqrt{2\beta^{-1}h_k}\Xi_k .$$

We firstly derive a bound for $W_2(\mu_k^{\Theta}, \hat{\mu}_k^{\boldsymbol{\Theta}})$ (the definition of $\hat{\mu}_k^{\boldsymbol{\Theta}}$ is given in the last section). According to the proof of Lemma 4.4 in [Xu et al., 2018]

$$W_2(\mu_k^{\Theta}, \hat{\mu}_k^{\boldsymbol{\Theta}}) \leq kh(L_{\boldsymbol{\Theta}}\Gamma' + MC_4)\left((6+2\Gamma')\beta/B\right)^{1/2}$$

where $\Gamma' = 2(1 + 1/m_{\Theta})(Mb + 2M^2C_4^2 + Md/\beta)$ and C_4 is some positive constant independent of (T, M, h). By applying the facts that $W_1(\mu_k^{\Theta}, \hat{\mu}_k^{\boldsymbol{\Theta}}) \leq W_2(\mu_k^{\Theta}, \hat{\mu}_k^{\boldsymbol{\Theta}})$ and $W_1(\mu_k, \hat{\mu}_k) \leq \frac{1}{\sqrt{M}} W_1(\mu_k^{\Theta}, \hat{\mu}_k^{\boldsymbol{\Theta}})$ (see the proof of Lemma 12, similar result holds here), we get

$$W_1(\mu_T, \hat{\mu}_T) \leq Th(L_{\boldsymbol{\Theta}}\Gamma' + MC_4)\left((6+2\Gamma')\beta/(BM)\right)^{1/2} .$$

Since the definitions of $\mathcal{W}_1(\mu, \nu)$ and $\tilde{\mathcal{B}}(\mu, \nu)$ are given as:

$$\begin{aligned} W_1(\mu, \nu) &\triangleq \sup_{\|g\|_{lip} \leq 1} \left|\mathbb{E}_{\boldsymbol{\theta}\sim\mu}[g(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim\nu}[g(\boldsymbol{\theta})]\right| \\ \tilde{\mathcal{B}}(\mu, \nu) &\triangleq \left|\mathbb{E}_{\boldsymbol{\theta}\sim\mu}[f(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim\nu}[f(\boldsymbol{\theta})]\right| \ , \end{aligned}$$

it is easily seen that $\tilde{\mathcal{B}}(\mu_T, \hat{\mu}_K) \leq L_f W_1(\mu_T, \hat{\mu}_T)$, which finishes the proof. ■

M Discussion on the complexity of the proposed SPOS

The complexity of an algorithm mainly refers to its time complexity (corresponding to the number of iterations in our method *i.e.* T) and space complexity (corresponding to the number of particles used in our method *i.e.* M). Hence the complexity of our method can be well explored with our work, since our non-asymptotic convergence theory is developed w.r.t. both the number of particles *i.e.* M and iterations *i.e.* T. Their relationship (tradeoff) is discussed further in the experiments. Moreover, by comparing (9) with (3) , one can easily find that our space complexity is exactly the same as SVGD and our computational time in each iteration is almost the same as SVGD with an extra addition operation. However, it is worth noting that our method have much better performance in practice with no "pitfall" verified by both our theory and experiments.

N Comparison with Related Work

Firstly, our proposed framework SPOS is different from the recently proposed particle-optimization sampling framework [Chen et al., 2018], in the sense that we solve the nonlinear PDE (6) stochastically. For example they deterministically solve the equation in (6) $\partial \nu_\tau = \beta^{-1}\nabla_{\boldsymbol{\theta}} \cdot \nabla_{\boldsymbol{\theta}}\nu_\tau$ approximately using blob method adopted from [Carrillo et al., 2017].

Secondly, our method is also distinguishable to existing work on granular media equations such as [Durmus et al., 2018]. The work about the granular media equations mainly focuses on the following PDE:

$$\partial_\tau \nu_\tau = \nabla_{\boldsymbol{\theta}} \cdot \left(\nu_\tau \beta^{-1} F(\boldsymbol{\theta}) + \nu_\tau \left(\nabla K * \nu_\tau(\boldsymbol{\theta})\right) + \beta^{-1}\nabla_{\boldsymbol{\theta}}\nu_\tau\right) \ , \tag{50}$$

whereas our framework focuses on the following one:

$$\begin{aligned} \partial_\tau \nu_\tau = &\nabla_{\boldsymbol{\theta}} \cdot \left(\nu_\tau \beta^{-1} F(\boldsymbol{\theta}) + \nu_\tau \left(E_{Y\sim\nu_\tau} K(\boldsymbol{\theta} - Y)F(Y)\right.\right. \\ &\left.\left. - \nabla K * \nu_\tau(\boldsymbol{\theta})\right) + \beta^{-1}\nabla_{\boldsymbol{\theta}}\nu_\tau\right) \ . \end{aligned} \tag{51}$$

The extra term $\nu_\tau \left(E_{Y\sim\nu_\tau} K(\boldsymbol{\theta} - Y)F(Y)\right)$ in our framework makes the analysis much more challenging. The main differences between our work and [Durmus et al., 2018] including related work are summarized below:

- Formulations are different. The extra term $E_{Y\sim\mu_\tau} K(\boldsymbol{\theta} - Y)F(Y)$ cannot be combined with the $F(\boldsymbol{\theta})$ term in (50) in [Durmus et al., 2018]. This is because function $F(\boldsymbol{\theta})$ **itself** is a function independent of τ; while $E_{Y\sim\mu_\tau} K(\boldsymbol{\theta} - Y)F(Y)$ depends on both $\boldsymbol{\theta}$ and τ. This makes our problem much more difficult.

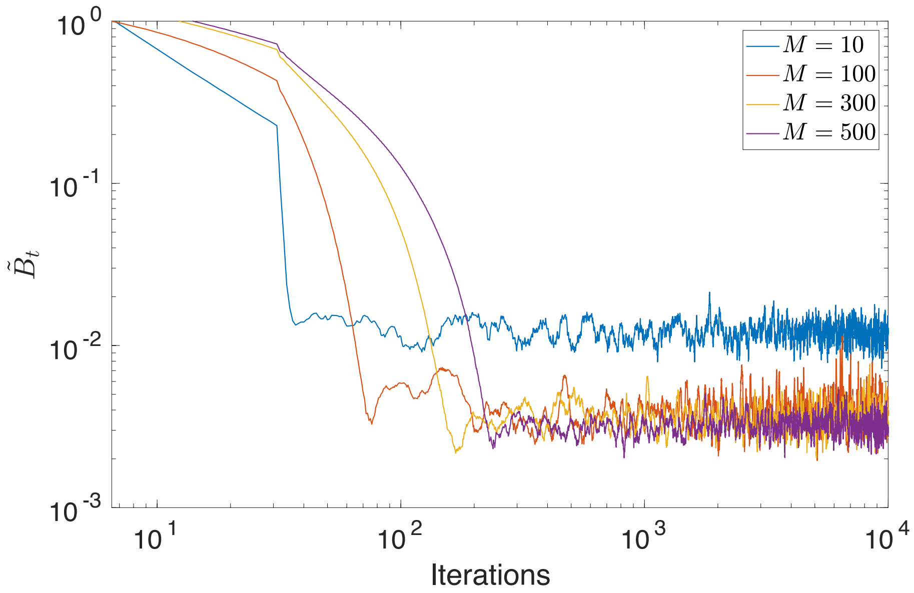

Figure 5: Errors versus #iterations on a simple posterior Gaussian model.

- Assumptions are different. For example, the analysis on granular media equations in [Cattiaux et al., 2008] requires that F satisfies a special condition $C(\mathbb{A}, \alpha)$, which is a strong condition impractical to be satisfied in our case; And [Durmus et al., 2018] adopts different assumptions from ours with a different goal.

- For the Euler integrator, [Durmus et al., 2018] does not consider an Euler solution. Furthermore, our sampling method needs "stochastic gradient" *i.e.* $G_k^{(i)} \triangleq \frac{N}{B_k} \sum_{q \in \mathcal{I}_k} F_q(\theta_k^{(i)})$ in (9) for computational feasibility, which is quite different from the former work on particle-SDE such as [Malrieu, 2003, Cattiaux et al., 2008]. Few of the former work on particle-SDE considered the stochastic gradient issue.

To sum up, the main purpose of our paper is to provide a non-asymptotic analysis of our method instead of improving the former work on a certain type of PDE. This is also the reason why we said that parts of our proof techniques are based on those for analyzing granular media equations.

O Extra Experiments

O.1 Posterior sampling of a Gaussian model

We further follow [Chen et al., 2015] and consider a relatively more complex Gaussian model for posterior sampling: $x_i \sim \mathcal{N}(\theta, 1), \theta \sim \mathcal{N}(0, 1)$, where 1000 data samples $\{x_i\}$ are generated. We adopt the same setting as above. The posterior average $\mathbb{E}_{\theta \sim p(\theta | \{x_i\})}[f(\theta)]$ endows an explicit expression. Figure 5 plots the error versus the running iterations for different particle sizes. It is observed that at the beginning, the errors for the ones with less particles decrease faster than those with more particles. This is reflected in the overall bound given in Theorem 9, which are dominated by the bound in Theorem 7 (indicating larger M results in larger errors at the beginning). When more running time/iterations are given, the impact of the exponentially-decaying term in Theorem 6 could be ignored. We also observe a trend of increasing errors when number of iterations are large enough, which is not drawn in the figure for simplicity.

O.2 Toy Experiments

We compare the proposed SPOS with other popular methods such as SVGD and standard SGLD on four mutil-mode toy examples. We aim to sample from four unnormalized 2D densities $p(z)/\exp\{U(z)\}$, with the functional form provided in [Rezende and Mohamed, 2015]. We optimize/sample 50 and 2000 particles to approximate the target distributions. The results are illustrated in Figure 6 and Figure 7, respectively.

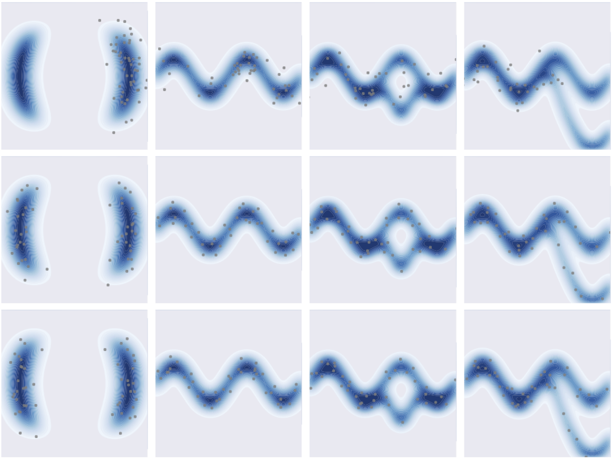

Figure 6: Illustration of different algorithms on toy distributions. Dots are the final particles; the blue regions represent ground true densities. Each column is a distribution case. First row: standard SGLD; Second row: SVGD; Third row: SPOS.

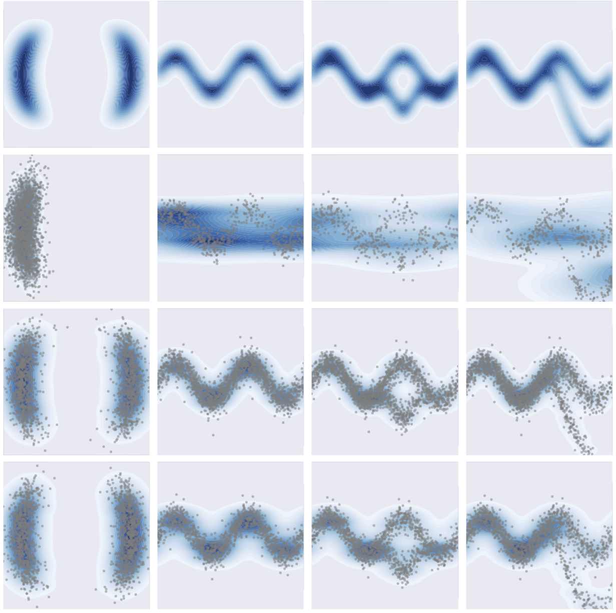

Figure 7: Illustration of different algorithms on toy distributions. Dots are the final particles; the blue regions represent densities estimated by the particles. Each column is a distribution case. First row: ground true densities; Second row: standard SGLD; Third row: SVGD; Fourth row: SPOS.

O.3 More details on Bayesian neural networks for regression

The Bayesian DNNs are used to model weight uncertainty of neural networks, an important topic that has been well explored [Hernández-Lobato and Adams, 2015, Blundell et al., 2015, Li et al., 2016, Louizos and Welling, 2016]. We assign simple isotropic Gaussian priors to the weights, and perform posterior sampling with different methods. For SVGD and SPOS methods, we use a RBF kernel $K(\boldsymbol{\theta}, \boldsymbol{\theta}') = \exp(-\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2/\eta^2)$, with the bandwidth set to $\eta = \texttt{med}^2/\log M$. Here $\texttt{med}$ is the median of the pairwise distance between particles. We use a single-layer BNN for regression tasks. Following [Li et al., 2015], 10 UCI public datasets are considered: 100 hidden units for 2 large datasets (Protein and YearPredict), and 50 hidden units for the other 8 small datasets. Following [Zhang et al., 2018b], we repeat the experiments 20 times with batchsize 100 for all datasets except for Protein and YearPredict, which we repeat 5 times and once with batchsize 1000. The datasets are randomly split into 90% training and 10% testing. For a fair comparison, we use the same split of data (train, val and test) for the three methods. The test results are reported on the best model on the validation set. We adopt the

Table 2: Averaged predictions with standard deviations in terms of RMSE and log-likelihood on test sets.

	Test RMSE			Test Log likelihood		
Dataset	SGLD	SVGD	SPOS	SGLD	SVGD	SPOS
Boston_Housing	3.114 ± 0.144	2.961 ± 0.109	**2.829 ± 0.126**	−2.633 ± 0.083	-2.591 ± 0.029	**−2.532 ± 0.082**
Concrete	5.508 ± 0.275	5.157 ± 0.082	**5.071 ± 0.1495**	−3.133 ± 0.087	-3.247 ± 0.01	**−3.062 ± 0.037**
Energy	0.842 ± 0.060	1.291 ± 0.029	**0.752 ± 0.0285**	−1.268 ± 0.143	-1.534 ± 0.026	**−1.158 ± 0.073**
Kin8nm	0.080 ± 0.001	0.090 ± 0.001	**0.079 ± 0.001**	1.080 ± 0.025	0.986 ± 0.004	**1.092 ± 0.013**
Naval	0.004 ± 0.000	0.004 ± 0.000	**0.004 ± 0.000**	4.127 ± 0.028	4.032 ± 0.008	**4.145 ± 0.02**
CCPP	4.059 ± 0.080	4.127 ± 0.027	**3.939 ± 0.0495**	−2.823 ± 0.039	-2.843 ± 0.006	**−2.794 ± 0.025**
Winequality	0.632 ± 0.022	0.604 ± 0.007	**0.598 ± 0.014**	−0.962 ± 0.067	-0.926 ± 0.009	**−0.911 ± 0.041**
Yacht	1.183 ± 0.263	1.597 ± 0.099	**0.84 ± 0.0865**	−1.680 ± 0.393	-1.818 ± 0.06	**−1.446 ± 0.121**
Protein	4.281 ± 0.011	4.392 ± 0.015	**4.254 ± 0.005**	−2.877 ± 0.002	-2.905 ± 0.010	**−2.876 ± 0.009**
YearPredict	8.707± NA	8.684 ± NA	**8.681 ± NA**	-3.582 ± NA	-3.580 ± NA	**−3.576 ± NA**

Table 3: Classification error of FNN on MNIST.

Method	Test Error	
	400-400	800-800
SPOS	**1.32%**	**1.24%**
SVGD	1.56%	1.47%
SGLD	1.64%	1.41%
RMSprop	1.59%	1.43%
RMSspectral	1.65%	1.56%
SGD	1.72%	1.47%
BPB, Gaussian	1.82%	1.99%
SGD, dropout	1.51%	1.33%

root mean squared error (RMSE) and test log-likelihood as the evaluation criteria. The experimental results are shown in Table 2, from which we can see the proposed SPOS outperforms SVGD and other existing methods presented in [Zhang et al., 2018b] (results not shown due to space limit), achieving state-of-the-art results.

O.3.1 Bayesian Neural Networks for MNIST classification

We perform the classification tasks on the standard MNIST dataset. A two-layer MLP 784-X-X-10 with ReLU activation function is used, with X being the number of hidden units for each layer. The training epoch is set to 100. The test errors are reported in Table 3. Surprisingly, the proposed SPOS outperforms other algorithms such as SVGD at a significant level, though it is just a simple modification of SVGD by adding in random Gaussian noise. This is partly due to the fact that our SPOS algorithm can jump out of local modes efficiently, as explained in Section 2.2.

O.4 Bayesian exploration in deep RL

We denote the policy as $\pi_{\boldsymbol{\theta}}(\mathbf{a} \,|\, \mathbf{s})$ parameterized by $\boldsymbol{\theta}$ with prior distribution $p(\boldsymbol{\theta})$, where $\mathbf{a}$ represent the action variable, and $\mathbf{s}$ the state variable. According to [Liu et al., 2017], learning the optimal policy corresponds to calculating the following posterior distribution for $\boldsymbol{\theta}$: $q(\boldsymbol{\theta}) \propto \exp(J(\boldsymbol{\theta})/\alpha)p(\boldsymbol{\theta})$, where $J(\boldsymbol{\theta})$ denotes the expected cumulative reward under the policy with parameter $\boldsymbol{\theta}$ and α a hyperparameter. Consequently, $\boldsymbol{\theta}$ could be updated by drawing samples from $q(\boldsymbol{\theta})$ with the proposed SPOS. We denote this method as SPOS-PG. In addition, when drawing samples with SVGD, the resulting algorithm is called Stein variational policy gradient (SVPG) [Liu et al., 2017]. Note in implementation, the term $J(\boldsymbol{\theta})$ can be approximated with REINFORCE [Williams, 1992] or advantage actor critic [Schulman et al., 2015], which we will investigate in our experiments.

The policy is parameterized as a two-layer (25-10 hidden units) neural network with `tanh` as the activation function. The maximal length of horizon is set to 500. We use a sample size of 10000 for policy gradient estimation, and $M = 16$, $\alpha = 10$. For the simplest task, Cartpole, all agents are trained for 100 episodes; whereas they are trained up to 1,000 episodes for the other two tasks. The average reward versus number of episodes are plotted in Figure 8. It is observed that our SPOS-PG obtains much larger average rewards and smaller variance compared to SVPG, though the convergence behaviors are similar in the simplest Carpole task.

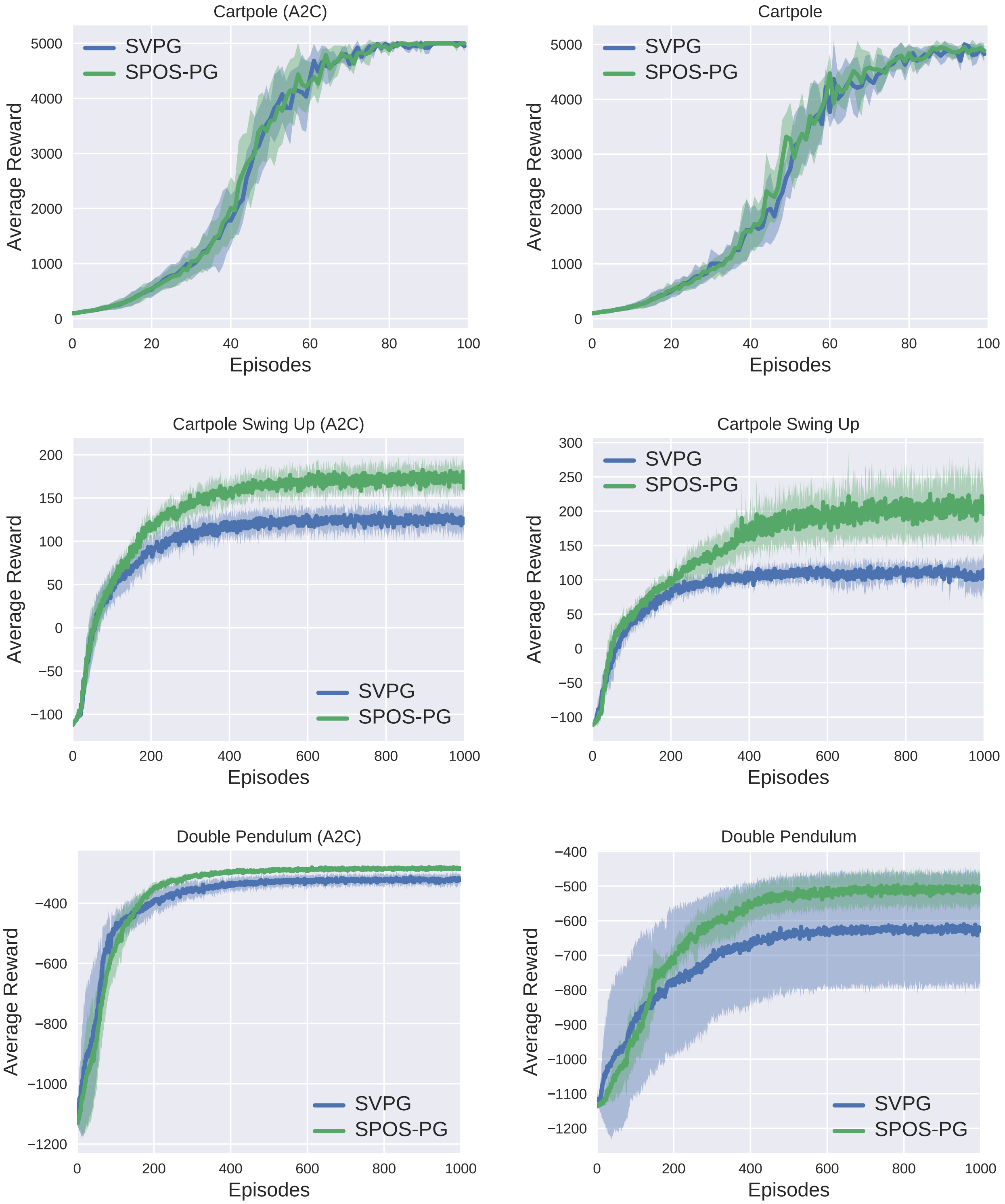

Figure 8: Policy learning with Bayesian exploration in policy-gradient methods on six scenarios with SVPG and SPOS-PG.