# Supplementary Material for Learning Interaction Kernels for Agent Systems on Riemannian Manifolds

## A. Preliminaries

In this work, $\mathcal{M}$ is a connected, smooth, and geodesically complete $d$-dimensional Riemannian manifold with Riemannian metric $g$. For details regarding the basic definitions of Riemannian manifolds, geodesics, Riemannian distances, exponential maps, cut loci, and injectivity radii, please see (Lee, 2003; do Carmo, 1976). We will discuss how to find the minimal geodesic and the Riemannian distance between any two points on the two prototypical manifolds used in our numerical algorithms: the two-dimensional sphere ($\mathbb{S}^2$) and the Poincaré Disk ($\mathbb{PD}$).

### A.1. Riemannian Geometry on the $2D$ Sphere

The $2D$ Sphere ($\mathbb{S}^2$) of radius $r$ and centered at the origin can be isometrically embedded in $\mathbb{R}^3$ in the natural way, i.e., $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^2 \subset \mathbb{R}^3$. Then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^2$, the Riemannian distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$d_{\mathcal{M}}(\boldsymbol{x}, \boldsymbol{y}) = r \cdot \theta, \quad \theta = \text{acos}\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|}\right).$$

The minimal geodesic between $\boldsymbol{x}$ and $\boldsymbol{y}$ is the piece of the arc on the great circle of $\mathbb{S}^2$ with the smallest length, assuming $\boldsymbol{x}$ and $\boldsymbol{y}$ are not in each others' cut locus, i.e. diametrically opposed. The unit vector on the minimal geodesic from $\boldsymbol{x}$ to $\boldsymbol{y}$, denoted as $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y})$, can be computed as follows

$$\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{y} - \boldsymbol{x} - \text{Proj}_{-\boldsymbol{x}}(\boldsymbol{y} - \boldsymbol{x})}{\left\|\boldsymbol{y} - \boldsymbol{x} - \text{Proj}_{-\boldsymbol{x}}(\boldsymbol{y} - \boldsymbol{x})\right\|}.$$

Here $\text{Proj}_{\boldsymbol{u}}(\boldsymbol{w})$ is the projection of $\boldsymbol{w}$ onto $\boldsymbol{u}$.

### A.2. Riemannian Geometry on the Poincaré Disk

For any two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{PD}$ on the Poincaré Disk ($\mathbb{PD}$) where $\mathbb{PD} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^2 \text{ s.t. } \|\boldsymbol{x}\| < 1\}$, the Riemannian metric, written in the standard coordinates of $\mathbb{R}^2$, is given by

$$g_{i,j}(\boldsymbol{x}) = \frac{4\delta_{i,j}}{(1 - \|\boldsymbol{x}\|^2)^2}, \quad \boldsymbol{x} \in \mathbb{PD},$$

with $\delta_{i,j}$ being the Kronecker delta, and the corresponding Riemannian distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$d_{\mathcal{M}}(\boldsymbol{x}, \boldsymbol{y}) = \text{acosh}\left(1 + \frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{(1 - \|\boldsymbol{x}\|^2)(1 - \|\boldsymbol{y}\|^2)}\right).$$

The minimal geodesics between $\boldsymbol{x}$ and $\boldsymbol{y}$ are either straight line segments if $\boldsymbol{x}$ and $\boldsymbol{y}$ are on a line through the origin or circular arc perpendicular to the boundary. For the straight line segment case, we have the unit vector on the minimal geodesic from $\boldsymbol{x}$ to $\boldsymbol{y}$, denoted as $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y})$, computed as follows: we identify the vector $\boldsymbol{y} - \boldsymbol{x}$, computed in $\mathbb{R}^2$ as a tangent vector in $T_{\boldsymbol{x}}\mathcal{M}$, then normalize it to obtain $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{y} - \boldsymbol{x}}{\|\boldsymbol{y} - \boldsymbol{x}\|_{T_{\boldsymbol{x}}\mathcal{M}}}$. For the perpendicular arc case, we first find the inverse $\boldsymbol{y}'$ of $\boldsymbol{y}$ w.r.t to the unit disk (in $\mathbb{R}^2$); then we use the three points $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{y}'$ to find the center $\boldsymbol{o}'$ of the circle passing through $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{y}'$. Then the unit tangent vector on the geodesic from $\boldsymbol{x}$ to $\boldsymbol{y}$ is computed as follows: , we compute $\boldsymbol{y} - \boldsymbol{x} - \text{Proj}_{\boldsymbol{o}' - \boldsymbol{x}}(\boldsymbol{y} - \boldsymbol{x})$ in $\mathbb{R}^2$ (with the Euclidean metric), then identify it as a tangent vector in $T_{\boldsymbol{x}}\mathcal{M}$, and normalize it:

$$\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{y} - \boldsymbol{x} - \text{Proj}_{\boldsymbol{o}' - \boldsymbol{x}}(\boldsymbol{y} - \boldsymbol{x})}{\left\|\boldsymbol{y} - \boldsymbol{x} - \text{Proj}_{\boldsymbol{o}' - \boldsymbol{x}}(\boldsymbol{y} - \boldsymbol{x})\right\|_{T_{\boldsymbol{x}}\mathcal{M}}}.$$

## B. Learning Theory: Foundation

In this section, we present the theoretical foundation needed to prove the theorems presented in the main body. We follow the ideas presented in (Lu et al., 2019b) with similar strategies presented in (Cucker & Smale, 2002; Györfi et al., 2006). We begin with the following assumption.

**Assumption 1.** *$\mathcal{H}$ is a compact (in $L^\infty$-norm) and convex subset of $L^2([0,R])$, such that every $\varphi \in \mathcal{H}$ is bounded above by some constant $S_0 \geq S$, i.e. $\|\varphi\|_{L^\infty([0,R])} \leq S_0$; moreover $\varphi$ is smooth enough to ensure the existence and uniqueness of solutions of*

$$\dot{\boldsymbol{x}}_i(t) = \frac{1}{N}\sum_{i'=1}^{N} \phi(d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t)))\boldsymbol{w}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t)), \qquad i = 1, \ldots, N. \tag{1}$$

*for $t \in [0,T]$, i.e. $\varphi \in \mathcal{H} \cap \mathcal{K}_{R,S_0}$.*

Another important observation is that since $\phi \in \mathcal{K}_{R,S}$ and $T$ is finite, the distribution of $\boldsymbol{x}_i(t)$'s does not blow up over $[0,T]$ ensuring that the $\boldsymbol{x}_i(t)$'s have bounded distance from the $\boldsymbol{x}_i(0)$'s. In fact, let $R_0$ be the maximum Riemannian distance between any pair of agents at $t = 0$, then

$$\max_{i,i'=1,\ldots,N} r_{i,i'}(t) = \max_{i,i'=1,\ldots,N} d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t)) \leq R_0 + TRS, \quad \text{for } t \in [0,T].$$

Hence the $\boldsymbol{x}_i(t)$'s live in a compact (w.r.t to the $d_{\mathcal{M}}$ metric) ball around the $\boldsymbol{x}_i(0)$'s, denoted as $\mathcal{B}_{\mathcal{M}}(\boldsymbol{X}_0, R_1)$ where $R_1 = R_0 + TRS$. Recall the definition of the loss functional used to find the estimator, namely $\widehat{\phi}_{L,M,\mathcal{H}}$ to the unknown interaction kernel $\phi$, give by

$$\mathcal{E}_{L,M,\mathcal{M}}(\varphi) := \frac{1}{ML}\sum_{l,m=1}^{L,M} \left\|\dot{\boldsymbol{X}}_{t_l}^m - \boldsymbol{f}_\varphi^{\mathrm{c}}(\boldsymbol{X}_{t_l}^m)\right\|_{T_{\boldsymbol{X}_{t_l}^m}\mathcal{M}^N}^2 . \tag{2}$$

Further recall that the estimator is defined as $\widehat{\phi}_{L,M,\mathcal{H}} := \underset{\varphi \in \mathcal{H}}{\arg\min}\, \mathcal{E}_{L,M,\mathcal{M}}(\varphi)$. When $M \to \infty$, we obtain the following loss functional (by the law of large numbers).

$$\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi) := \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_{\boldsymbol{X}_0 \sim \mu_0(\mathcal{M}^N)}\Big[\left\|\dot{\boldsymbol{X}}_{t_l} - \boldsymbol{f}_\varphi^{\mathrm{c}}(\boldsymbol{X}_{t_l})\right\|_{T_{\boldsymbol{X}_{t_l}}\mathcal{M}^N}^2\Big]. \tag{3}$$

The minimizer of $\mathcal{E}_{L,\infty,\mathcal{M}}$ over $\mathcal{H}$ is defined as $\widehat{\phi}_{L,\infty,\mathcal{H}}$, which is closely related to $\widehat{\phi}_{L,M,\mathcal{H}}$ (in the $M \to \infty$ sense). And they are close to $\phi$, when we establish the following condition on $\mathcal{H}$.

**Definition B.1** (Geometric Coercivity condition)**.** *The geometric evolution system in (1) with initial condition sampled from $\mu_0(\mathcal{M}^N)$ on $\mathcal{M}^N$ is said to satisfy the geometric coercivity condition on the admissible hypothesis space $\mathcal{H}$ if there exists a constant $c_{L,N,\mathcal{H},\mathcal{M}} > 0$ such that for any $\varphi \in \mathcal{H}$ with $\varphi(\cdot)\cdot \in L^2(\rho_{T,\mathcal{M}}^L)$, the following inequality holds:*

$$c_{L,N,\mathcal{H},\mathcal{M}} \|\varphi(\cdot)\cdot\|_{L^2(\rho_{T,\mathcal{M}}^L)}^2 \leq \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_{\boldsymbol{X}_0 \sim \mu_0(\mathcal{M}^N)}\Big[\|\boldsymbol{f}_\varphi^c(\boldsymbol{X}_{t_l})\|_{T_{\boldsymbol{X}_{t_l}}\mathcal{M}^N}^2\Big]. \tag{4}$$

From this condition, we can derive the following theorem.

**Theorem B.1.** *Let $\phi \in L^2([0,R])$, and $\mathcal{H}$ a compact (w.r.t the $L^\infty$ norm) and convex subset of $L^2([0,R])$ such that the geometric coercivity condition (4) holds with a constant $c_{L,N,\mathcal{H},\mathcal{M}}$. Then, for $\widehat{\phi}_{L,M,\mathcal{H}}$, estimated by minimizing (2) on the trajectory data generated by (1), the following inequality*

$$\left\|\widehat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot - \phi(\cdot)\cdot\right\|_{L^2(\rho_{T,\mathcal{M}}^L)}^2 \leq \frac{2}{c_{L,N,\mathcal{H},\mathcal{M}}}\Big(\epsilon + \inf_{\varphi \in \mathcal{H}} \|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_{T,\mathcal{M}}^L)}^2\Big) \tag{5}$$

*holds with probability at least $1 - \tau$, when $M \geq \frac{1152 S_0^2 R^2}{\epsilon c_{L,N,\mathcal{H},\mathcal{M}}}\Big(\ln(\mathcal{N}(\mathcal{H}, \frac{\epsilon}{48 S_0 R^2})) + \ln(\frac{1}{\tau})\Big)$. Here $\mathcal{N}(\mathcal{U}, \epsilon)$ is the covering number of a set $\mathcal{U}$ with open balls of radius $\epsilon$ w.r.t the $L^\infty$-norm.*

Using this concentration result, we can get the strong consistency of our estimators under mild hypotheses.

**Theorem B.2.** *For a family of compact (w.r.t. the $L^\infty$ norm) convex subsets, $\{\mathcal{H}_M\}_{M=1}^\infty$, of $L^2([0,R])$, when the following conditions hold, (i) $\cup_M \mathcal{H}_M$ is compact in $L^\infty$; (ii) the geometric coercivity condition,* (B.1)*, holds on $\cup_M \mathcal{H}_M$; (iii) $\inf_{\varphi\in\mathcal{H}_M} \|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})} \overset{M\to\infty}{\longrightarrow} 0$, then*

$$\lim_{M\to\infty} \left\| \widehat{\phi}_{L,M,\mathcal{H}_M}(\cdot)\cdot - \phi(\cdot)\cdot \right\|_{L^2(\rho^L_{T,\mathcal{M}})} = 0 \quad a.s. \tag{6}$$

*This theorem establishes the almost sure convergence of our estimator to the true interaction kernel as $M \to \infty$.*

### B.1. Concentration and Consistency

Our first step is to establish the consistency of the estimator for the true kernel $\phi$ of the system. Note that $\mathcal{H}$ can be embedded as a compact (in $L^\infty$ sense) set of $L^2(\rho^L_{T,\mathcal{M}})$. We establish a strong consistency result on our estimators of the form,

$$\lim_{M\to\infty} \left\| \widehat{\phi}_{L,M}(\cdot)\cdot - \phi(\cdot)\cdot \right\|_{L^2(\rho^L_{T,\mathcal{M}})} = 0, \ a.s.$$

Our discussions of consistency under the $L^2-$norm on manifolds can be regarded as a natural extension from the case on Euclidean Space in (Lu et al., 2019b). We define the following loss functional of the vectorized system, $\boldsymbol{X}_t$

$$\begin{aligned}\mathcal{E}_{\boldsymbol{X}_t}(\varphi) &:= \frac{1}{N}\sum_{i=1}^N \left\| \frac{1}{N}\sum_{i'=1}^N (\phi_{ii',t} - \varphi_{ii',t})\boldsymbol{w}_{ii',t} \right\|^2_{T_{\boldsymbol{x}_i(t)}\mathcal{M}} \\ &= \frac{1}{N}\sum_{i=1}^N \langle \frac{1}{N}\sum_{i'=1}^N (\phi_{ii',t} - \varphi_{ii',t})\boldsymbol{w}_{ii',t}, \frac{1}{N}\sum_{i''=1}^N (\phi_{ii'',t} - \varphi_{ii'',t})\boldsymbol{w}_{ii'',t}\rangle_{g(\boldsymbol{x}_i(t))}.\end{aligned} \tag{7}$$

Here we take $\boldsymbol{w}_{ii',t} = d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))\boldsymbol{v}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))$ and $\phi_{ii',t} = \phi(d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t)))$; similarly for $\varphi_{ii',t}$. Now we can see that

$$\mathcal{E}_{L,M,\mathcal{M}}(\varphi) = \frac{1}{LM}\sum_{l,m=1}^{L,M} \mathcal{E}_{\boldsymbol{X}^m_{t_l}}(\varphi).$$

When $M\to\infty$, this functional converges to, by the law of large numbers,

$$\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi) = \frac{1}{L}\sum_{l=1}^L \mathbb{E}_{\boldsymbol{X}_0\sim\mu_0(\mathcal{M}^N)}\mathcal{E}_{\boldsymbol{X}_{t_l}}(\varphi)\,.$$

We are ready to summarize some basic properties of $\mathcal{E}_{\boldsymbol{X}_t}(\varphi)$.

**Proposition 1.** *For $\varphi_1, \varphi_2 \in \mathcal{H}$, we have*

$$\left|\mathcal{E}_{\boldsymbol{X}_t}(\varphi_1) - \mathcal{E}_{\boldsymbol{X}_t}(\varphi_2)\right| \le \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})} \|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})}\,. \tag{8}$$

*Here we define the probability measure, $\hat{\rho}^t_{\mathcal{M}}(r) := \frac{1}{N^2}\sum_{i,i'=1}^N \delta_{d_{\mathcal{M}}(\boldsymbol{x}_i(t),\boldsymbol{x}_{i'}(t))}(r)$.*

*Proof.* Let $\varphi_1, \varphi_2 \in \mathcal{H}$, and define $\varphi^1_{ii',t} := \varphi_1(d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t)))$, similarly for $\varphi^2_{ii',t}$. Moreover, let $r_{ii',t} := d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))$ and $\boldsymbol{w}_{ii',t} := d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))\boldsymbol{v}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))$. Immediately, we have

$$\|\boldsymbol{w}_{ii',t}\|_{T_{\boldsymbol{x}_i(t)}\mathcal{M}} \le r_{ii',t},$$

since $\boldsymbol{v}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))$ has either length 1 or 0. Next, using Jensen's inequality, we have

$$\begin{aligned}
\left|\mathcal{E}_{\boldsymbol{X}_t}(\varphi_1) - \mathcal{E}_{\boldsymbol{X}_t}(\varphi_2)\right| &= \Big|\frac{1}{N}\sum_{i=1}^{N}\langle\frac{1}{N}\sum_{i'=1}^{N}(\varphi^1_{ii',t} - \varphi^2_{ii',t})\boldsymbol{w}_{ii',t}, \frac{1}{N}\sum_{i''=1}^{N}(2\phi_{ii'',t} - \varphi^1_{ii'',t} - \varphi^2_{ii'',t})\boldsymbol{w}_{ii'',t}\rangle_{g(\boldsymbol{x}_i(t))}\Big| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{N}\sum_{i'=1}^{N}(\varphi^1_{ii',t} - \varphi^2_{ii',t})\boldsymbol{w}_{ii',t}\right\|_{T_{\boldsymbol{x}_i(t)}\mathcal{M}}\left\|\frac{1}{N}\sum_{i''=1}^{N}(2\phi_{ii'',t} - \varphi^1_{ii'',t} - \varphi^2_{ii'',t})\boldsymbol{w}_{ii'',t}\right\|_{T_{\boldsymbol{x}_i(t)}\mathcal{M}} \\
&\leq \sqrt{\frac{1}{N^2}\sum_{i,i'=1}^{N}(\varphi^1_{ii',t} - \varphi^2_{ii',t})r^2_{ii',t}}\sqrt{\frac{1}{N^2}\sum_{i,i''=1}^{N}(2\phi_{ii'',t} - \varphi^1_{ii',t} - \varphi^2_{ii',t})r^2_{ii'',t}} \\
&\leq \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot)\|_{L^2(\hat{\rho}^t_{\mathcal{M}})}\,\|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})},
\end{aligned}$$

where $\hat{\rho}^t_{\mathcal{M}}(r) = \frac{1}{N^2}\sum_{i,i'=1}^{N}\delta_{r_{ii',t}}(r)$. $\square$

With Proposition 1 proven, we get the following proposition establishing the continuity of our error functionals.

**Proposition 2.** *For $\varphi_1, \varphi_2 \in \mathcal{H}$, we have the inequalities*

$$\begin{aligned}
\left|\mathcal{E}_{L,M,\mathcal{M}}(\varphi_1) - \mathcal{E}_{L,M,\mathcal{M}}(\varphi_2)\right| &\leq \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^\infty}\,\|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^\infty} \\
\left|\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi_1) - \mathcal{E}_{L,\infty,\mathcal{M}}(\varphi_2)\right| &\leq \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})}\,\|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})}.
\end{aligned} \tag{9}$$

*Proof.* Using the results from Prop. 1, and defining $\hat{\rho}^L_{T,\mathcal{M}} \coloneqq \frac{1}{L}\sum_{l=1}^{L}\hat{\rho}^{t_l}_{\mathcal{M}}$, we have

$$\begin{aligned}
\Big|\frac{1}{L}\sum_{l=1}^{L}\mathcal{E}_{\boldsymbol{X}_{t_l}}(\varphi_1) - \frac{1}{L}\sum_{l=1}^{L}\mathcal{E}_{\boldsymbol{X}_{t_l}}(\varphi_2)\Big| &\leq \frac{1}{L}\sum_{l=1}^{L}\left|\mathcal{E}_{\boldsymbol{X}_{t_l}}(\varphi_1) - \mathcal{E}_{\boldsymbol{X}_{t_l}}(\varphi_2)\right| \\
&< \frac{1}{L}\sum_{l=1}^{L}\|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})}\,\|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})} \\
&\leq \sqrt{\frac{1}{L}\sum_{1=1}^{L}\|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})}}\sqrt{\frac{1}{L}\sum_{l=1}^{L}\|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})}} \\
&= \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^L_{T,\mathcal{M}})}\,\|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^L_{T,\mathcal{M}})}
\end{aligned}$$

Next, we have

$$\begin{aligned}
\left|\mathcal{E}_{L,M,\mathcal{M}}(\varphi_1) - \mathcal{E}_{L,M,\mathcal{M}}(\varphi_2)\right| &\leq \frac{1}{M}\sum_{m=1}^{M}\Big|\frac{1}{L}\sum_{l=1}^{L}\mathcal{E}_{\boldsymbol{X}^m_{t_l}}(\varphi_1) - \frac{1}{L}\sum_{l=1}^{L}\mathcal{E}_{\boldsymbol{X}^m_{t_l}}(\varphi_2)\Big| \\
&\leq \frac{1}{M}\sum_{m=1}^{M}\|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^L_{T,\mathcal{M}})}\,\|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^L_{T,\mathcal{M}})} \\
&\leq \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^\infty}\,\|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^\infty} \\
&\leq R^2\,\|\varphi_1 - \varphi_2\|_{L^\infty}\,\|2\phi - \varphi_1 - \varphi_2\|_{L^\infty}.
\end{aligned}$$

Meanwhile, taking $M \to \infty$ for $\left|\mathcal{E}_{L,M,\mathcal{M}}(\varphi_1) - \mathcal{E}_{L,M,\mathcal{M}}(\varphi_2)\right|$, we obtain

$$\left|\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi_1) - \mathcal{E}_{L,\infty,\mathcal{M}}(\varphi_2)\right| \leq \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})}\,\|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})},$$

where $\rho^L_{T,\mathcal{M}} = \mathbb{E}_{\boldsymbol{X}_0\sim\mu_0(\mathcal{M}^N)}[\hat{\rho}^L_{T,\mathcal{M}}]$. $\square$

As a further derivation, we observe that for any $\varphi \in \mathcal{H} \subset L^2([0,R])$, we have that $\max_{r\in[0,R]}\left|\varphi(\cdot)\cdot\right| \leq R\max_{r\in[0,R]}\left|\varphi(\cdot)\right|$, so we obtain the following Corollary:

**Corollary B.3.** *For $\varphi \in \mathcal{H}$, define*

$$\mathcal{L}_M(\psi) := \mathcal{E}_{L,\infty,\mathcal{M}}(\varphi) - \mathcal{E}_{L,M,\mathcal{M}}(\varphi),$$

*then for any $\varphi_1, \varphi_2 \in \mathcal{H}$, we have*

$$\left|\mathcal{L}_M(\varphi_1) - \mathcal{L}_M(\varphi_2)\right| \leq 2R^2 \left\|\varphi_1 - \varphi_2\right\|_{L^\infty} \left\|2\phi - \varphi_1 - \varphi_2\right\|_{L^\infty}.$$

Now we can consider the distance between the minimizer of the error functional $\mathcal{E}_{L,\infty,\mathcal{M}}$ over $\mathcal{H}$ and any other $\varphi \in \mathcal{H}$. Let

$$\widehat{\phi}_{L,\infty,\mathcal{H}} = \underset{\varphi\in\mathcal{H}}{\arg\min}\, \mathcal{E}_{L,\infty,\mathcal{M}}(\varphi).$$

From the geometric coercivity condition and the convexity of $\mathcal{H}$, we obtain

**Proposition 3.** *For any $\varphi \in \mathcal{H}$,*

$$\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi) - \mathcal{E}_{L,\infty,\mathcal{M}}(\widehat{\phi}_{L,\infty,\mathcal{H}}) \geq c_{L,N,\mathcal{H},\mathcal{M}} \left\|\varphi(\cdot)\cdot - \widehat{\phi}_{L,\infty,\mathcal{H}}(\cdot)\cdot\right\|_{L^2(\rho^L_{T,\mathcal{M}})}. \tag{10}$$

We now define the defect function $\mathcal{D}_{L,M,\mathcal{H}}(\varphi) := \mathcal{E}_{L,M,\mathcal{M}}(\varphi) - \mathcal{E}_{L,M,\mathcal{M}}(\widehat{\phi}_{L,\infty,\mathcal{H}})$, and define

$$\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi) := \lim_{M\to\infty} \mathcal{D}_{L,M,\mathcal{H}}(\varphi) = \mathcal{E}_{L,\infty,\mathcal{H}}(\varphi) - \mathcal{E}_{L,\infty,\mathcal{M}}(\widehat{\phi}_{L,\infty,\mathcal{H}}).$$

Then, we show that we can uniformly bound $\frac{\mathcal{D}_{L,\infty,\mathcal{H}}(\cdot) - \mathcal{D}_{L,M,\mathcal{H}}(\cdot)}{\mathcal{D}_{L,\infty,\mathcal{H}}(\cdot)+\epsilon}$ on $\mathcal{H}$ with high probability,

**Proposition 4.** *For any $\epsilon > 0$ and $\alpha \in (0,1)$, we have*

$$\mathbb{P}_{\mu_0(\mathcal{M}^N)}\left(\sup_{\varphi\in\mathcal{H}} \frac{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi) - \mathcal{D}_{L,M,\mathcal{H}}(\varphi)}{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi)+\epsilon} \geq 3\alpha\right) \leq \mathcal{N}\Big(\mathcal{H}, \frac{\alpha\epsilon}{8S_0R^2}\Big) \exp\left(-\frac{c_{L,N,\mathcal{H},\mathcal{M}}\alpha^2 M\epsilon}{32S_0^2}\right)$$

*where $\mathcal{N}(U,r)$ is the covering number of set $U$ with open balls of radius $r$ w.r.t the $L^\infty-$norm.*

The proof of Proposition 4 uses the following Lemma similar to Lemma 19 in (Lu et al., 2019b),

**Lemma B.4.** *For any $\epsilon > 0$ and $\alpha \in (0,1)$, if $\varphi_1 \in \mathcal{H}$ satisfies*

$$\frac{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi_1) - \mathcal{D}_{L,M,\mathcal{H}}(\varphi_1)}{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi_1)+\epsilon} < \alpha$$

*then for any $\varphi_2 \in \mathcal{H}$ s.t. $\left\|\varphi_1 - \varphi_2\right\|_{L^\infty} \leq r_0 = \frac{\alpha\epsilon}{8S_0R^2}$, we have*

$$\frac{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi_2) - \mathcal{D}_{L,M,\mathcal{H}}(\varphi_2)}{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi_2)+\epsilon} < 3\alpha$$

Using the results we have just established, the proofs of theorems B.1 and B.2 now follow similarly to the analogous results in (Lu et al., 2019b;a; Miller et al., 2020).

### B.2. Rate of Convergence

Using these results, we establish the convergence rate of $\widehat{\phi}_{L,M,\mathcal{H}}$ to $\phi$ as $M$ increases.

**Theorem B.5.** *Let $\mu_0(\mathcal{M}^N)$ be the distribution of the initial conditions of trajectories, and $\mathcal{H}_M = \mathcal{B}_n$ with $n \asymp (M/\log M)^{\frac{1}{2s+1}}$, where $\mathcal{B}_n$ is the central ball of $\mathcal{L}_n$ with radius $c_1 + S$, and the linear space $\mathcal{L}_n \subseteq L^\infty([0,R])$ satisfies the dimension and approximation conditions below,*

$$dim(\mathcal{L}_n) \leq c_0 n \quad \text{and} \quad \inf_{\varphi\in\mathcal{L}_n} \left\|\varphi - \phi\right\|_{L^\infty} \leq c_1 n^{-s}$$

*for some constants $c_0, c_1, s > 0$. Suppose that the geometric coercivity condition holds on $\mathcal{L} := \cup_n \mathcal{L}_n$ with constant $c_{L,N,\mathcal{L},\mathcal{M}}$. Then there exists some constant $C(S,R,c_0,c_1)$ such that*

$$\mathbb{E}\left[\left\|\widehat{\phi}_{L,M,\mathcal{H}_M}(\cdot)\cdot - \phi(\cdot)\cdot\right\|_{L^2(\rho^L_{T,\mathcal{M}})}\right] \leq \frac{C(S,R,c_0,c_1)}{c_{L,N,\mathcal{L},\mathcal{M}}} \Big(\frac{\log M}{M}\Big)^{\frac{s}{2s+1}}.$$

The proof of the theorem uses the results above, which took into account the geometry of $\mathcal{M}$, while closely following the ideas in (Lu et al., 2019b) and their further development in (Lu et al., 2019a; Miller et al., 2020), and is therefore omitted.

### B.3. Trajectory Estimation Error

Recall the following theorem on the trajectory estimator error:

**Theorem B.6.** *Let $\phi \in \mathcal{K}_{R,S}$ and $\widehat{\phi} \in \mathcal{K}_{R,S_0}$, for some $S_0 \geq S$. Suppose that $\boldsymbol{X}_{[0,T]}$ and $\hat{\boldsymbol{X}}_{[0,T]}$ are solutions of (1) w.r.t to $\phi$ and $\widehat{\phi}$, respectively, for $t \in [0,T]$, with $\hat{\boldsymbol{X}}_0 = \boldsymbol{X}_0$. Then the following inequalities hold:*

$$d_{traj,\mathcal{M}^N}\left(\boldsymbol{X}_{[0,T]}, \hat{\boldsymbol{X}}_{[0,T]}\right)^2 \leq 4C(\mathcal{M},T)T\exp(64T^2S_0^2)\left\|\dot{\boldsymbol{X}}_t - \boldsymbol{f}^c_{\widehat{\phi}}(\boldsymbol{X}_t)\right\|^2_{T_{\boldsymbol{X}_t}M^N}, \tag{11}$$

*and*

$$\mathbb{E}_{\boldsymbol{X}_0\sim\mu_0(\mathcal{M}^N)}\left[d_{traj,\mathcal{M}^N}\left(\boldsymbol{X}_{[0,T]}, \hat{\boldsymbol{X}}_{[0,T]}\right)^2\right] \leq 4C(\mathcal{M},T)T^2\exp(64T^2S_0^2)\left\|\phi(\cdot)\cdot - \widehat{\phi}(\cdot)\cdot\right\|^2_{L^2(\rho_{T,\mathcal{M}})}, \tag{12}$$

*where $C(\mathcal{M},T)$ is a positive constant depending only on geometric properties of $\mathcal{M}$ and on $T$, but may be chosen independent of $T$ if $\mathcal{M}$ is compact.*

It states two different estimates of the trajectory estimation error. First, it bounds the system trajectory error for any one single initial condition; second, it bounds the expectation of the worst trajectory estimation error on time interval $[0,T]$ among all different initial conditions.

*Proof of Theorem B.6.* Assume that $\phi \in \mathcal{K}_{R,S}$, $\widehat{\phi} \in \mathcal{K}_{R,S_0}$, and $\boldsymbol{X}_t, \hat{\boldsymbol{X}}_t$ are two system states, at some $t \in [0,T]$, generated by $\phi, \widehat{\phi}$ with the same initial conditions at $t = 0$. Next, we assume that $\mathcal{M}$ is isometrically embedded in $\mathbb{R}^{d'}$ (at least one such embedding exists, by Nash's embedding theorem), via a map $\mathcal{I} : \mathcal{M} \to \mathbb{R}^{d'}$. From now on, we will identify $\boldsymbol{x}_i$ with $\mathcal{I}\boldsymbol{x}_i$. Then for any $t \in [0,T]$, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \|\boldsymbol{x}_i(t) - \hat{\boldsymbol{x}}_i(t)\|^2_{\mathbb{R}^{d'}} &= \frac{1}{N}\sum_{i=1}^N \left\|\int_{s=0}^t (\dot{\boldsymbol{x}}_i(s) - \dot{\hat{\boldsymbol{x}}}_i(s))\,ds\right\|^2_{\mathbb{R}^{d'}} \leq \frac{1}{N}\sum_{i=1}^N t\int_{s=0}^t \left\|\dot{\boldsymbol{x}}_i(s) - \dot{\hat{\boldsymbol{x}}}_i(s)\right\|^2_{\mathbb{R}^{d'}}\,ds \\ &\leq \frac{T}{N}\sum_{i=1}^N \int_{s=0}^t \left\|\dot{\boldsymbol{x}}_i(s) - \dot{\hat{\boldsymbol{x}}}_i(s)\right\|^2_{\mathbb{R}^{d'}}\,ds.\end{aligned}$$

Define the function $F^{\mathcal{M}}_{\varphi}(\boldsymbol{x},\cdot) : \mathcal{M} \to T_{\boldsymbol{x}}\mathcal{M}$ for every $\boldsymbol{x} \in \mathcal{M}$ as $F^{\mathcal{M}}_{\varphi}(\boldsymbol{x},\cdot) := \varphi(d_{\mathcal{M}}(\boldsymbol{x},\cdot))\boldsymbol{w}(\boldsymbol{x},\cdot)$. Let $F^{\mathcal{M}}_{\varphi,ii',t} = F^{\mathcal{M}}_{\varphi}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))$ and $F^{\mathcal{M}}_{\varphi,\hat{i}\hat{i}',t} = F^{\mathcal{M}}_{\varphi}(\hat{\boldsymbol{x}}_i(t), \hat{\boldsymbol{x}}_{i'}(t))$. Then

$$\begin{aligned}\sum_{i=1}^N \int_{s=0}^t \left\|\dot{\boldsymbol{x}}_i(s) - \dot{\hat{\boldsymbol{x}}}_i(s)\right\|^2_{\mathbb{R}^{d'}}\,ds &= \sum_{i=1}^N \int_{s=0}^t \left\|\dot{\boldsymbol{x}}_i(s) - \frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\widehat{\phi},\hat{i}\hat{i}',s}\right\|^2_{\mathbb{R}^{d'}}\,ds \\ &\leq 2\sum_{i=1}^N \int_{s=0}^t \Big(\left\|\dot{\boldsymbol{x}}_i(s) - \frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\widehat{\phi},ii',s}\right\|^2_{\mathbb{R}^{d'}} + \left\|\frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\widehat{\phi},ii',s} - \frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\widehat{\phi},\hat{i}\hat{i}',s}\right\|^2_{\mathbb{R}^{d'}}\Big)\,ds \\ &= 2\sum_{i=1}^N \int_{s=0}^t \Big(\left\|\dot{\boldsymbol{x}}_i(s) - \frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\widehat{\phi},ii',s}\right\|^2_{\mathbb{R}^{d'}} + I(s)\Big)\,ds\,.\end{aligned}$$

Next,

$$\begin{aligned}I(s) &= \left\|\frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\widehat{\phi},ii',s} - \frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\widehat{\phi},\hat{i}\hat{i}',s}\right\|^2_{\mathbb{R}^{d'}} = \frac{1}{N^2}\left\|\sum_{i'=1}^N \big(F^{\mathcal{M}}_{\widehat{\phi},ii',s} - F^{\mathcal{M}}_{\widehat{\phi},i\hat{i}',s} + F^{\mathcal{M}}_{\widehat{\phi},i\hat{i}',s} - F^{\mathcal{M}}_{\widehat{\phi},\hat{i}\hat{i}',s}\big)\right\|^2_{\mathbb{R}^{d'}} \\ &\leq \frac{2}{N^2}\Big(\left\|\sum_{i'=1}^N \big(F^{\mathcal{M}}_{\widehat{\phi},ii',s} - F^{\mathcal{M}}_{\widehat{\phi},i\hat{i}',s}\big)\right\|^2_{\mathbb{R}^{d'}} + \left\|\sum_{i'=1}^N \big(F^{\mathcal{M}}_{\widehat{\phi},i\hat{i}',s} - F^{\mathcal{M}}_{\widehat{\phi},\hat{i}\hat{i}',s}\big)\right\|^2_{\mathbb{R}^{d'}}\Big)\,.\end{aligned}$$

Since $\widehat{\phi} \in \mathcal{K}_{R,S_0}$, $F^{\mathcal{M}}_{\widehat{\phi}}$ is Lipschitz in each of its arguments; moreover, $\max_{r\in[0,R]} |\widehat{\phi}| \leq S_0$, so that $\mathrm{Lip}(F^{\mathcal{M}}_{\widehat{\phi}}(\boldsymbol{x},\cdot))$, $\mathrm{Lip}(F^{\mathcal{M}}_{\widehat{\phi}}(\cdot,\boldsymbol{x})) \leq 2S_0$. Therefore,

$$\begin{aligned}
I(s) &\leq \frac{2}{N^2}\Big(2\mathrm{Lip}(F^{\mathcal{M}}_{\widehat{\phi}}(\boldsymbol{x}_i(s),\cdot))^2 \sum_{i'=1}^N \|\boldsymbol{x}_{i'}(s)-\hat{\boldsymbol{x}}_{i'}(s)\|^2_{\mathbb{R}^{d'}} + 2\sum_{i'=1}^N \mathrm{Lip}(F^{\mathcal{M}}_{\widehat{\phi}}(\cdot,\hat{\boldsymbol{x}}_{i'}(s)))^2 \|\boldsymbol{x}_i(s)-\hat{\boldsymbol{x}}_i(s)\|^2_{\mathbb{R}^{d'}}\Big)\\
&\leq \frac{4}{N^2}\mathrm{Lip}(F^{\mathcal{M}}_{\widehat{\phi}}(\boldsymbol{x}_i(s),\cdot))^2 \sum_{i'=1}^N \|\boldsymbol{x}_{i'}(s)-\hat{\boldsymbol{x}}_{i'}(s)\|^2_{\mathbb{R}^{d'}} + \frac{4}{N^2}\sum_{i'=1}^N \mathrm{Lip}(F^{\mathcal{M}}_{\widehat{\phi}}(\cdot,\hat{\boldsymbol{x}}_{i'}(s)))^2 \|\boldsymbol{x}_{i'}(s)-\hat{\boldsymbol{x}}_{i'}(s)\|^2_{\mathbb{R}^{d'}}\\
&\leq \frac{16S_0^2}{N^2}\sum_{i'=1}^N \|\boldsymbol{x}_{i'}(s)-\hat{\boldsymbol{x}}_{i'}(s)\|^2_{\mathbb{R}^{d'}} + \frac{16S_0^2}{N^2}\sum_{i'=1}^N \|\boldsymbol{x}_{i'}(s)-\hat{\boldsymbol{x}}_{i'}(s)\|^2_{\mathbb{R}^{d'}}\\
&\leq \frac{32S_0^2}{N^2}\sum_{i'=1}^N \|\boldsymbol{x}_{i'}(s)-\hat{\boldsymbol{x}}_{i'}(s)\|^2_{\mathbb{R}^{d'}}\ .
\end{aligned}$$

Putting these results together, we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \|\boldsymbol{x}_i(t)-\hat{\boldsymbol{x}}_i(t)\|^2_{\mathbb{R}^{d'}} &\leq \frac{2T}{N}\sum_{i=1}^N \int_{s=0}^t \Big( \Big\|\dot{\boldsymbol{x}}_i(s) - \frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\widehat{\phi},ii',s}\Big\|^2_{\mathbb{R}^{d'}} + \frac{32S_0^2}{N^2}\sum_{i'=1}^N \|\boldsymbol{x}_{i'}(s)-\hat{\boldsymbol{x}}_{i'}(s)\|^2_{\mathbb{R}^{d'}}\Big)\, ds\\
&= \frac{64TS_0^2}{N}\sum_{i=1}^N \|\boldsymbol{x}_i(t)-\hat{\boldsymbol{x}}_i(t)\|^2_{\mathbb{R}^{d'}} + \frac{2T}{N}\sum_{i=1}^N\int_{s=0}^t \Big\|\dot{\boldsymbol{x}}_i(s) - \frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\widehat{\phi},ii',s}\Big\|^2_{\mathbb{R}^{d'}}\, ds.
\end{aligned}$$

By Grönwall's inequality, we have

$$\frac{1}{N}\sum_{i=1}^N \|\boldsymbol{x}_i(t)-\hat{\boldsymbol{x}}_i(t)\|^2_{\mathbb{R}^{d'}} \leq \frac{2T}{N}\exp(64T^2S_0^2)\sum_{i=1}^N\int_{s=0}^t \Big\|\dot{\boldsymbol{x}}_i(s) - \frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\widehat{\phi},ii',s}\Big\|^2_{\mathbb{R}^{d'}}\, ds.$$

Recall that $T$ is small, hence the solution $\boldsymbol{X}_t$ and $\hat{\boldsymbol{X}}_t$ live in a compact neighborhood of the initial condition, $\boldsymbol{X}_0 = \hat{\boldsymbol{X}}_0 \in \mathcal{M}^N$; i.e. $\boldsymbol{X}_t, \hat{\boldsymbol{X}}_t \in \mathcal{B}_{\mathcal{M}}(\boldsymbol{X}_0, R_2)$ with $R_2 = R_0 + TRS_0$. From the compactness of (the closure of) this set, and via the embedding $\mathcal{I}$, we deduce that there exists a constant $C_1(\mathcal{M},\mathcal{I},T)$ such that

$$d_{\mathcal{M}}(\boldsymbol{x}_i(t),\hat{\boldsymbol{x}}_i(t)) \leq C_1(\mathcal{M},\mathcal{I},T)\,\|\boldsymbol{x}_i(t)-\hat{\boldsymbol{x}}_i(t)\|_{\mathbb{R}^{d'}}\ , \quad \text{for } t\in[0,T].$$

Since $\mathcal{I}$ is isometric, for $\boldsymbol{u} \in T_{\boldsymbol{x}}\mathcal{M}$ we have $\|d\mathcal{I}(\boldsymbol{u})\|_{\mathbb{R}^{d'}} = \|\boldsymbol{u}\|_{T_{\boldsymbol{x}}\mathcal{M}}$. Using both the bounds above, we have

$$\begin{aligned}
d_{\mathcal{M}}(\boldsymbol{X}_t,\hat{\boldsymbol{X}}_t)^2 = \frac{1}{N}\sum_{i=1}^N d_{\mathcal{M}}(\boldsymbol{x}_i(t),\hat{\boldsymbol{x}}_i(t))^2 &\leq \frac{C_1(\mathcal{M},\mathcal{I},T)^2}{N}\sum_{i=1}^N \|\boldsymbol{x}_i(t)-\hat{\boldsymbol{x}}_i(t)\|^2_{\mathbb{R}^{d'}}\\
&\leq \frac{2C_1(\mathcal{M},\mathcal{I},T)^2T\exp(64T^2S_0^2)}{N}\sum_{i=1}^N\int_{s=0}^t \Big\|\dot{\boldsymbol{x}}_i(s) - \frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\widehat{\phi},ii',s}\Big\|^2_{\mathbb{R}^{d'}}\, ds.\\
&= \frac{2C_1(\mathcal{M},\mathcal{I},T)^2T\exp(64T^2S_0^2)}{N}\sum_{i=1}^N\int_{s=0}^t \Big\|\dot{\boldsymbol{x}}_i(s) - \frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\widehat{\phi},ii',s}\Big\|^2_{T_{\boldsymbol{x}_i(s)}\mathcal{M}}\, ds\\
&= 2C_1(\mathcal{M},\mathcal{I},T)^2T\exp(64T^2S_0^2)\int_{s=0}^t \Big\|\dot{\boldsymbol{X}}_s - \boldsymbol{f}^{\mathrm{c}}_{\widehat{\phi}}(\boldsymbol{X}_s)\Big\|^2_{T_{\boldsymbol{X}_s}\mathcal{M}^N}\, ds
\end{aligned}$$

Letting

$$C(\mathcal{M},T) := \inf_{\text{all isometric embeddings } \mathcal{I}} C_1(\mathcal{M},\mathcal{I},T)^2\,,$$

and choosing an isometric embedding $\mathcal{I}$ which gives a value at most twice the infimum, we obtain

$$d_{\mathcal{M}}(\boldsymbol{X}_t,\hat{\boldsymbol{X}}_t)^2 \leq 4TC(\mathcal{M},T)\exp(64T^2S_0^2)\int_{s=0}^t \Big\|\dot{\boldsymbol{X}}_s - \boldsymbol{f}^{\mathrm{c}}_{\widehat{\phi}}(\boldsymbol{X}_s)\Big\|^2_{T_{\boldsymbol{X}}\mathcal{M}^N}\, ds.$$

Now, take $\phi$ to be the true interaction kernel, and $\widehat{\phi}$ the estimator of $\phi$ by our learning approach, by Prop. 1 we have that

$$\frac{1}{T}\int_{t=0}^{T}\left\|\dot{\boldsymbol{X}}_s-\boldsymbol{f}^{\mathrm{c}}_{\widehat{\phi}}(\boldsymbol{X}_s)\right\|^2_{T_{\boldsymbol{X}}\mathcal{M}^N}\,dt\le\left\|\phi(\cdot)\cdot-\widehat{\phi}(\cdot)\cdot\right\|^2_{L^2(\rho_{T,\mathcal{M}})}.$$

Together with (11), recalling that $\hat{\boldsymbol{X}}_0=\boldsymbol{X}_0$ and $\boldsymbol{X}_0\sim\mu_0(\mathcal{M}^N)$, we have the desired result that

$$\mathbb{E}_{\boldsymbol{X}_0\sim\mu_0(\mathcal{M}^N)}\left[d_{\text{traj},\mathcal{M}}(\boldsymbol{X}_{[0,T]},\hat{\boldsymbol{X}}_{[0,T]})^2\right]\le 4T^2C(\mathcal{M},T)\exp(64T^2S_0^2)\mathbb{E}_{\boldsymbol{X}_0\sim\mu_0(\mathcal{M}^N)}\left\|\phi(\cdot)\cdot-\widehat{\phi}(\cdot)\cdot\right\|^2_{L^2(\rho_{T,\mathcal{M}})}.$$

□

## C. Numerical Implementations

If the trajectory data, $\{\boldsymbol{x}_i^m(t_l),\dot{\boldsymbol{x}}_i^m(t_l)\}_{i,l,m=1}^{N,L,M}$, is given by the user, we use the following geometry-based algorithm to find the minimizer of (2). First, we construct a finite dimensional subspace of the hypothesis space, i.e. $\mathcal{H}_M\subset\mathcal{H}$, where $\mathcal{H}_M$ with dimension $\dim(\mathcal{H}_M)=n=n(M)\approx\mathcal{O}(M^{\frac{1}{3}})$ is a space of clamped B-spline functions[1] supported on $[R^{\text{obs}}_{\min},R^{\text{obs}}_{\max}]$ with $R^{\text{obs}}_{\min}/R^{\text{obs}}_{\max}$ being the minimum/maximum interaction radius computed from the observation data. Hence the test functions can be expressed as linear combination of the basis functions of $\mathcal{H}_M$, i.e., $\varphi(r)=\sum_{\eta=1}^{n}\alpha_\eta\psi_\eta(r)$ with $\{\psi_\eta\}_{\eta=1}^{n}$ being a basis for $\mathcal{H}_M$. Next, we use either a local chart $\mathcal{U}:\mathcal{M}\to\mathbb{R}^d$ or a natural embedding $\mathcal{I}:\mathcal{M}\to\mathbb{R}^{d'}$, such that $\boldsymbol{x}_i\in\mathcal{M}$ can be expressed using either local coordinates in $\mathbb{R}^d$ (as in the $\mathbb{PD}$ case) or global coordinates in $\mathbb{R}^{d'}$ (as in the $\mathbb{S}^2$ case). The computation of $\langle\cdot,\cdot\rangle_{g(\boldsymbol{x})}$ will be based on the choice of the local chart, or on the embedding, accordingly. Then, we define a basis matrix, $\Psi^m\in(T_{\boldsymbol{X}^m_{t_1}}\mathcal{M}^N\times\cdots\times T_{\boldsymbol{X}^m_{t_L}}\mathcal{M}^N)^n$, whose columns are

$$\Psi^m(:,\eta)=\Psi^m_\eta=\frac{1}{\sqrt{N}}\begin{bmatrix}\boldsymbol{f}^{\mathrm{c}}_{\psi_\eta}(\boldsymbol{X}^m_{t_1})\\\vdots\\\boldsymbol{f}^{\mathrm{c}}_{\psi_\eta}(\boldsymbol{X}^m_{t_L})\end{bmatrix}\in T_{\boldsymbol{X}^m_{t_1}}\mathcal{M}^N\times\cdots\times T_{\boldsymbol{X}^m_{t_L}}\mathcal{M}^N,$$

recall

$$\boldsymbol{f}^{\mathrm{c}}_{\varphi}(\boldsymbol{X}_t)=\begin{bmatrix}\vdots\\\frac{1}{N}\sum_{i'=1}^{N}\varphi(d_{\mathcal{M}}(\boldsymbol{x}_i(t),\boldsymbol{x}_{i'}(t)))\boldsymbol{w}(\boldsymbol{x}_i(t),\boldsymbol{x}_{i'}(t))\\\vdots\end{bmatrix}\in T_{\boldsymbol{X}_t}\mathcal{M}^N.$$

Next, we define the derivative vector, $\vec{d}^m\in T_{\boldsymbol{X}^m_{t_1}}\mathcal{M}^N\times\cdots\times T_{\boldsymbol{X}^m_{t_L}}\mathcal{M}^N$, as follows,

$$\vec{d}^m=\frac{1}{\sqrt{N}}\begin{bmatrix}\dot{\boldsymbol{X}}^m_{t_1}\\\vdots\\\dot{\boldsymbol{X}}^m_{t_L}\end{bmatrix}.$$

Then, we define the learning matrix $A_M\in\mathbb{R}^{n\times n}$ as follows

$$A_M(\eta,\eta')=\frac{1}{LM}\sum_{m=1}^{m}\langle\Psi^m_\eta,\Psi^m_{\eta'}\rangle_G,\quad\text{for }\eta,\eta'=1,\dots,n.$$

Here the inner product $\langle\cdot,\cdot\rangle_G$ on $\Psi^m_\eta\in T_{\boldsymbol{X}^m_{t_1}}\mathcal{M}^N\times\cdots\times T_{\boldsymbol{X}^m_{t_L}}\mathcal{M}^N$ is defined as

$$\langle\Psi^m_\eta,\Psi^m_{\eta'}\rangle_G=\sum_{l=1}^{L}\langle\boldsymbol{f}^{\mathrm{c}}_{\psi_\eta}(\boldsymbol{X}^m_{t_l}),\boldsymbol{f}^{\mathrm{c}}_{\psi_{\eta'}}(\boldsymbol{X}^m_{t_l})\rangle_{g^{\mathcal{M}^N}(\boldsymbol{X}^m_l)}.$$

Next for the learning right hand side, $\vec{b}_M\in\mathbb{R}^{n\times 1}$, we have

$$\vec{b}_M(\eta)=\frac{1}{LM}\sum_{m=1}^{m}\langle\vec{d},\Psi^m_\eta\rangle_G,\quad\text{for }\eta=1,\dots,n$$

[1] Other type of basis functions can be considered, such as piecewise polynomials, Fourier, etc., provided they satisfy the approximation assumptions in the main theorem.

Therefore, the minimization of (2) over $\mathcal{H}_M$ can be rewritten as

$$A_M\vec{\alpha} = \vec{b}_M, \quad \vec{\alpha} = \begin{bmatrix}\alpha_1\\ \vdots \\ \alpha_n\end{bmatrix} \in \mathbb{R}^{n\times 1}.$$

$A_M$ is symmetric positive definite (guaranteed by the geometric coercivity condition), hence we can solve the linear system to obtain $\hat{\vec{\alpha}}$, and assemble

$$\widehat{\phi}(r) = \sum_{\eta=1}^{n} \hat{\alpha}_\eta \psi_\eta(r).$$

In order to produce unique solution of (1) using $\widehat{\phi}$, we smooth out $\widehat{\phi}$ for the evolution of the dynamics.

If the trajectory data is not given, we will generate it using a Geometric Numerical Integrator, which is a fourth order Backward Differentiation Formula (BDF) of fixed time step size $h$ combined with a projection scheme. For details see (Hairer et al., 2006). Once a reasonable evolution of the dynamics is obtained, we observe it at $0 = t_1 < \ldots < t_L = T$ to obtain a set of trajectory data, and use it as training data to input to the learning algorithm. The observation times do not need to be aligned with the numerical integration times, i.e. where numerical solution of $\{\boldsymbol{x}_i^m(t), \dot{\boldsymbol{x}}_i^m(t)\}_{i,m=1}^{N,M}$ is obtained at $\{t_{l'}\}_{l'=1}^{L'}$ (except for $t_1 = 0$ and $t_{L'} = T$). When $t_l$ does not land on one of the numerical integration time points, a continuous extension method is used to interpolate the numerical solution at $t_l$.

### C.1. Computational Complexity

The total computational cost for solving the learning problem is: $MLN^2 + MLdn^2 + n^3$ with $MLN^2$ for computing pairwise distances, $MLdn^2$ for assembling $A_M$ and $\vec{b}_M$, and $n^3$ for solving $A_M\vec{\alpha} = \vec{b}_M$. When choosing the optimal $n = n_* \approx (\frac{M}{\log M})^{\frac{1}{2s+1}} \approx M^{\frac{1}{3}}$ ($s = 1$ for $C^1$ functions) as per Thm. B.5, we have comp. time $= MLN^2 + MLdM^{\frac{2}{3}} + M = \mathcal{O}(M^{\frac{5}{3}})$. The computational bottleneck comes from the assembly of $A_M$ and $\vec{b}_M$. However, since we can parallelize our learning approach in $m$, the updated computing time in the parallel regime is comp. time $= \mathcal{O}\Big(\Big(\frac{M}{\text{num. cores}}\Big)^{\frac{5}{3}}\Big)$. The total storage for the algorithm is $MLNd$ floating-point numbers for the trajectory data, albeit one does not need to hold all of the trajectory data in memory. The algorithm can process the data from one trajectory at a time, requiring $LNd$. Once the linear system, $A_M\vec{\alpha} = \vec{b}_M$, is assembled, the algorithm just needs to hold roughly $n^2$ floating-point numbers in memory. When we use the optimal number of basis functions, i.e. $n_* = M^{\frac{1}{3}}$, the memory used is $O(M^{\frac{2}{3}})$.

## D. Numerical Experiments

We consider three prototypical first order dynamics, Opinion Dynamics (OD), Lennard-Jones Dynamics (LJD), and Predator-Swarm dynamics (PS1), on two different manifolds, the $2D$ sphere ($\mathbb{S}^2$ centered at the origin with radius $\frac{5}{\pi}$) and the Poincaré disk ($\mathbb{PD}$, unit disk centered at the origin, with the hyperbolic metric). The two prototypical manifolds are chosen because $\mathbb{S}^2$ and $\mathbb{PD}$ are model spaces with constant positive and negative curvature, respectively. We conduct extensive experiments on the aforementioned six different scenarios to demonstrate the performance of our learning approach for dynamics evolving on manifolds. We report the results in terms of function estimation errors and trajectory estimation errors, and discuss in detail the learning performance of the estimators.

The setup of the numerical experiments is as follows. We generate a set of $M_\rho$ different initial conditions, and evolve the various dynamics of $N$ agents for $t \in [0, T]$ using a Geometric Numerical Integrator with a uniform time step $h$ (for details see section C); then we observe each dynamics at equidistant times, i.e. $0 = t_1 < \ldots < t_L = T$, to obtain a set of trajectory data, $\{\boldsymbol{x}_i^m(t_l), \dot{\boldsymbol{x}}_i^m(t_l)\}_{i,l,m=1}^{N,L,M_\rho}$, to approximate the "true" probability distribution $\rho_{T,\mathcal{M}}^L$. From this set of pre-generated trajectory data, we randomly choose a subset of $M \ll M_\rho$ of them to be used as training data for the learning simulation. The hypothesis space where the estimator is learned is generated as a set of $n$ first-degree clamped B-spline basis functions built on a uniform partition of the learning interval $[R_{\min}^{\text{obs}}, R_{\max}^{\text{obs}}]$, with $R_{\min}^{\text{obs}}$ and $R_{\max}^{\text{obs}}$ being the minimum and maximum interaction radii computed from the training and trajectory data, respectively. Once an estimator, denoted as $\widehat{\phi}$, is obtained, we report the estimation error, $\phi(\cdot)\cdot - \widehat{\phi}(\cdot)\cdot$, using

$$\|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{\text{Rel.}L^2(\rho_{T,\mathcal{M}})} := \frac{\|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_{T,\mathcal{M}})}}{\|\phi(\cdot)\cdot\|_{L^2(\rho_{T,\mathcal{M}})}}; \tag{13}$$

and the trajectory estimation error

$$d_{\text{trj}}(\boldsymbol{X}^m_{[0,T]}, \hat{\boldsymbol{X}}^m_{[0,T]})^2 := \sup_{t\in[0,T]} \frac{\sum_i d_{\mathcal{M}}(\boldsymbol{x}_i^m(t), \hat{\boldsymbol{x}}_i^m(t))^2}{N} \tag{14}$$

between, the true and estimated dynamics, evolved using $\phi$ or $\widehat{\phi}$ with the same initial conditions for $t \in [0, T]$ respectively, and observed at the same observation times $0 = t_1 < \ldots < t_L = T$, over both the training initial conditions and another set of $M$ randomly chosen initial conditions. Moreover, the above learning procedure is run 10 times independently in order to generate empirical error bars. We will report the errors in the form of mean $\pm$ std. Visual comparisons of $\phi$ versus $\widehat{\phi}$, and $\boldsymbol{X}$ versus $\hat{\boldsymbol{X}}$ will be shown, and discussions of learning results will be presented in each subsection.

Table 1 shows the values of the common parameters shared by all six experiments.

| $M_\rho$ | $N$ | $L$ | $M$ | Num. of Learning Trials | $R_{\mathcal{M}}$ on $\mathbb{S}^2$ | $R_{\mathcal{M}}$ on $\mathbb{PD}$ |
|---|---|---|---|---|---|---|
| 3000 | 20 | 500 | 500 | 10 | 5 | $\infty$ |

*Table 1.* Values of the parameters shared by the six experiments

Moreover, section A shows the details on how to calculate the geodesic direction and the Riemannian distance between any two points on $\mathbb{S}^2$ and $\mathbb{PD}$. The distribution of the initial conditions, $\mu_0(\mathcal{M}^N)$, is given as follows: uniform on $\mathcal{M} = \mathbb{S}^2$; whereas uniform on an open ball (centered at origin with radius $r_0$) for the $\mathbb{PD}$ case with $r_0$ given as follows.

$$r_0 = \left(2 + \frac{1}{\cosh(5) - 1} - \sqrt{\frac{4}{\cosh(5) - 1} + \frac{1}{(\cosh(5) - 1)^2}}\right)/2.$$

This radius is used so that the maximum distance between any pair of agents on the Poincaré disk is 5. PS1 will have different setup for the initial conditions, which will be discussed in section D.4.

### D.1. Computing Platform

We use a computing workstation with an AMD Ryzen 9 3900X CPU (which has 12 computing cores), and available 128 GB memory, running CentOS 7. All 6 experiments are ran in the MATLAB (R2020$a$) environment with parallel mode enabled and a parallel pool of 12 workers. Such parallel mode is used in each experiment for the computation of $\rho^L_{T,\mathcal{M}}$, learning, and trajectory error estimation. Detailed report of the running time for the experiments is provided in the result section of each experiment.

### D.2. Opinion Dynamics

We first choose opinion dynamics, which is used to model simple interactions of opinions (Aydoğdu et al., 2017; Weisbuch et al., 2003) as well as choreography (Caponigro et al., 2014). We consider the generalization of this dynamics to take place on two different manifolds: the $2D$ sphere ($\mathbb{S}^2$) and the Poincaré disk ($\mathbb{PD}$). We consider the interaction kernel

$$\phi(r) := \begin{cases} 1, & 0 \le r < \frac{1}{\sqrt{2}} - 0.01 \\ a_1 r^3 + b_1 r^2 + c_1 r + d_1, & \frac{1}{\sqrt{2}} - 0.01 \le r < \frac{1}{\sqrt{2}} \\ 0.1, & \frac{1}{\sqrt{2}} \le r < 0.99 \\ a_2 r^3 + b_2 r^2 + c_2 r + d_2, & 0.99 \le r < 1 \\ 0, & \text{otherwise} \end{cases}$$

The parameters, i.e. $(a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2)$, are chosen so that $\phi \in C^1([0, 1])$. Table 2 shows the values of the parameters needed for the learning simulation.

| $n_{\mathbb{S}^2}$ | $n_{\mathbb{PD}}$ | $T$ | $h$ |
|---|---|---|---|
| 51 | 69 | 10 | 0.01 |

*Table 2.* Test Parameters for OD.

**Results for the $\mathbb{S}^2$ case:** Fig. 1 shows the comparison between $\phi$ and its estimator $\widehat{\phi}$ learned from the trajectory data.

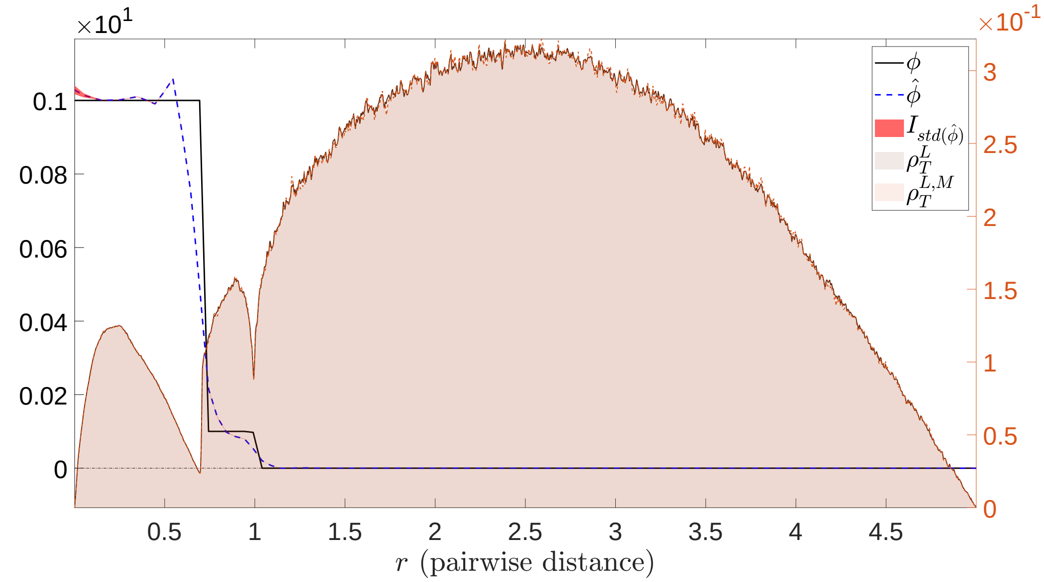


*Figure 1.* (OD on $\mathbb{S}^2$) Comparison of $\phi$ and $\widehat{\phi}$, with the relative error being $1.894 \cdot 10^{-1} \pm 3.1 \cdot 10^{-4}$ (calculated using (13)). The true interaction kernel is shown in a black solid line, whereas the mean estimated interaction kernel is shown in a blue dashed line with its std interval, i.e. mean$(\widehat{\phi}) \pm$ std$(\widehat{\phi})$, region shaded in red. Shown in the background is the comparison of the approximate $\rho^L_{T,\mathcal{M}}$ versus the empirical $\rho^{L,M}_{T,\mathcal{M}}$.

As it is shown in Fig. 1, the estimator is able to capture the compact support of the $\phi$ from the trajectory data. Fig. 2 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

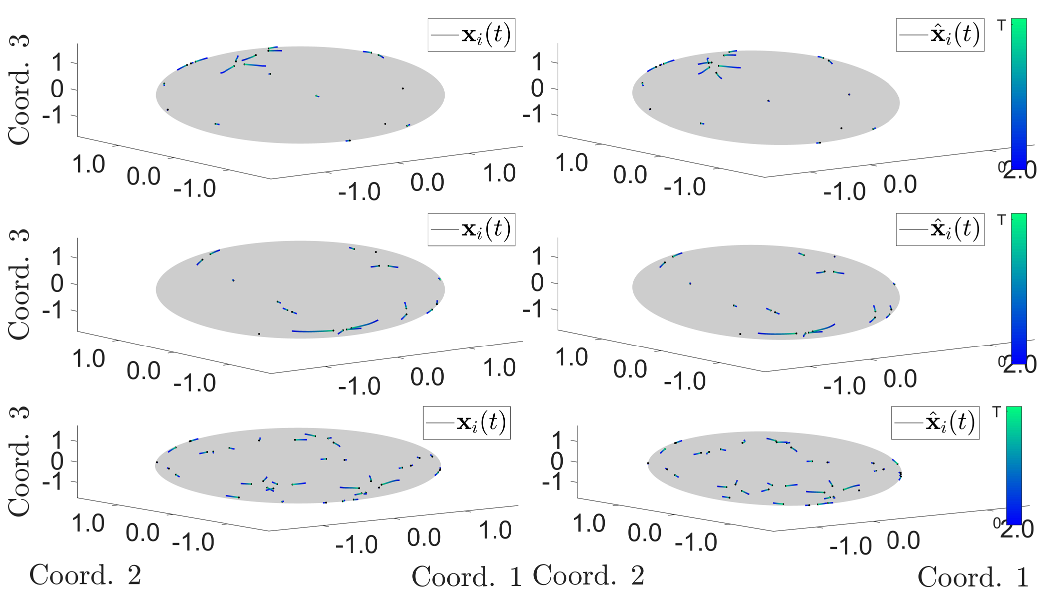


*Figure 2.* (OD on $\mathbb{S}^2$) Comparison of $\boldsymbol{X}$ (generated by $\phi$) and $\hat{\boldsymbol{X}}$ (generated by $\widehat{\phi}$), with the errors reported in table 3. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a randomly chosen initial condition. **Bottom**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a new initial condition with bigger $N = 40$. The color of the trajectory indicates the flow of time, from deep blue (at $t = 0$) to light green (at $t = T$).

A quantitative comparison of the trajectory estimation errors is shown in Table 3.

| | $[0, T]$ |
|---|---|
| $\text{mean}_{\text{IC}}$: Training ICs | $8.8 \cdot 10^{-2} \pm 1.7 \cdot 10^{-3}$ |
| $\text{std}_{\text{IC}}$: Training ICs | $5.9 \cdot 10^{-2} \pm 1.5 \cdot 10^{-3}$ |
| $\text{mean}_{\text{IC}}$: Random ICs | $9.0 \cdot 10^{-2} \pm 1.6 \cdot 10^{-3}$ |
| $\text{std}_{\text{IC}}$: Random ICs | $6.0 \cdot 10^{-2} \pm 1.7 \cdot 10^{-3}$ |

*Table 3.* (OD on $\mathbb{S}^2$) trajectory estimation errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu_0(\mathcal{M}^N)$ (second set of two rows). $\text{mean}_{\text{IC}}$ and $\text{std}_{\text{IC}}$ are the mean and standard deviation of the trajectory errors calculated using (14).

We also report the condition number and the smallest eigenvalue of the learning matrix $A$ to indirectly verify the geometric coercivity condition in table 4.

| | |
|---|---|
| Condition Number | $1.8 \cdot 10^5 \pm 1.4 \cdot 10^4$ |
| Smallest Eigenvalue | $1.09 \cdot 10^{-7} \pm 9.0 \cdot 10^{-9}$ |

*Table 4.* (OD on $\mathbb{S}^2$) Information from the learning matrix $A$.

It took $1.41 \cdot 10^4$ seconds to generate $\rho^L_{T,\mathcal{M}}$ and $4.76 \cdot 10^4$ seconds to run 10 learning simulations, with $1.44 \cdot 10^3$ seconds spent on learning the estimated interactions (on average, it took $1.44 \cdot 10^2 \pm 3.1$ seconds to run one estimation), and $4.61 \cdot 10^4$ seconds spent on computing the trajectory error estimates (on average, it took $4.61 \cdot 10^3 \pm 20.0$ seconds to run one set of trajectory error estimation).

**Results for the $\mathbb{PD}$ case:** Fig. 3 shows the comparison between the $C^1$ version of $\phi$ and its estimator $\hat{\phi}$ learned from the trajectory data.

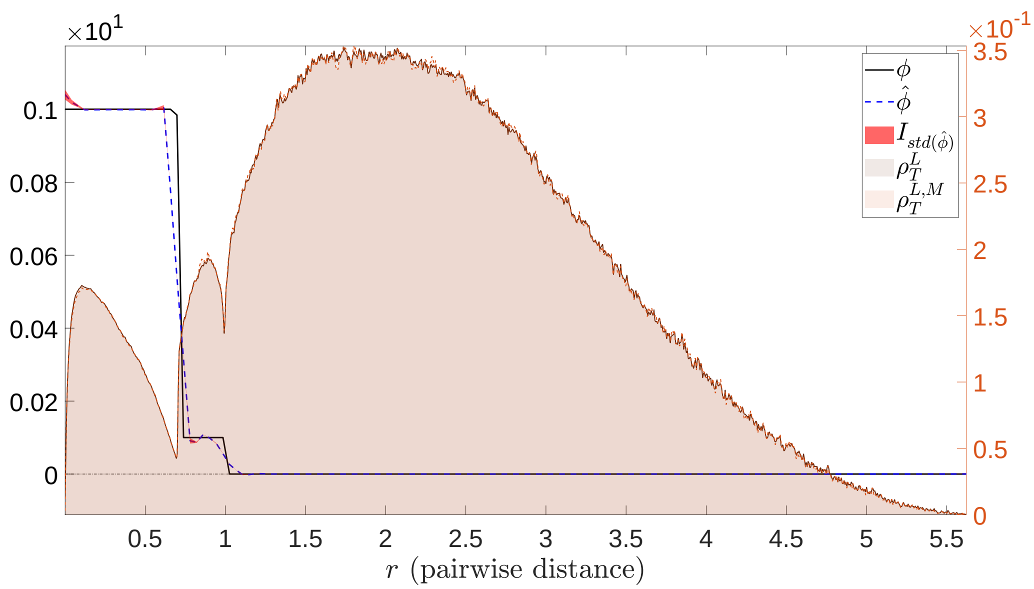


*Figure 3.* (OD on $\mathbb{PD}$) Comparison of $\phi$ and $\widehat{\phi}$, with the relative error being $2.114 \cdot 10^{-1} \pm 5.0 \cdot 10^{-4}$ (calculated using (13)). The true interaction kernel is shown in a black solid line, whereas the mean estimated interaction kernel is shown in a blue dashed line with its std interval, i.e. mean($\widehat{\phi}$) $\pm$ std($\widehat{\phi}$), region shaded in red. Shown in the background is the comparison of the approximate $\rho^L_{T,\mathcal{M}}$ versus the empirical $\rho^{L,M}_{T,\mathcal{M}}$.

As it is shown in Fig. 3, the estimator is able to capture the compact support of the $\phi$ from the trajectory data. Fig. 4 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

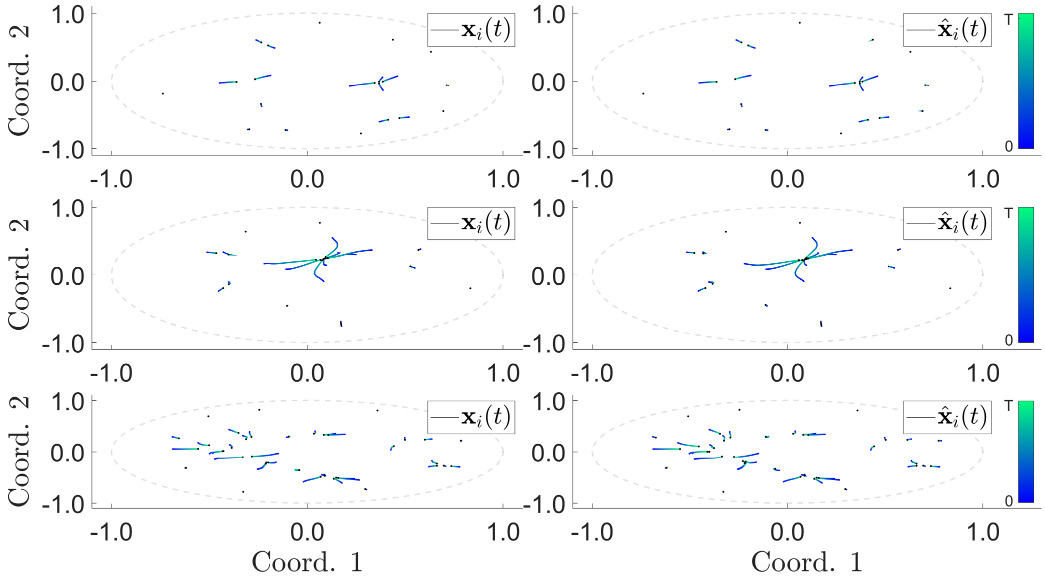


*Figure 4.* (OD on $\mathbb{PD}$ ) Comparison of $\boldsymbol{X}$ (generated by $\phi$) and $\hat{\boldsymbol{X}}$ (generated by $\widehat{\phi}$), with the errors reported in table 5. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a randomly chosen initial condition. **Bottom**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a new initial condition with bigger $N = 40$. The color of the trajectory indicates the flow of time, from deep blue (at $t = 0$) to light green (at $t = T$).

As shown in Fig. 3, around $r = \frac{1}{\sqrt{2}}$, the estimator $\widehat{\phi}$ produces values bigger than that from $\phi$, leading to stronger influence, hence the merging of cluster happening in the predicted trajectories in the second row of Fig. 4. As demonstrated by the average prediction error on trajectories, this is a relatively rare event, occurring for only certain initial conditions. A quantitative comparison of the trajectory estimation errors is shown in Table 5.

| | $[0, T]$ |
|---|---|
| $\text{mean}_{\text{IC}}$: Training ICs | $2.53 \cdot 10^{-1} \pm 7.2 \cdot 10^{-3}$ |
| $\text{std}_{\text{IC}}$: Training ICs | $1.90 \cdot 10^{-1} \pm 6.5 \cdot 10^{-3}$ |
| $\text{mean}_{\text{IC}}$: Random ICs | $2.55 \cdot 10^{-1} \pm 9.7 \cdot 10^{-3}$ |
| $\text{std}_{\text{IC}}$: Random ICs | $1.89 \cdot 10^{-1} \pm 5.9 \cdot 10^{-3}$ |

*Table 5.* (OD on $\mathbb{PD}$ ) trajectory estimation errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu_0(\mathcal{M}^N)$ (second set of two rows). $\text{mean}_{\text{IC}}$ and $\text{std}_{\text{IC}}$ are the mean and standard deviation of the trajectory errors calculated using (14).

We also report the condition number and the smallest eigenvalue of the learning matrix $A$ to indirectly verify the geometric coercivity condition in table 6.

| | |
|---|---|
| Condition Number | $4.9 \cdot 10^{5} \pm 1.5 \cdot 10^{4}$ |
| Smallest Eigenvalue | $5.3 \cdot 10^{-6} \pm 1.2 \cdot 10^{-7}$ |

*Table 6.* (OD on $\mathbb{PD}$ ) Information from the learning matrix $A$.

It took $1.33 \cdot 10^4$ seconds to generate $\rho^L_{T,\mathcal{M}}$ and $4.06 \cdot 10^4$ seconds to run 10 learning simulations, with $1.23 \cdot 10^3$ seconds spent on learning the estimated interactions (on average, it took $1.23 \cdot 10^2 \pm 1.1$ seconds to run one estimation), and $3.93 \cdot 10^4$ seconds spent on computing the trajectory error estimates (on average, it took $3.93 \cdot 10^3 \pm 82.1$ seconds to run one set of trajectory error estimation).

### D.3. Lennard-Jones Dynamics

The second first-order model considered here is induced from a special energy functional, the so-called Lennard-Jones energy potential. This first-order model, the Lennard-Jones Dynamics (LJD), is a simplified version of the second-order dynamics used in molecular dynamics. The energy function, $U_{\mathrm{LJ}}$, is given by

$$U_{\mathrm{LJ}}(r) := 4\varepsilon\Big(\Big(\frac{\sigma}{r}\Big)^{12} - \Big(\frac{\sigma}{r}\Big)^{6}\Big).$$

Here $\varepsilon$ is the depth of the potential well, $\sigma$ is the distance when $U$ is zero, and $r$ is the distance between any pair of agents. We set $\varepsilon = 10$ and $\sigma = 1$. The corresponding interaction kernel $\phi$, derived from this potential, is

$$\phi_{\mathrm{LJ}}(r) := \frac{U'_{\mathrm{LJ}}(r)}{r} = 24\frac{\varepsilon}{\sigma^2}\Big(\Big(\frac{\sigma}{r}\Big)^{8} - 2\Big(\frac{\sigma}{r}\Big)^{14}\Big).$$

We shall use a slightly modified version of $\phi_{\mathrm{LJ}}$:

$$\phi(r) := \begin{cases} \phi_{\mathrm{LJ}}(1) - \phi'_{\mathrm{LJ}}(1)/4, & 0 \le r < \frac{1}{2} \\ \phi'_{\mathrm{LJ}}(1)r^2 - \phi'_{\mathrm{LJ}}(1)r + \phi_{\mathrm{LJ}}(1), & \frac{1}{2} \le r < 1 \\ \phi_{\mathrm{LJ}}(r), & 1 \le r < 0.99R_{\mathcal{M}} \\ a_3r^3 + b_3r^2 + c_3r + d_3, & 0.99R_{\mathcal{M}} \le r < R_{\mathcal{M}} \\ 0, & R_{\mathcal{M}} \le r. \end{cases}$$

The parameters, $(a_3, b_3, c_3, d_3)$, are chosen so that $\phi \in C^1([0, R_{\mathcal{M}}])$ when $R_{\mathcal{M}} < \infty$; otherwise $\phi(r) = \phi_{\mathrm{LJ}}(r)$ for $r \ge 1$. Table 7 shows the values of the parameters needed for the learning simulation.

| $n_{\mathbb{S}^2}$ | $n_{\mathbb{PD}}$ | $T$ | $h$ |
|---|---|---|---|
| 51 | 69 | $10^{-3}$ | $10^{-6}$ |

*Table 7.* Test Parameters for LJD.

**Results for the $\mathbb{S}^2$ case:** Fig. 5 shows the comparison between $\phi$ and its estimator $\widehat{\phi}$ learned from the trajectory data.

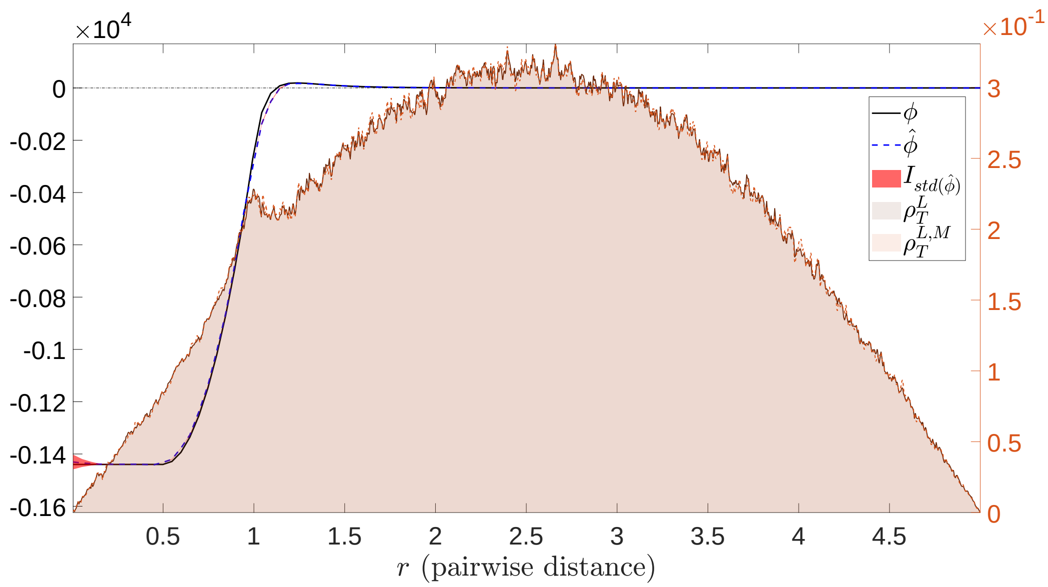


*Figure 5.* (LJD on $\mathbb{S}^2$) Comparison of $\phi$ and $\widehat{\phi}$, with the relative error being $3.65 \cdot 10^{-2} \pm 2.7 \cdot 10^{-4}$ (calculated using (13)). The true interaction kernel is shown in a black solid line, whereas the mean estimated interaction kernel is shown in a blue dashed line with its std interval, i.e. mean$(\widehat{\phi}) \pm$ std$(\widehat{\phi})$, region shaded in red. Shown in the background is the comparison of the approximate $\rho_T^L$ versus the empirical $\rho_T^{L,M}$.

Fig. 6 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

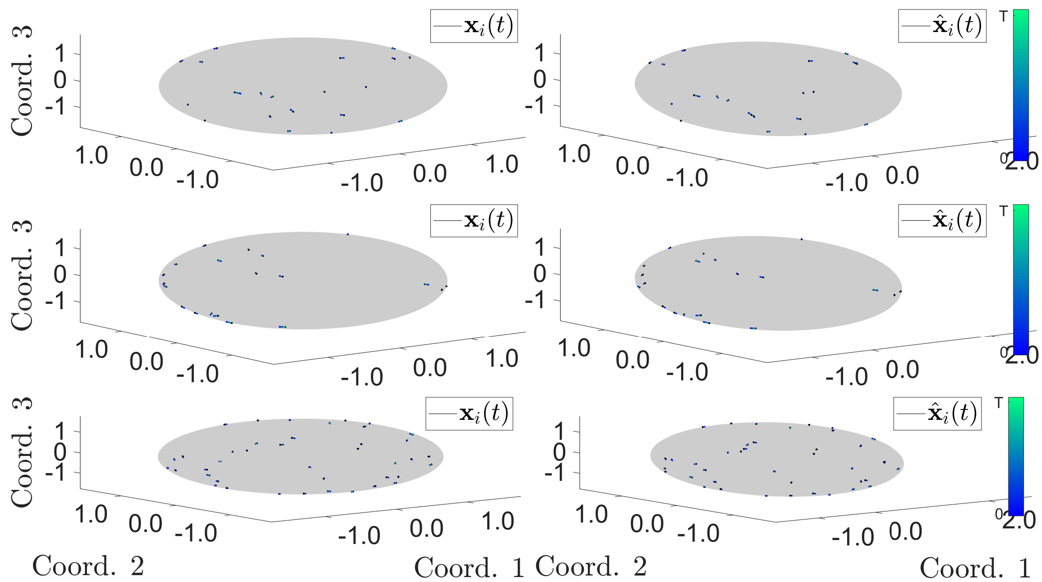


*Figure 6.* (LJD on $\mathbb{S}^2$) Comparison of $\boldsymbol{X}$ (generated by $\phi$) and $\hat{\boldsymbol{X}}$ (generated by $\widehat{\phi}$), with the errors reported in table 8. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a randomly chosen initial condition. **Bottom**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a new initial condition with bigger $N = 40$. The color of the trajectory indicates the flow of time, from deep blue (at $t = 0$) to light green (at $t = T$).

A quantitative comparison of the trajectory estimation errors is shown in Table 8.

| | $[0, T]$ |
|---|---|
| $\text{mean}_{\text{IC}}$: Training ICs | $2.88 \cdot 10^{-3} \pm 2.5 \cdot 10^{-5}$ |
| $\text{std}_{\text{IC}}$: Training ICs | $6.1 \cdot 10^{-4} \pm 1.8 \cdot 10^{-5}$ |
| $\text{mean}_{\text{IC}}$: Random ICs | $2.88 \cdot 10^{-3} \pm 3.2 \cdot 10^{-5}$ |
| $\text{std}_{\text{IC}}$: Random ICs | $6.0 \cdot 10^{-4} \pm 1.8 \cdot 10^{-5}$ |

*Table 8.* (LJD on $\mathbb{S}^2$) trajectory estimation errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu_0(\mathcal{M}^N)$ (second set of two rows). The trajectory estimation errors is calculated using (13).

We also report the condition number and the smallest eigenvalue of the learning matrix $A$ to indirectly verify the geometric coercivity condition in table 9.

| | |
|---|---|
| Condition Number | $6 \cdot 10^5 \pm 1.5 \cdot 10^5$ |
| Smallest Eigenvalue | $2.4 \cdot 10^{-8} \pm 6.2 \cdot 10^{-9}$ |

*Table 9.* (LJD on $\mathbb{S}^2$) Information from the learning matrix $A$.

It took $2.43 \cdot 10^4$ seconds to generate $\rho_{T,\mathcal{M}}^L$ and $7.14 \cdot 10^4$ seconds to run 10 learning simulations, with $1.72 \cdot 10^3$ seconds spent on learning the estimated interactions (on average, it took $1.72 \cdot 10^2 \pm 2.5$ seconds to run one estimation), and $6.96 \cdot 10^4$ seconds spent on computing the trajectory error estimates (on average, it took $6.96 \cdot 10^3 \pm 35.9$ seconds to run one set of trajectory error estimation).

**Results for the $\mathbb{PD}$ case:** Fig. 7 shows the comparison between $\phi$ and its estimator $\widehat{\phi}$ learned from the trajectory data.

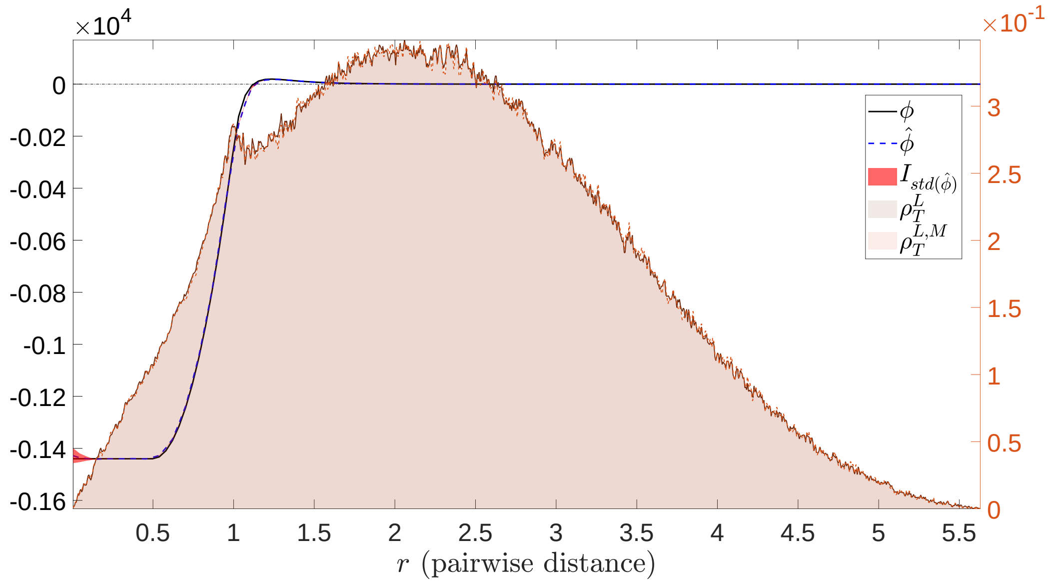


*Figure 7.* (LJD on $\mathbb{PD}$ ) Comparison of $\phi$ and $\widehat{\phi}$, with the relative error being $2.52 \cdot 10^{-2} \pm 3.6 \cdot 10^{-4}$ (calculated using (13)). The true interaction kernel is shown in a black solid line, whereas the mean estimated interaction kernel is shown in a blue dashed line with its std interval, i.e. mean$(\widehat{\phi}) \pm$ std$(\widehat{\phi})$, region shaded in red. Shown in the background is the comparison of the approximate $\rho_T^L$ versus the empirical $\rho_T^{L,M}$.

Fig. 8 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

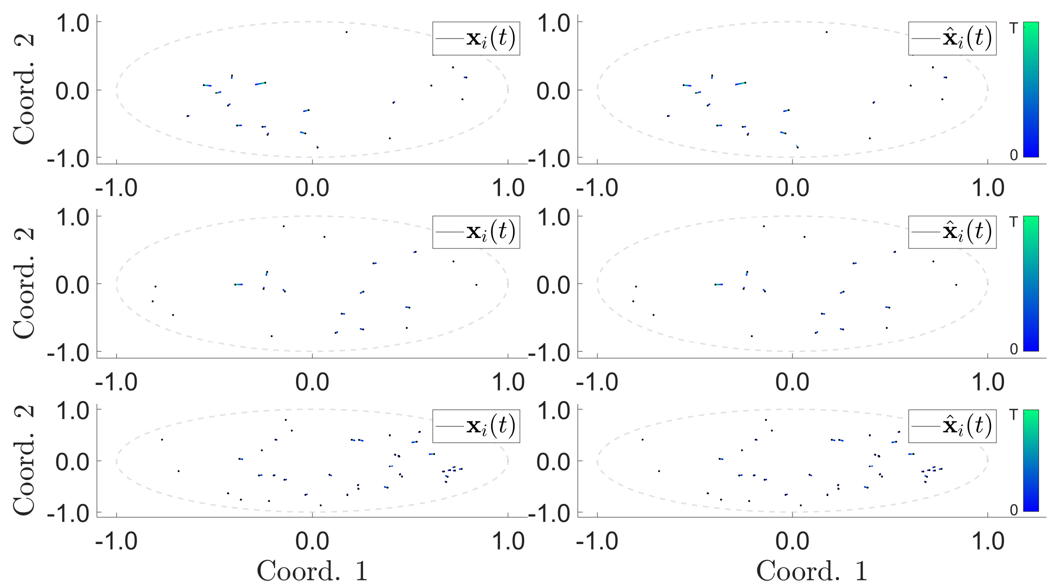


*Figure 8.* (LJD on $\mathbb{PD}$ ) Comparison of $\boldsymbol{X}$ (generated by $\phi$) and $\hat{\boldsymbol{X}}$ (generated by $\widehat{\phi}$), with the errors reported in table 10. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a randomly chosen initial condition. **Bottom**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a new initial condition with bigger $N = 40$. The color of the trajectory indicates the flow of time, from deep blue (at $t = 0$) to light green (at $t = T$).

A quantitative comparison of the trajectory estimation errors is shown in Table 10.

| | $[0,T]$ |
|---|---|
| $\text{mean}_{\text{IC}}$: Training ICs | $2.27 \cdot 10^{-3} \pm 4.0 \cdot 10^{-5}$ |
| $\text{std}_{\text{IC}}$: Training ICs | $5.6 \cdot 10^{-4} \pm 1.7 \cdot 10^{-5}$ |
| $\text{mean}_{\text{IC}}$: Random ICs | $2.28 \cdot 10^{-3} \pm 3.8 \cdot 10^{-5}$ |
| $\text{std}_{\text{IC}}$: Random ICs | $5.6 \cdot 10^{-4} \pm 1.6 \cdot 10^{-5}$ |

*Table 10.* (LJD on $\mathbb{PD}$ ) trajectory estimation errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu_0(\mathcal{M}^N)$ (second set of two rows). $\text{mean}_{\text{IC}}$ and $\text{std}_{\text{IC}}$ are the mean and standard deviation of the trajectory errors calculated using (14).

We also report the condition number and the smallest eigenvalue of the learning matrix $A$ to indirectly verify the geometric coercivity condition in table 11.

| | |
|---|---|
| Condition Number | $6 \cdot 10^{6} \pm 1.9 \cdot 10^{6}$ |
| Smallest Eigenvalue | $1.7 \cdot 10^{-8} \pm 6.6 \cdot 10^{-9}$ |

*Table 11.* (LJD on $\mathbb{PD}$ ) Information from the learning matrix $A$.

It took $1.51 \cdot 10^4$ seconds to generate $\rho_{T,\mathcal{M}}^L$ and $6.23 \cdot 10^4$ seconds to run 10 learning simulations, with $1.20 \cdot 10^3$ seconds spent on learning the estimated interactions (on average, it took $1.20 \cdot 10^2 \pm 9.4$ seconds to run one estimation), and $6.10 \cdot 10^4$ seconds spent on computing the trajectory error estimates (on average, it took $6 \cdot 10^3 \pm 1.3 \cdot 10^3$ seconds to run one set of trajectory error estimation).

### D.4. Predator-Swarm Dynamics

The third first-order model considered here is a heterogeneous agent system, which is used to model interactions between multiple types of animals (Chen & Kolokolnikov, 2013; Olson et al., 2016) or agents (need ref.). The learning theory presented in this work is described for homogeneous agent systems, but the theory and the corresponding algorithms extend naturally to heterogeneous agent systems in a manner analogous to (Lu et al., 2019a; Miller et al., 2020).

We consider here a system of a single predator versus a group of preys, namely the Predator-Swarm Dynamics (PS1), discussed in (Chen & Kolokolnikov, 2013). The preys are in type 1, and the single predator is in type 2. We have multiple interaction kernels, depending on the types of agents in each interacting pair: $\phi_{kk'}$ defines the influence of agents in type $k'$ on agents in type $k$, for $k, k' = 1, 2$. The interaction kernels are given as follows.

$$\phi_{11}(r) := \begin{cases} \frac{2}{0.01^3}(r - 0.01) + (1 - \frac{1}{0.01^2}) & 0 < r \leq 0.01 \\ 1 - \frac{1}{r^2} & 0.01 < r \leq 0.99R_{\mathcal{M}} \\ a_{1,1}r^3 + b_{1,1}r^2 + c_{1,1}r + d_{1,1}, & 0.99R_{\mathcal{M}} \leq r < R_{\mathcal{M}} \\ 0, & R_{\mathcal{M}} \leq r \end{cases}$$

The parameters, $(a_{1,1}, b_{1,1}, c_{1,1}, d_{1,1})$, are chosen so that $\phi_{11}(r) \in C^1([0, R_{\mathcal{M}}])$ when $R_{\mathcal{M}} < \infty$; otherwise $\phi_{11}(r) = 1 - \frac{1}{r^2}$ for $r \geq 0.01$;

$$\phi_{12}(r) := \begin{cases} \frac{4}{0.01^3}(r - 0.01) + \frac{-2}{0.01^2}) & 0 < r \leq 0.01 \\ \frac{-2}{r^2} & 0.01 < r \leq 0.99R_{\mathcal{M}} \\ a_{1,2}r^3 + b_{1,2}r^2 + c_{1,2}r + d_{1,2}, & 0.99R_{\mathcal{M}} \leq r < R_{\mathcal{M}} \\ 0, & R_{\mathcal{M}} \leq r \end{cases}$$

The parameters, $(a_{1,2}, b_{1,2}, c_{1,2}, d_{1,2})$, are chosen so that $\phi_{12}(r) \in C^1([0, R_{\mathcal{M}}])$ when $R_{\mathcal{M}} < \infty$; otherwise $\phi_{12}(r) = \frac{-2}{r^2}$ for $r \geq 0.01$;

$$\phi_{21}(r) := \begin{cases} \frac{-10.5}{0.01^4}(r - 0.01) + \frac{3.5}{0.01^3}) & 0 < r \leq 0.01 \\ \frac{3.5}{r^3} & 0.01 < r \leq 0.99R_{\mathcal{M}} \\ a_{2,1}r^3 + b_{2,1}r^2 + c_{2,1}r + d_{2,1}, & 0.99R_{\mathcal{M}} \leq r < R_{\mathcal{M}} \\ 0, & R_{\mathcal{M}} \leq r \end{cases}$$

The parameters, $(a_{2,1}, b_{2,1}, c_{2,1}, d_{2,1})$, are chosen so that $\phi_{21}(r) \in C^1([0, R_{\mathcal{M}}])$ when $R_{\mathcal{M}} < \infty$; otherwise $\phi_{21}(r) = \frac{3.5}{r^3}$ for $r \geq 0.01$; then $\phi_{22} \equiv 0$, since there is only one predator. We set $T = 0.5$ and $h = 10^{-4}$ for the two $PS1$ models.

**Results for the $\mathbb{S}^2$ case:** In order to produce more interesting interactions, we choose the distribution of the initial condition to be as follows. The setting will start from $\mathbb{R}^2$ first. The position of the predator is randomly chosen uniformly within a circular disk of radius $0.1$ centered at the origin of $\mathbb{R}^2$. The remaining $N-1$ agents will be prey and chosen uniformly at random within an annulus of radii $0.3$ and $0.8$, centered at the origin. Then these positions will mapped through a stereographic projection (where the origin of $\mathbb{R}^2$ is the south pole of $\mathbb{S}^2$) back to $\mathbb{S}^2$. When back on $\mathbb{S}^2$, the position of the predator is moved via parallel transport to a random location on $\mathbb{S}^2$, and the rest of the preys are moved using the same map, so that the relative position between each pair of agents is not changed.

Table 12 shows the number of basis functions, namely $n_{kk'}$'s, for each estimator $\widehat{\phi}_{kk'}$ for $k, k' = 1, 2$, and their corresponding degrees, $p_{k,k'}$'s, for the Clamped B-spline basis.

| $n_{1,1}$ | $n_{1,2}$ | $n_{2,1}$ | $n_{2,2}$ |
|---|---|---|---|
| 50 | 37 | 37 | 1 |
| $p_{1,1}$ | $p_{1,2}$ | $p_{2,1}$ | $p_{2,2}$ |
| 1 | 1 | 1 | 0 |

*Table 12.* (PS1 on $\mathbb{S}^2$ ) Number of basis functions.

Fig. 11 shows the comparison between $\phi_{kk'}$ and its estimators $\widehat{\phi}_{kk'}$ learned from the trajectory data.

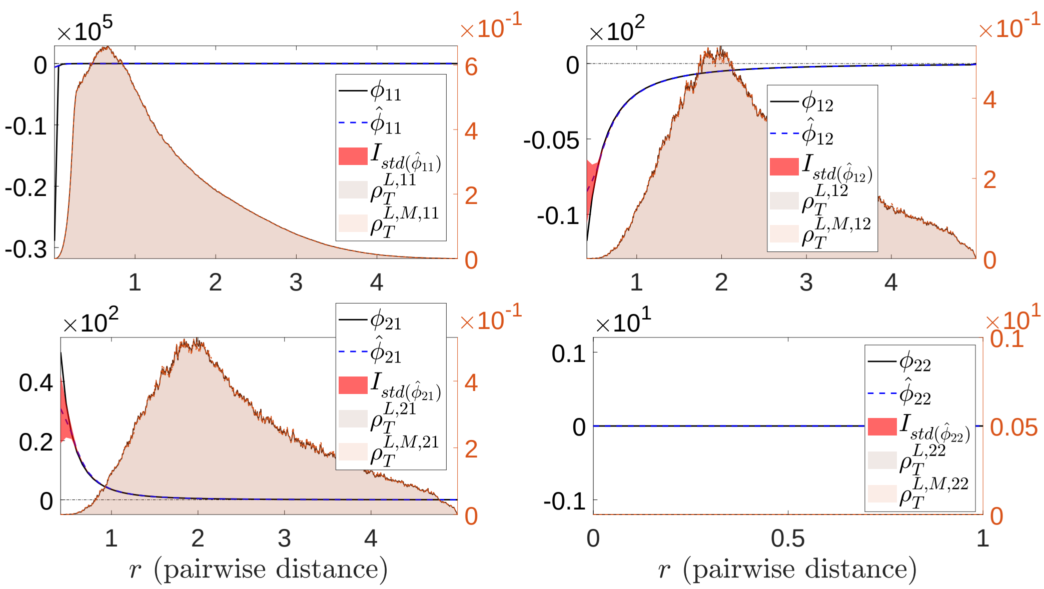


*Figure 9.* (PS1 on $\mathbb{S}^2$ ) Comparison of $\phi_{kk'}$ and $\widehat{\phi}_{k,k'}$, with the relative errors shown in table 17. The true interaction kernels are shown in black solid lines, whereas the mean estimated interaction kernel are shown in blue dashed lines with their corresponding std interval, i.e. mean$(\widehat{\phi}_{kk'}) \pm$ std$(\widehat{\phi}_{kk'})$, regions shaded in red. Shown in the background is the comparison of the approximate $\rho_T^{L,kk'}$ versus the empirical $\rho_T^{L,M,kk'}$. Notice that $\rho_T^{L,12}/\rho_T^{L,M,12}$ and $\rho_T^{L,12}/\rho_T^{L,M,21}$ are the same distributions.

| $\mathrm{Err}_{1,1}$ | $\mathrm{Err}_{1,2}$ | $\mathrm{Err}_{2,1}$ | $\mathrm{Err}_{2,2}$ |
|---|---|---|---|
| $2.98 \cdot 10^{-1} \pm 5.9 \cdot 10^{-3}$ | $8.4 \cdot 10^{-3} \pm 3.0 \cdot 10^{-4}$ | $2.5 \cdot 10^{-2} \pm 1.6 \cdot 10^{-3}$ | 0 |

*Table 13.* (PS1 on $\mathbb{S}^2$ ) Relative estimation errors calculated using (13).

Fig. 10 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

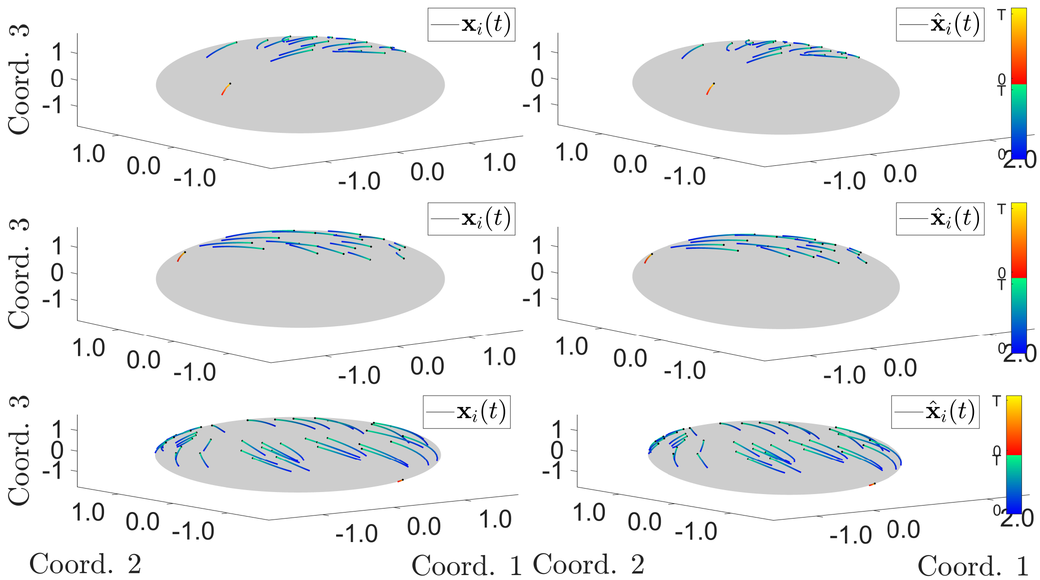


*Figure 10.* (PS1 on $\mathbb{S}^2$ ) Comparison of $\boldsymbol{X}$ (generated by $\phi_{k,k'}$'s) and $\hat{\boldsymbol{X}}$ (generated by $\widehat{\phi}_{k,k'}$'s), with the errors reported in table 14. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a randomly chosen initial condition. **Bottom**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a new initial condition with bigger $N = 40$. The color of the trajectory indicates the flow of time, from deep blue/bright red (at $t = 0$) to light green/light yellow (at $t = T$). The blue/green combination is assigned to the preys; whereas the red/yellow comb for the predator.

A quantitative comparison of the trajectory estimation errors is shown in Table 18.

| | $[0, T]$ |
|---|---|
| $\text{mean}_{\text{IC}}$: Training ICs | $2.36 \cdot 10^{-2} \pm 9.8 \cdot 10^{-4}$ |
| $\text{std}_{\text{IC}}$: Training ICs | $1.9 \cdot 10^{-2} \pm 1.5 \cdot 10^{-4}$ |
| $\text{mean}_{\text{IC}}$: Random ICs | $2.40 \cdot 10^{-2} \pm 8.1 \cdot 10^{-4}$ |
| $\text{std}_{\text{IC}}$: Random ICs | $2.3 \cdot 10^{-3} \pm 6.1 \cdot 10^{-3}$ |

*Table 14.* (PS1 on $\mathbb{S}^2$ ) trajectory estimation errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu_0(\mathcal{M}^N)$ (second set of two rows). $\text{mean}_{\text{IC}}$ and $\text{std}_{\text{IC}}$ are the mean and standard deviation of the trajectory errors calculated using (14).

We also report the condition number and the smallest eigenvalue of the learning matrix $A$ to indirectly verify the geometric coercivity condition in table 19.

| | |
|---|---|
| Condition Number for $A_1$ | $2.2 \cdot 10^{7} \pm 1.8 \cdot 10^{6}$ |
| Smallest Eigenvalue for $A_1$ | $1.28 \cdot 10^{-8} \pm 8.5 \cdot 10^{-10}$ |
| Condition Number for $A_2$ | $2.9 \cdot 10^{5} \pm 2.2 \cdot 10^{5}$ |
| Smallest Eigenvalue for $A_2$ | $9 \cdot 10^{-7} \pm 5.7 \cdot 10^{-7}$ |

*Table 15.* (PS1 on $\mathbb{S}^2$ ) Information from the learning matrix $A_k$'s.

The matrix $A_1$ is used to obtain the estimators, $\widehat{\phi}_{1,1}$ and $\widehat{\phi}_{1,2}$; whereas $A_2$ is used to obtain $\widehat{\phi}_{2,1}$ and $\widehat{\phi}_{2,2}$. Since there is one single predator, we set $\widehat{\phi}_{2,2}$ to zero. It took $9.77 \cdot 10^4$ seconds to generate $\rho^L_{T,\mathcal{M}}$ and $4.01 \cdot 10^5$ seconds to run 10 learning simulations, with $1.66 \cdot 10^3$ seconds spent on learning the estimated interactions (on average, it took $1.66 \cdot 10^2 \pm 4.6$ seconds to run one estimation), and $4.05 \cdot 10^5$ seconds spent on computing the trajectory error estimates (on average, it took $4.0 \cdot 10^4 \pm 7.1 \cdot 10^3$ seconds to run one set of trajectory error estimation).

**Results for the $\mathbb{PD}$ case:** In order to produce more interesting interactions, we choose the distribution of the initial condition to be as follows: the predator is randomly placed in a circle centered at the origin with radius $r_1$, given as follows

$$r_0 = \left(2 + \frac{1}{\cosh(0.5) - 1} - \sqrt{\frac{4}{\cosh(0.5) - 1} + \frac{1}{(\cosh(0.5) - 1)^2}}\right)/2,$$

so that the agents are at most $0.5$ distance away from each other; then the group of preys (Swarm) will be randomly and uniformly placed on an annulus centered at the origin with radii,$(R_1, r_1)$, given as follows

$$r_1 = \left(2 + \frac{1}{\cosh(1) - 1} - \sqrt{\frac{4}{\cosh(1) - 1} + \frac{1}{(\cosh(1) - 1)^2}}\right)/2$$

and

$$R_1 = \left(2 + \frac{1}{\cosh(2) - 1} - \sqrt{\frac{4}{\cosh(2) - 1} + \frac{1}{(\cosh(2) - 1)^2}}\right)/2;$$

so that the group of preys are surrounding the single predator. Table 16 shows the number of basis functions, namely $n_{kk'}$'s, for each estimator $\widehat{\phi}_{kk'}$ for $k, k' = 1, 2$, and their corresponding degrees, $p_{k,k'}$'s, for the Clamped B-spline basis.

| $n_{1,1}$ | $n_{1,2}$ | $n_{2,1}$ | $n_{2,2}$ |
|---|---|---|---|
| 68 | 43 | 43 | 1 |
| $p_{1,1}$ | $p_{1,2}$ | $p_{2,1}$ | $p_{2,2}$ |
| 1 | 1 | 1 | 0 |

*Table 16.* (PS1 on $\mathbb{PD}$ ) Number of basis functions.

Fig. 11 shows the comparison between $\phi_{kk'}$ and its estimators $\widehat{\phi}_{kk'}$ learned from the trajectory data.

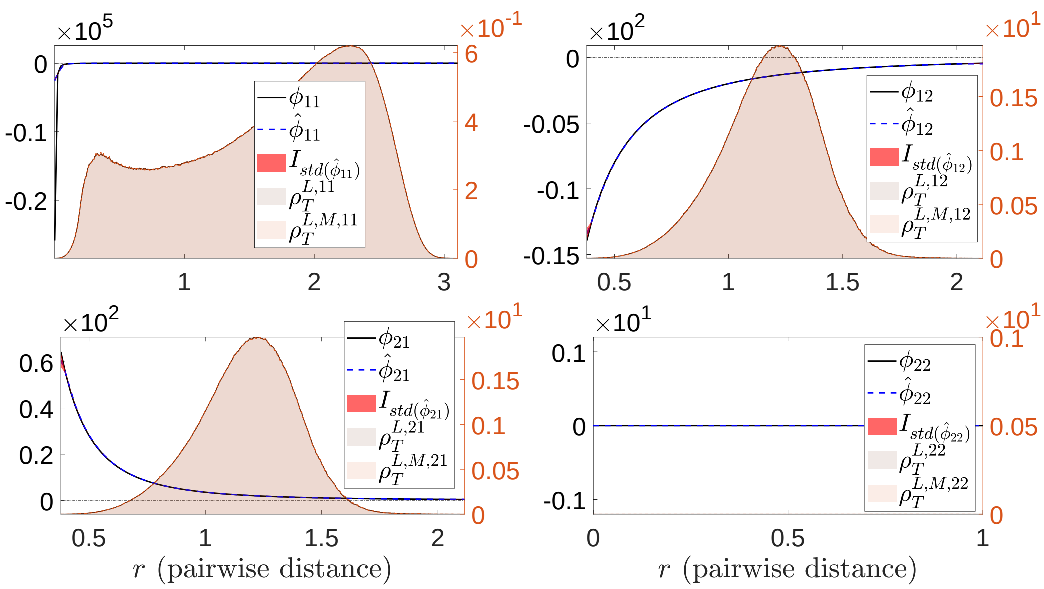


*Figure 11.* (PS1 on $\mathbb{PD}$ ) Comparison of $\phi_{kk'}$ and $\widehat{\phi}_{k,k'}$, with the relative errors shown in table 17. The true interaction kernels are shown in black solid lines, whereas the mean estimated interaction kernel are shown in blue dashed lines with their corresponding std interval, i.e. mean$(\widehat{\phi}_{kk'}) \pm$ std$(\widehat{\phi}_{kk'})$, regions shaded in red. Shown in the background is the comparison of the approximate $\rho_T^{L,kk'}$ versus the empirical $\rho_T^{L,M,kk'}$. Notice that $\rho_T^{L,12}$/$\rho_T^{L,M,12}$ and $\rho_T^{L,12}$/$\rho_T^{L,M,21}$ are the same distributions.

| $\text{Err}_{1,1}$ | $\text{Err}_{1,2}$ | $\text{Err}_{2,1}$ | $\text{Err}_{2,2}$ |
|---|---|---|---|
| $9.0 \cdot 10^{-2} \pm 2.6 \cdot 10^{-3}$ | $1.34 \cdot 10^{-3} \pm 8.8 \cdot 10^{-5}$ | $3.6 \cdot 10^{-3} \pm 2.4 \cdot 10^{-4}$ | 0 |

*Table 17.* (PS1 on $\mathbb{PD}$ ) Relative estimation errors calculated using (13).

Fig. 12 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

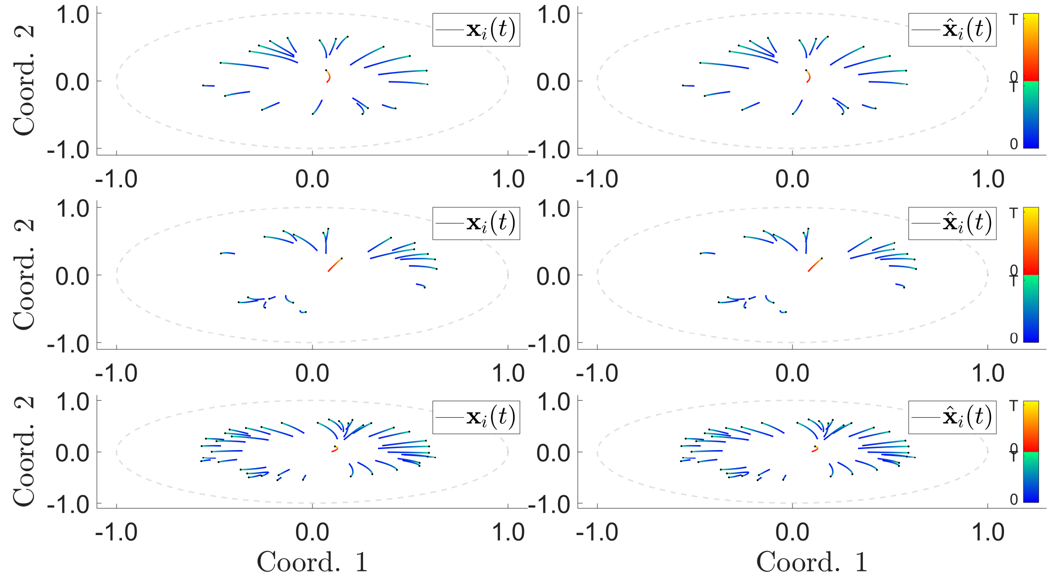


*Figure 12.* (PS1 on $\mathbb{PD}$ ) Comparison of $\boldsymbol{X}$ (generated by $\phi_{k,k'}$'s) and $\hat{\boldsymbol{X}}$ (generated by $\widehat{\phi}_{k,k'}$'s), with the errors reported in table 18. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a randomly chosen initial condition. **Bottom**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a new initial condition with bigger $N = 40$. The color of the trajectory indicates the flow of time, from deep blue/bright red (at $t = 0$) to light green/light yellow (at $t = T$). The blue/green combination is assigned to the preys; whereas the red/yellow comb for the predator.

A quantitative comparison of the trajectory estimation errors is shown in Table 18.

| | $[0, T]$ |
|---|---|
| $\text{mean}_{\text{IC}}$: Training ICs | $4.8 \cdot 10^{-3} \pm 1.2 \cdot 10^{-4}$ |
| $\text{std}_{\text{IC}}$: Training ICs | $2.3 \cdot 10^{-3} \pm 3.0 \cdot 10^{-4}$ |
| $\text{mean}_{\text{IC}}$: Random ICs | $4.8 \cdot 10^{-3} \pm 1.2 \cdot 10^{-4}$ |
| $\text{std}_{\text{IC}}$: Random ICs | $2.5 \cdot 10^{-3} \pm 3.9 \cdot 10^{-3}$ |

*Table 18.* (PS1 on $\mathbb{PD}$ ) trajectory estimation errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu_0(\mathcal{M}^N)$ (second set of two rows). $\text{mean}_{\text{IC}}$ and $\text{std}_{\text{IC}}$ are the mean and standard deviation of the trajectory errors calculated using (14).

We also report the condition number and the smallest eigenvalue of the learning matrix $A$ to indirectly verify the geometric coercivity condition in table 19.

| | |
|---|---|
| Condition Number for $A_1$ | $2.3 \cdot 10^{9} \pm 4.7 \cdot 10^{8}$ |
| Smallest Eigenvalue for $A_1$ | $7 \cdot 10^{-11} \pm 1.7 \cdot 10^{-11}$ |
| Condition Number for $A_2$ | $5 \cdot 10^{5} \pm 3.1 \cdot 10^{5}$ |
| Smallest Eigenvalue for $A_2$ | $4 \cdot 10^{-8} \pm 2.9 \cdot 10^{-8}$ |

*Table 19.* (PS1 on $\mathbb{PD}$ ) Information from the learning matrix $A_k$'s.

The matrix $A_1$ is used to obtain the estimators, $\widehat{\phi}_{1,1}$ and $\widehat{\phi}_{1,2}$; whereas $A_2$ is used to obtain $\widehat{\phi}_{2,1}$ and $\widehat{\phi}_{2,2}$. Since there is one single predator, we set $\widehat{\phi}_{2,2}$ to zero. It took $7.37 \cdot 10^4$ seconds to generate $\rho^L_{T,\mathcal{M}}$ and $2.49 \cdot 10^5$ seconds to run 10 learning simulations, with $1.25 \cdot 10^3$ seconds spent on learning the estimated interactions (on average, it took $1.25 \cdot 10^2 \pm 1.5$ seconds to run one estimation), and $2.48 \cdot 10^5$ seconds spent on computing the trajectory error estimates (on average, it took $2.48 \cdot 10^4 \pm 2.3 \cdot 10^2$ seconds to run one set of trajectory error estimation).

## References

Aydoğdu, A., McQuade, S. T., and Duteil, N. P. Opinion dynamics on a general compact Riemannian manifold. *Networks and Heterogeneous Media*, 12:489, 2017. ISSN 1556-1801. doi: 10.3934/nhm.2017021. URL http://aimsciences.org//article/id/abcc2983-446b-4399-bd12-a509b0d061e8.

Caponigro, M., Lai, A., and Piccoli, B. A nonlinear model of opinion formation on the sphere. *Discrete and Continuous Dynamical Systems*, 35, 05 2014. doi: 10.3934/dcds.2015.35.4241.

Chen, Y. and Kolokolnikov, T. A minimal model of predator-swarm interactions. *J. R. Soc. Interface*, 11:20131208, 2013.

Cucker, F. and Smale, S. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39 (1):1–49, 2002.

do Carmo, M. P. *Diferential Geometry of Curves and Surfaces*. Prentice-Hall, 1976.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer New York, 2006. ISBN 9780387224428. URL https://books.google.at/books?id=3NERBwAAQBAJ.

Hairer, E., Lubich, C., and Wanner, G. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer, 2006.

Lee, J. M. *Introduction to Smooth Manifolds*. Springer, 2003.

Lu, F., Maggioni, M., and Tang, S. Learning interaction kernels in heterogeneous systems of agents from multiple trajectories, 2019a. URL https://arxiv.org/abs/1910.04832.

Lu, F., Zhong, M., Tang, S., and Maggioni, M. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proceedings of the National Academy of Sciences of the United States of America*, 116(29):14424–14433, 2019b. ISSN 10916490. doi: 10.1073/pnas.1822012116.

Miller, J., Tang, S., Zhong, M., and Maggioni, M. Learning theory for inferring interaction kernels in second-order interacting agent systems, 2020. URL https://arxiv.org/abs/2010.03729.

Olson, R., Hintze, A., Dyer, F., Moore, J., and Adami, C. Exploring the coevolution of predator and prey morphology and behavior. *Proceedings of the Artificial Life Conference 2016*, 2016. doi: 10.7551/978-0-262-33936-0-ch045. URL http://dx.doi.org/10.7551/978-0-262-33936-0-ch045.

Weisbuch, G., Deffuant, G., Amblard, F., and Nadal, J.-P. Interacting agents and continuous opinions dynamics. In Cowan, R. and Jonard, N. (eds.), *Heterogenous Agents, Interactions and Economic Performance*, pp. 225–242, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-642-55651-7.