Learning Interaction Kernels for Agent Systems on Riemannian Manifolds

Mauro Maggioni [1 2] **Jason Miller** [1] **Hongda Qiu** [1] **Ming Zhong** [1]

Abstract

Interacting agent and particle systems are extensively used to model complex phenomena in science and engineering. We consider the problem of learning interaction kernels in these dynamical systems constrained to evolve on Riemannian manifolds from given trajectory data. The models we consider are based on interaction kernels depending on pairwise Riemannian distances between agents, with agents interacting locally along the direction of the shortest geodesic connecting them. We show that our estimators converge at a rate that is independent of the dimension of the state space, and derive bounds on the trajectory estimation error, on the manifold, between the observed and estimated dynamics. We demonstrate the performance of our estimator on two classical first order interacting systems: Opinion Dynamics and a Predator-Swarm system, with each system constrained on two prototypical manifolds, the 2-dimensional sphere and the Poincaré disk model of hyperbolic space.

1. Introduction

Dynamical systems of interacting agents, where "agents" may represent atoms, particles, neurons, cells, animals, people, robots, planets, etc..., are an important modeling tool in many disciplines, including Physics, Biology, Chemistry, Economics and Social Sciences. It is a fundamental challenge to learn the governing equations of these systems. Often, agents are either associated with state variables which belong to non-Euclidean spaces, e.g., phase variables considered in various Kuramoto models (Kuramoto, 1975; Strogatz, 2000), or constrained to move on non-Euclidean spaces, for example (Ahn et al., 2020). This has motivated a growing body of research considering interacting agent systems on various manifolds (Lee et al., 2018; Caponigro et al., 2014; Sarlette & Sepulchre, 2008), including opinion dynamics (Aydoğdu et al., 2017), flocking models (Ahn et al., 2020) and a classical aggregation model (C. Fetecau & Zhang, 2019). Further recent approaches for interacting agents on manifolds include (Yang et al., 2020; Soize & Ghanem, 2020).

In this work, we offer a nonparametric and inverse-problem-based learning approach to infer the governing structure of interacting agent dynamics, in the form of $\dot{\boldsymbol{X}}_t = \boldsymbol{f}(\boldsymbol{X}_t)$, constrained on Riemannian manifolds, from observations of trajectories. Our method is different from others introduced to learn ODEs/PDEs from observations, that aim to infer $\boldsymbol{f}$ directly, and would be cursed by the high-dimension of the state space of $\boldsymbol{X}$ (Lu et al., 2019b). Instead, we exploit the form of the function $\boldsymbol{f}$, special to the class of interacting agent systems under consideration, which is determined by an interaction kernel function ϕ of one variable only, and learn ϕ, with minimal assumptions on ϕ. By exploiting invariance of the equations under permutation of the agents as well as the radial symmetry of ϕ, we are able to avoid the curse of dimensionality. We also demonstrate how our approach can perform transfer learning in section 5.

The research on inferring a suitable dynamical system of interacting agents from observation data has been a long-standing problem in science and engineering; see (Lukeman et al., 2010; Katz et al., 2011; Cui et al., 2014; Tran & Ward, 2017) and references therein. Many recent approaches in machine learning have been developed for inferring general dynamical systems, including multistep methods (Keller & Du, 2019), optimization (Wróbel et al., 2013), sparse regression (Brunton et al., 2016; Rudy et al., 2017; Schaeffer et al., 2013), Bayesian regression (Zhang & Lin, 2018), and deep learning (Raissi et al., 2018; Rudy et al., 2019). In a different direction, the generalization of traditional machine learning algorithms in Euclidean settings to Riemannian manifolds, and the development of new algorithms designed to work on Riemannian manifolds, has been attracting increased attention; for example in variational calculus (Soize & Ghanem, 2020), reinforcement learning (Riccio et al., 2018), deep learning (Chen et al., 2020) and theoretical CS (Monte-Alto et al., 2020).

Let $(\mathcal{M}, g)$ be a connected, smooth, and geodesically-

---

[1] Department of Applied Mathematics & Statistics, Johns Hopkins University [2] Department of Mathematics, Department of Applied Mathematics & Statistics, Mathematical Institute for Data Science, Johns Hopkins University.. Correspondence to: Ming Zhong <mzhong5@jhu.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

complete d-dimensional Riemannian manifold, with the Riemannian distance denoted by $d_{\mathcal{M}}$. Consider N interacting agents, each represented by a state vector $\boldsymbol{x}_i(t) \in \mathcal{M}$. Their dynamics is governed by the following first order dynamical system, where ϕ, the *interaction kernel*, is the object of our inference: for each $i = 1, \ldots, N$,

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

and $t \in [0, T]$. Here $\boldsymbol{w}(\boldsymbol{z}_1, \boldsymbol{z}_2)$, for $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{M}$, is a weight vector pointing in the tangent direction at $\boldsymbol{z}_1$ to the shortest geodesic from $\boldsymbol{z}_1$ to $\boldsymbol{z}_2$. For this to make sense, we restrict our attention to local interactions, e.g. by assuming that ϕ is compactly supported in a sufficiently small interval $[0, R]$, so that length-minimizing geodesics exist uniquely. We discuss the well-posedness of this model in greater detail in section 2.1, where we emphasize that this model is derived naturally as a gradient system with a special potential energy depending on pairwise Riemannian distances.

With $(\mathcal{M}, g)$ known to us, our observations consist of $\{\boldsymbol{x}_i^m(t_l), \dot{\boldsymbol{x}}_i^m(t_l)\}_{i,l,m=1}^{N,L,M}$ with $0 = t_1 < \cdots < t_L = T$, L being the number of observations made in time, M being the number of trajectories, and each $(\boldsymbol{x}_i^m(0))_{i=1}^N \in \mathcal{M}^N$ is drawn i.i.d from a probability measure $\mu_0(\mathcal{M}^N)$. We construct an estimator $\widehat{\phi}_{L,M,\mathcal{H}}$ of ϕ, close to ϕ in an appropriate L^2 sense, and generating a system in the form (1) with trajectories close to those of the original system (with the same initial condition); it is defined as

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

Here $\mathcal{H}$ is a function space containing suitable approximations to ϕ and $\mathcal{E}_{L,M,\mathcal{M}}$ is a least squares loss functional built from the trajectory data, which takes into account the geometry of $(\mathcal{M}, g)$. Having established a geometry-dependent coercivity condition that ensures, among other things, the recoverability of ϕ, our theory shows that the convergence rate (in M) of our estimator to the true interaction kernel is independent of the dimension Nd of the observation data, and is the same as the minimax rate for 1-dimensional nonparametric regression:

$$\mathbb{E}\Big[\Big\|\widehat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot - \phi(\cdot)\cdot\Big\|_{L^2(\rho_{T,\mathcal{M}}^L)}\Big] \lesssim \Big(\frac{\log M}{M}\Big)^{\frac{1}{3}}.$$

where the expectation is with respect to the initial condition distributed as described above, ϕ is assumed to be 1-time differentiable, $\rho_{T,\mathcal{M}}^L$ is a dynamics-adapted probability measure which captures the distribution of pairwise Riemannian distances, and the implicit constant depends on $\mathcal{M}$.

We also establish bounds on trajectory predictions: let $\hat{\boldsymbol{X}}_{[0,T]}, \boldsymbol{X}_{[0,T]}$ be trajectories evolved with the interaction kernels $\widehat{\phi}_{L,M,\mathcal{H}}$ and ϕ respectively, started at the same initial condition, then:

$$\mathbb{E}\Big[d_{\text{trj}}(\boldsymbol{X}_{[0,T]}, \hat{\boldsymbol{X}}_{[0,T]})^2\Big] \lesssim \Big\|\phi(\cdot)\cdot - \hat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot\Big\|^2_{L^2(\rho_{T,\mathcal{M}})},$$

where d_{trj} is a natural geometry-based distance on trajectories. As M grows, the norm on the right hand side converges at the rate above, yielding convergence of the trajectories. We demonstrate the performance of our estimators on an opinion dynamics and a predator-swarm model, each constrained on two model manifolds: the two-dimensional sphere $\mathbb{S}^2$ and the Poincaré disk.

2. Model Equations

In this section we introduce the governing equations which we use to model interacting agents constrained on Riemannian manifolds, and discuss the properties of the dynamics. Table 1 shows a list of definitions of the common terms used throughout this paper.

Variable	Definition
$(\mathcal{M}, g)$	Riemannian Manifold with metric g
$T_{\boldsymbol{x}}\mathcal{M}$	Tangent plane to $\mathcal{M}$ at $\boldsymbol{x}$
$\langle \cdot, \cdot \rangle_{g(\boldsymbol{x})}$, $\langle \cdot, \cdot \rangle_g$	Inner product on $T_{\boldsymbol{x}}\mathcal{M}$
$\|\boldsymbol{v}\|_{T_{\boldsymbol{x}}\mathcal{M}}$, $\|\boldsymbol{v}\|_g$	Length of $\boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{M}$ induced by $g(\boldsymbol{x})$
$d_{\mathcal{M}}(\cdot, \cdot)$	Geodesic distance induced by g

Table 1. Notation for first-order models.

2.1. Main model

In order to motivate the choice of the model equations we use, we begin with a geometric gradient flow model of an interacting agent system. Consider a system of N interacting agents, with each agent described by a state vector $\boldsymbol{x}_i(t)$ on a d-dimensional connected, smooth, and geodesically complete Riemannian manifold $\mathcal{M}$ with metric g. The change of the state vectors seeks to decrease a system energy E:

$$\frac{d\boldsymbol{x}_i(t)}{dt} = -\partial_{\boldsymbol{x}_i} E(\boldsymbol{x}_1(t), \ldots, \boldsymbol{x}_N(t)), \qquad i = 1, \ldots, N.$$

Our first key assumption is that E takes the special form

$$E(\boldsymbol{x}_1(t), \ldots, \boldsymbol{x}_N(t)) = \frac{1}{N}\sum_{i'=1}^{N} U(d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))^2),$$

for some $U : \mathbb{R}^+ \to \mathbb{R}$ with $U(0) = 0$, and $d_{\mathcal{M}}(\cdot, \cdot)$ the geodesic distance on $(\mathcal{M}, g)$. Simplifying, and omitting from the notation the dependency on t of $\dot{\boldsymbol{x}}_i$ and $\boldsymbol{x}_i$, we obtain the first-order geometric evolution equation,

$$\dot{\boldsymbol{x}}_i = \frac{1}{N}\sum_{i'=1}^{N} \phi(d_{\mathcal{M}}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}))\boldsymbol{w}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}), \tag{2}$$

for $i = 1, \ldots, N$. We call $\phi(r) := 2U'(r^2)$ the *interaction kernel*. We have let $\boldsymbol{w}(\boldsymbol{z}_1, \boldsymbol{z}_2) := d_{\mathcal{M}}(\boldsymbol{z}_1, \boldsymbol{z}_2)\boldsymbol{v}(\boldsymbol{z}_1, \boldsymbol{z}_2)$ for

$\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{M}$, with $\boldsymbol{v}(\boldsymbol{z}_1, \boldsymbol{z}_2)$ being, for $\boldsymbol{z}_2 \neq \boldsymbol{z}_1$, the unit vector (i.e. $\|\boldsymbol{v}\|_{T_{\boldsymbol{z}_1}\mathcal{M}} = 1$) tangent at $\boldsymbol{z}_1$ to the minimizing geodesic from $\boldsymbol{z}_1$ to $\boldsymbol{z}_2$ if $\boldsymbol{z}_2$ is not in the cut locus of $\boldsymbol{z}_1$, and equal to $\mathbf{0}$ otherwise. In order to guarantee existence and uniqueness of a solution for (2) over the time interval $[0, T]$, we make a further assumption that ϕ belongs to

$$\mathcal{K}_{R,S} := \{\varphi \in C^1([0, R]) \Big| \, \|\varphi\|_{L^\infty} + \|\varphi'\|_{L^\infty} \leq S\},$$

for some constant $S > 0$. Here, R is smaller than the global injectivity radius of $\mathcal{M}$, and $L^\infty = L^\infty([0, R])$. With this assumption, the possible discountinuity of $\boldsymbol{v}(\boldsymbol{z}_1, \boldsymbol{z}_2)$ due to either $\boldsymbol{z}_2 \to \boldsymbol{z}_1$ or $\boldsymbol{z}_2$ tends to a point in the cut locus of $\boldsymbol{z}_1$ is canceled by the multiplication by $d_\mathcal{M}(\boldsymbol{z}_1, \boldsymbol{z}_2) \to 0$ in the former case, and $\phi(d_\mathcal{M}(\boldsymbol{z}_1, \boldsymbol{z}_2)) \to 0$ in the latter case. Therefore, the ODE system in (2) has a Lipschitz right-hand side, and thus it has a unique solution existing for $t \in [0, T]$ see (Hairer et al., 2006).

Using this geometric gradient flow point of view, the form of the equations and the radial symmetry of the interaction kernels are naturally pre-determined by the energy potential. This approach seems to us natural and geometric; for different approaches see (Aydoğdu et al., 2017; Caponigro et al., 2014). Note that in the case of $\mathcal{M} = \mathbb{R}^d$ with the Euclidean metric, we have $d_\mathcal{M}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|$ and $\boldsymbol{v}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \frac{\boldsymbol{x}_{i'} - \boldsymbol{x}_i}{\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|}$, and we recover the Euclidean space models used in (Bongini et al., 2017; Lu et al., 2019b) and the many works referenced therein. Moreover, our learning method still applies to models with different definitions of the weight vector, e.g. $\boldsymbol{w}(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$, as long as $\boldsymbol{w}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) \in T_{\boldsymbol{x}_i}\mathcal{M}$.

3. Learning Framework

We are given a set of trajectory data of the form $\{\boldsymbol{x}_i^m(t_l), \dot{\boldsymbol{x}}_i^m(t_l)\}_{i,l,m=1}^{N,L,M}$, for $0 = t_1 < \ldots < t_L = T$, with the initial conditions $\{\boldsymbol{x}_i^m(0)\}_{i=1}^N$ being i.i.d from a distribution $\mu_0(\mathcal{M})$. The objective is to construct an estimator $\widehat{\phi}_{L,M,\mathcal{H}}$ of the interaction kernel ϕ.

Before we describe the construction of our estimator, we introduce some vectorized notations. We let, in $\mathcal{M}^N := \mathcal{M} \times \cdots \times \mathcal{M}$,

$$\boldsymbol{X}_{t_l}^m := \begin{bmatrix} \vdots \\ \boldsymbol{x}_i^m(t_l) \\ \vdots \end{bmatrix} \quad \text{and} \quad \boldsymbol{X} := \begin{bmatrix} \vdots \\ \boldsymbol{x}_i \\ \vdots \end{bmatrix},$$

where $(\mathcal{M}^N, g_\mathcal{M}^N)$ is the canonical product of Riemannian manifolds with product Riemannian metric given by,

$$\left\langle \begin{bmatrix} \vdots \\ \boldsymbol{u}_i \\ \vdots \end{bmatrix}, \begin{bmatrix} \vdots \\ \boldsymbol{z}_i \\ \vdots \end{bmatrix} \right\rangle_{g_\mathcal{M}^N(\boldsymbol{X})} := \frac{1}{N} \sum_{i=1}^N \langle \boldsymbol{u}_i, \boldsymbol{z}_i \rangle_{g(\boldsymbol{x}_i)},$$

for $\boldsymbol{u}_i, \boldsymbol{z}_i \in T_{\boldsymbol{x}_i}\mathcal{M}$. The initial conditions, $\boldsymbol{X}_0^m$ are drawn **i.i.d.** from $\mu_0(\mathcal{M}^N)$. Finally, $\boldsymbol{f}_\phi$ is the vector field on $\mathcal{M}^N$ (i.e. $\boldsymbol{f}_\phi(\boldsymbol{X}) \in T_{\boldsymbol{X}}\mathcal{M}^N$ for $\boldsymbol{X} \in \mathcal{M}^N$), given by

$$\boldsymbol{f}_\phi(\boldsymbol{X}_{t_l}^m) := \begin{bmatrix} \vdots \\ \frac{1}{N}\sum_{i'=1}^N \phi(d_\mathcal{M}(\boldsymbol{x}_i^m(t_l), \boldsymbol{x}_{i'}^m(t_l)))\boldsymbol{w}(\boldsymbol{x}_i^m(t_l), \boldsymbol{x}_{i'}^m(t_l)) \\ \vdots \end{bmatrix},$$

The system of equations (2) can then be rewritten, for each $m = 1, \ldots, M$, as $\dot{\boldsymbol{X}}_t^m = \boldsymbol{f}_\phi(\boldsymbol{X}_t^m)$.

3.1. Geometric Loss Functionals

In order to simplify the presentation, we assume that the observation times, i.e. $\{t_l\}_{l=1}^L$, are equispaced in $[0, T]$ (the general case is similar). We begin with the definition of the hypothesis space $\mathcal{H}$, over which we shall minimize an error functional to obtain an estimator of ϕ.

Definition 3.1. *An admissible hypothesis space $\mathcal{H}$ is a compact (in L^∞-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 φ is smooth enough to ensure the existence and uniqueness of solutions of (2) for $t \in [0, T]$, i.e. $\varphi \in \mathcal{H} \cap \mathcal{K}_{R,S_0}$.*

For a function $\varphi \in \mathcal{H}$, we define the loss functional

$$\mathcal{E}_{L,M,\mathcal{M}}(\varphi) := \frac{1}{ML} \sum_{l,m=1}^{L,M} \left\| \dot{\boldsymbol{X}}_{t_l}^m - \boldsymbol{f}_\varphi(\boldsymbol{X}_{t_l}^m) \right\|_g^2, \tag{3}$$

where the norm $\|\cdot\|_g$ in $T_{\boldsymbol{X}_{t_l}^m}\mathcal{M}^N$ can be written as

$$\left\| \dot{\boldsymbol{X}}_{t_l}^m - \boldsymbol{f}_\varphi(\boldsymbol{X}_{t_l}^m) \right\|_g^2 =$$

$$\frac{1}{N} \sum_{i=1}^N \left\| \dot{\boldsymbol{x}}_{i,t_l}^m - \frac{1}{N} \sum_{i'=1}^N \varphi(r_{ii',t_l}^m) \boldsymbol{w}_{ii',t_l}^m \right\|_{T_{\boldsymbol{x}_i^m(t_l)}\mathcal{M}}^2,$$

with $\dot{\boldsymbol{x}}_{i,t_l}^m := \dot{\boldsymbol{x}}_i^m(t_l)$, $r_{ii',t_l}^m := d_\mathcal{M}(\boldsymbol{x}_i^m(t_l), \boldsymbol{x}_{i'}^m(t_l))$, and $\boldsymbol{w}_{ii',t_l}^m := \boldsymbol{w}(\boldsymbol{x}_i^m(t_l), \boldsymbol{x}_{i'}^m(t_l))$. This loss functional is nonnegative, and reaches 0 when φ is equal to the (true) interaction kernel ϕ if $\phi \in \mathcal{H} \cap \mathcal{K}_{R,S}$. Given that $\mathcal{H}$ is compact and convex and $\mathcal{E}_{L,M,\mathcal{M}}$ is continuous on $\mathcal{H}$, the minimizer of $\mathcal{E}_{L,M,\mathcal{M}}$ exists and is unique. We define it to be our estimator:

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

As $M \to \infty$, since each trajectory has i.i.d. ICs, by the law of large numbers, we have $\mathcal{E}_{L,M,\mathcal{M}} \to \mathcal{E}_{L,\infty,\mathcal{M}}$, with

$$\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi) := \frac{1}{L} \sum_{l=1}^L \mathbb{E}_{\boldsymbol{X}_0 \sim \mu_0(\mathcal{M}^N)} \left[\left\| \dot{\boldsymbol{X}}_{t_l} - \boldsymbol{f}_\varphi(\boldsymbol{X}_{t_l}) \right\|_g^2 \right]. \tag{4}$$

Since $\mathcal{E}_{L,\infty,\mathcal{M}}$ is continuous on $\mathcal{H}$, the minimization of $\mathcal{E}_{L,\infty,\mathcal{M}}$ over $\mathcal{H}$ is well-posed and it has a unique minimizer $\widehat{\phi}_{L,\infty,\mathcal{H}} := \mathrm{argmin}_{\varphi\in\mathcal{H}}\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi)$. Much of our theoretical work establishes the relationship between the estimator $\widehat{\phi}_{L,M,\mathcal{H}}$, the closely related (in the infinite sample limit $M\to\infty$) $\widehat{\phi}_{L,\infty,\mathcal{H}}$, and the true interaction kernel ϕ.

3.2. Performance Measures

We introduce a suitable normed function space in which to compare the estimator $\widehat{\phi}_{L,M,\mathcal{H}}$ with the true interaction kernel ϕ. We also measure performance in terms of trajectory estimation error based on a distance between trajectories generated from the true dynamics (evolved using ϕ with some initial condition $\boldsymbol{X}_0 \sim \mu_0(\mathcal{M}^N)$) and the estimated dynamics (evolved using the estimated interaction kernel $\widehat{\phi}_{L,M,\mathcal{H}}$, and with the same initial condition, i.e. $\boldsymbol{X}_0$).

3.2.1. ESTIMATION ERROR

First we introduce a probability measure $\rho_{T,\mathcal{M}}$ on $\mathbb{R}_+$, that is used to define a norm to measure the error of the estimator, derived from the loss functionals (given by (3) and (4)), that reflects the distribution of pairwise data given by the dynamics as well as the geometry of the manifold $\mathcal{M}$:

$$\rho_{T,\mathcal{M}}(r) := \frac{1}{\binom{N}{2}}\mathbb{E}\Big[\frac{1}{T}\int_0^T\sum_{i,i'}\delta_{d_{\mathcal{M}}(\boldsymbol{x}_i(t),\boldsymbol{x}_{i'}(t))}(r)\,dt\Big],$$

where δ is the Dirac measure. Note that $\mathbb{E}$ is w.r.t $\boldsymbol{X}_0 \sim \mu_0(\mathcal{M}^N)$. In words, this measure is obtained by averaging δ-functions having mass at any pairwise distances in any trajectory, over all initial conditions drawn from $\mu_0(\mathcal{M}^N)$, over all pairs of agents and all times. A time-discretized version is given by:

$$\rho^L_{T,\mathcal{M}}(r) := \frac{1}{L\binom{N}{2}}\mathbb{E}\Big[\sum_{l=1}^{L}\sum_{1\le i<i'\le N}\delta_{d_{\mathcal{M}}(\boldsymbol{x}_i(t_l),\boldsymbol{x}_{i'}(t_l))}(r)\Big].$$

Note that $\mathbb{E}$ is w.r.t $\boldsymbol{X}_0 \sim \mu_0(\mathcal{M}^N)$. The two probability measures defined above appear naturally in the proofs for the convergence rate of the estimator. From observational data we compute the empirical version:

$$\rho^{L,M}_{T,\mathcal{M}}(r) := \frac{1}{ML\binom{N}{2}}\sum_{l,m=1}^{L,M}\sum_{1\le i<i'\le N}\delta_{d_{\mathcal{M}}(\boldsymbol{x}_i(t_l),\boldsymbol{x}_{i'}(t_l))}(r).$$

The geometry of $\mathcal{M}$ is incorporated in these three measures by the presence of geodesic distances. The norm

$$\|\varphi(\cdot)\cdot\|^2_{L^2(\rho_{T,\mathcal{M}})} := \int_{r=0}^{\infty}\big|\varphi(r)r\big|^2\,d\rho_{T,\mathcal{M}}(r)$$

is used to define the estimation error: $||\widehat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot-\phi(\cdot)\cdot||_{L^2(\rho_{T,\mathcal{M}})}$. We also use a relative version of this error, to enable a meaningful comparison across different interaction kernels:

$$\|\varphi(\cdot)\cdot-\phi(\cdot)\cdot\|_{\mathrm{Rel.}L^2(\rho_{T,\mathcal{M}})} := \frac{\|\varphi(\cdot)\cdot-\phi(\cdot)\cdot\|_{L^2(\rho_{T,\mathcal{M}})}}{\|\phi(\cdot)\cdot\|_{L^2(\rho_{T,\mathcal{M}})}}. \tag{5}$$

3.2.2. TRAJECTORY ESTIMATION ERROR

Let $\boldsymbol{X}^m_{[0,T]} := (\boldsymbol{X}^m_t)_{t\in[0,T]}$ be the trajectory generated by the m^{th} initial condition, $\boldsymbol{X}^m_0$. The trajectory estimation error between $\boldsymbol{X}^m_{[0,T]}$ and $\hat{\boldsymbol{X}}^m_{[0,T]}$, evolved using, the unknown interaction kernel ϕ and, respectively, the estimated one, $\widehat{\phi}$, with the same initial condition, is given by

$$d_{\mathrm{trj}}(\boldsymbol{X}^m_{[0,T]},\hat{\boldsymbol{X}}^m_{[0,T]})^2 := \sup_{t\in[0,T]}\frac{\sum_i d_{\mathcal{M}}(\boldsymbol{x}^m_i(t),\hat{\boldsymbol{x}}^m_i(t))^2}{N}. \tag{6}$$

This quantity is random with the initial conditions, hence we report the mean and standard deviation of these trajectory errors over a (large) number of initial conditions sampled i.i.d. from $\mu_0(\mathcal{M}^N)$; and the errors are denoted as $\mathrm{mean}_{\mathrm{IC}}$ and $\mathrm{std}_{\mathrm{IC}}$ respectively.

3.3. Algorithm and Computational Complexity

Algorithm[1] 1 shows the detailed steps on how to construct the estimator to ϕ given the observation data. We emphasize that our estimator, and the learning theory we develop, do not dependent on a particular choice of basis. In our examples we choose Clamped B-splines due to their regularity and approximation-theoretic properties.

Assuming a finite dimensional subspace of $\mathcal{H}$, i.e. $\mathcal{H}_M \subset \mathcal{H}$ with $\dim(\mathcal{H}_M) = n(M)$, we are able to re-write the minimization problem of (3) over $\mathcal{H}_M$ as a linear system, i.e. $A_M\vec{\alpha} = \vec{b}_M$ with $A_M \in R^{n\times n}$ and $\vec{b}_M \in \mathbb{R}^{n\times 1}$; for details, see the Sec. $C.1$. in SI. Moreover, this linear system is well conditioned, ensured by the geometric coercivity condition.

The total computational cost for solving the learning problem is of $\mathcal{O}(M^{\frac{5}{3}})$ when 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. 4.2 is used. 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 $= \mathcal{O}((\frac{M}{\text{num. cores}})^{5/3})$.

---

[1] Implementation of the algorithm can be found on `https://github.com/MingZhongCodes/LearningDynamics`, which also includes code to reproduce the results presented here.

Algorithm 1 Learning Algorithm

```
**Input:** data $\{\boldsymbol{x}_i^m(t_l), \dot{\boldsymbol{x}}_i^m(t_l)\}_{i,l,m=1}^{N,L,M}$
Compute $R^{\text{obs}}_{\{\min,\max\}} = \{\min,\max\}_{i,i',l,m} d_{\mathcal{M}}(\boldsymbol{x}_i^m(t_l), \boldsymbol{x}_{i'}^m(t_l))$
Choose a type of basis functions, e.g., clamped B-spline
Construct basis of $\mathcal{H}_M$, e.g. $\{\psi_\eta\}_{\eta=1}^n$, on the uniform partition of $[R^{\text{obs}}_{\min}, R^{\text{obs}}_{\max}]$
Choose either a local chart $\mathcal{U} : \mathcal{M} \to \mathbb{R}^d$ or a natural embedding $\mathcal{I} : \mathcal{M} \to \mathbb{R}^{d'}$
Construct $\Psi^m \in (T_{\boldsymbol{X}^m_{t_1}}\mathcal{M}^N \times \cdots \times T_{\boldsymbol{X}^m_{t_L}}\mathcal{M}^N)^n$ and $\vec{d}^m \in T_{\boldsymbol{X}^m_{t_1}}\mathcal{M}^N \times \cdots \times T_{\boldsymbol{X}^m_{t_L}}\mathcal{M}^N$:
    $\Psi^m(:,\eta) := \Psi^m_\eta = \frac{1}{\sqrt{N}} \begin{bmatrix} \boldsymbol{f}_\phi(\boldsymbol{X}^m_{t_1}) \\ \vdots \\ \boldsymbol{f}_\phi(\boldsymbol{X}^m_{t_L}) \end{bmatrix}, \ \vec{d}^m := \frac{1}{\sqrt{N}} \begin{bmatrix} \dot{\boldsymbol{X}}^m_{t_1} \\ \vdots \\ \dot{\boldsymbol{X}}^m_{t_L} \end{bmatrix}$
Define $\langle \cdot, \cdot \rangle_G$ on $\Psi^m_\eta \in T_{\boldsymbol{X}^m_{t_1}}\mathcal{M}^N \times \cdots \times T_{\boldsymbol{X}^m_{t_L}}\mathcal{M}^N$ as
    $\langle \Psi^m_\eta, \Psi^m_{\eta'} \rangle_G = \sum_{l=1}^L \langle \boldsymbol{f}_\phi(\boldsymbol{X}^m_{t_l}), \boldsymbol{f}_\phi(\boldsymbol{X}^m_{t_l}) \rangle_{g^{\mathcal{M}^N}(\boldsymbol{X}^m_l)}$
Assemble $A_M(\eta,\eta') = \frac{1}{LM}\sum_{m=1}^M \langle \Psi^m_\eta, \Psi^m_{\eta'} \rangle_G \in \mathbb{R}^{n\times n}$.
Assemble $\vec{b}_M(\eta) = \frac{1}{LM}\sum_{m=1}^M \langle \vec{d}, \Psi^m_\eta \rangle_G \in \mathbb{R}^{n\times 1}$.
Solve $A_M \vec{\alpha} = \vec{b}_M$ for $\hat{\vec{\alpha}} \in \mathbb{R}^n$.
Assemble $\widehat{\phi} = \sum_{\eta=1}^n \hat{\alpha}_\eta \psi_\eta$.
```

4. Learning Theory

We present in this section the major results, including the convergence of the estimator $\widehat{\phi}_{L,M,\mathcal{H}}$ to ϕ at the optimal learning rate, and bounding the trajectory estimation error between the true and estimated dynamics (evolved using $\widehat{\phi}_{L,M,\mathcal{H}}$), with corresponding proofs in Sec. B in the SI.

4.1. Learnability: geometric coercivity condition

We establish a geometry-adapted coercivity condition, extending that of (Bongini et al., 2017; Lu et al., 2019b) to the Riemannian setting, which will guarantee the uniqueness of the minimizer of $\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi)$, and show that $\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi)$ controls the $\|\cdot\|_{L^2(\rho_{T,\mathcal{M}})}$ distance between the minimizer and the true interaction kernel.

Definition 4.1 (Geometric Coercivity condition)**.** *The geometric evolution system in* (2) *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* $c \equiv c_{L,N,\mathcal{H},\mathcal{M}} > 0$ *such that for any* $\varphi \in \mathcal{H}$ *with* $\varphi(\cdot)\cdot \in L^2(\rho^L_{T,\mathcal{M}})$ *we have*

$$c\,\|\varphi(\cdot)\cdot\|^2_{L^2(\rho^L_{T,\mathcal{M}})} \leq \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}\Big[\,\|\boldsymbol{f}_\varphi(\boldsymbol{X}_{t_l})\|^2_{T_{\boldsymbol{X}_{t_l}}\mathcal{M}^N}\Big].$$

Here and in what follows, $\mathbb{E}$ is taken, as usual, w.r.t $\boldsymbol{X}_0 \sim \mu_0(\mathcal{M}^N)$; unless otherwise indicated. In order to simplify the argument on how this geometric coercivity condition controls the distance between $\widehat{\phi}_{L,\infty,\mathcal{H}}$ and ϕ, we introduce the inner product on $L^2 = L^2(\rho^L_{T,\mathcal{M}})$ defined as

$$\langle\!\langle \varphi_1, \varphi_2 \rangle\!\rangle_{L^2} := \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}\Big[\langle \boldsymbol{f}_{\varphi_1}(\boldsymbol{X}_{t_l}), \boldsymbol{f}_{\varphi_2}(\boldsymbol{X}_{t_l}) \rangle_{T_{\boldsymbol{X}_{t_l}}\mathcal{M}^N}\Big].$$

Then the geometric coercivity condition can be rewritten as

$$c_{L,N,\mathcal{H},\mathcal{M}}\,\|\varphi(\cdot)\cdot\|^2_{L^2(\rho^L_{T,\mathcal{M}})} \leq \langle\!\langle \varphi, \varphi \rangle\!\rangle_{L^2(\rho^L_{T,\mathcal{M}})},$$

and since the loss function from (4) can be written as $\mathcal{E}_{L,\infty,\mathcal{H}}(\varphi) = \langle\!\langle \varphi - \phi, \varphi - \phi \rangle\!\rangle$, this implies

$$c_{L,N,\mathcal{H},\mathcal{M}}\,\|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|^2_{L^2(\rho^L_{T,\mathcal{M}})} \leq \mathcal{E}_{L,\infty,\mathcal{H}}(\varphi).$$

Hence when $\mathcal{E}_{L,\infty,\mathcal{H}}(\varphi)$ is small, $\|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})}$ is also small; hence if we construct a sequence of minimizers of $\mathcal{E}_{L,\infty,\mathcal{H}}$ over increasing $\mathcal{H}$ with decreasing $\mathcal{E}_{L,\infty,\mathcal{H}}$ values, the convergence of $\widehat{\phi}_{L,\infty,\mathcal{H}}$ to ϕ can be established.

4.2. Concentration and Consistency

The first theorem bounds, with high probability, the difference between the estimator $\widehat{\phi}_{L,M,\mathcal{H}}$ and the true interaction kernel ϕ, which makes apparent the trade-off between the $L^2(\rho^L_{T,\mathcal{M}})$-distance between ϕ and $\mathcal{H}$ (approximation error), and M the number of trajectories needed for achieving the desired accuracy. Here $\mathcal{N}(\mathcal{U},\epsilon)$ is the covering number of a set $\mathcal{U}$ with open balls of radius ϵ w.r.t the L^∞-norm.

Theorem 4.1. *Let* $\phi \in L^2([0,R])$*, and* $\mathcal{H}$ *an admissible hypothesis space such that the geometric coercivity condition holds with a constant* $c_{L,N,\mathcal{H},\mathcal{M}}$*. Then,* $\widehat{\phi}_{L,M,\mathcal{H}}$*, minimizer of* (3) *on the trajectory data generated by* (2)*, satisfies*

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

with probability at least $1-\tau$*, when* $M \geq \frac{1152 S_0^2 R^2}{\epsilon c_{L,N,\mathcal{H},\mathcal{M}}}(\ln \mathcal{N}(\mathcal{H}, \frac{\epsilon}{48 S_0 R^2}) + \ln \frac{1}{\tau})$.

This quantifies the usual bias-variance tradeoff in our setting: on the one hand, with a large hypothesis space, the quantity $\inf_{\varphi\in\mathcal{H}} \|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})}$ could be made small. On the other hand, we wish to have the right number of samples to make the variance of the estimator small, by controlling the covering number of the hypothesis space $\mathcal{H}$.

4.3. Convergence Rate

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

Theorem 4.2. *Let $\mu_0(\mathcal{M}^N)$ be the distribution of the initial conditions of trajectories, and $\mathcal{H}_M = \mathcal{B}_n$ with $n = n_* \asymp (M/\log M)^{\frac{1}{2s+1}}$, 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*

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

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

$$\mathbb{E}_{\boldsymbol{X}_0 \sim \mu_0(\mathcal{M}^N)} \Big[\Big\| \widehat{\phi}_{L,M,\mathcal{H}_M}(\cdot)\cdot - \phi(\cdot)\cdot \Big\|_{L^2(\rho^L_{T,\mathcal{M}})} \Big] \leq \frac{C(S,R,c_0,c_1)}{c_{L,N,\mathcal{L},\mathcal{M}}} \Big(\frac{\log M}{M} \Big)^{\frac{s}{2s+1}}.$$

The constant s is tied closely to the regularity of ϕ, and it plays an important role in the convergence rate. For example, when $\phi \in C^1$, we can take $s = 1$ with linear spaces of first degree piecewise polynomials, we end up with a $M^{\frac{1}{3}}$ learning rate. The rate is the same as the minimax rate for nonparametric regression with noise in one dimension (up to the logarithmic factor), and in particular it is independent of the dimension $D = Nd$ of the state space. Empirical results suggest that at least in some cases, when L grows, i.e. each trajectory is sampled at more points, then the estimators improve; this is however not captured by our bound.

4.4. Trajectory Estimation Error

We have established the convergence of the estimator $\widehat{\phi}_{L,M,\mathcal{H}}$ to the true interaction kernel ϕ. We now establish the convergence of the trajectories of the estimated dynamics, evolved using $\widehat{\phi}_{L,M,\mathcal{H}}$, to the observed trajectories.

Theorem 4.3. *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 (2) w.r.t to ϕ and $\widehat{\phi}$, respectively, for $t \in [0,T]$, with $\hat{\boldsymbol{X}}_0 = \boldsymbol{X}_0$. Then we have the following inequality,*

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

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

While these bounds are mainly useful for small times T, given the exponential dependence on T of the bounds, they can be overly pessimistic. It may also happen that the predicted trajectories are not accurate in terms of agent positions, but they maintain, and even predict from initial conditions, large-scale, emergent properties of the original system, such as flocking of birds of milling of fish (Zhong et al., 2020). We suspect this can hold also in the manifold setting, albeit in ways that are affected by geometric properties of the manifold.

5. Numerical Experiments

We consider two prototypical first order dynamics, Opinion Dynamics (OD) and Predator-Swarm dynamics (PS1), each 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). These are model spaces with constant positive and negative curvature, respectively. We conduct extensive experiments on these four scenarios to demonstrate the performance of the estimators both in terms of the estimation errors (approximating ϕ's) and trajectory estimator errors (estimating the observed dynamics) over $[0,T]$.

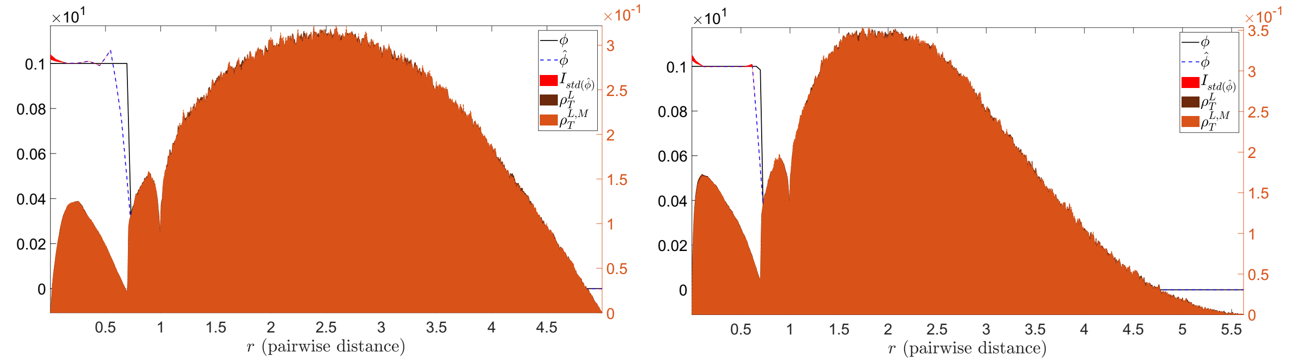
For each type of dynamics, on each of the two model manifolds, we visualize trajectories of the system, with a random initial condition (i.e. not in the training set), driven by ϕ and $\widehat{\phi}$. We also augment the system by adding new agents: without any re-learning, thus we can transfer $\widehat{\phi}$ to drive this augmented system (with $N = 40$ in our examples), and will visualize the trajectories (again, started from a new random initial condition). We also report on the (relative) estimation error of the interaction kernel, as defined in (5), and on the trajectory errors, defined in (6).

For each system of $N = 20$ agents, we take $M = 500$ and $L = 500$ to generate the training data. For each $\mathcal{H}_M$, we use first-degree clamped B-splines as the basis functions with $\dim(\mathcal{H}_M) = \mathcal{O}(n_*) = \mathcal{O}((\frac{ML}{\log(ML)})^{\frac{1}{3}})$. We use a geometric numerical integrator (Hairer, 2001) (4^{th} order Backward Differentiation Formula with a projection scheme) for the evolution of the dynamics. For details, see Sec. C in the SI.

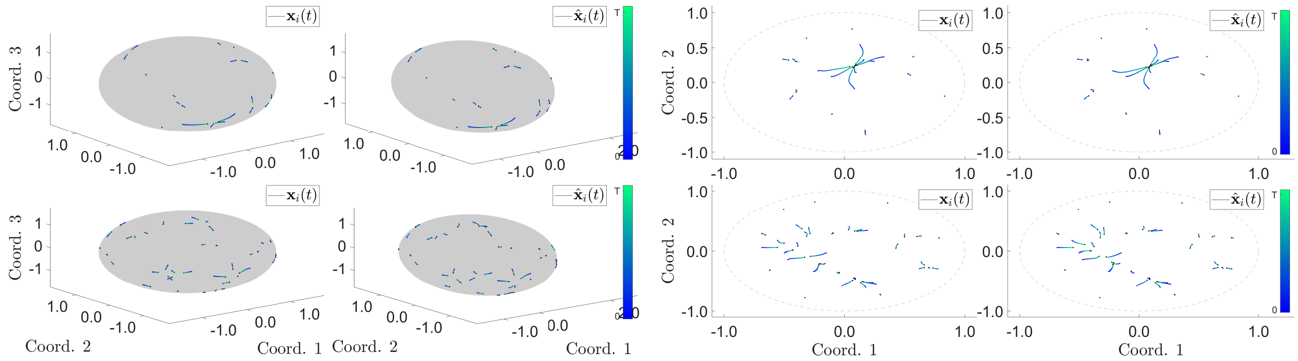
OD	$[0,T]$
$\text{mean}_{\text{IC}}^{\mathbb{S}^2}$: Training ICs	$8.8 \cdot 10^{-2} \pm 1.7 \cdot 10^{-3}$
$\text{mean}_{\text{IC}}^{\mathbb{S}^2}$: Random ICs	$9.0 \cdot 10^{-2} \pm 1.6 \cdot 10^{-3}$
$\text{mean}_{\text{IC}}^{\mathbb{PD}}$: Training ICs	$1.08 \cdot 10^{-1} \pm 1.6 \cdot 10^{-3}$
$\text{mean}_{\text{IC}}^{\mathbb{PD}}$: Random ICs	$1.08 \cdot 10^{-1} \pm 2.6 \cdot 10^{-3}$

Table 2. (Dynamics on $\mathbb{S}^2$ or $\mathbb{PD}$) mean_{IC} is the mean of the trajectory errors over M initial conditions (ICs), as defined in eq.(6).

Opinion Dynamics (OD) is used to model simple interactions of opinions (Aydoğdu et al., 2017; Weisbuch et al., 2003) as well as choreography (Caponigro et al., 2014). In fig.1 we display trajectories of the system on the two model manifolds. The relative error of the estimator $\widehat{\phi}$ for OD on $\mathbb{S}^2$ is $1.894 \cdot 10^{-1} \pm 3.1 \cdot 10^{-4}$, whereas for OD on $\mathbb{PD}$ is $1.935 \cdot 10^{-1} \pm 9.5 \cdot 10^{-4}$, both are calculated using (5). The errors for trajectory prediction are reported in table 2.



(a) OD on $\mathbb{S}^2$ **(b)** OD on $\mathbb{PD}$



(c) OD on $\mathbb{S}^2$ **(d)** OD on $\mathbb{PD}$

Figure 1. **Top**: comparison of ϕ and $\widehat{\phi}$. The true interaction kernel is shown with a black solid line, whereas the mean estimated interaction kernel is shown with 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 $\rho_{T,\mathcal{M}}^{L}$ versus the empirical $\rho_{T,\mathcal{M}}^{L,M}$. **Bottom**: comparison of trajectories $\boldsymbol{X}_{[0,T]}$ and $\hat{\boldsymbol{X}}_{[0,T]}$. The trajectories $\boldsymbol{X}_{[0,T]}$ and $\hat{\boldsymbol{X}}_{[0,T]}$ are generated by the interaction kernel ϕ or $\widehat{\phi}$, respectively, with the same initial conditions. In the first row, trajectories are started from a randomly chosen initial condition. In the second row, trajectories are generated for a new system, with $N = 40$ agents. The colors along the trajectories indicate time, from deep blue (at $t = 0$) to light green (at $t = T$).

$\mathrm{Err}^{\mathbb{S}^2}_{1,1} = 2.98 \cdot 10^{-1} \pm 5.9 \cdot 10^{-3}$	$\mathrm{Err}^{\mathbb{S}^2}_{1,2} = 8.4 \cdot 10^{-3} \pm 3.0 \cdot 10^{-4}$
$\mathrm{Err}^{\mathbb{S}^2}_{2,1} = 2.5 \cdot 10^{-2} \pm 1.6 \cdot 10^{-3}$	$\mathrm{Err}^{\mathbb{S}^2}_{2,2} = 0$
$\mathrm{Err}^{\mathbb{PD}}_{1,1} = 9.0 \cdot 10^{-2} \pm 2.6 \cdot 10^{-3}$	$\mathrm{Err}^{\mathbb{PD}}_{1,2} = 1.34 \cdot 10^{-3} \pm 8.8 \cdot 10^{-5}$
$\mathrm{Err}^{\mathbb{PD}}_{2,1} = 3.6 \cdot 10^{-3} \pm 2.4 \cdot 10^{-4}$	$\mathrm{Err}^{\mathbb{PD}}_{2,2} = 0$

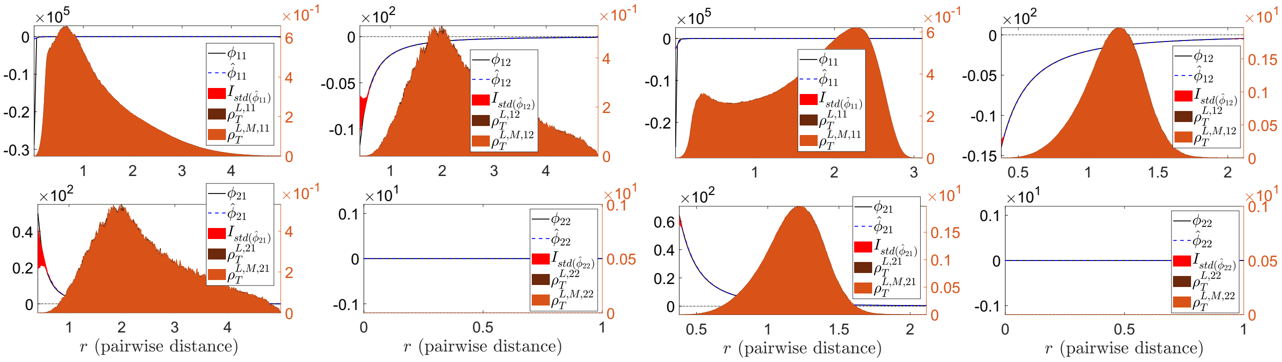
Table 3. (PS1 on $\mathbb{S}^2$ or $\mathbb{PD}$) Relative estimation errors for $\widehat{\phi}$.

Predator-Swarm System (PS1): this is a heterogeneous agent system, which is used to model interactions between multiple types of animals (Chen & Kolokolnikov, 2013; Olson et al., 2016). The learning theory presented in section 4 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). In this case, there are K^2 different interaction kernels, one $\phi_{k,k'}$ for each (directed) interaction between agents of type k and agents of type k'. In our example here there are two types, {prey, predator}, and therefore 4 interaction kernels; however there is only one predator, so the interaction kernel predator-predator is 0. The results are visualized in fig.2. The (relative) errors of the estimators are in table 3. The errors for trajectory prediction are reported in table 4.

PS1	$[0, T]$
$\mathrm{mean}^{\mathbb{S}^2}_{\mathrm{IC}}$: Training ICs	$2.36 \cdot 10^{-2} \pm 9.8 \cdot 10^{-4}$
$\mathrm{mean}^{\mathbb{S}^2}_{\mathrm{IC}}$: Random ICs	$2.40 \cdot 10^{-2} \pm 8.1 \cdot 10^{-4}$
$\mathrm{mean}^{\mathbb{PD}}_{\mathrm{IC}}$: Training ICs	$4.8 \cdot 10^{-3} \pm 1.2 \cdot 10^{-4}$
$\mathrm{mean}^{\mathbb{PD}}_{\mathrm{IC}}$: Random ICs	$4.8 \cdot 10^{-3} \pm 1.2 \cdot 10^{-4}$

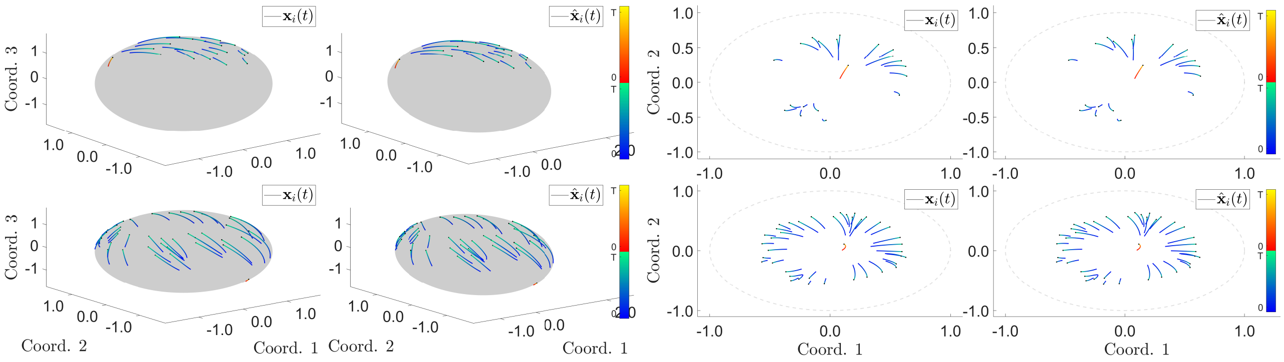
Table 4. As in table 2, but for the PS1 system.

Discussion: As shown in the figures and tables in this section, the estimators not only provide close approximation to their corresponding interaction kernels ϕ's, but also capture additional information about the true interaction laws, e.g. the support. The accuracy on the trajectories is consistent with the theory, and the lack of overfitting and the ability to generalize well to predicting trajectories started at new random initial conditions, which in general are very far from any of the initial conditions in the training data, given the



(a) PS1 on $\mathbb{S}^2$

(b) PS1 on $\mathbb{PD}$



(c) PS1 on $\mathbb{S}^2$

(d) PS1 on $\mathbb{PD}$

Figure 2. **Top**: comparison of $\phi_{k,k'}$ and $\widehat{\phi}_{k,k'}$. The true interaction kernels are shown with black solid lines, whereas the mean estimated interaction kernels are shown with blue dashed lines with their corresponding std interval regions shaded in red. Shown in the background is the comparison of the approximate $\rho_{T,\mathcal{M}}^{L,kk'}$ versus the empirical $\rho_{T,\mathcal{M}}^{L,M,kk'}$. 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. **Bottom**: comparison of trajectories $\boldsymbol{X}_{[0,T]}$ and $\hat{\boldsymbol{X}}_{[0,T]}$. The trajectories $\boldsymbol{X}_{[0,T]}$ and $\hat{\boldsymbol{X}}_{[0,T]}$ are generated by the interaction kernels, $\{\phi_{k,k'}\}_{k,k'=1}^K$ and $\{\widehat{\phi}_{k,k'}\}_{k,k'=1}^K$, respectively, with the same initial conditions. The two rows use a similar setup as in the OD case. The colors along the trajectories indicate time, from deep blue/bright red (at $t = 0$) to light green/light yellow (at $t = T$). The blue/green combo is assigned to the preys; whereas the red/yellow combo to the predator.

high-dimensionality of the state space, demonstrates the effectiveness of our approach. This is made possible because we have taken advantage of the symmetries in the system, in particular invariance of the governing equations under permutations of the agents (of the same type, in the case of heterogeneous agent systems, such as PS1), and radial symmetry of the interaction kernels. Further invariances, when the number of agents increases, make it possible to re-use the interaction kernel estimated on a system of N agents to predict trajectories of a system with the same interaction kernel, but a different number of agents, which of course has a state space of a different dimension. This simple example of transfer learning would not be possible for general-purpose techniques that directly estimate the r.h.s. of the system of ODEs.

6. Conclusion

We have considered the problem of estimating the dynamics of a particular yet widely used set of dynamical systems, consisting of interacting agents on Riemannian manifolds. These are driven by a first-order system of ODEs on the manifold, with a typically very high-dimensional state space $\mathcal{M}^N$, where N is the (typically large) number of agents. We constructed estimators that converge optimally and avoid the curse of dimensionality, by exploiting the multiple symmetries in these systems. Extensions to more complex systems of interacting agents may be considered, in particular to second-order systems, which will require the use of parallel transport on $\mathcal{M}$, to more general interaction kernels, depending on other variables beyond pairwise distances, as well as to systems interacting with a varying environment.

7. Acknowledgment

MM is grateful for partial support from NSF-1837991, NSF-1913243, NSF-1934979, NSF-Simons-2031985, FA9550-20-1-0288, ARO W911NF-18-C-0082, and to the Simons Foundation for the Simons Fellowship for the year '20-'21. Prisma Analytics, Inc. provided computing equipment and support. JM for support from NIH - T32GM11999.

MM and MZ designed the research; all authors jointly wrote the manuscript; HQ derived theoretical results together with JM and MZ; MZ developed algorithms and applications; JM and MZ analyzed the data.

8. Addressing Reviewers' Comments

We thank the reviewers for providing such detailed reviews and feedback on our paper. Due to the page limit, we are not be able to provide detailed responses to every comment; instead we address three groups of reviews briefly and highlight the most important issues and how we are addressing them. **To all reviewers**: we have fixed the typos, and made the corresponding cosmetic changes, including using vector graphics, repositioning figures and tables, etc. We have added a section, namely Sec. $D.1.$, in the Supplementary Information (SI) to discuss the computing platform used to run the simulations. The software package to reproduce the results shown in this paper will be made available online on GitHub (starting on June 10^{th}); and a link to the software package is also added in Sec. 3.3. We encourage the reviewers to check out Sec. D in SI for detailed discussion on how we set up the experiments and important learning results, as well as the computing time needed to run our experiments demonstrating the efficiency of our learning methods. Our paper strives to keep a delicate balance of theory and empirical findings.

To reviewers #5, #8, and #9: We have made the changes to comply with most of your comments in order to make the paper more accessible. We have already responded in our first response letters to the major issues and we sincerely appreciate the detailed reviews and feedback. We also encourage the reviewers to briefly go through the Sec. D in SI for a detailed background introduction of the different dynamical systems examined in the paper.

To reviewers #6, #7: We have gone through the introduction and hopefully cleared any possible confusion. We have also merged sections 3.3 and 3.4, and improved their clarity, so that the main idea of the computational complexity stands out. A more detailed description of computational complexity is now added as Sec. $C.1.$ in SI. The overall organization of the paper has been re-examined, and it has been improved for a cleaner presentation.

To Meta Review: we have gone through the paper and improved its overall organization, i.e. clean up the notations/organization/structure of our paper. As for baseline comparisons, we have pointed out in the introduction, as it had been already done in (Lu et al., 2019b), that most of the current methods (sparse approximation such as SINDy, neural network, etc.) have trouble dealing with the curse of dimensionality from the observation data, as they infer directly the right hand side of the ODE, $\dot{\boldsymbol{X}}_t = \boldsymbol{f}(\boldsymbol{X}_t)$. Our method, however, exploits the innate structure of the ODE systems (e.g. invariances and symmetries), hence our method is able to avoid the curse of dimensionality from the observation data, and perform transfer of learning readily. We have substantially improved notational clarity, and enhanced the readability for an ML venue.

References

Ahn, H., Ha, S.-Y., Park, H., and Shim, W. Emergent behaviors of Cucker-Smale flocks on the hyperboloid. 2020. URL `https://arxiv.org/abs/2007.02556`.

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`.

Bongini, M., Fornasier, M., Hansen, M., and Maggioni, M. Inferring interaction rules from observations of evolutive systems I: The variational approach. *Mathematical Models and Methos in Applied Sciences*, 27(5):909 – 951, 2017.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15): 3932–3937, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1517384113. URL `https://www.pnas.org/content/113/15/3932`.

C. Fetecau, R. and Zhang, B. Self-organization on Riemannian manifolds. *Journal of Geometric Mechanics*, 11(3):397–426, 2019. ISSN 1941-4897. doi: 10.3934/jgm.2019020. URL `http://dx.doi.org/10.3934/jgm.2019020`.

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, M., Liu, H., Liao, W., and Zhao, T. Doubly robust off-policy learning on low-dimensional manifolds by deep neural networks, 2020. URL `https://arxiv.org/abs/2011.01797`.

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

Cui, T., Marzouk, Y., and Willcox, K. Data-driven model reduction for the Bayesian solution of inverse problems. *International Journal for Numerical Methods in Engineering*, 102(5):966 – 990, 2014.

Hairer, E. Geometric integration of ordinary differential equations on manifolds. *BIT Numerical Mathematics*, 41 (5):996 – 1007, 2001.

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

Katz, Y., Tunstrom, K., Ioannou, C., Huepe, C., and Couzin, I. Inferring the structure and dynamics of interactions in schooling fish. *Proceedings of the National Academy of Sciences of the United States of America*, 108:18720–8725, 2011.

Keller, R. and Du, Q. Discovery of dynamics using linear multistep methods, 2019. URL `https://arxiv.org/abs/1912.12728`.

Kuramoto, Y. Lecture notes in physics. In *Internatonal Symposium on Mathematical Problems in Theoretical Physics*, pp. 420. Springer-Verlag, 1975.

Lee, T., Leok, M., and McClamroch, N. H. *Global Formations of Lagrangian and Hamiltonian Dynamics on Manifolds: A Geometric Approach to Modeling and Analysis*. Springer, 2018.

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.

Lukeman, R., Li, Y., and Edelstein-Keshet, L. Inferring individual rules from collective behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 107:12576 – 12580, 2010.

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`.

Monte-Alto, H. H. L. C., Morveli-Espinoza, M., and Tacla, C. A. Multi-agent systems based on contextual defeasible logic considering focus, 2020. URL `https://arxiv.org/abs/2010.00168`.

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`.

Raissi, M., Perdikaris, P., and Karniadakis, G. Multistep neural networks for data-driven discovery of nonlinear dynamical systems. *arXiv preprint arXiv:1801.01236*, 2018. URL `https://arxiv.org/abs/1801.01236`.

Riccio, F., Capobianco, R., and Nardi, D. DOP: deep optimistic planning with approximate value function evaluation. *CoRR*, abs/1803.08501, 2018. URL `http://arxiv.org/abs/1803.08501`.

Rudy, H., Kutz, N., and Brunton, S. Deep learning of dynamics and signal-noise decomposition with time-stepping constraints. *Journal of Computational Physics*, 2019.

Rudy, S., Brunton, S., Proctor, J., and Kutz, N. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017. doi: 10.1126/sciadv.1602614. URL `http://advances.sciencemag.org/content/3/4/e1602614`.

Sarlette, A. and Sepulchre, R. Consensus optimization on manifolds. *SIAM Journal on Control and Optimization*, 48, 12 2008. doi: 10.1137/060673400.

Schaeffer, H., Caflisch, R., Hauck, C., and Osher, S. Sparse dynamics for partial differential equations. *Proceedings of the National Academy of Sciences of the United States of America*, 110(17):6634–6639, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1302752110. URL `https://www.pnas.org/content/110/17/6634`.

Soize, C. and Ghanem, R. Probabilistic learning on manifolds constrained by nonlinear partial differential equations for small datasets, 2020. URL `https://arxiv.org/abs/2010.14324`.

Strogatz, S. H. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D*, 143:1 – 20, 2000.

Tran, G. and Ward, R. Exact recovery of chaotic systems from highly corrupted data. *Multiscale Modeling & Simulation*, 15(3):1108 – 1129, 2017.

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.

Wróbel, K., Torba, P., Paszyński, M., and Byrski, A. Evolutionary multi-agent computing in inverse problems. *Computer Science*, 14, 2013.

Yang, S., Wong, S. W. K., and Kou, S. C. Inference of dynamic systems from noisy and sparse data via manifold-constrained Gaussian processes, 2020. URL `https://arxiv.org/abs/2009.07444`.

Zhang, S. and Lin, G. Robust data-driven discovery of governing physical laws with error bars. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2217):20180305, 2018.

Zhong, M., Miller, J., and Maggioni, M. Data-driven discovery of emergent behaviors in collective dynamics. *Physica D: Nonlinear Phenomena*, pp. 132542, 2020. ISSN 0167-2789. doi: https://doi.org/10.1016/j.physd.2020.132542. URL `http://www.sciencedirect.com/science/article/pii/S0167278919308152`.

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

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

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

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

Here $\operatorname{Proj}_{\boldsymbol{u}}(\boldsymbol{w})$ 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 $\mathbb{PD} := \{\boldsymbol{x} \in \mathbb{R}^2 \text{ s.t. } \|\boldsymbol{x}\| < 1\}$, 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

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

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 $\boldsymbol{y} - \boldsymbol{x} - \operatorname{Proj}_{\boldsymbol{o}'-\boldsymbol{x}}(\boldsymbol{y} - \boldsymbol{x})$ in $\mathbb{R}^2$ (with the Euclidean metric), then identify it as a tangent vector in $T_{\boldsymbol{x}}\mathcal{M}$, and normalize it:

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

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^∞-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 φ 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 ϕ, give by

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

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).

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

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 ϕ, 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:*

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

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^∞ 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*

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

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 ϵ w.r.t the L^∞-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^∞ 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^∞; (ii) the geometric coercivity condition, (B.1), holds on $\cup_M \mathcal{H}_M$; (iii) $\inf\limits_{\varphi\in\mathcal{H}_M} \|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})} \overset{M\to\infty}{\longrightarrow} 0$, then*

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

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 ϕ of the system. Note that $\mathcal{H}$ can be embedded as a compact (in L^∞ 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 $\hat{\rho}^t_{\mathcal{M}}(r) = \frac{1}{N^2}\sum_{i,i'=1}^{N}\delta_{r_{ii',t}}(r)$. □

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

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*

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

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 $\mathcal{D}_{L,M,\mathcal{H}}(\varphi) \coloneqq \mathcal{E}_{L,M,\mathcal{M}}(\varphi) - \mathcal{E}_{L,M,\mathcal{M}}(\widehat{\phi}_{L,\infty,\mathcal{H}})$, and define

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

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 ϕ 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,*

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

for some constants $c_0, c_1, s > 0$. Suppose that the geometric coercivity condition holds on $\mathcal{L} \coloneqq \cup_n \mathcal{L}_n$ 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 ϕ and $\widehat{\phi}$, respectively, for $t \in [0,T]$, with $\hat{\boldsymbol{X}}_0 = \boldsymbol{X}_0$. Then the following inequalities hold:*

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

and

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

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, $\max_{r\in[0,R]} \left|\widehat{\phi}\right| \le S_0$, so that $\mathrm{Lip}(F^{\mathcal{M}}_{\widehat{\phi}}(\boldsymbol{x},\cdot))$, $\mathrm{Lip}(F^{\mathcal{M}}_{\widehat{\phi}}(\cdot,\boldsymbol{x})) \le 2S_0$. Therefore,

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

Putting these results together, we have

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

By Grönwall's inequality, we have

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

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

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

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

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

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

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

Now, take ϕ to be the true interaction kernel, and $\widehat{\phi}$ the estimator of ϕ 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_ρ 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 $[R_{\text{min}}^{\text{obs}}, R_{\text{max}}^{\text{obs}}]$, with $R_{\text{min}}^{\text{obs}}$ and $R_{\text{max}}^{\text{obs}}$ 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 ϕ 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 ± std. Visual comparisons of ϕ 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_ρ	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	∞

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 (R2020a) 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 ϕ and its estimator $\widehat{\phi}$ learned from the trajectory data.

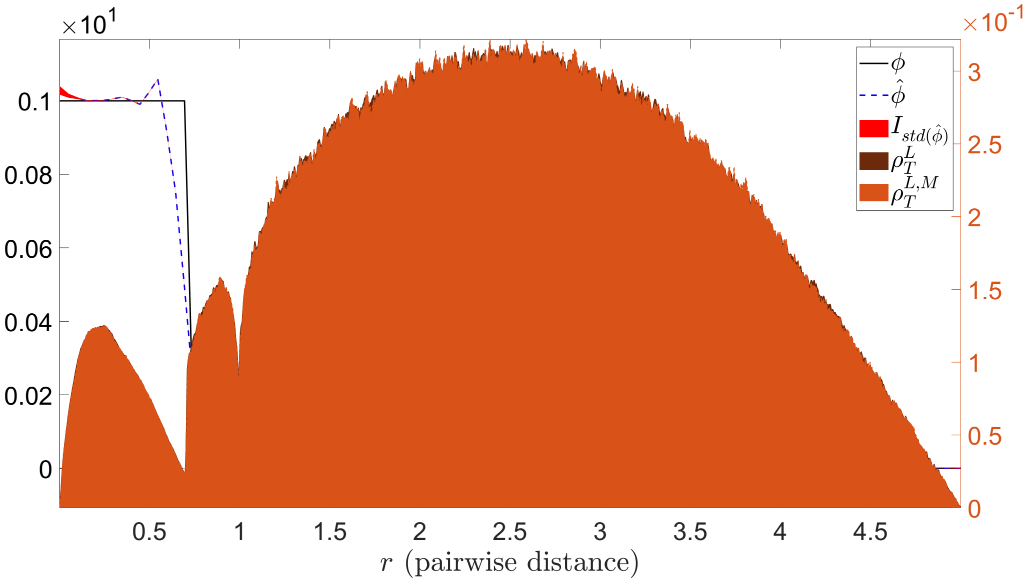


Figure 1. (OD on $\mathbb{S}^2$) Comparison of ϕ and $\widehat{\phi}$, with the relative error being $1.894 \cdot 10^{-1} \pm 3.1 \cdot 10^{-4}$ (calculated using (13)). The true interaction kernel is shown in a black solid line, whereas the mean estimated interaction kernel is shown in a blue dashed line with its std interval, i.e. mean($\widehat{\phi}$) ± 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 ϕ from the trajectory data. Fig. 2 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

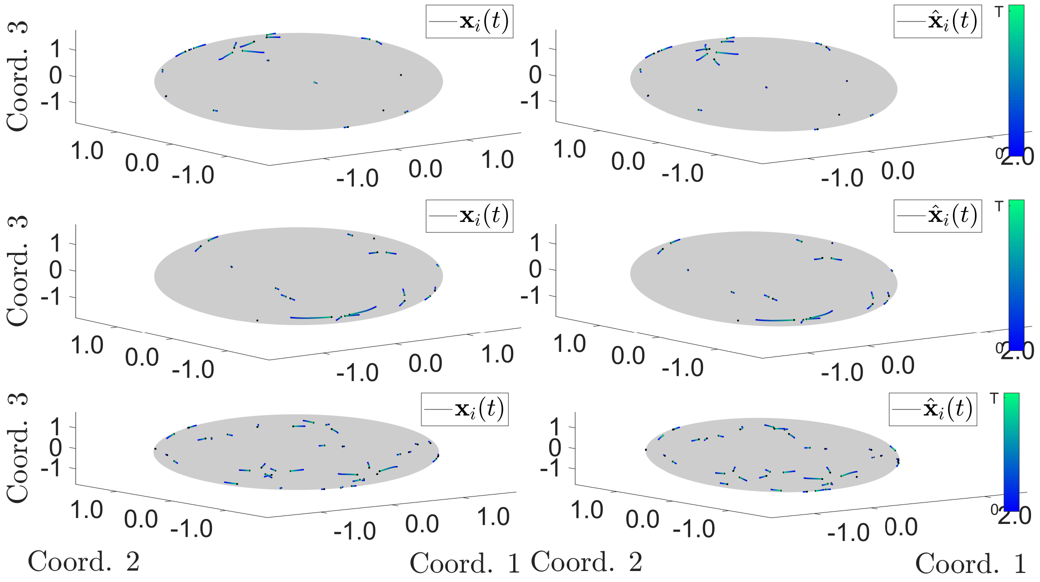


Figure 2. (OD on $\mathbb{S}^2$) Comparison of $\boldsymbol{X}$ (generated by ϕ) and $\hat{\boldsymbol{X}}$ (generated by $\widehat{\phi}$), with the errors reported in table 3. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **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.

	$[0, T]$
mean$_{\text{IC}}$: Training ICs	$8.8 \cdot 10^{-2} \pm 1.7 \cdot 10^{-3}$
std$_{\text{IC}}$: Training ICs	$5.9 \cdot 10^{-2} \pm 1.5 \cdot 10^{-3}$
mean$_{\text{IC}}$: Random ICs	$9.0 \cdot 10^{-2} \pm 1.6 \cdot 10^{-3}$
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). mean$_{\text{IC}}$ and 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 ϕ and its estimator $\hat{\phi}$ learned from the trajectory data.

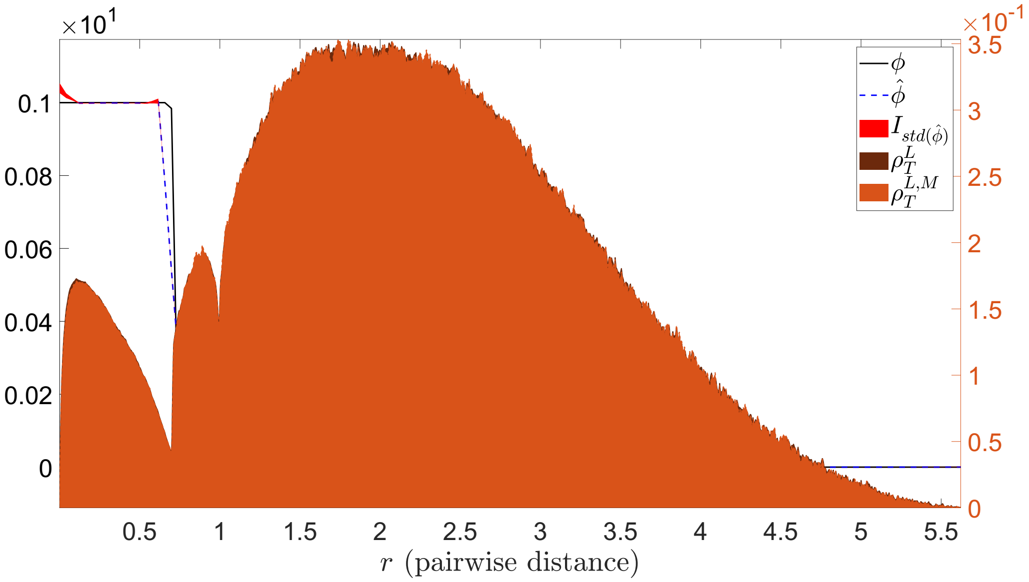


Figure 3. (OD on $\mathbb{PD}$) Comparison of ϕ and $\widehat{\phi}$, with the relative error being $2.114 \cdot 10^{-1} \pm 5.0 \cdot 10^{-4}$ (calculated using (13)). The true interaction kernel is shown in a black solid line, whereas the mean estimated interaction kernel is shown in a blue dashed line with its std interval, i.e. mean($\widehat{\phi}$) ± 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 ϕ from the trajectory data. Fig. 4 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

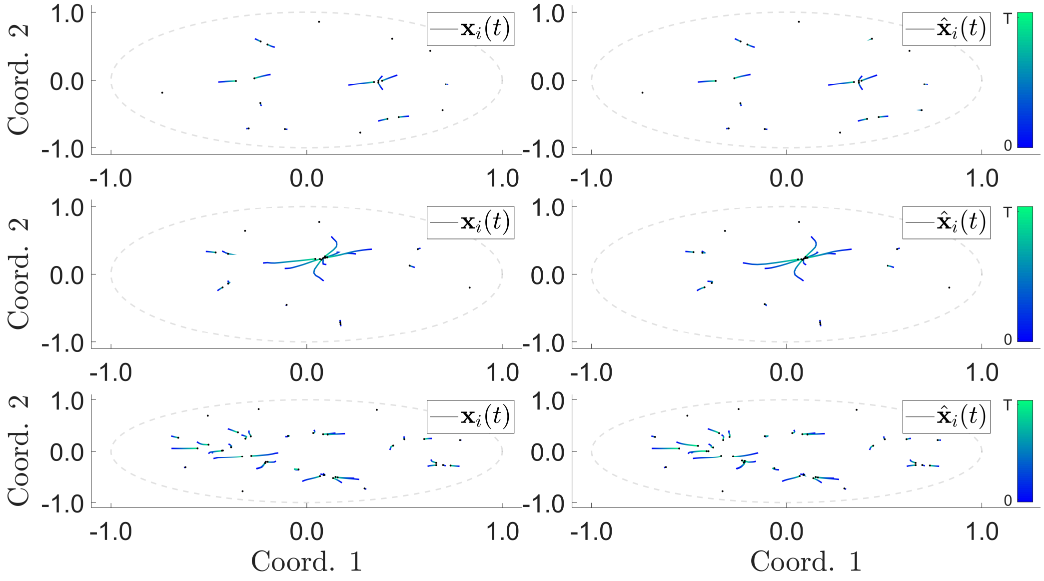


Figure 4. (OD on $\mathbb{PD}$) Comparison of $\boldsymbol{X}$ (generated by ϕ) and $\hat{\boldsymbol{X}}$ (generated by $\widehat{\phi}$), with the errors reported in table 5. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ 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 ϕ, 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.

	$[0, T]$
mean$_{\text{IC}}$: Training ICs	$2.53 \cdot 10^{-1} \pm 7.2 \cdot 10^{-3}$
std$_{\text{IC}}$: Training ICs	$1.90 \cdot 10^{-1} \pm 6.5 \cdot 10^{-3}$
mean$_{\text{IC}}$: Random ICs	$2.55 \cdot 10^{-1} \pm 9.7 \cdot 10^{-3}$
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). mean$_{\text{IC}}$ and 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 $\rho_{T,\mathcal{M}}^L$ 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_{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 ε is the depth of the potential well, σ 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 ϕ, 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 ϕ_{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 ϕ and its estimator $\widehat{\phi}$ learned from the trajectory data.

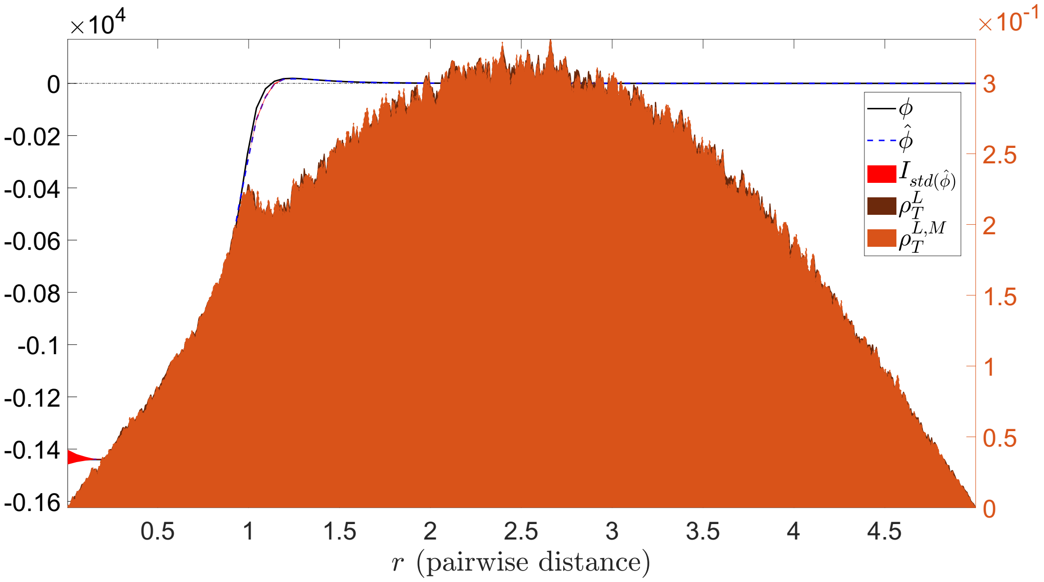


Figure 5. (LJD on $\mathbb{S}^2$) Comparison of ϕ and $\widehat{\phi}$, with the relative error being $3.65 \cdot 10^{-2} \pm 2.7 \cdot 10^{-4}$ (calculated using (13)). The true interaction kernel is shown in a black solid line, whereas the mean estimated interaction kernel is shown in a blue dashed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

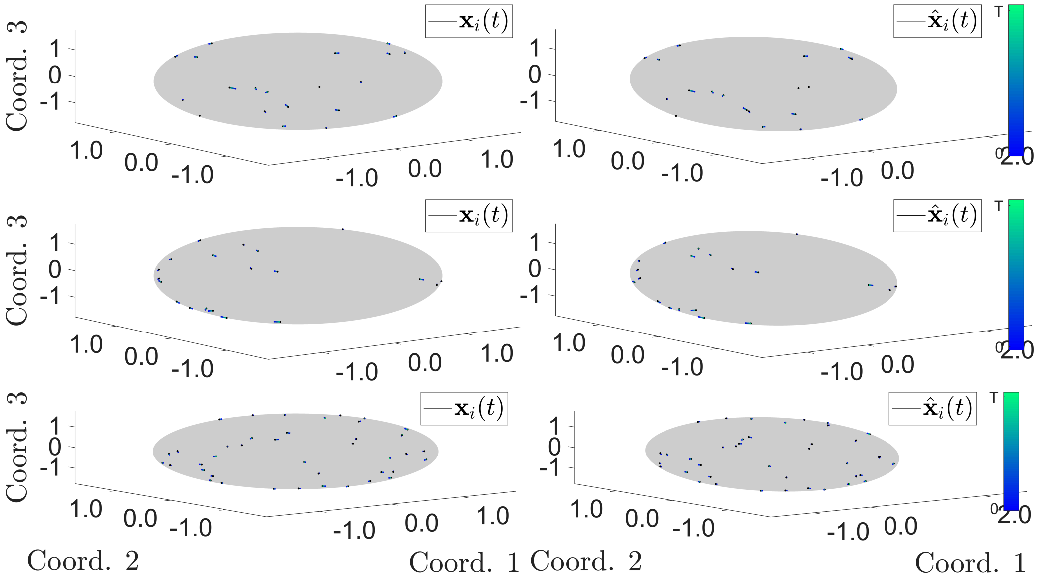


Figure 6. (LJD on $\mathbb{S}^2$) Comparison of $\boldsymbol{X}$ (generated by ϕ) and $\hat{\boldsymbol{X}}$ (generated by $\widehat{\phi}$), with the errors reported in table 8. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **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.

	$[0, T]$
mean_{IC}: Training ICs	$2.88 \cdot 10^{-3} \pm 2.5 \cdot 10^{-5}$
std_{IC}: Training ICs	$6.1 \cdot 10^{-4} \pm 1.8 \cdot 10^{-5}$
mean_{IC}: Random ICs	$2.88 \cdot 10^{-3} \pm 3.2 \cdot 10^{-5}$
std_{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 ϕ and its estimator $\widehat{\phi}$ learned from the trajectory data.

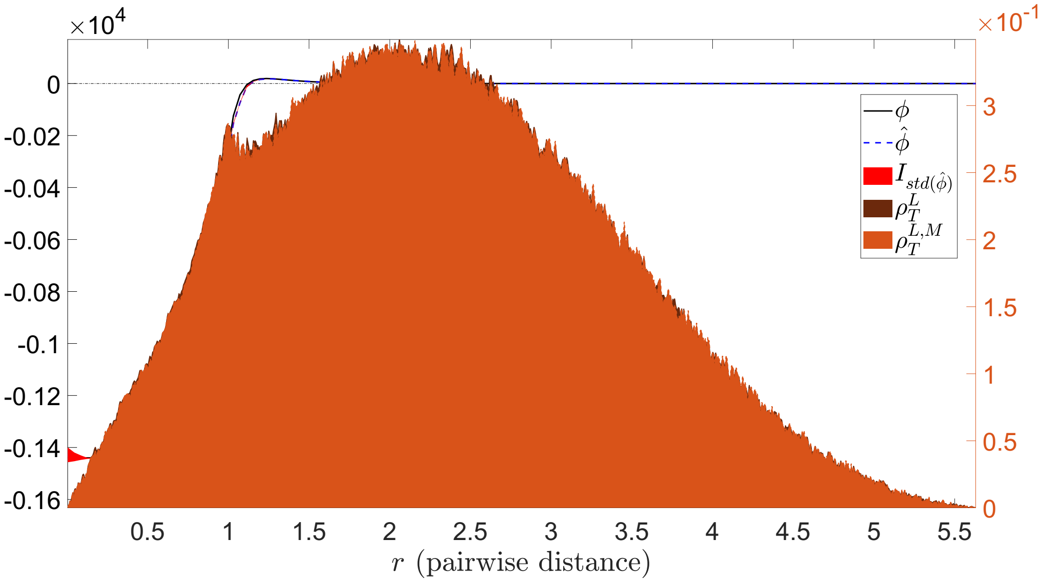


Figure 7. (LJD on $\mathbb{PD}$) Comparison of ϕ and $\widehat{\phi}$, with the relative error being $2.52 \cdot 10^{-2} \pm 3.6 \cdot 10^{-4}$ (calculated using (13)). The true interaction kernel is shown in a black solid line, whereas the mean estimated interaction kernel is shown in a blue dashed 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 ρ_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.

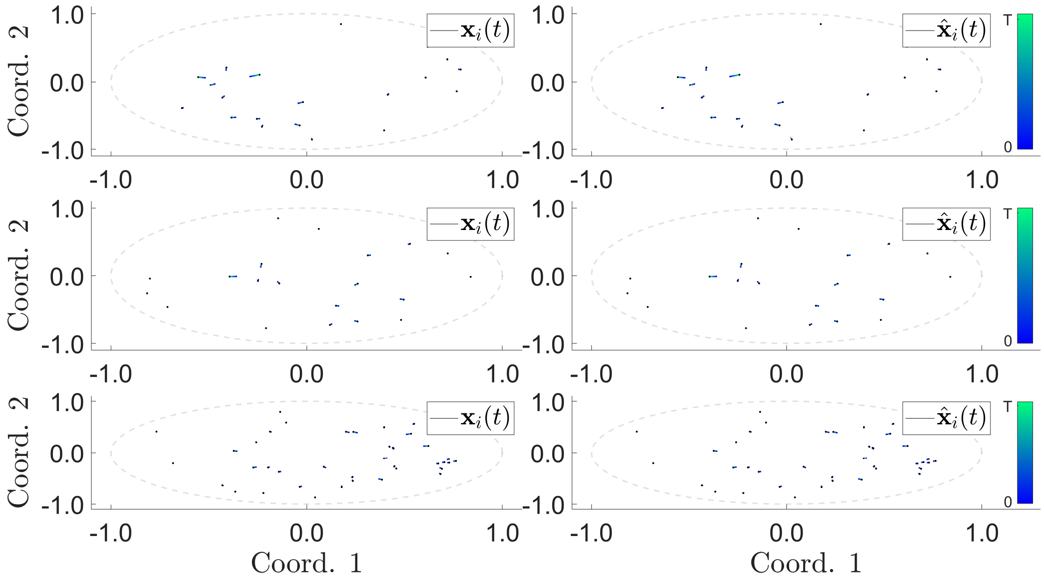


Figure 8. (LJD on $\mathbb{PD}$) Comparison of $\boldsymbol{X}$ (generated by ϕ) and $\hat{\boldsymbol{X}}$ (generated by $\widehat{\phi}$), with the errors reported in table 10. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **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.

	$[0, T]$
mean_{IC}: Training ICs	$2.27 \cdot 10^{-3} \pm 4.0 \cdot 10^{-5}$
std_{IC}: Training ICs	$5.6 \cdot 10^{-4} \pm 1.7 \cdot 10^{-5}$
mean_{IC}: Random ICs	$2.28 \cdot 10^{-3} \pm 3.8 \cdot 10^{-5}$
std_{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). mean_{IC} and std_{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.

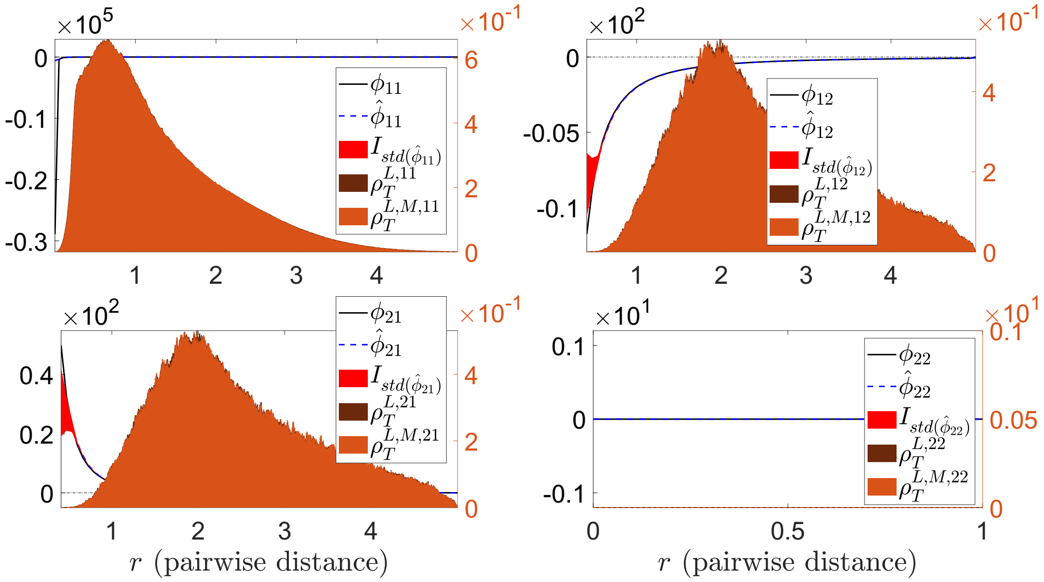


Figure 9. (PS1 on $\mathbb{S}^2$) Comparison of $\phi_{kk'}$ and $\widehat{\phi}_{k,k'}$, with the relative errors shown in table 17. The true interaction kernels are shown in black solid lines, whereas the mean estimated interaction kernel are shown in blue dashed lines with their corresponding std 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.

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

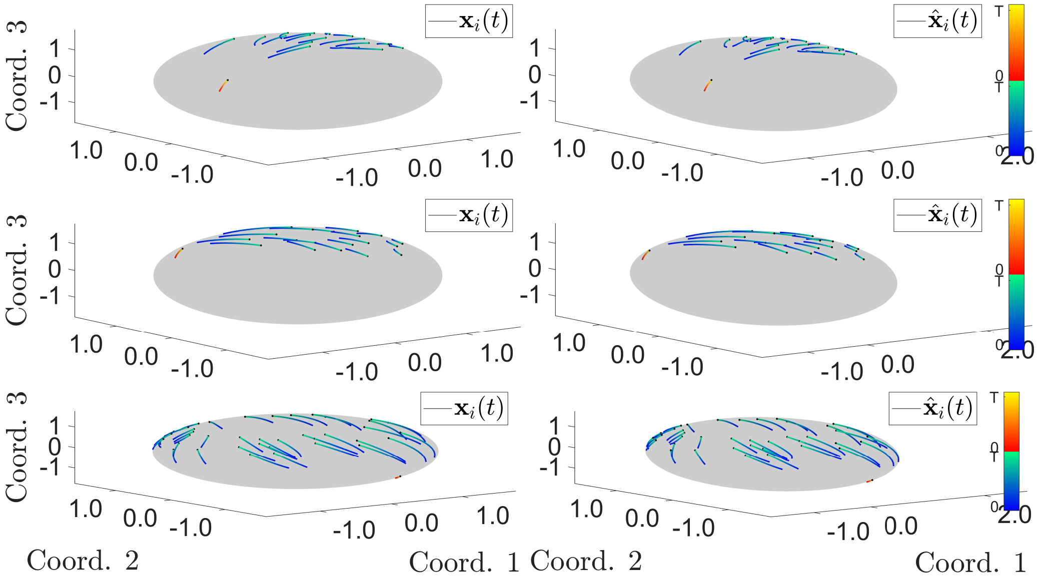


Figure 10. (PS1 on $\mathbb{S}^2$) Comparison of $\boldsymbol{X}$ (generated by $\phi_{k,k'}$'s) and $\hat{\boldsymbol{X}}$ (generated by $\widehat{\phi}_{k,k'}$'s), with the errors reported in table 14. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken 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.

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

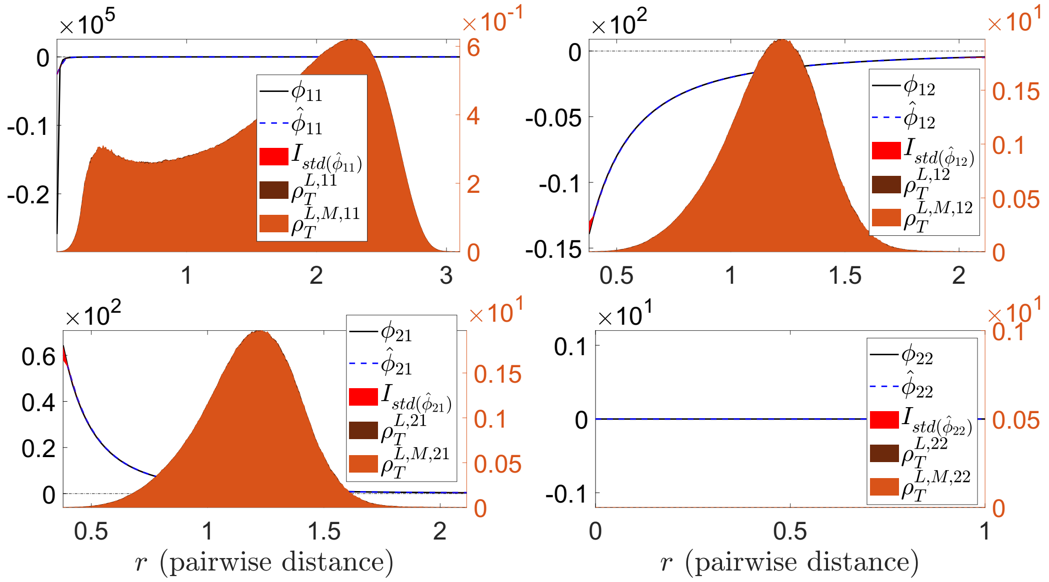


Figure 11. (PS1 on $\mathbb{PD}$) Comparison of $\phi_{kk'}$ and $\widehat{\phi}_{k,k'}$, with the relative errors shown in table 17. The true interaction kernels are shown in black solid lines, whereas the mean estimated interaction kernel are shown in blue dashed lines with their corresponding std 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.

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

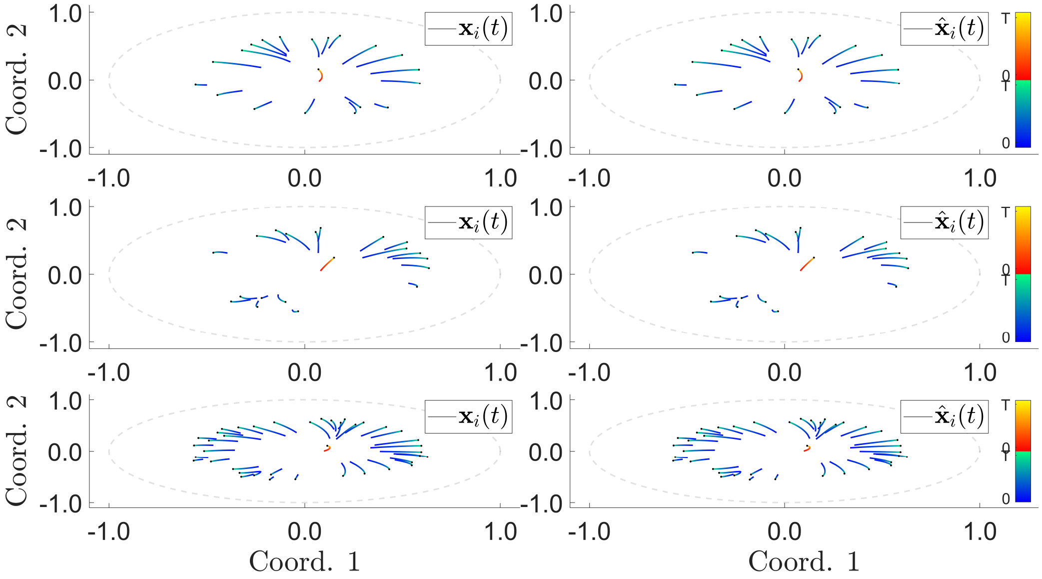


Figure 12. (PS1 on $\mathbb{PD}$) Comparison of $\boldsymbol{X}$ (generated by $\phi_{k,k'}$'s) and $\hat{\boldsymbol{X}}$ (generated by $\widehat{\phi}_{k,k'}$'s), with the errors reported in table 18. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken 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.

	$[0, T]$
mean_{IC}: Training ICs	$4.8 \cdot 10^{-3} \pm 1.2 \cdot 10^{-4}$
std_{IC}: Training ICs	$2.3 \cdot 10^{-3} \pm 3.0 \cdot 10^{-4}$
mean_{IC}: Random ICs	$4.8 \cdot 10^{-3} \pm 1.2 \cdot 10^{-4}$
std_{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). mean_{IC} and std_{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.