Learning theory for inferring interaction kernels in second-order interacting agent systems

Jason Miller$^{2,3}$ · Sui Tang$^{1,4}$ · Ming Zhong$^5$ · Mauro Maggioni$^{1,2,3}$

Received: 16 September 2022 / Accepted: 12 April 2023 / Published online: 13 June 2023
© The Author(s) 2023

Abstract
Modeling the complex interactions of systems of particles or agents is a fundamental problem across the sciences, from physics and biology, to economics and social sciences. In this work, we consider second-order, heterogeneous, multivariable models of interacting agents or particles, within simple environments. We describe a nonparametric inference framework to efficiently estimate the latent interaction kernels which drive these dynamical systems. We develop a learning theory which establishes strong consistency and optimal nonparametric min–max rates of convergence for the estimators, as well as provably accurate predicted trajectories. The optimal rates only depend on intrinsic dimension of interactions, which is typically much smaller than the ambient dimension. Our arguments are based on a coercivity condition which ensures that the interaction kernels can be estimated in stable fashion. The numerical algorithm presented to build the estimators is parallelizable, performs well on high-dimensional problems, and its performance is tested on a variety of complex dynamical systems.

Communicated by Mark Iwen.

✉ Sui Tang
suitang@math.ucsb.edu
Jason Miller
jason.miller@jhu.edu
Ming Zhong
mzhong3@iit.edu
Mauro Maggioni
mauromaggionijhu@icloud.com

1 Departments of Mathematics, University of California, Santa Barbara, Isla Vista, USA
2 Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, USA
3 Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA
4 University of California, Santa Barbara, 552 University Rd, Isla Vista, CA 93117, USA
5 Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616, USA
Keywords  Machine learning · Dynamical systems · Interacting particle systems · Agent-based dynamics · Inverse problems · Regularized least squares · Nonparametric statistics

Mathematics Subject Classification  62G08 · 70F17 · 65FXX

Contents

1 Introduction ............................................. 2
2 Model description .......................................... 5
3 Inference problem and learning approach .............................. 8
   3.1 Preliminaries and notation .................................. 8
   3.2 Problem setting ......................................... 9
   3.3 Loss functionals ......................................... 10
   3.4 Overview of contributions .................................... 10
   3.5 Comparison with existing work ................................. 11
   3.6 Function spaces ......................................... 12
4 Learning theory ........................................... 14
   4.1 Probability measures and weighted $L^2$ for measuring learning performance ......... 14
   Notational summary ......................................... 17
   4.2 Identifiability of kernels from data ........................................ 17
   4.3 Consistency and optimal convergence rate of estimators ..................................... 19
      4.3.1 Concentration ....................................... 19
      4.3.2 Consistency ........................................ 20
   4.4 Rate of convergence ....................................... 20
      4.4.1 Examples of convergence rates in prototype systems .................. 23
   4.5 Performance of trajectory prediction .............................. 24
5 Algorithm for constructing the interaction kernel estimators .................... 25
6 Applications ............................................. 26
   6.1 Learning results for anticipation dynamics with $U(r) = \frac{r^p}{p}$ .................. 28
7 Conclusion and further directions .................................. 29
Appendix A: Continuity of the error functionals ............................ 31
   Alignment and energy based kernels ........................................ 32
   Environment interaction kernels .......................................... 34
   A.1 Uniqueness of minimizers over a compact convex space .................... 35
   A.2 Uniform estimates on defect functions ..................................... 36
   A.3 Concentration .......................................... 37
Appendix B: Existence, uniqueness and properties of the measures ............... 44
   B.1 Well-posedness of second-order heterogeneous systems .................. 44
   B.2 Properties of measures ........................................ 45
Appendix C: Additional performance measures ............................ 46
Appendix D: Numerical algorithm ................................... 46
Appendix E: Control of trajectory error ................................ 50
References ................................................ 55

1 Introduction

Physical, biological, and social systems across all scales of complexity and size can often be described as dynamical systems written in terms of interacting agents (e.g. particles, cells, humans, planets,…). Rich theories have been developed to explain the collective behavior of these interacting agents across many fields including astronomy,
particle physics, economics, social science, and biology. Examples include predator–prey systems, molecular dynamics, coupled harmonic oscillators, flocking birds or milling fish, human social interactions, and celestial mechanics, to name a few. In order to encompass many of these examples, we will consider a rather general family of second-order, heterogeneous (the agents can be of different types), interacting (the acceleration of an agent is a function of properties of the other agents) agent systems that includes external forces, masses of the agents, multivariable interaction kernels, and an additional environment variable that is a dynamical property of the agent (for example, a firefly having its luminescence varying in time). We propose a learning approach that combines machine learning and dynamical systems in order to provide highly accurate dynamical models of the observation data from these systems.

The model and learning framework presented in Sects. 2–4 includes a very large number of relevant systems and allows for their modeling. Clustering of opinions [8, 21, 41, 53] is a simple first-order case that exhibits clustering. Flocking of birds [22, 23, 26] can be modeled as the behavior of a second-order system that exhibits an emergent shared velocity of all agents. Milling of fish [1, 2, 19, 20] may be modeled as a large-time behavior of a second-order system (in 2 or 3-dimensions), with a non-collective force from the environment. A model of oscillators (fireflies) that sync and swarm together, and have their dynamics governed by their positions and a phase variable $\xi$, was studied by [54–56, 66]. There are also models that include both energy and alignment interaction kernels, a particular case of this is the anticipation dynamics model from [65], which we also consider in this work. These dynamics exhibit a wide range of emergent behaviors, and as shown in [4, 18, 23, 35, 53, 67, 72], the behaviors can be studied when the governing equations are known. However, if the equations are not known and the data consists of only trajectories, we still wish to develop a model that can make accurate predictions of the trajectories and discover a dynamical form that accurately reflects their emergent properties. To achieve this, we present a provably optimal learning algorithm that is accurate, captures emergent behavior for large time, and, by exploiting the structure of the collective dynamical system, avoids the curse of dimensionality.

Our learning approach discovers the governing laws of a particular subset of dynamical systems of the form,

$$\dot{Y}(t) = F_{\phi^E A, \phi^\xi}(Y(t)), \quad Y(0) = Y_0 \in \mathbb{R}^D, \quad t \in [0, T].$$

The learning problem is to infer the right hand side function $F_{\phi^E A, \phi^\xi}$ from observations \(\{Y^{(m)}_{t_l}, \dot{Y}^{(m)}_{t_l}\}_{m=1,l=1}^{M,L}\) of the dynamical system, where \(m\) indexes different trajectories, started from initial conditions (IC) sampled i.i.d. from a measure $\mu^Y$ on the state space. Here \(M\) is the total number of trajectories observed, with each trajectory forming a single observation \(M\) plays a fundamental role in the learning theory where we study convergence as \(M\) varies; \(L\) refers to the number of observations at different times along each trajectory. Throughout this work, \(m\) will index the trajectories \(1, \ldots, M\) and \(l\) will index the points in time \(1, \ldots, L\). The main difficulties in establishing an effective theory of learning $F_{\phi^E A, \phi^\xi}$ are the curse of dimensionality caused by the dimension of \(Y\), which is \(D = N(2d + 1)\), where \(N\) is the number of agents, \(d\) the
dimension of physical space; and the dependence of the observation data, for example \( Y(t_{i+1}) \) is a deterministic function of \( Y(t_i) \).

We present a learning approach based on exploiting the structure of collective dynamical systems and nonparametric estimation techniques (see [6, 24, 31, 36, 69]) where we recover the interaction kernels \( \phi^{EA}, \phi^{A} \). A simplified form of our model equations, generalizing the first order models, is derived from Newton’s second law and given by: for \( i = 1, \ldots, N \)

\[
m_{i} \ddot{x}_{i}(t) = F^{X}(x_{i}, \dot{x}_{i}) + \frac{1}{N} \sum_{i' = 1}^{N} \phi^{E}(\|x_{i'}(t) - x_{i}(t)\|)(x_{i'}(t) - x_{i}(t))
\]

\[
\quad + \phi^{A}(\|x_{i'}(t) - x_{i}(t)\|)(\dot{x}_{i'}(t) - \dot{x}_{i}(t)).
\]  

(1.1)

Here, \( m_{i} \) is the mass of the \( i \)th agent, \( x_{i} \) is its position, \( F^{X} \) is a non-collective force, and \( \phi^{E}, \phi^{A} : \mathbb{R}^{+} \rightarrow \mathbb{R} \) are known as the interaction kernels.

To use the trajectory data to derive estimators, we consider appropriate hypothesis spaces in which to build our estimators, measures adapted to the dynamics, norms, and other performance metrics, and ultimately an inverse problem built from these tools. More specifically, let \( \hat{\phi}^{EA} \) denote the direct sum of the kernels \( \hat{\phi}^{E} \oplus \hat{\phi}^{A} \) (for the notation, see Sect. 3), and define our estimator as

\[
\hat{\phi}^{EA} := \arg \min_{\phi^{EA} \in \mathcal{H}^{EA}} \mathcal{E}_{M}^{EA}(\phi^{EA}),
\]

where \( \mathcal{E}_{M}^{EA} \) is an empirical error functional depending on the observation data, \( \mathcal{H}^{EA} \) is a hypothesis space to search for our estimators, and based on the form of the error functional the estimator is calculated as the solution of a constrained least squares problem. Once we have obtained this estimated interaction kernel, we want to study its properties as a function of the amount of trajectory data we receive, which is the \( M \) trajectories sampled from different initial conditions from the same underlying system, each consisting of \( L \) time observations along the trajectory. Here we study properties of the error functional, establish the uniqueness of its minimizers, and use the probability measures to define a dynamics-adapted norm to measure the error of our estimators over the hypothesis spaces. In comparing the estimators to the true interaction kernels, we first establish concentration estimates over the hypothesis space.

Our first main result is the strong asymptotic consistency of our learned estimators, as the number \( M \) of trajectories increases, which for the model (1.1) yields:

\[
\lim_{M \to \infty} \| \hat{\phi}^{EA} - \phi^{EA} \|_{L^{2}(\rho_{T}^{E,A,L})} = 0 \quad \text{with probability one,}
\]

(1.2)

where \( \rho_{T}^{E,A,L} \) is a dynamics-adapted probability measure on pairwise distances, and we use a weighted \( L^{2} \) space (see Sect. 4, particularly (4.3)); see Sect. 3 for the required definitions and Sect. 4.3.2 for the full theorem. In fact, we also prove a stronger result that provides the rate of convergence. We achieve the minimax rate of convergence.
for any number of effective variables $\nu$ in the interaction kernels. See Sect. 4.4 for the full theorem (see Sect. 3 for relevant definitions) which is given by:

$$\mathbb{E}_{\mu} \left[ \| \hat{\phi}^{EA} - \phi^{EA} \|_{L^2(\rho_T^{EA,I})}^2 \right] \leq C \left( \frac{\log M}{M} \right)^{\frac{2\nu}{s+\nu}}. \tag{1.3}$$

In the case of model (1.1), we have $\nu = 1$. Our result recovers the results for first-order systems [45, 47].

This means that our estimators converge at the same rate in $M$ as the best possible estimator (up to a logarithmic factor) one could construct when the initial conditions are randomly sampled from some underlying initial condition distribution denoted $\mu^Y$ throughout this work (see Sect. 4.3).

To solve the inverse problem, we give a detailed discussion of an essential link between these three aspects, the notion of coercivity of the system—detailed in Sect. 4.2. Coercivity plays a key role in the approximation properties, the algorithm design, and the learning theory. We also present numerical examples, see also the detailed numerical study in [75], which help to explain why the particular norms we define are the right choice, as well as show excellent performance on complex dynamical systems, in Sect. 6.

Our paper is structured as follows. The first part of the paper describes the model, learning framework, inference problem, and the basic tools needed for the learning theory. These ideas are all explained in detail in Sects. 2–4. If one wishes to quickly jump to the theoretical sections, and then refer back to the definitions as needed, we have provided Tables 1, 3 which explains the model equations and outlines the definitions and concepts needed for the learning theory and general theoretical results, respectively. The theoretical part of the paper (Sects. 4.2–4.5) discusses fundamental questions of identifiability and solvability of the inverse problem, consistency, and rate of convergence of the estimators, and the ability to control trajectory error of the evolved trajectories using our estimators. Some key highlights of our theoretical contributions are described in Sect. 3.4, with full details in the corresponding sections. Lastly, we consider applications in Sect. 6, as well as have many additional proofs and details in appendices E-D.

### 2 Model description

In order to motivate the choice of second-order models considered in this paper, we begin our discussion with a simple second-order model derived from classical mechanics. Let us consider a closed system of $N$ homogeneous agents (or particles) equipped with a certain type of Lagrangian $L(t)$ in the form

$$L(t) = \frac{1}{2} \sum_{i=1}^{N} m_i \| \dot{x}_i(t) \|^2 - \frac{1}{2N} \sum_{i,i'=1}^{N} U(\| x_{i'}(t) - x_i(t) \|), \quad i = 1, \ldots, N.$$
Here $U$ is a potential energy depending on pairwise distance. From the Lagrange equation, $\frac{d}{dt} \partial_t L = \partial_x L$, we obtain the second-order collective dynamics model

$$m_i \ddot{x}_i(t) = \frac{1}{N} \sum_{i'=1}^{N} \phi^E(||x_{i'}(t) - x_i(t)||)(x_{i'}(t) - x_i(t)), \quad i = 1, \ldots, N. \quad (2.1)$$

Here, $\phi^E(r) = \frac{U'(r)}{r}$ represents an energy-based interaction between agents. We are assuming a regularity condition on $\phi^E$, i.e. $\phi^E(0)0 = 0$. For example, the choice $U(r) = \frac{N G m_i m_j}{r}$ corresponds to Newton’s gravity model.

In order to incorporate a wider spectrum of behaviors, we add alignment-based interactions, which enable the alignment of velocities (so that short-range repulsion, mid-range alignment, and long range attraction are all present), auxiliary state variables describing internal states of agents (emotion, excitation, phases, etc.), and non-collective forces (interaction with the environment). We also allow for heterogeneous systems, consisting of agents belonging to $K$ disjoint types: in this case we partition agents in $K$ disjoint subsets $\{C_k\}_{k=1}^{K}$, with $N_k$ being the number of agents of type $k$, grouped in the index subset $C_k$. These systems will be modeled by equations of the form

$$\begin{align*}
\left\{ \begin{array}{l}
m_i \ddot{x}_i(t) = F^E(x_i(t), \dot{x}_i(t), \xi_i(t)) + \sum_{i' = 1}^{N} \frac{1}{N_{k_i}} \left[ \phi^E_{k_i k_{i'}}(r_{ii'}(t), s_{ii'}^E(t))(x_{i'}(t) - x_i(t)) \\
+ \phi^A_{k_i k_{i'}}(r_{ii'}(t), s_{ii'}^A(t))(\dot{x}_{i'}(t) - \dot{x}_i(t)) \right] \\
\dot{\xi}_i(t) = F^\xi(x_i(t), \dot{x}_i(t), \xi_i(t)) + \sum_{i' = 1}^{N} \frac{1}{N_{k_i}} \phi^\xi_{k_i k_{i'}}(r_{ii'}(t), s_{ii'}^\xi(t))(\xi_{i'}(t) - \xi_i(t)) \end{array} \right. 
\end{align*}$$

for $i = 1, \ldots, N$, where $k_i \in \{1, \ldots, K\}$ is the type of the agent $i$. The interaction kernels $\phi^E_{kk'}, \phi^A_{kk'}, \phi^\xi_{kk'}$ are in general different for each directed pair of interacting agent types; moreover, they not only depend on the pairwise distance $r_{ii'}(t) = ||x_{i'}(t) - x_i(t)||$, but also on other (known) pairwise features, $s^E_{ii'}, s^A_{ii'}, s^\xi_{ii'}$, which are vector-valued functions of $x_i(t), \dot{x}_i(t), \xi_i(t), x_{i'}(t), \dot{x}_{i'}(t), \xi_{i'}(t)$. For example, the interactions between birds or fish may depend on the field of vision, not just the distance between pairs of birds or fish. We will often suppress the explicit dependence on time $t$ when it is clear from the context. These feature maps are modeled as being the composition of a vector-valued feature map $F$ shared among all types of agent, and a projection, for each ordered pair of agent types, onto a subset of coordinates in the range of this map—see Table 1. The unknowns, for which we will construct estimators, in these equations, are the functions $\phi^E_{k_i k_{i'}}, \phi^A_{k_i k_{i'}}$ and $\phi^\xi_{k_i k_{i'}}$; everything else is assumed given.

We note that in what follows, the notation $\{E, A, \xi\}$ attached to a map/function/etc. means that there is one of those maps/functions/etc. for each element in the set $\{E, A, \xi\}$. It is a convenient way to avoid excessive repetition of similar definitions.

The specific instances of the feature map $F$ together with corresponding projections $\pi_{kk'}^{E,A,\xi}$ include a variety of systems that have found a wide range of applications in
Table 1 Notation for the variables in (2.2)

| Variable | Definition |
|----------|------------|
| $i, i'$  | Index of agent, from $1, \ldots, N$ |
| $m_i$    | Mass of agent $i$ |
| $x_i(t), \dot{x}_i(t), \ddot{x}_i(t) \in \mathbb{R}^d$ | Position/velocity/acceleration vector of agent $i$ at time $t$ |
| $\xi_i, \dot{\xi}_i$ | Auxiliary variable, and its derivative |
| $\|\cdot\|_E$ | Euclidean norm in $\mathbb{R}^d$ |
| $K$      | Number of agent types |
| $k, k'$  | Index for agent types, ranging in $\{1, \ldots, K\}$ |
| $N_k$    | Number of agents in type $k$ |
| $N$      | The total number of agents and $N = \sum_{i=1}^K N_k$. |
| $\hat{k}_i$ | Element of $\{1, \ldots, K\}$ indicating the type of agent $i$ |
| $C_k$    | Subset of $\{1, \ldots, N\}$ consisting of indices of the agents of type $k$ |
| $\phi_{kk'}$ | Influence of any agent of type $k$ onto any agent of type $k'$ |
| $F^\dot{x}, F^\xi$ | Non-collective forces affecting $\dot{x}_i$ and $\dot{\xi}_i$, respectively |
| $\phi_{E}^E, \phi_A^A, \phi^\xi$ | Energy, alignment, and environment-based interaction kernels respectively |
| $\mathcal{F}$ | Common Feature Map (CFM), $\mathcal{F}(x, \dot{x}, \xi, x', \dot{x}') : \mathbb{R}^{4d+2} \to \mathbb{R}^p$ |
| $\pi_{kk'}^{[E, A, \xi]}$ | Projection map onto subset of features from the range of the CFM |
| $s_{kk'}^{[E, A, \xi]}$ | Feature map, $\pi_{kk'}^{[E, A, \xi]} \circ \mathcal{F}(x, \dot{x}, \xi, x', \dot{x}') : \mathbb{R}^{4d+2} \to \mathbb{R}^{p(k, k')}$ |
| $s_{ii'}^{[E, A, \xi]}(t)$ | Feature evaluation, $s_{ii'}^{[E, A, \xi]}(x_i(t), \dot{x}_i(t), \xi_i(t), x_{i'}(t), \dot{x}_{i'}(t), \xi_{i'}(t)) \in \mathbb{R}^{p(k, k')}$ |

physics, biology, ecology, and social science; see the examples in the Table 2 below. We assume that the function $\mathcal{F}$ is Lipschitz, and known, together with all the $\pi_{kk'}^{[E, A, \xi]}$s. The Lipschitz assumption is sufficient to ensure the well-posedness of the system and will also be used to control the trajectory error, and of course implies that the feature maps $s_{kk'}^{[E, A, \xi]}$ are all Lipschitz. The function $\mathcal{F}$ is a uniform way to collect all of the different variables (functions of the inputs) used across any of the $(k, k')$ pairs over all of the $E, A, \xi$ functions in the system. This uniformity is helpful when discussing the rate of convergence, among other places. Examples of where this generality matters emerge naturally, say when one has a different number of variables across interaction kernels for different pairs $(k, k')$, or when the energy and alignment kernels depend on $r$ and then additional but distinct other variables. From this uniform set of variables, we then project to arrive at the relevant function $s_{kk'}^{[E, A, \xi]}$ for each pair (and each of the elements of the wildcard). Last, we can then evaluate this map at the specific pair of agents $(i, i')$, that leads to the feature evaluation, $s_{ii'}^{[E, A, \xi]}$ which is the expression used in the model equation (2.2).

The model class (2.2) is quite large. We will consider several different concrete example in Sect. 6.1. We summarize how those examples, and others, map to the model class in Table 2, with a shaded (respectively: empty) cell indicating that the
Table 2  Summary of the models studied in this work and in [45, 47, 49, 75]. The black boxes show the existence of the terms

| Model                                           | Properties |
|-------------------------------------------------|------------|
| Anticipation Dynamics [71]                      | φE, φA, m1, F^α, φ^ξ, K > 1, s^E, s^A, s^ξ, V, V^ξ |
| Celestial Mechanics [53]                        | 1          |
| Cucker-Smale [29]                               | 1          |
| Fish Milling 2D [19]                            | 1          |
| Fish Milling 3D [19]                            | 1          |
| Flocking w. Ext. Poten. [71]                    | 1          |
| Phototaxis [40]                                 | 1          |
| Predator-Swarm (2nd Order) [17]                 | 1          |
| Lennard-Jones [51]                              | 1          |
| Opinion Dynamics [58]                           | 1          |
| Predator-Swarm (1st Order) [17]                 | 1          |
| Synchronized Oscillator [81]                    | 2, 2       |

model has (respectively: has not) that characteristic. A numeric value indicates this is the number of unique variables, V, V^ξ used within the EA or ξ portions of the system. The number of these unique variables specifies the dimension in the minimax convergence rate, see Sect. 4.4.

3 Inference problem and learning approach

In this section, we first introduce the problem of inferring the interaction kernels from observations of trajectory data and give a brief review and generalization of the learning approach proposed in the works [47] and [75].

3.1 Preliminaries and notation

We vectorize the model in (2.2) in order to obtain a more compact description. We let v_i(t) :=  \dot{x}_i(t)  and

\[
X_t := \begin{bmatrix} x_1(t) \\ \vdots \\ x_N(t) \end{bmatrix} \in \mathbb{R}^{Nd}, \quad V_t := \begin{bmatrix} v_1(t) \\ \vdots \\ v_N(t) \end{bmatrix} \in \mathbb{R}^{Nd}, \quad \Xi_t := \begin{bmatrix} \xi_1(t) \\ \vdots \\ \xi_N(t) \end{bmatrix} \in \mathbb{R}^N.
\]

We introduce the weighted norm

\[
\|Z\|_S^2 := \sum_{i=1}^{N} \frac{1}{N_k_i} \|z_i\|^2
\]

for Z = [z_1^T, \ldots, z_N^T]^T with each z_i ∈ \mathbb{R}^d or \mathbb{R}. Here \|·\| is the same norm used in the construction of pairwise distance data for the interaction kernels (typically,
the Euclidean norm. The weight factor $1/N_k$ is introduced so that different types of agents of different types are overall weighted equally, which is important in the estimation phase, especially in the case when the number of agents of different types is highly non-uniform. The model (2.2) becomes

$$
\begin{align*}
\ddot{m} \circ \dot{X}_t &= f^{\text{nc}, \dot{x}}(X_t, V_t, \Xi_t) + f^{\phi E}(X_t, V_t, \Xi_t) + f^{\phi A}(X_t, V_t, \Xi_t) \\
\dot{\Xi}_t &= f^{\text{nc}, \dot{\Xi}}(X_t, V_t, \Xi_t) + f^{\phi E}(X_t, V_t, \Xi_t).
\end{align*}
$$

Here $\vec{m} = [m_1, \ldots, m_N]^T \in \mathbb{R}^N$, $\circ$ is the Hadamard product, and we use boldface fonts to denote the vectorized form of our estimators (with some once-for-all-fixed ordering of the pairs $(k, k')$, $k, k' = 1, \ldots, K$):

$$
\phi^E = [\phi^E_{kk'}]_{k,k'=1}^K, \quad \phi^A = [\phi^A_{kk'}]_{k,k'=1}^K, \quad \phi^\xi = [\phi^\xi_{kk'}]_{k,k'=1}^K,
$$

and of the non-collective force:

$$
\begin{align*}
f^{\text{nc}, \dot{x}}(X, V, \Xi) := \left[ F^x(x_i, \dot{x}_i, \xi_i) \right]_{i=1,\ldots,N}^T, \\
f^{\phi E} := \left[ \sum_{i'=1}^N \frac{1}{N_k} \phi^E_{kk'}(r_{ii'}, s^E_{ii'})(x_{i'} - x_i) \right]_{i=1,\ldots,N}^T,
\end{align*}
$$

both of which are vectors in $\mathbb{R}^{Nd}$. We omit the analogous definitions for $f^{\phi A}$ and $f^{\phi \xi}$. We also use the shorthand:

$$
\phi^{EA} := \phi^E \oplus \phi^A,
$$

(3.3)

to denote the element of the direct sum of the function spaces containing $\phi^E, \phi^A$.

### 3.2 Problem setting

Our observation data is given by approximations of $\{Y^{(m)}(t), \dot{Y}^{(m)}(t)\}_{m=1,\ldots,M}$ for $0 = t_1 < t_2 < \cdots < t_L = T$. Here $Y_t = [y_1^T(t), \ldots, y_N^T(t)]$ and $y_i(t) = [x_i^T(t), \dot{x}_i^T(t), \xi_i(t)]^T$, and $m$ indexes the $M$ different trajectories, each generated by the system (2.1) with the unknown set of interaction kernels, i.e. $\phi^E, \phi^A, \phi^\xi$, with initial conditions $\{Y^{(m)}(0)\}_{m=1,\ldots,M}$ drawn i.i.d from $\mu^Y$, a probability measure defined on the space $\mathbb{R}^{N(2d+2)}$. We use a superscript $(m)$ to denote that the variable is calculated from the data from that $m^{th}$ trajectory. The objective is to construct estimators $\hat{\phi}^E, \hat{\phi}^A, \hat{\phi}^\xi$ the unknown interaction kernels given these observations.
3.3 Loss functionals

For simplicity, we only consider equidistant observation points: \( t_l - t_{l-1} = h \) for \( l = 2, \ldots, L \); the proposed estimator is easily extended to the case non-equispaced time points. Following and extending [45, 47, 75], we consider the empirical error functional (recall the shorthand notation (3.3))

\[
E^E_M(\phi^E) := \frac{1}{L M} \sum_{l=1, m=1}^{L, M} \| \ddot{X}^{(m)}_{t_l} - f^{nc, \dot{x}}(X^{(m)}_{t_l}, V^{(m)}_{t_l}, \Xi^{(m)}_{t_l}) \\
- f^{\phi^E}(X^{(m)}_{t_l}, V^{(m)}_{t_l}, \Xi^{(m)}_{t_l}) \|_S^2,
\]

\[
E^\xi_M(\phi^\xi) := \frac{1}{L M} \sum_{l=1, m=1}^{L, M} \| \ddot{\Xi}^{(m)}_{t_l} - f^{nc, \dot{\Xi}}(X^{(m)}_{t_l}, V^{(m)}_{t_l}, \Xi^{(m)}_{t_l}) \\
- f^{\phi^\xi}(X^{(m)}_{t_l}, V^{(m)}_{t_l}, \Xi^{(m)}_{t_l}) \|_S^2,
\]

(3.4)

where we recall \( \| \cdot \|_S \) in (3.1) and the introduction of this norm is to balance the contributions from different species. The estimators of interaction kernels are defined as the minimizers of the error functionals \( E^E_M \) and \( E^\xi_M \) over suitably chosen finite-dimensional function spaces \( \mathcal{H}^E \) and \( \mathcal{H}^\xi \):

\[
\hat{\phi}^E = \arg \min_{\phi^E \in \mathcal{H}^E} E^E_M(\phi^E), \quad \hat{\phi}^\xi = \arg \min_{\phi^\xi \in \mathcal{H}^\xi} E^\xi_M(\phi^\xi).
\]

(3.5)

3.4 Overview of contributions

We focus on the regime where \( L \) is fixed but \( M \to \infty \). We provide a learning theory that answers the fundamental questions:

- **Quantitative description of estimator errors.** We will introduce measures to describe how close the estimators are to the true interaction kernels, that lead to novel dynamics-adapted norms. See Sect. 4.
- **Identifiability of kernels.** We will establish the existence and uniqueness of the estimators as well as relate the solvability of our inverse problem to a fundamental coercivity property. See Sect. 4.2.
- **Consistency and optimal convergence rate of the estimators.** We prove theorems on the strong consistency and optimal minimax rates of convergence of the estimators, as the number of observe trajectories increases. These results exploit the separability of the learning on the energy and alignment from the learning on the environment variable. See Sect. 4.3.
- **Trajectory Prediction.** We prove a theorem that bounds the expected (over initial conditions) supremum error (over the entire time interval) of the trajectories obtained via the estimated interaction kernel is controlled by the norm of the difference between the true and estimated kernels, which in turn is controlled by the result above. This further justifies our choice of norms and estimation procedure. See Sect. 4.5.
Applications. Our generalized model demonstrates unique features that were not explored in previous work on estimating interaction kernels. Specifically, we showcase applications of the anticipation dynamics (AD) model introduced in [65] that go beyond the scope of previous research. Our numerical results support the effectiveness of our method and are in line with our theoretical findings. In Sect. 5 and D we discuss computation complexity and algorithmic implementation.

3.5 Comparison with existing work

Nonparametric inference of radial interaction kernels. We have studied the nonparametric inference of radial interaction kernels from trajectory data in special cases of model 2.2. In [10], a convergence study for first-order models of homogeneous agents was done for increasing \( N \), the number of agents. The estimation problem with \( N \) fixed, but the number of trajectories \( M \) varying, for first-order and second-order models of heterogeneous agents was numerically studied in [47] and learning theory on these first-order models was developed in [32, 42, 45]. A big data application to real celestial motion ephemerides is developed and discussed in [49]. In [75], we numerically examine the proposed inference approach on second-order systems, with particular emphasis on emergent collective behaviors.

Novelty of our work. In this paper, we provide a rigorous learning theory for second-order heterogeneous models 2.2 with interaction kernels depending on pairwise higher-dimensional features. Our second-order model equations cover the first-order models considered in [32, 45, 47, 75] as special cases, but they are a significantly larger class: even when written as a first-order system in more variables, they are a strict generalization of the previous first-order models. Furthermore, the dynamical characteristics produced by second-order models are much richer and can model more complicated collective motions and emergent behavior of the agents.

Our theoretical results also focus on the joint learning of \( \phi^E, \phi^A \) that takes into account their natural weighted direct sum structure that is described in the following sections, whereas the previous work’s learning theories are on single \( \phi^E \)'s. We carefully discuss the identifiability and separability of \( \phi^E \) and \( \phi^A \) from the sum. In general, the current theoretical framework is not able to conclusively show that \( \phi^E \) and \( \phi^A \) can be learned separately. However, we show a structured sum of \( \phi^E \) and \( \phi^A \) can be learned at an optimal rate only depending on the intrinsic dimension, which is sufficient to conclude the force field on the whole state space \( \mathbb{R}^{2Nd+N} \) can be learned without curse of ambient dimensionality. This is also demonstrated in various numerical experiments.

Other works on data-driven discovery of collective dynamics. The majority of the earliest work in inferring interaction kernels in systems of the type (1.1), (2.2) occurred in the Physics literature, going back to the works of Newton. From the viewpoint of purely data-driven analysis of the equations, requiring limited or no physical reasoning, foundational work on estimating interaction laws includes [40, 48]. One can also refer to [50] for the recent development of the Weak SINDy algorithm that leverages the weak form of the differential equation and sparse parametric regression.
with applications to cellular dynamics [51]. In these works, the interaction kernels are assumed to be in the span of a known family of functions and parameters are estimated. In statistics, the problem of parameter estimation in dynamical systems from observations is classical, e.g. [12, 15, 43, 59, 71]. The question of identifiability of the parameter emerges, see e.g. [28, 52]. Our work is closely related to this viewpoint but our parameter is now infinite-dimensional; identifiability is discussed in Sect. 4.2.

Another highly active area of research in recent years is on stochastic interacting agent systems. The maximum likelihood approach is the most frequently studied approach in recent works, including the parameter estimation [7, 16, 34, 39, 64] and nonparametric estimation of drift in stochastic McKean–Vlasov equation [30, 33, 73], and radial interaction kernel learning in [46].

**Machine learning of dynamical system.** A vast literature exists in the context of learning dynamical systems [3, 5, 9, 13, 14, 37, 38, 44, 57, 58, 60, 62, 74]. There are many techniques which can be used to tackle the high-dimensionality of the data set: sparsity assumptions [11, 14, 61, 68], dimension reduction, reduced-order modeling. The dependent nature of the data prevents traditional regression-based approaches, see the discussion in [45], but many of the approaches above successfully address this. Our work, however, exploits the interacting-agent structure of collective dynamical systems, which is driven by a collection of two-body interactions where each interaction depends only on pairwise data between the states of agents, as in (1.1). With this structure in mind, we are able to reduce the ambient dimension of the data $N(2d + 1)$ to the dimension of the variables in the interaction kernels, which is independent of $N$. We also naturally incorporate the dependence in the data in an appropriate manner by considering trajectories generated from different initial conditions.

### 3.6 Function spaces

We begin by describing some basic ideas about measures and function spaces. Consider a compact or precompact set $\mathcal{U} \subset \mathbb{R}^p$ for some $p$; the infinity norm is defined as $\|h\|_\infty := \text{ess sup}_{x \in \mathcal{U}} |h(x)|$, and $L^\infty(\mathcal{U})$ as the space of real valued functions defined on $\mathcal{U}$ with finite $\infty$-norm. A key function space we need to consider is, $C^{k,\alpha}(\mathcal{U})$, for $k \in \mathbb{N}, 0 < \alpha \leq 1$, defined as the space of compactly supported, $k$-times continuously differentiable functions with a $k$-th derivative that is Hölder continuous of order $\alpha$. We can then consider vectorizations of these spaces over agent types as

$$L^\infty(\mathcal{U}) := \bigoplus_{k, k', 1 \leq k, k' \leq K} L^\infty(\mathcal{U}), \quad \text{endowed with the norm } \|f\|_\infty := \max_{k, k'} \|f_{kk'}\|_\infty, \forall f \in L^\infty(\mathcal{U}).$$

Similarly, we consider direct sums of measures, with corresponding vectorized function spaces, in particular $L^2$ (see Sect. 4.1).

We now define a suitable function class for the interaction kernels in the model (2.2). A simple model is that the agents get farther and farther apart, they eventually should have no influence on each other. For each pair $(k, k')$, $k, k' = 1, \ldots, K$ we
define the admissible space

\[
K_{kk'}^{\{E, A, \xi\}} := L^\infty([R_{kk'}^{\min}, R_{kk'}^{\max}] \times S_{kk'}^{\{E, A, \xi\}}), \quad K^{\{E, A, \xi\}} := \bigoplus_{k, k'=1, 1} K_{kk'}^{\{E, A, \xi\}},
\]

(3.6)

where we remind the reader that the \(\{E, A, \xi\}\) notation means, in this case, that there is an admissible space for each element of the set \(\{E, A, \xi\}\). Here, \(R_{kk'}^{\min}, R_{kk'}^{\max}\) are the minimum or maximum, respectively, possible interaction radius for agents in \(C_{k'}\) influencing agents in \(C_k\). Similarly, \(S_{kk'}^{E}, S_{kk'}^{A}, S_{kk'}^{\xi}\) are compact sets in \(R_{kk'}^{pE}, R_{kk'}^{pA}, R_{kk'}^{p\xi}\) which contain the ranges of the feature maps, \(s_{kk'}^{E}, s_{kk'}^{A}\) and \(s_{kk'}^{\xi}\). We will also need the sets:

\[
S^{\{E, A, \xi\}} := \prod_{k, k'} S_{kk'}^{\{E, A, \xi\}}, \quad R := \prod_{k, k'} [R_{kk'}^{\min}, R_{kk'}^{\max}], \quad R := \max_{k, k'} R_{kk'}^{\max}.
\]

(3.7)

Notice that all interaction kernels are supported on the interval of pairwise distance \([0, R]\).

We denote the distribution of the initial conditions by \(\mu^Y\). This measure is unknown and is the source of randomness in our system. It is a product measure of three measures \(\mu^X, \mu^V, \mu^\xi\), all also unknown, that represent the distribution on the initial positions, velocities, and environment variables, respectively. Specifically, we define,

\[
\mu^Y := \begin{bmatrix} \mu^X \\ \mu^V \\ \mu^\xi \end{bmatrix}
\]

(3.8)

It reflects that we will observe trajectories which start at different initial conditions, but that evolve from the same dynamical system. For example, in our numerical experiments we will choose \(\mu^Y\) to be uniform over a system-dependent compact set.

We let

\[
R_{\dot{x}} := \sup_{Y(0) \sim \mu^Y} \sup_{t \in [0, T]} \max_{i,i'} \| \dot{x}_i(t) - \dot{x}_{i'}(t) \|,
\]

(3.9)

\[
R_\xi := \sup_{Y(0) \sim \mu^Y} \sup_{t \in [0, T]} \max_{i,i'} \| \xi_i(t) - \xi_{i'}(t) \|,
\]

(3.10)

and we assume that both of these quantities are finite. A sufficient condition is that the measures \(\mu^V, \mu^\xi\) (specifying the distribution of the initial conditions on the velocities and the environment variable are compactly supported, which follows by the assumptions on the interaction kernels below and that we only consider finite final time \(T\).
First, we define the following vectorized function spaces, which we call *admissible sets*,

$$\mathcal{K}_{S(E,A,\xi)}^{\{E,A,\xi\}} := \left\{ \phi_{kk'}^{\{E,A,\xi\}} \in C^{0,1} \left( [R_{kk'}^{\text{min}}, R_{kk'}^{\text{max}}] \times S_{kk'}^{\{E,A,\xi\}} \right), \right. $$

$$\left. \| \phi_{kk'}^{\{E,A,\xi\}} \|_\infty + \text{Lip} \left[ \phi_{kk'}^{\{E,A,\xi\}} \right] \leq S_{\{E,A,\xi\}}, \text{ for all } k, k' = 1, \ldots, K, \right\}. \hspace{1cm} (3.11)$$

We will assume that the interaction kernels are in corresponding admissible sets:

$$\phi^E \in \mathcal{K}_E^E, \quad \phi^A \in \mathcal{K}_A^A, \quad \phi^\xi \in \mathcal{K}_\xi^\xi. \hspace{1cm} (3.12)$$

The admissibility assumptions (3.12) allow us to establish properties such as existence and uniqueness of solutions to (2.2) as well as to have control on the trajectory errors in finite time $[0, T]$. It further allows us to show regularity and absolute continuity with respect to Lebesgue measure of the appropriate performance measures defined in Sect. 4.1.

When estimating the $EA$ part of the system, we will consider the direct sum admissible space, for $S_{EA} \geq \max\{S_E, S_A\}$,

$$\mathcal{K}_{S_{EA}}^{EA} := \mathcal{K}_E^E \oplus \mathcal{K}_A^A \hspace{1cm} (3.13)$$

In the learning approach, we will consider hypothesis spaces that we will search in order to estimate the various interaction kernels. The hypothesis spaces corresponding to $\{\phi_{kk'}^{\{E,A,\xi\}}\}$ are denoted as $\{\mathcal{H}_{kk'}^{\{E,A,\xi\}}\}$ and we vectorize them as,

$$\mathcal{H}_{\{E,A,\xi\}}^{kk'} := \bigoplus_{k,k'=1,1}^{K,K} \mathcal{H}_{kk'}^{\{E,A,\xi\}}. \hspace{1cm} (3.14)$$

Analogous to our simplified notation for $\phi^{EA}, \varphi^{EA}$ described in (3.3), we define the direct sum of the hypothesis spaces as,

$$\mathcal{H}^{EA} := \mathcal{H}^E \oplus \mathcal{H}^A \hspace{1cm} (3.15)$$

We will consider specific choices for the hypothesis spaces during the learning theory and numerical algorithm sections.

### 4 Learning theory

#### 4.1 Probability measures and weighted $L^2$ for measuring learning performance

The interaction kernels depend on $(r, \dot{r}, s^E, s^A, s^\xi)$, and to measure distances between estimated interaction kernels and true interaction kernels, we consider a natural set.
of probability measures and corresponding weighted $L^2$ spaces. These generalize the constructions of [10, 45, 47]. For each interacting pair $(k, k')$, we let
\[ \rho_{T}^{E.A,k,k'}(r, s^E, \hat{r}, s^A) := \mathbb{E}_{Y_0 \sim \mu^Y} \frac{1}{TN_{kk'}} \int_{t=0}^{T} \sum_{i \in C_k, i' \in C_{k'}} \delta_{ii',t}(r, s^E, \hat{r}, s^A) \, dt \]

\[ \rho_{T}^{E.A,L,k,k'}(r, s^E, \hat{r}, s^A) := \mathbb{E}_{Y_0 \sim \mu^Y} \frac{1}{LN_{kk'}} \sum_{l=1}^{L} \sum_{i \in C_k, i' \in C_{k'}} \delta_{ii',tl}(r, s^E, \hat{r}, s^A) \] (4.1)

\[ \rho_{T}^{E.A,L,M,k,k'}(r, s^E, \hat{r}, s^A) := \frac{1}{MLN_{kk'}} \sum_{l,m=1}^{L,M} \sum_{i \in C_k} \delta_{ii',tm}(r, s^E, \hat{r}, s^A) \]

where \( N_{kk'} = N_k N_{k'} \) for \( k \neq k' \) and \( N_{kk'} = \binom{N_k}{2} \) for \( k = k' \), and we used the following shorthand notation for the Dirac measures:

\[ \delta_{ii',t}(r, s^E, \hat{r}, s^A) := \delta_{i\tilde{i}'(t),i\tilde{i}'(t),\hat{i}\tilde{i}'(t),i\tilde{i}'(t)}(r, s^E, \hat{r}, s^A) \]

\[ \delta_{ii',t,m}(r, s^E, \hat{r}, s^A) := \delta_{i\tilde{i}'(m),i\tilde{i}'(m),\hat{i}\tilde{i}'(m),i\tilde{i}'(m)}(r, s^E, \hat{r}, s^A) . \]

The measure \( \rho_{T}^{E.A.L,k,k'} \) is the discrete counterpart of \( \rho_{T}^{E.A,k,k'} \) with the continuous average over \([0, T]\) replaced by the average over the observation times \( 0 = t_1 < \ldots < t_L = T \). \( \rho_{T}^{E.A.L,M,k,k'} \) can be computed from observations and converges to \( \rho_{T}^{E.A.L,k,k'} \) as \( M \to \infty \).

We also consider the marginal distributions

\[ \rho_{T}^{E,k,k'}(r, s^E) := \int_{\hat{r}} \int_{s^A} \rho_{T}^{E.A,k,k'}(r, s^E, \hat{r}, s^A) \, ds^A \, d\hat{r} \]

\[ \rho_{T}^{A,k,k'}(r, \hat{r}, s^A) := \int_{s^E} \rho_{T}^{E.A,k,k'}(r, s^E, \hat{r}, s^A) \, ds^E \] (4.2)

and \( \rho_{T}^{E.L,k,k'}(r, s^E), \rho_{T}^{E.L,M,k,k'}(r, s^E), \rho_{T}^{A.L,k,k'}(\hat{r}, s^A), \rho_{T}^{A.L,M,k,k'}(\hat{r}, \hat{r}, s^A) \) defined analogously as above. The empirical measures, \( \rho_{T}^{E.L,M,k,k'}, \rho_{T}^{A.L,M,k,k'} \), are the ones used in the actual algorithm to quantify the learning performances of the estimators \( \hat{\phi}_{kk'}^E \) and \( \hat{\phi}_{kk'}^A \), respectively. They are also crucial in discussing the separability of \( \hat{\phi}_{kk'}^E \) and \( \hat{\phi}_{kk'}^A \).

For ease of notation, we introduce the following measures to handle the heterogeneity of the system, and which are used to describe error over all of the pairs \((k, k')\).

\[ \rho_{T}^{E.A.L} = \bigoplus_{k,k'=1,1}^{K,K} \rho_{T}^{E.A,L,k,k'} \]

\[ \rho_{T}^{E.A} = \bigoplus_{k,k'=1,1}^{K,K} \rho_{T}^{E.A,k,k'} \]

\[ L^2 \left( \rho_{T}^{E.A.L} \right) = \bigoplus_{k,k'=1,1}^{K,K} L^2 \left( \rho_{T}^{E.A,L,k,k'} \right) \] (4.3)
Similar definitions apply for measures related to learning the $\xi$-based interaction kernels, see supplement C. We discuss some key properties of the measures in supplement B.

We now discuss the performance measures for the estimated interaction kernels. We use weighted $L^2$-norms (with mild abuse of notation, we omit the weight from the notation) based on measures introduced above (with analogous definitions for the measures corresponding to finite $L^2$) that are adapted to the underlying dynamics:

$$
\| \widehat{\phi}_{E,k,k'}^E - \phi_{E,k,k'}^E \|_{L^2(\rho_{E,k,k'}^T)}^2 := \int_{(r,s^E)} (\widehat{\phi}_{E,k,k'}^E(r, s^E) - \phi_{E,k,k'}^E(r, s^E))^2 r^2 \, d\rho_{E,k,k'}^T
$$

$$
\| \widehat{\phi}_{A,k,k'}^A - \phi_{A,k,k'}^A \|_{L^2(\rho_{A,k,k'}^T)}^2 := \int_{(r,\dot{r},s^A)} (\widehat{\phi}_{A,k,k'}^A(r, \dot{r}, s^A) - \phi_{A,k,k'}^A(r, \dot{r}, s^A))^2 \dot{r}^2 \, d\rho_{A,k,k'}^T
$$

$$
\| \widehat{\phi}_{E,k,k'}^E - \phi_{E,k,k'}^E \|_{L^2(\rho_{E,k,k'}^T)}^2 := \int_{(r,s^E,\dot{r},s^A)} \left[ (\widehat{\phi}_{E,k,k'}^E(r, s^E) - \phi_{E,k,k'}^E(r, s^E)) r \\
+ (\widehat{\phi}_{A,k,k'}^A(r, \dot{r}, s^A) - \phi_{A,k,k'}^A(r, \dot{r}, s^A)) \dot{r} \right]^2 \, d\rho_{E,k,k'}^T
$$

(4.4)

Our learning theory focuses on minimizing the difference between $\widehat{\phi}_{E,k,k'}^E \oplus \widehat{\phi}_{E,k,k'}^A$ and $\phi_{E,k,k'}^E \oplus \phi_{E,k,k'}^A$, in the joint norm given by (4.4). As long as the joint norm is small, our estimators produce faithful approximations of the right hand side function of the original system and trajectories. However, it does not necessarily imply that both $\widehat{\phi}_{E,k,k'}^E - \phi_{E,k,k'}^E$’s and $\widehat{\phi}_{A,k,k'}^A - \phi_{A,k,k'}^A$’s are small in their corresponding energy- and alignment-based norms, since the joint norm is a weaker norm. It would be interesting to study if there is any equivalence between these two norms, but the problem appears to be quite delicate. The theoretical investigation is still ongoing.

Now, we have all the tools needed to establish a theoretical framework: dynamics induced probability measures, performance measurements in appropriate norms, and loss functionals. These will allow us to discuss the convergence properties of our estimators. Full details of the numerical algorithm are given in supplement D.

**Notational summary**

A summary of the learning theory notation introduced in Sects. 3.1, 3, and the notation above, is given below in Table 3.

### 4.2 Identifiability of kernels from data

In this section we introduce a technical condition, called coercivity condition, on the dynamical system that relates to the well-posedness (solvability and uniqueness of the solution) of the inverse problem and plays a key role in the learning theory. It generalizes the previous work [42, 45, 47]. In fact, for the second-order systems considered here, we will have two coercivity conditions, one for the energy and alignment terms and the other one for the $\xi$ variable. These conditions ensure that the minimizers to the error functionals are unique, and second that when the expected error functional is
Table 3  Notation used throughout the paper

| Notation | Definition | Ref |
|----------|------------|-----|
| $M$      | Number of trajectories | Sect. 1 |
| $L$      | Number of times in $[0, T]$ for each trajectory | Sect. 2 |
| $Y(t)$  | Full state space vector containing $X_t, V_t, \Omega_t$ | Sect. 2 |
| $\{E, A, \xi\}$ | Wildcard, means the notation applies for all 3 variables | Sect. 2 |
| $\|X\|_S$ | $\sum_{j=1}^{N} \frac{1}{N_j} \|x_j\|^2$ | (3.1) |
| $\mu^Y$ | Distribution on the initial conditions $Y(0)$ | Sect. 3.6 |
| $\phi^{[E,A,\xi]} = (\phi^{[E,A,\xi]}_{kk'})_{k,k'}$ | Vectorized true $E,A,\xi$ interaction kernels | (3.2) |
| $\varphi^{[E,A,\xi]} = (\varphi^{[E,A,\xi]}_{kk'})_{k,k'}$ | Shorthand denoting energy and alignment part of system | (3.3) |
| $E_A$    | Represents the joint function $\phi^E \oplus \phi^A \in \mathcal{H}^{E_A}$ | (3.3) |
| $\|Y\|^2_Y$ | $\|X\|^2_S + \|V\|^2_S + \|\Omega\|^2_S$. | (4.14) |
| $S^{[E,A,\xi]}_k$ | $\prod_{k,k'} \mathcal{S}^{[E,A,\xi]}_{kk'}$ | (3.7) |
| $R$      | Admissible spaces for the $E,A,\xi$ kernels | (3.11) |
| $\mathcal{K}^{[E,A,\xi]}_S$ | $\mathcal{K}^{[E,A,\xi]}_S$ | (3.14) |
| $\mathcal{H}^{[E,A,\xi]}_{k,k'}$ | $\mathcal{H}^{[E,A,\xi]}_{k,k'}$ | (3.14) |
| $\mathcal{H}^{[E,A,\xi]}_k$ | $\oplus_{k,k'} \mathcal{H}^{[E,A,\xi]}_{k,k'}$ | (3.15) |
| $\mathcal{E}^{E_A}(-), \mathcal{E}^{\xi}(-)$ | Empirical $E_A, \xi$ error functional | Sect. 3.3 |
| $\mathcal{H}^{E_A} \ominus_\mathcal{E}^{E_A} \mathcal{H}^{\xi}$ | $\mathcal{H}^{E_A} \ominus_\mathcal{E}^{E_A} \mathcal{H}^{\xi}$ | (3.5) |
| $\mathcal{H}^{E_A} \ominus_\mathcal{E}^{E_A} \mathcal{H}^{\xi}$ | Basis for $\mathcal{H}^{[E,A,\xi]}_{k,k'}$ | (3.5) |
| $\rho^{E_A,L} \cdot \rho^{\xi}_{T}$ | Measure for $E_A, \xi$ discrete in time, infinite trajectories | (4.3) |
| $\rho^{E_A,L} \cdot \rho^{\xi}_{T}$ | Measure for $E_A, \xi$ discrete in time, infinite trajectories | (4.3) |
| $L^2 (\rho^{E_A,L})$ | Coercivity constant on the $\mathcal{H}^{E_A}, \mathcal{H}^{\xi}$ hypothesis spaces | Definition 1 |
| $\mathcal{H}^{E_A,C} \cdot \mathcal{H}^{\xi}$ | Hypothesis spaces on $E_A, \xi$ depending on $M$ | Sect. 4.3 |
| $\mathcal{H}^{E_A,M} \cdot \mathcal{H}^{\xi}$ | Dimension for minimax convergence rates | (4.9) |
| $A^{E_A,M}, A^{\xi}_M$ | Learning matrices for the inverse problem | Sect. 6 |
| $N(\mathcal{H}, \delta)$ | $\delta$-covering number, under the $\infty$-norm, of a set $\mathcal{H}$ | [70] |
small, then the distance from the estimator to the true kernels is small in the appropriate \( \rho_T \) norm.

**Definition 1** *(Coercivity condition)* For the dynamical system (2.2) observed at time instants \( 0 = t_1 < t_2 < \cdots < t_L = T \) and with initial condition distributed \( \mu^Y \) on \( \mathbb{R}^{(2d+1)N} \), it satisfies the coercivity condition on the hypothesis space \( \mathcal{H}^{E,A} \) with constant \( c_{\mathcal{H}^{E,A}} \) if

\[
\begin{align*}
  c_{\mathcal{H}^{E,A}} := \inf_{\phi_{E,A} \in \mathcal{H}^{E,A} \setminus \{0\}} & \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_{\mu^Y} \left[ \left\| f_{\phi_{E,A}}(X_{t_l}, V_{t_l}, \Xi_{t_l}) \right\|_S^2 \right] \geq 0. 
\end{align*}
\]

Similarly, the system satisfies the coercivity condition on the hypothesis space \( \mathcal{H}^\xi \) with constant \( c_{\mathcal{H}^\xi} \) if

\[
\begin{align*}
  c_{\mathcal{H}^\xi} := \inf_{\phi^{\xi} \in \mathcal{H}^{\xi} \setminus \{0\}} & \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_{\mu^Y} \left[ \left\| f_{\phi^{\xi}}(X_{t_l}, V_{t_l}, \Xi_{t_l}) \right\|_S^2 \right] \geq 0. 
\end{align*}
\]

Analogous definitions holds for continuous observations over the time interval \([0, T]\), by replacing the average over observations at discrete times with an integral average over \([0, T]\).

**4.3 Consistency and optimal convergence rate of estimators**

**4.3.1 Concentration**

Our first main result is a concentration estimate that relates the coercivity condition to an appropriate bias-variance tradeoff in our setting. Let \( \mathcal{N}(\mathcal{H}, \delta) \) be the \( \delta \)-covering number, with respect to the \( \infty \)-norm, of the set \( \mathcal{H} \).

**Theorem 2** *(Concentration)* Suppose that \( \phi^{E,A,\xi} \in \mathcal{K}^{E,A,\xi} \). Consider a convex, compact (with respect to the \( \infty \)-norm) hypothesis spaces

\[
\mathcal{H}^{E,A}_M \subset L^\infty(\mathbb{R} \times S^E) \oplus L^\infty(\mathbb{R} \times S^A), \quad \mathcal{H}^\xi_M \subset L^\infty(\mathbb{R} \times S^\xi),
\]

bounded above by \( S_0 \geq \max\{S_E, S_A, S_\xi\} \) respectively. Additionally, assume that the coercivity conditions (4.5), (4.6) hold on \( \mathcal{H}^{E,A}_M \) and \( \mathcal{H}^\xi_M \), respectively.

Then for all \( \epsilon > 0 \), with probability (with respect to \( \mu^Y \)) at least \( 1 - \delta \), we have the estimates

\[
\begin{align*}
  c_{\mathcal{H}^{E,A}_M} \left\| \hat{\phi}^{E,A}_M - \phi^{E,A} \right\|_{L^2(\rho^E, L)}^2 & \leq 2 \inf_{\phi^{E,A} \in \mathcal{H}^{E,A}_M} \left\| \phi^{E,A} - \phi^{E,A} \right\|_{L^2(\rho^E, L)}^2 + 2 \epsilon, \\
  c_{\mathcal{H}^\xi_M} \left\| \hat{\phi}^{\xi}_M - \phi^{\xi} \right\|_{L^2(\rho^\xi, L)}^2 & \leq 2 \inf_{\phi^{\xi} \in \mathcal{H}^\xi_M} \left\| \phi^{\xi} - \phi^{\xi} \right\|_{L^2(\rho^\xi, L)}^2 + 2 \epsilon, 
\end{align*}
\]

Birkhäuser
provided that, for the first bound to hold,
\[ M \geq \frac{1152S_0^2 \max\{R, R_\xi\}^2 K^4}{\epsilon C_{\mathcal{H}_M^{EA}}} \left( \log\left( \mathcal{N}\left( \mathcal{H}_M^{EA}, \frac{\epsilon}{48.50 \max\{R, R_\xi\}^2 K^4} \right) \right) + \log\left( \frac{1}{\delta} \right) \right), \]
and similarly for the second inequality, using \( \mathcal{H}_M^{\xi} \).

### 4.3.2 Consistency

In the regime where \( M \to \infty \), we can choose a sequence of \( \mathcal{H}_M^{EA} \)'s such that the approximation error goes to 0. This enables us to control the infimum on the right hand side of (4.7).

**Theorem 3** (Strong consistency) Suppose that
\[ \{ \mathcal{H}_M^{EA} \}_{M=1}^{\infty} \subset L^\infty(\mathbb{R} \times S^E) \oplus L^\infty(\mathbb{R} \times S^A) \]
is a family of compact and convex subsets such that the approximation error goes to zero,
\[ \inf_{\phi^{EA} \in \mathcal{H}_M^{EA}} \| \phi^{EA} - \phi^{EA} \|_\infty \xrightarrow{M \to \infty} 0. \]
Further suppose that the coercivity condition holds on \( \bigcup_M \mathcal{H}_M^{EA} \), and that \( \bigcup_M \mathcal{H}_M^{EA} \) is compact in \( L^\infty(\mathbb{R} \times S^E) \oplus L^\infty(\mathbb{R} \times S^A) \). Then the estimator is strongly consistent with respect to the \( L^2(\rho_{T,EA,L}) \) norm:
\[ \lim_{M \to \infty} \| \hat{\phi}_M^{EA} - \phi^{EA} \|_{L^2(\rho_{T,EA,L})} = 0 \quad \text{with probability one.} \]
An analogous consistency result holds for the estimator in the \( \xi \) variable.

These two results together provide a consistency result on the full estimation of the triple \( (\hat{\phi}_M^{\xi}, \hat{\phi}_M^{E}, \hat{\phi}_M^{A}) \) and thus consistency of our estimation procedure on the full system (2.2).

### 4.4 Rate of convergence

Theorem 2 highlights the classical bias-variance tradeoff in our setting. Given data collected from \( M \) trajectories, we would like to choose the best hypothesis space to maximize the accuracy of the estimators. On the one hand, we would like the hypothesis space \( \mathcal{H}_M^{EA} \) to be large so that the bias
\[ \inf_{\phi^{EA} \in \mathcal{H}_M^{EA}} \| \phi^{EA} - \phi^{EA} \|_{L^2(\rho_{T,EA,L})}^2, \quad \text{or} \quad \inf_{\phi^{EA} \in \mathcal{H}_M^{EA}} \| \phi^{EA} - \phi^{EA} \|_\infty^2, \]

\( \odot \) Birkhäuser
is small. Simultaneously, we would like $\mathcal{H}_{M}^{E\cdot A}$ to be small enough so that the covering number $\mathcal{N}(\mathcal{H}_{M}^{E\cdot A}, \epsilon)$ is small. Just as in nonparametric regression, our rate of convergence depends on a regularity condition of the true interaction kernels and corresponding approximation properties of the hypothesis space, as is demonstrated in the following theorem. We establish the optimal (up to a log factor) min–max rate of convergence by choosing a hypothesis space of an optimal dimension as a function of the sample size $M$.

The dimension of the space supporting $\rho_{L}^{E\cdot A}$ is typically large: it is equal to $1 + \sum_{k, k'} p_{E}(k, k') + \sum_{(k, k')} p_{A}(k, k')$, see Table 1 for the definition of the $p(k, k')$. However we can exploit the structure of the system in such a way that our convergence rates only depend on the maximum number of unique variables in a pair $(k, k')$ across the $E$, $A$ portions of the system. A similar result holds for the $\rho_{L}^{E\cdot A}$ and its convergence rate. For the system (2.2), consider $V_{E}, kk'$ to be the number of distinct variables in the function $\phi_{kk'}^{E}$, similarly we define $V_{A}, kk'$, $V_{\xi}, kk'$, more precisely, and recalling the notation in Table 1:

\begin{align}
V_{E}, kk' &= 1 + p_{E}(k, k'), \\
V_{A}, kk' &= 1 + p_{A}(k, k'), \\
V_{\xi}, kk' &= 1 + p_{\xi}(k, k').
\end{align} (4.8)

Using these, we get the dimensions needed for the minimax rates:

\begin{align}
\mathcal{V} := \max_{k, k'} \{ V_{E}, kk', V_{A}, kk' \}, \\
\mathcal{V}_{\xi} := \max_{k, k'} V_{\xi}, kk'.
\end{align} (4.9)

The dimension for the minimax rates on the energy and alignment inference is given by $\mathcal{V}$, representing the maximum number of unique variables used in any one of the $\phi_{kk'}^{E}, \phi_{kk'}^{A}$ pairs. Analogously, $\mathcal{V}_{\xi}$ is used for the minimax convergence rate for the inference of $\phi_{kk'}^{\xi}$. As an extreme example, consider a problem with 10 different types of agents, leading to 100 distinct interaction kernels, each depending on $r$ and one additional variable that is unique for each function. In this case, we only pay the 2 dimensional rate, rather than the 101-dimensional rate in the ambient space of the 101 unique variables, although the heterogeneity affects the constant in the convergence rate. We note that we are not predicting the number of variables nor their form: these are assumed known. In Table 2 we report the values of $\mathcal{V}$ and $\mathcal{V}_{\xi}$ for a variety of prototypical systems. We are now ready to state our main result:

**Theorem 4 (Rate of convergence)** Let $\hat{\phi}_{M}^{E\cdot A} := \hat{\phi}_{M}^{E} \oplus \hat{\phi}_{M}^{A}$ denote the minimizer of the empirical error functional $\mathcal{E}_{M}^{E\cdot A}$ (defined in (3.4)) over the hypothesis space $\mathcal{H}_{M}^{E\cdot A}$.

(a) Let the hypothesis space be chosen as the direct sum of the admissible spaces, namely $\mathcal{H}_{M}^{E\cdot A} = K_{SE}^{E} \oplus K_{SA}^{A}$, and assume that the coercivity condition (4.5) holds on $\mathcal{H}_{M}^{E\cdot A}$. Then, there exists a constant $C$ depending only on $K$, $S_{E\cdot A}$, $R$, $R_{\xi}$ such that

$$
\mathbb{E}_{\mu_{Y}} \left[ \left\| \hat{\phi}_{M}^{E\cdot A} - \phi^{E\cdot A}_{M} \right\|_{L^{2}(\rho_{L}^{E\cdot A})}^{2} \right] \leq \frac{C}{c_{\mathcal{H}_{M}^{E\cdot A}}} M^{-\frac{1}{\mathcal{V} + 1}}.
$$

\[\Box\]
(b) Assume that \( \{L_n\}_{n=1}^\infty \) is a sequence of finite-dimensional linear subspaces of \( L^\infty(R \times S^E) \oplus L^\infty(R \times S^A) \) satisfying the dimension and approximation constraints
\[
\dim(L_n) \leq c_0 K^2 n^V, \quad \inf_{\varphi^{EA} \in L_n} \| \varphi^{EA} - \phi^{EA} \|_\infty \leq c_1 n^{-s}, \tag{4.10}
\]
for some fixed constants \( c_0, c_1 \) representing dimension-independent approximation characteristics of the linear subspaces, and \( s > 0 \) related to the regularity of the interaction kernels. The value \( n \) can be thought of as the number of basis functions along each of the \( (k, k') \) axes for each \( (k, k') \). Suppose the coercivity condition holds true on the set \( L := \bigcup_n L_n \), and let \( c^{EA}_L \) be the coercivity constant of \( L \).

Define \( B_n \) to be the closed ball centered at the origin of radius \( (c_1 + S_{EA}) \) in \( L_n \).

If we choose the hypothesis space as \( H_M = B_{k(M)} \), where \( k(M) \propto \left( \frac{M}{\log M} \right)^{\frac{1}{2s+1}} \), then there exists a constant \( C \) depending on \( K, R, R_\xi, S_{EA}, c_0, c_1, s \) such that we achieve the convergence rate,
\[
\mathbb{E}_{\mu^Y} \left[ \| \hat{\phi}^{EA}_M - \phi^{EA} \|_{L^2(\rho^{EA,L}_T)}^2 \right] \leq \frac{C}{c^{EA}_L} \left( \frac{\log M}{M} \right)^{\frac{2s}{2s+1}}. \tag{4.11}
\]

(c) under the corresponding assumptions as in (a), there exists a constant \( C \) depending only on \( K, S_{\xi}, R \) such that
\[
\mathbb{E}_{\mu^Y} \left[ \| \hat{\phi}^{\xi}_M - \phi^{\xi} \|_{L^2(\rho^{\xi,L}_T)}^2 \right] \leq \frac{C}{c^{\xi}_H} M^{-\frac{1}{2s+1}}. \tag{4.12}
\]

(d) under the corresponding assumptions as in (b), there exists a constant \( C \) depending only on \( K, R, S_{\xi}, c_0, c_1, s \) such that, and for \( c^{\xi}_H \) the coercivity constant of the corresponding linear space,
\[
\mathbb{E}_{\mu^Y} \left[ \| \hat{\phi}^{\xi}_M - \phi^{\xi} \|_{L^2(\rho^{\xi,L}_T)}^2 \right] \leq \frac{C}{c^{\xi}_H} \left( \frac{\log M}{M} \right)^{\frac{2s}{2s+1}}. \tag{4.12}
\]

We in fact prove bounds not only in expectation, but also with high probability, for every fixed large-enough \( M \), as the proof will show. In addition, we remark that in the case that the coercivity constant is independent of the number of agents \( N \), and not only will the convergence rate of our estimators be independent of the dimension \((2d + 1)N\) of the phase space, but even the constants in front of the rate term are independent of \( N \). We have proven this could be true for some initial conditions \([45]\) for first order systems. The analysis could be extend to some special cases of second-order systems we considered in this paper. In general, we do expect that the coercivity constant is dependent on \( N \). The empirical numerical experiments on some second-order systems \([75]\) support the idea that the coercivity condition is satisfied by large classes of second-order systems, we leave it as a future work for further theoretical investigation.
In both theorems, the convergence rates \( \frac{2s}{2s+d} \) and \( \frac{2s}{2s+d} V \) coincide with the minimax rate of convergence \( \frac{2s}{2s+d} \) for nonparametric regression in the corresponding dimension \( d \) — up to the logarithmic factor. (This logarithmic factor may be removable (using, e.g., the techniques in Chapter 11–15 of [36]), but with additional complexity of the proofs.) Achieving the same rate of convergence as if we had observed the noisy values of the interaction kernels directly, rather than through the dynamics, demonstrates the optimality of our approach. The strong consistency results show the asymptotic optimality of our method, and for wide classes of systems the assumptions of the theorems apply. Specifically, for part (b) of the theorems, the dimension and approximation conditions can be explicitly achieved by piecewise polynomials or splines appropriately adapted to the regularity of the interaction kernel. In the conditions of theorem 4, \( n \) can be the number of partitions along each axis of the variables in \( V \). Then, by using multivariate splines or piecewise polynomials we will have a fixed constant \( c_0 \) (corresponding to the number of parameters to estimate for each function) times \( K n^V \) as the dimension of the linear space. Furthermore, by standard approximation theory results, see [63] (Chapters 12, 13), [27], [29], for \( s \) the regularity of the interaction kernels we achieve the desired approximation condition with piecewise polynomials of degree \( \lfloor s \rfloor \). In our admissible spaces we have \( s = 1 \), note that the rate of convergence is faster if we have a kernel of higher regularity.

### 4.4.1 Examples of convergence rates in prototype systems

We next briefly examine the convergence rate on a few systems of interest. Recall that in Table 2 we have as the final two columns the values \( V, V^\xi \), which dictate the rate of convergence of our estimators in each system. Some specific highlights:

- For Anticipation Dynamics (AD) (see (6.2)), we have \( V = 2 \) unique variables shared across both of energy and alignment based kernels. So we learn at the 2-dimensional rate.

- For the Synchronized Oscillator dynamics (see (11) in [75]), each agent is indexed by \( i, \xi_i \) is its phase, \( x_i \) is (as usual) its position, \( \omega_i \) is the fixed natural frequency, \( v_i \) is the fixed self-propulsion velocity. The dynamics of \( x_i \) and \( \xi_i \) are governed by the following equations,

\[
\begin{align*}
\dot{x}_i & = v_i + \frac{1}{N} \sum_{i'=1}^{N} \left( \frac{x_i' - x_i}{\|x_i' - x_i\|} \left( A + J \cos(\xi_i' - \xi_i) \right) - B \frac{x_i' - x_i}{\|x_i' - x_i\|^2} \right), \\
\dot{\xi}_i & = \omega_i + K \sum_{i'=1}^{N} \frac{\sin(\xi_i' - \xi_i)}{\|x_i' - x_i\|}.
\end{align*}
\]  

\tag{4.13}

where \( A, J, B, K \) are constants. Formatting (4.13) into our model, we can see that we achieve the 2-dimensional optimal learning rate on each of the \( EA \) and \( \xi \) portions (rather than a 4-dimensional rate) due to the decoupled nature of the system; similarly we only pay the 1-dimensional rate twice for the Phototaxis system (see section D of SI in [47]). This is a key reason for splitting our learning theory between \( EA \)- and \( \xi \)-interaction kernels and accounting for shared and non-shared variables.
• The rates of convergence of our estimators for all previously-studied first-order systems (see [45, 47, 75]) can be derived from Theorem 4.

One downside of the results above is the lack of dependence on $L$: it seems natural that finer time samples in each trajectory should improve the results, at least up to a point. Indeed, the numerical experiments of [45, 47, 75] demonstrate that more data in $L$ may indeed be helpful to improve the performance. One technique used in [75] for very long trajectory data (large $L$, medium to small $M$) is to split each trajectory into larger $M$ with smaller $L$ in each. The dependence on the number of agents $N$ is not the objective of this work; it was considered [10] in the case of first-order systems; but further study in this mean-field regime is of interest to the authors and work is ongoing.

### 4.5 Performance of trajectory prediction

Once estimators $\hat{\phi}^{E,A}, \hat{\phi}^\xi$ are obtained, a natural question is the accuracy of the evolved trajectory based on these estimated kernels. We compare the observed trajectories to the estimated trajectories evolved from the same initial conditions but with the estimated interaction kernels. Recall that $Y_t = [X_t^T, V_t^T, \Xi_t^T]^T$ be the trajectory from dynamics generated by the true and unknown interaction kernels with initial condition $Y_0$, and $\hat{Y}_t = [\hat{X}_t^T, \hat{V}_t^T, \hat{\Xi}_t^T]^T$ be the trajectory from dynamics generated, with the same initial condition $\hat{Y}_0 = Y_0$, by the interaction kernels estimated from observations at times $\{t_i\}_{i=1}^L$. We let

$$\|Y_t - \hat{Y}_t\|_2^2 := \|X_t - \hat{X}_t\|^2_S + \|V_t - \hat{V}_t\|^2_S + \|\Xi_t - \hat{\Xi}_t\|^2_S.$$  \hspace{1cm} (4.14)

The next theorem shows that the error in prediction is (i) bounded trajectory-wise by a continuous-time version of the error functional, and (ii) bounded on average by the $L^2(\rho^{E,A}_T)$, $L^2(\rho^\xi_T)$, respectively, error of the estimator. This further validates the effectiveness of our error functional and $L^2(\rho_T)$-metric to assess the quality of the estimator. In particular, this emphasizes that although the system is a coupled system of ODE’s, our decoupled learning procedure with our choice of norm will lead to control of the expected supremum error as long as we minimize the $L^2(\rho^{E,A}_T)$, $L^2(\rho^\xi_T)$ norms in obtaining our estimators.

**Theorem 5** Suppose that $\hat{\phi}^E \in \mathcal{K}^E_{S_E}$, $\hat{\phi}^A \in \mathcal{K}^A_{S_A}$ and $\hat{\phi}^\xi \in \mathcal{K}^\xi_{S_T}$. Denote by $\hat{Y}(t)$ and $Y(t)$ the solutions of the systems with kernels $\hat{\phi}^E = (\hat{\phi}^E_{kk'})_{k,k'=1}^{K,K}$, $\hat{\phi}^A = (\hat{\phi}^A_{kk'})_{k,k'=1}^{K,K}$, and $\hat{\phi}^\xi = (\hat{\phi}^\xi_{kk'})_{k,k'=1}^{K,K}$ and $\phi^E$, $\phi^A$, $\phi^\xi$ respectively, both with the same initial condition. Then

$$\sup_{t \in [0,T]} \|\hat{Y}_t - Y_t\|^2_\mathcal{Y} \leq g(T) \left[ 2T^2 \int_{u=0}^T \int_{s=0}^u \|\hat{X}_s\| \right.$$\hspace{1cm} $- f^{nc,\hat{\phi}}(X_s, V_s, \Xi_s) - f^{\phi^{E,A}}(X_s, V_s, \Xi_s) \|_{\mathcal{S}} \| \] ds du

\(\Box\) Birkhäuser
where \( g(T) = 1 + (1 + B_1)T \exp(A_1 T + T^2/2) \). The constants are \( A_1 = 2T(8K P + L + 8QK + L^\xi) \) and \( B_1 = 2T^2(8K P + L) \), with any unspecified constants made precise in the proof and only depending on the Lipschitz constants of the noncollective forces and the feature maps, as well as the values \( S^E, S^A, S^\xi \) coming from the admissible spaces. It is bounded on average, with respect to the initial distribution \( \mu^1 \), by

\[
\mathbb{E}_{\mu^1} \left[ \sup_{t \in [0,T]} \| \hat{Y}_t - Y_t \|_{\mathcal{V}}^2 \right] \leq g(T) \left( (T^2 K^2 + TK^2) \| \phi^{EA} - \phi^{EA} \|^2_{L^2(\rho^{EA}_T)} + TK^2 \| \phi^{\xi} - \phi^{\xi} \|^2_{L^2(\rho^{\xi}_T)} \right)
\]

(4.15)

with the measures \( \rho^{EA}_T, \rho^{\xi}_T \) defined in \((4.1), (C.1)\). Expression (4.15) shows that by minimizing the right hand side, we can control the expected \( \mathcal{V} \)-supremum error of the estimated trajectories.

We postpone the somewhat lengthy proof to Appendix E.

## 5 Algorithm for constructing the interaction kernel estimators

Let \( \mathcal{H}^{E}_{kk'} \) be a finite dimensional function space of dimension \( n^{E}_{kk'} \) with basis functions given by piecewise polynomials whose degree will be chosen later (other type of basis functions are also possible, e.g., clamped B-splines as shown in [47]). It is built on uniform partitions of \([R_{kk'}^{\min, \text{obs}}, R_{kk'}^{\max, \text{obs}}]\) where \( R_{kk'}^{\min, \text{obs}} / R_{kk'}^{\max, \text{obs}} \) is the minimum/maximum interacting radius for agents in type \( k' \) influencing agents in type \( k \), derived from the observation data. Similar construction is done for \( \mathcal{H}^{A} \) with dimension \( n^{A}_{kk'} \). We write the candidate \( \varphi^{E}_{kk'}, \varphi^{A}_{kk'} \) as a linear combination of the basis functions:

\[
\varphi^{E}_{kk'}(r, s^E) = \sum_{n^{E}_{kk'} = 1}^{n^{E}_{kk'}} \alpha_{k,k',n_{kk'}^{E}}^{E} \psi_{k,k',n_{kk'}^{E}}^{x}(r, s^E),
\]

\[
\varphi^{A}_{kk'}(r, \hat{r}, s^A) = \sum_{n^{A}_{kk'} = 1}^{n^{A}_{kk'}} \alpha_{k,k',n_{kk'}^{A}}^{A} \psi_{k,k',n_{kk'}^{A}}^{x}(r, \hat{r}, s^A).
\]

Substituting this linear combination back into (3.4), we obtain a system of linear equations,

\[
A^{EA}_M \varphi^{EA} = \tilde{b}^{EA}_M,
\]
and minimizing the empirical loss functional corresponds to solving this system in the last square sense. Here, $\bar{\alpha}_{EA}^{E} \in \mathbb{R}^{n_{EA}} = [\bar{\alpha}_{E}^{E} \bar{\alpha}_{A}^{A}]^{T}$ with $\bar{\alpha}_{E}^{E}$ and $\bar{\alpha}_{A}^{A}$ being the collection of $\alpha_{k,k',\eta}^{E}$ or $\alpha_{k,k',\eta}^{A}$, respectively. Moreover, $A_{EA}^{E} \in \mathbb{R}^{n_{EA} \times n_{EA}}$ and $b_{EA}^{E} \in \mathbb{R}^{n_{EA}}$. See Sec. D for full details.

The total computational complexity is detailed as follows: $MLN^{2}$ for computing pairwise data, $MLd(n_{EA}^{2})^{2}$ for constructing the learning matrix and right hand side vector, and $(n_{EA}^{3})^{3}$ for solving the linear system, hence the total computing time is $MLN^{2} + MLd(n_{EA}^{2})^{2} + (n_{EA}^{3})^{3}$. Assuming that a tensor-grid construction of $\mathcal{H}_{kk'}^{E}$ is used, and the optimal $n_{*}$ with

$$n_{*} = \mathcal{O}\left(\left(\frac{M}{\log M}\right)^{\frac{1}{2s+V}}\right) \approx M^{\frac{1}{2s+V}}.$$ 

is used in each dimension, we have $n_{*}^{E} = 2n_{*}^{V} \approx 2M^{\frac{V}{2s+V}}$; then we obtain the total computing time in terms of $M$ as follows

$$\text{Comp. Time} = MLN^{2} + 4LdM^{\frac{2V}{2s+V} + 1} + 8M^{\frac{3V}{2s+V}}$$

In the special case of $s = 1$ (Lipschitz functions) and $V = 1$, we have

$$\text{Comp. Time} = MLN^{2} + 4LdM^{\frac{2}{3} + 1} + 8M \approx M^{\frac{2}{3} + 1}.$$ 

It is slightly super-linear in $M$.

Similar computational complexity analysis on solving $A_{E}^{E} \bar{\alpha}_{E}^{E} = \bar{b}_{E}^{E}$ also shows that the computational cost is slightly super linear in $M$ when $s = 1$ and $V = 1$.

The overall memory storage needed for the learning problem is $MLN(d(5 + n_{EA}^{E} + n_{E}^{E}) + 3)$, with $MLN(4d + 2)$ for storing the trajectory data, $MLN d(n_{EA}^{E} + n_{E}^{E})$ (here $n_{EA}^{E} = n_{E}^{E} + n_{A}^{E}$) for learning matrices, and $MLN(d + 1)$ for right hand side vectors. Hence if $M \gg \mathcal{O}(1)$, we can consider parallelization in $m$ in order to reduce the overhead memory, ending up with $M_{\text{per core}} \left( LN(d(5 + n_{EA}^{E} + n_{E}^{E}) + 3) \right)$ with $M_{\text{per core}} = \frac{M}{n_{\text{cores}}}$. The final storage of $A$ and $b$ only needs $n_{EA}^{E}(n_{EA}^{E} + 1) + n_{E}^{E}(n_{E}^{E} + 1)$.

### 6 Applications

Our learning theory, as well as measures, norms, functionals etc. can be applied to study all the examples considered in the works [45, 47, 75]. These examples, particularly those of [75], can thus be considered as applications of the theoretical results as well as of the algorithm in Sect. D. We choose to study one new dynamics, which is not considered in [47, 75] since they exhibit some unique features of our generalized model. In particular, we choose them due to their special form of having both energy-based and alignment-based interactions. It is called the anticipation dynamics (AD) model in [65]. Table 4 shows the value of learning parameters for these dynamics.
Table 4  Values of parameters for the learning

| $M_\rho$ | $L$ | $T_f$ | $T$ | $\mu^x$ | $\mu^\dot{x}$ | Num. of learning trials |
|----------|-----|------|----|---------|---------|------------------------|
| 2000     | 500 | 10   | 5  | Unif.([0, 5]^2) | Unif.([0, 5]^2) | 10                      |

The setup of the learning experiment is as follows. We use $M_\rho$ different initial conditions to evolve the dynamics from 0 to $T$ for the sole purpose of obtaining a good approximation to $\rho_{T,E}^M, \rho_{T,E}^L, \rho_{T,A}^L$. Then we use another set of $M$ ($M = 500$ for FwEP and $M = 750$ for AD) initial conditions to generate training data to learn the corresponding $\phi^E$ and $\phi^A$ from the empirical distributions, $\rho_{T,M,E}^L, \rho_{T,E}^E, \rho_{T,E}^A$, etc. We report the relative learning errors calculated via (4.4) for $\hat{\phi}^E \oplus \hat{\phi}^A$, (4.4) for $\hat{\phi}^E$, and (4.4) for $\hat{\phi}^A$, along with pictorial comparison of those interaction kernels as well as a visualization on the pairwise data which is used to learn the estimated kernels. Then we evolve the dynamics either from the training set of $M$ initial conditions or another set of $M$ randomly chosen initial conditions with $\phi^E \oplus \phi^A$ and $\hat{\phi}^E \oplus \hat{\phi}^A$ from 0 to $T_f > T$, and report the trajectory errors calculated using (6.1) on $y$ (the whole system), and for $x$ (the position) and $v$ (the velocity). Again, pictorial comparison of the trajectories are also shown. We report the trajectory errors over $[0, T]$ and $[T, T_f]$. The learning results are shown in the following sections. We consider a related norm on the trajectory $Y_{[0,T]} = \{ Y_{t_l} \}_{l=1}^L (0 = t_1 < \cdots < t_L = T)$:

$$\| Y_{[0,T]} - \hat{Y}_{[0,T]} \|_{\text{traj}} = \max_{l=1,...,L} \| Y(t_l) - \hat{Y}(t_l) \|_{\mathcal{Y}}. \quad (6.1)$$

We also consider a relative version, invariant under changes of units of measure:

$$\| Y_{[0,T]} - \hat{Y}_{[0,T]} \|_{\text{traj}}^* = \frac{\| Y_{[0,T]} - \hat{Y}_{[0,T]} \|_{\text{traj}}}{\| Y_{[0,T]} \|_{\text{traj}}}. \quad$$

Lastly, we report errors between $X_{[0,T]}$ and $\hat{X}_{[0,T]}$:

$$\| X_{[0,T]} - \hat{X}_{[0,T]} \|_{\mathcal{S}^*} = \max_{l=1,...,L} \frac{\| X(t_l) - \hat{X}(t_l) \|_{\mathcal{S}}}{\max_{l=1,...,L} \{ \| X(t_l) \|_{\mathcal{S}} \}}. \quad$$

Similar re-scaled norms are used for the difference between $V_{[0,T]}$ and $\hat{V}_{[0,T]}$, and for the difference between $\Xi_{[0,T]}$ and $\hat{\Xi}_{[0,T]}$.

---

1 We use the built-in MATLAB integrating routine, *ode15s*, with relative tolerance at $10^{-8}$ and absolute tolerance at $10^{-11}$.  

Birkhäuser
6.1 Learning results for anticipation dynamics with $U(r) = \frac{r^p}{p}$

The energy-based interactions are constants in the FwEP models, if we want to consider more complicated models, i.e., interactions depending on pairwise distance and more, the AD models are suitable candidates. The dynamics of the AD model is given as follows,

$$\ddot{x}_i = \frac{1}{N} \sum_{i' = 1, i' \neq i}^{N} \frac{\tau U'(||x_{i'} - x_i||)}{||x_{i'} - x_i||} (\dot{x}_{i'} - \dot{x}_i) + \frac{1}{N} \sum_{i' = 1, i' \neq i}^{N} \left\{ \frac{-\tau U'(||x_{i'} - x_i||)(x_{i'} - x_i) \cdot (\dot{x}_{i'} - \dot{x}_i)}{||x_{i'} - x_i||^3} + \frac{U'(||x_{i'} - x_i||)}{||x_{i'} - x_i||^2} \right\} (x_{i'} - x_i).$$

(6.2)

Here $\tau$ measures the amount (in time) of anticipation. In order to fit the model into our learning regime, we take

$$\phi^A(r) := \frac{\tau U'(r)}{r} \quad \text{and} \quad \phi^E(r, s) := -\frac{\tau U'(r)s}{r^3} + \frac{\tau U''(r)s}{r^2} + \frac{U'(r)}{r}.$$ 

Here we have no $\xi_i$, $K = 1$, $m_i = 1$, and

$$s^E_{i, i'} = s^A_{i, i'} := (x_{i'} - x_i) \cdot (\dot{x}_{i'} - \dot{x}_i).$$

We also use $\tau = 0.1$.

It is shown in [65] that if $U''$ is bounded when $r \to \infty$ with $U(0) = U'(0) = 0$, then unconditionally flocking would occur. We take $U(r) = \frac{r^p}{p}$ for $1 < p \leq 2$, then the system would show unconditional flocking. We choose $p = 1.5$ for our learning trials.\(^2\) We use a tensor grid of 1st degree piecewise standard polynomials with $n^E = 282$ for learning $\phi^E(r, s)$, then a set of 1st degree piecewise standard polynomials with $n^A = 138$ for learning $\phi^A(r)$. For the energy-based interactions we have the following results.

As is shown in Fig. 1b, the concentration of pairwise distance data is away from 0, making the estimation of the behavior of $\phi^E(r, s)$ at $r$ close to 0 extremely difficult, meanwhile, since $\phi^E$ is also weighted by the pairwise difference, $x_{i'} - x_i$, and at $r_{i, i'}$ close to 0, the information is also lost. Next, we present the alignment-based interaction kernels in Fig. 2a.

As shown in Fig. 2, the behavior of $\phi^A$ at $r = 0$ is learned accurately. Less accurate is the estimation of $\phi^A$ for large $r$: since the agents have aligned their velocities, the weight $v_{i'} - v_i$ is close to a zero vector. The overall learning performance for estimating $\phi^A$ is better compared to that of estimating $\phi^E$. The $\hat{\phi}^E \oplus \hat{\phi}^A$ error is:

\(^2\) $p = 2$ induces constant forces on the dynamics.
(a) \( U(r) = \frac{1.5}{1.5} : \phi^E \) vs. \( \hat{\phi}^E \), Err: \( 6 \cdot 10^{-1} \) \( 2.6 \)

(b) \( U(r) = \frac{1.5}{1.5} : \rho^E_{T} \) vs. \( \rho^E_{T} \).

Fig. 1 The lines shown in blue are the estimated interaction kernels, and the lines shown in black are the true interaction kernels. The colored areas shown in the background are the learned distributions of pairwise distance data.

(a) \( U(r) = \frac{1.5}{1.5} : \phi^A \) vs. \( \hat{\phi}^A \). Err: \( 1.7 \cdot 10^{-1} \) \( 3.9 \cdot 10^{-2} \).

(b) \( U(r) = \frac{1.5}{1.5} : \rho^A_{T} \) vs. \( \rho^A_{T} \).

Fig. 2 The lines shown in blue are the estimated interaction kernels, and the lines shown in black are the true interaction kernels. The colored areas shown in the background are the learned distributions of pairwise distance data.

\( 6 \cdot 10^{-1} \pm 3.0 \cdot 10^{-1} \). The comparison of trajectories between the true kernels (LHS) and the estimators (RHS) is shown in Fig. 3.

As shown in Fig. 3, visually, there is no difference between the true dynamics and the estimated dynamics. We offer more quantitative insight into the difference between the two in Table 5.

We maintain a 3-digit relative accuracy in estimating the position/velocity of the agents, even though for the interaction kernels, we are only able to maintain a 1-digit relative accuracy.

7 Conclusion and further directions

We have described a second-order model of interacting agents that incorporates multiple agent types, an environment, external forces, and multivariable interaction kernels. The inference procedure described exploits the structure of the system to achieve a learning rate that only depends on the dimension of the interaction kernels, which is much smaller than the full ambient dimension \((2d + 1)N\). Our estimators are strongly…
consistent, and in fact have learning rates that are min–max optimal within the non-parametric class, under mild assumptions on the interaction kernels and the system. We described how one can relate the expected supremum error of the trajectories for the system driven by the estimated interaction kernels to the difference between the true interaction kernels and the estimated ones—this result gives strong support to the use of our weighted $L^2$ norms as the correct way to measure performance and derive estimators. A detailed discussion of the full numerical algorithm, including the inverse problem derived from data and a coercivity condition to ensure learnability, along with complex examples, were presented and we showed how the formulation presented covers a very wide range of systems coming from many disciplines.
There are various ways that one could build on this work to handle different systems and for many of these further directions, the theoretical framework, techniques, and theorems presented here would be directly useful. In particular, one could consider second-order stochastic systems or a similar system but on a manifold, more complex environments, having more unknowns within the model beyond just the interaction kernels (say estimating the non-collective forces as well), identifying the best feature maps to model the data, and considering semiparametric problems where there are hidden parameters within the interaction kernels or other parts of the model that we wish to estimate along with the interaction kernels. The generality of the model and its broad coverage of models across the sciences, together with the scalability and performance of the algorithm, could inspire new models—both explicit equations and nonparametric estimators learned from data—which are theoretically justified and highly practical.

**Acknowledgements**  MM is grateful for discussions with Fei Lu and Yannis Kevrekidis, and for partial support from NSF-1837991, NSF-1913243, NSF-1934979, NSF-Simons-2031985, AFOSR-FA9550-17-1-0280 and FA9550-20-1-0288, ARO W911NF-18-C-0082, and to the Simons Foundation for the Simons Fellowship for the year 20-21; ST was partially supported by Regents Junior Faculty fellowship, Faculty Early Career Acceleration grant sponsored by University of California Santa Barbara, Hellman Family Faculty Fellowship, and the NSF under Award No. DMS-2111303; JM for support from NIH-T32GM11999; MZ for support from NSF-AoF-2225507. Please direct correspondence to JM and ST. All authors jointly designed research and wrote the manuscript; JM and ST derived theoretical results; MZ developed algorithms and applications; JM and MZ analyzed data.

**Data Availability**  Data and software package can be fond on: https://github.com/mingjzhong/LearningDynamics/tree/master/LearningDynamics2ndOrder.

**Open Access**  This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

**Appendix A: Continuity of the error functionals**

For any \( t \in [0, T] \), consider the two random variables,

\[
\mathcal{E}^{E_A}_{X_t} (\varphi^{E_A}) = \left\| \dot{X}_t - \mathbf{f}^{\text{nc}} (X_t, \bar{V}_t, \Xi_t) - \mathbf{f}^{\varphi^E} (X_t, \bar{V}_t, \Xi_t) - \mathbf{f}^{\varphi^A} (X_t, \bar{V}_t, \Xi_t) \right\|_S^2 \quad (A.1)
\]

\[
\mathcal{E}^{\varphi^E}_{\Xi_t} (\varphi^E) = \left\| \Xi_t - \mathbf{f}^{\text{nc}} (X_t, \bar{V}_t, \Xi_t) - \mathbf{f}^{\varphi^E} (X_t, \bar{V}_t, \Xi_t) \right\|_S^2 \quad (A.2)
\]
These will be used in various places throughout the technical proofs and easily relate to the error functionals
\[ E_{\infty}(\phi^E) := \mathbb{E}_\mu \frac{1}{L} \sum_{l=1}^{L} \| \tilde{X}_{t_l} - f_{\text{nc}, \xi}(X_{t_l}, V_{t_l}, \Xi_{t_l}) - f_{\phi^E}(X_{t_l}, V_{t_l}, \Xi_{t_l}) \|_S^2 \]  
\[ E_{\infty}(\phi^E) := \mathbb{E}_\mu \frac{1}{L} \sum_{l=1}^{L} \| \tilde{X}_{t_l} - f_{\text{nc}, \xi}(X_{t_l}, V_{t_l}, \Xi_{t_l}) - f_{\phi^E}(X_{t_l}, V_{t_l}, \Xi_{t_l}) \|_S^2 \]  
(A.3)

which by the Strong Law of Large Numbers satisfy
\[ E_{\infty}(\phi^E) = \lim_{M \to \infty} E_{\infty}(\phi^E) \quad \text{and} \quad E_{\infty}(\phi^E) = \lim_{M \to \infty} E_{\infty}(\phi^E). \]

Indeed
\[ E_{\infty}(\phi^E) = \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_\mu \left[ E_{\infty}(\phi^E) \right], \quad E_{\infty}(\phi^E) = \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_\mu \left[ E_{\infty}(\phi^E) \right]. \]

We begin by establishing basic continuity results for our error functionals over the hypothesis space. The specific structure of the governing equations plays a critical role in the analysis. We adapt the analysis in [45] to second-order systems. We list the key proofs and skip the proof of main technical lemmas.

**Alignment and energy based kernels**

**Proposition 6** Recall that $\phi^E$ are the true interaction kernels. For $\hat{\phi}^E$, $\hat{\phi}^E$ the true and empirical error functionals are bounded as follows,
\[
|E_{\infty}(\hat{\phi}^E) - E_{\infty}(\hat{\phi}^E)| \leq K^2 \| \hat{\phi}^E - \hat{\phi}^E \|_{L^2(\rho^E_L, L)} \| 2 \hat{\phi}^E - \hat{\phi}^E - \hat{\phi}^E \|_{L^2(\rho^E_L, L)} \]  
\[ |E_{M}(\hat{\phi}^E) - E_{M}(\hat{\phi}^E)| \leq K^4 \max\{R, R'\}^2 \| \hat{\phi}^E - \hat{\phi}^E \|_{L^2(\rho^E_L, L)} \| 2 \hat{\phi}^E - \hat{\phi}^E - \hat{\phi}^E \|_{L^2(\rho^E_L, L)} \]  
(A.4)

Recall the definitions of $R, R'$ in Eqs. (3.7), and (3.9).

**Proof** Using Jensen’s inequality,
\[
|E_{\infty}(\hat{\phi}^E) - E_{\infty}(\hat{\phi}^E)| = \left| \sum_{k=1}^{K} \frac{1}{N_k} \sum_{c_k} \left( \sum_{k'=1}^{K'} \frac{1}{N_{k'}} \sum_{c_{k'}} \left( (\hat{\phi}_{kk'}^E - \hat{\phi}_{kk'}^E)(r_{ii'})(s_{ii'}^E) + (\hat{\phi}_{kk'}^A - \hat{\phi}_{kk'}^A)(r_{ii'})(s_{ii'}^A) \right) \right) \right|. \]
\[
\left| \sum_{k'=1}^{K} \frac{1}{N_{k'}} \sum_{i' \in C_{k'}} (2\phi_{kk'}^{E} - \tilde{\phi}_{kk'}^{E})(r_{ii'}, s_{ii'}^{E})r_{ii'} + (2\phi_{kk'}^{A} - \tilde{\phi}_{kk'}^{A})(r_{ii'}, s_{ii'}^{A})r_{ii'} \right|
\]

\[
\leq \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{i \in C_{k}} \sum_{i' \in C_{k'}} \frac{1}{N_{k}} \left\| \frac{1}{N_{k'}} \sum_{i' \in C_{k'}} (\tilde{\phi}_{kk'}^{E} - \tilde{\phi}_{kk'}^{E})(r_{ii'}, s_{ii'}^{E})r_{ii'} \right\|
\]

\[
\times \left| \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{i \in C_{k}} \sum_{i' \in C_{k'}} \frac{1}{N_{k}N_{k'}} \left( (2\phi_{kk'}^{E} - \tilde{\phi}_{kk'}^{E})(r_{ii'}, s_{ii'}^{E})r_{ii'} + (2\phi_{kk'}^{A} - \tilde{\phi}_{kk'}^{A})(r_{ii'}, s_{ii'}^{A})r_{ii'} \right) \right| \times \frac{1}{N_{k}N_{k'}} \sum_{i \in C_{k}} \sum_{i' \in C_{k'}} \left( (\tilde{\phi}_{kk'}^{E} - \tilde{\phi}_{kk'}^{E})(r_{ii'}, s_{ii'}^{E})r_{ii'} + (\tilde{\phi}_{kk'}^{A} - \tilde{\phi}_{kk'}^{A})(r_{ii'}, s_{ii'}^{A})r_{ii'} \right) \right|^2
\]

\[
\leq K^2 \left\| \tilde{\phi}^{E} - \phi^{E} \right\|_{L^2(\tilde{\rho}_T^{k,k'})} \left\| 2\phi^{E} - \tilde{\phi}^{E} \right\|_{L^2(\tilde{\rho}_T^{k,k'})}
\]

where

\[
\hat{\rho}_T^{l,k,k'}(r, s^E, \hat{r}, s^A) = \frac{1}{L N_{k,k'}} \sum_{l=1}^{L} \sum_{i \in C_{k}, i' \in C_{k'}} \delta_{r_{i,l}}(t), s_{i,l}^{E}(t), r_{i',l}(t), s_{i',l}^{A}(t) (r, s^E, \hat{r}, s^A)
\]

and \(\hat{\rho}_T^{l} = \bigoplus_{k,k'=1}^{K} \hat{\rho}_T^{l,k,k'}\). Therefore, we have that

\[
\left| \frac{1}{L} \sum_{l=1}^{L} \mathcal{E}_{X_l(t)}^{EA}(\phi^{EA}) - \frac{1}{L} \sum_{l=1}^{L} \mathcal{E}_{X_l(t)}^{EA}(\tilde{\phi}^{EA}) \right| \leq \frac{1}{L} \sum_{l=1}^{L} \mathcal{E}_{X_l(t)}^{EA}(\tilde{\phi}^{EA}) - \mathcal{E}_{X_l(t)}^{EA}(\phi^{EA})
\]

\[
\leq \frac{K^2}{L} \sum_{l=1}^{L} \left\| \tilde{\phi}^{EA} - \phi^{EA} \right\|_{L^2(\tilde{\rho}_T^{l})} \left\| 2\phi^{EA} - \tilde{\phi}^{EA} \right\|_{L^2(\tilde{\rho}_T^{l})}
\]

\[
= K^2 \left\| \tilde{\phi}^{EA} - \phi^{EA} \right\|_{L^2(\tilde{\rho}_T^{l})} \left\| 2\phi^{EA} - \tilde{\phi}^{EA} \right\|_{L^2(\tilde{\rho}_T^{l})}
\]

\[
\leq K^4 \max \{ R, R_{i} \} \left\| \tilde{\phi}^{EA} - \phi^{EA} \right\|_{\infty} \left\| 2\phi^{EA} - \tilde{\phi}^{EA} \right\|_{\infty}
\]
Taking the expectation with respect to $\mu^Y$ on each side of (A.8) we get the first inequality. The second inequality follows by noticing that,

$$|E_M^{E}(\widehat{\phi}^{E}) - E_M^{E}(\widehat{\phi}^{E})| \leq \frac{1}{M} \sum_{m=1}^{M} \left| \frac{1}{L} \sum_{l=1}^{L} E_{X_l^{(m)}}^{E}(\widehat{\phi}^{E}) - \frac{1}{L} \sum_{l=1}^{L} E_{X_l^{(m)}}^{E}(\widehat{\phi}^{E}) \right|.$$ 

\[\square\]

**Environment interaction kernels**

Here we show an analogous result to the alignment and energy result above. The techniques are similar and the result serves an identical purpose in the theory. Recall the definition of $R_\xi$ in (3.10).

**Proposition 7** For $\widehat{\phi}, \widehat{\phi} \in \mathcal{H}^E$, we have

$$|E_\infty^E(\widehat{\phi}) - E_\infty^E(\widehat{\phi})| \leq K^2\|\widehat{\phi} - \widehat{\phi}\|_{L^2(\rho_\xi)} \|2\phi - \widehat{\phi}\|_{L^2(\rho_\xi)} \tag{A.10}$$

$$|E_M^E(\widehat{\phi}) - E_M^E(\widehat{\phi})| \leq K^4R_\xi^2\|\widehat{\phi} - \widehat{\phi}\|_\infty \|2\phi - \widehat{\phi}\|_\infty \tag{A.11}$$

The following lemma can be immediately deduced using (A.4), (A.5).

**Lemma 8** For all $\phi^{E_A} \in \mathcal{H}^{E_A}$, define the defect function $L_M^{E_A}(\phi^{E_A})$ as

$$L_M^{E_A}(\phi^{E_A}) = E_\infty^{E_A}(\phi^{E_A}) - E_M^{E_A}(\phi^{E_A}). \tag{A.12}$$

Then, given two functions $\phi_1^{E_A}, \phi_2^{E_A} \in \mathcal{H}^{E_A}$, the defect function is bounded by

$$|L_M^{E_A}(\phi_1^{E_A}) - L_M^{E_A}(\phi_2^{E_A})| \leq 2K^4\max\{R, R_\xi\}^2\|\phi_1^{E_A} - \phi_2^{E_A}\|_\infty \|\phi_1^{E_A} + \phi_2^{E_A} - 2\phi^{E_A}\|_\infty$$

almost surely with respect to $\mu^Y$.

A similar lemma can be immediately deduced on the $\xi$ variable.

**Lemma 9** For all $\phi^{E_\xi} \in \mathcal{H}^{E_\xi}$, define the defect function $L_M^{E_\xi}(\phi^{E_\xi})$ as

$$L_M^{E_\xi}(\phi^{E_\xi}) = E_\infty^{E_\xi}(\phi^{E_\xi}) - E_M^{E_\xi}(\phi^{E_\xi}). \tag{A.13}$$

Then, given two functions $\phi_1^{E_\xi}, \phi_2^{E_\xi} \in \mathcal{H}^{E_\xi}$, the defect function is bounded by

$$|L_M^{E_\xi}(\phi_1^{E_\xi}) - L_M^{E_\xi}(\phi_2^{E_\xi})| \leq 2K^4R_\xi^2\|\phi_1^{E_\xi} - \phi_2^{E_\xi}\|_\infty \|\phi_1^{E_\xi} + \phi_2^{E_\xi} - 2\phi^{E_\xi}\|_\infty$$

almost surely with respect to $\mu^Y$.  

\[\copyright\] Birkhäuser
A.1 Uniqueness of minimizers over a compact convex space

Recall the energy and alignment bilinear functional \( \langle \cdot, \cdot \rangle_{EA} \)

\[
\langle \phi_{EA}^1, \phi_{EA}^2 \rangle_{EA} := \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_{\mu^{Y\{1\}}} \left[ \left( f_{\phi_{EA}^1}(X_{tl}, V_{tl}, Z_{tl}), f_{\phi_{EA}^2}(X_{tl}, V_{tl}, Z_{tl}) \right) \right].
\]

for any \( \phi_{EA}^1, \phi_{EA}^2 \in \mathcal{H}^{EA} \). The \( \mathcal{S} \)-inner product is the inner product induced by the \( \| \cdot \|_{\mathcal{S}} \) norm by the polarization identity, which holds as we are working in an \( L^2 \) space, so the parallelogram law holds. Then our coercivity condition (4.5). can be written in terms of this bilinear functional as: for all \( \phi_{EA} \in \mathcal{H}^{EA} \)

\[
c_{\mathcal{H}^{EA}} \| \phi_{EA} \|^2_{L^2(\rho_{EA,L}^T)} \leq \langle \phi_{EA}, \phi_{EA} \rangle.
\]

Proposition 10  Let the minimizer of the error functional be denoted

\[
\hat{\phi}_{L,\infty, \mathcal{H}^{EA}} := \phi_{L,\infty, \mathcal{H}^{EA}} \oplus \phi_{L,\infty, \mathcal{H}^{EA}} := \arg \min_{\phi_{EA} \in \mathcal{H}^{EA}} E_{\infty}(\phi_{EA});
\]

then for all \( \phi_{EA} \in \mathcal{H}^{EA} \), the difference of the error functional at this element of \( \mathcal{H}^{EA} \) and the minimizer is lower bounded as,

\[
E_{\infty}(\phi_{EA}) - E_{\infty}(\hat{\phi}_{L,\infty, \mathcal{H}^{EA}}) \geq c_{\mathcal{H}^{EA}} \| \phi_{EA} - \hat{\phi}_{L,\infty, \mathcal{H}^{EA}} \|^2_{L^2(\rho_{EA,L}^T)}. \tag{A.14}
\]

Thus, the minimizer of \( E_{\infty}^{EA} \) over \( \mathcal{H}^{EA} \) is unique in \( L^2(\rho_{EA,L}^T) \).

Proof  The proof leverages the convexity of the risk functional and hypothesis function space. The proof is the same with Proposition 18 in [45]. \( \square \)

Proposition 11  Let the minimizer of the error functional be denoted

\[
\hat{\phi}_{L,\infty, \mathcal{H}^{\xi}} := \arg \min_{\phi_{\xi} \in \mathcal{H}^{\xi}} E_{\infty}(\phi_{\xi});
\]

then for all \( \phi_{\xi} \in \mathcal{H}^{\xi} \), the difference of the error functional at this element of \( \mathcal{H}^{\xi} \) and the minimizer is lower bounded as,

\[
E_{\infty}(\phi_{\xi}) - E_{\infty}(\hat{\phi}_{L,\infty, \mathcal{H}^{\xi}}) \geq c_{\mathcal{H}^{\xi}} \| \phi_{\xi} - \hat{\phi}_{L,\infty, \mathcal{H}^{\xi}} \|^2_{L^2(\rho_{\xi,L}^T)}. \tag{A.15}
\]

Thus, the minimizer of \( E_{\infty}^{\xi} \) over \( \mathcal{H}^{\xi} \) is unique in \( L^2(\rho_{\xi,L}^T) \).
A.2 Uniform estimates on defect functions

We start this section by introducing normalized errors of the estimators. Denote the minimizer of $E^A_{\infty}(\cdot)$ over $\mathcal{H}^E$ by

$$\hat{\phi}^A_{L,\infty,\mathcal{H}^E} := \hat{\phi}^E_{L,\infty,\mathcal{H}^E} \oplus \hat{\phi}^A_\infty,\mathcal{H}^E = \arg\min_{\phi^A \in \mathcal{H}^E} E^A_{\infty}(\phi^A).$$  \hfill (A.16)

For any $\phi^A \in \mathcal{H}^E$, define the normalized errors as

$$D^\infty(\phi^A) := E^A_{\infty}(\phi^A) - E^A_{\infty}(\hat{\phi}^A_{L,\infty,\mathcal{H}^E}),$$  \hfill (A.17)

$$D^M(\phi^A) := E^A_{M}(\phi^A) - E^A_{M}(\hat{\phi}^A_{L,\infty,\mathcal{H}^E}).$$  \hfill (A.18)

These quantities capture the difference between the expected/empirical errors of the estimator and the function in the hypothesis space minimizing the expected error functional. We begin by proving a lemma that assumes the distance between the expected and empirical normalized errors are small for a given estimator. We then show that we have similar control on these distances for all points in a neighborhood of this particular one. This control enables us to apply a covering argument in the main proposition of this section due to the compactness of the hypothesis space.

Remark 12 Exactly analogous definitions hold for the $\xi$ variable and we will simply state the results in that case.

We skip the proof of Lemma 13, Lemma 14, as they follow in the same way as Lemma 19 in [45].

Lemma 13 For all $\epsilon > 0$ and $0 < \alpha < 1$, if the function $\phi^A_1 \in \mathcal{H}^E$ satisfies

$$\frac{D^\infty(\phi^A_1) - D^M(\phi^A_1)}{D^\infty(\phi^A_1) + \epsilon} < \alpha,$$

then for all $\phi^A_2 \in \mathcal{H}^E$ such that $\|\phi^A_1 - \phi^A_2\|_{\infty} \leq \frac{\alpha \epsilon}{8 S_{EA} \max\{R, R_1\}^4 K^4}$, where $S_{EA} = \max\{S_E, S_A\}$ we have

$$\frac{D^\infty(\phi^A_2) - D^M(\phi^A_2)}{D^\infty(\phi^A_2) + \epsilon} < 3\alpha.$$

Arguing in the same way as above, we can derive the lemma below using Eq. (A.11). We define $D^\xi_{\infty}, D^\xi_{M}$ similarly to (A.17), (A.18) using $E^\xi_{\infty}, E^\xi_{M}$ in the obvious way.

Lemma 14 For all $\epsilon > 0$ and $0 < \alpha < 1$, if the function $\phi^\xi_1 \in \mathcal{H}^\xi$ satisfies

$$\frac{D^\xi_{\infty}(\phi^\xi_1) - D^\xi_{M}(\phi^\xi_1)}{D^\xi_{\infty}(\phi^\xi_1) + \epsilon} < \alpha.$$
then for all $\phi^k \in H^k$ such that $\|\phi^1_1 - \phi^k_2\|_\infty \leq \frac{\alpha \epsilon}{8S_0 R^2 \xi K^4}$, we have, for $S_0 \geq S_\xi$,

$$\frac{D^k_\infty(\phi^k_2) - D^k_M(\phi^k_2)}{D^k_\infty(\phi^k_2) + \epsilon} < 3\alpha.$$ 

### A.3 Concentration

**Proposition 15** For all $\epsilon > 0$, $0 < \alpha < 1$, $\phi^{EA} \in H^{EA}$, the following concentration bound holds

$$P_{\mu^Y} \left\{ \frac{D_\infty(\phi^{EA}) - D_M(\phi^{EA})}{D_\infty(\phi^{EA}) + \epsilon} \geq \alpha \right\} \leq \exp \left( \frac{-c_{H^{EA}} \alpha^2 M \epsilon}{32 S_{EA}^2 K^4} \right)$$

**Proof** Consider the random variable $\Theta$ (with randomness coming from the random initial condition distributed $\mu^Y$), and to ease the notation let $\hat{\phi}^{EA} := \hat{\phi}_L, H^{EA},$ 

$$\Theta = \frac{1}{L} \sum_{l=1}^L (E^{EA}_{X(l)}(\phi^{EA}) - E^{EA}_{X(l)}(\hat{\phi}^{EA}))$$

The coercivity condition given in Definition 1, Proposition 10 and (A.4) allow us to bound the variance, denoted $\sigma^2$, of $\Theta$ as follows.

$$\sigma^2 \leq \mathbb{E}_{\mu^Y} \left[ \left| \frac{1}{L} \sum_{l=1}^L (E^{EA}_{X(l)}(\phi^{EA}) - E^{EA}_{X(l)}(\hat{\phi}^{EA})) \right|^2 \right]$$

$$\leq \frac{1}{L} \sum_{l=1}^L \mathbb{E}_{\mu^Y} \left[ \left| E^{EA}_{X(l)}(\phi^{EA}) - E^{EA}_{X(l)}(\hat{\phi}^{EA}) \right|^2 \right]$$

$$\leq K^4 \max \{R, R_\xi\}^2 \|\phi^{EA} - \hat{\phi}^{EA}\|^2_{L^2(\rho_L^T)} \|\phi^{EA} + \hat{\phi}^{EA} - 2\phi^{EA}\|^2_\infty$$

$$\leq \frac{K^4 \max \{R, R_\xi\}^2}{c_{H^{EA}}} (E^{EA}_\infty(\phi^{EA}) - E^{EA}_\infty(\hat{\phi}^{EA})) \|\phi^{EA} + \hat{\phi}^{EA} - 2\phi^{EA}\|^2_\infty$$

$$\leq \frac{16 S_{EA}^2 \max \{R, R_\xi\}^2 K^4}{c_{H^{EA}}} (E^{EA}_\infty(\phi^{EA}) - E^{EA}_\infty(\hat{\phi}^{EA}))$$

$$\leq \frac{16 S_{EA}^2 \max \{R, R_\xi\}^2 K^4}{c_{H^{EA}}} D_\infty(\phi^{EA}). \quad (A.19)$$

By applying Eq. (A.9) from the proof of Proposition 6, we have that $\Theta \leq 8 S_{EA}^2 \max \{R, R_\xi\}^2 K^4$ almost surely. We then apply the one-sided Bernstein inequality to $\Theta$ and recalling the definitions (A.1) together with the definitions of the normalized
errors in (A.17, A.18), we get that:

\[
P_{\mu^y} \left\{ \frac{D_\infty (\varphi^{EA}) - D_M (\varphi^{EA})}{D_\infty (\varphi^{EA}) + \epsilon} \geq \alpha \right\} \leq \exp \left( - \frac{\alpha^2 (D_\infty (\varphi^{EA}) + \epsilon)^2 M}{2 \left( \sigma^2 + \frac{8S_{EA}^2 \max \{ R, R_\xi \}^2 K^4 \alpha (D_\infty (\varphi^{EA}) + \epsilon) \} \right) \right).
\]

Now we provide a lower bound for the exponent to simplify the dependencies. We show that,

\[
\epsilon \cdot c_{\mathcal{H}^{EA}} \leq \frac{(D_\infty (\varphi^{EA}) + \epsilon)^2}{2 \left( \sigma^2 + \frac{8S_{EA}^2 \max \{ R, R_\xi \}^2 K^4 \alpha (D_\infty (\varphi^{EA}) + \epsilon) \} \right)},
\]

or equivalently,

\[
\epsilon \cdot c_{\mathcal{H}^{EA}} \leq \frac{8S_{EA}^2 \max \{ R, R_\xi \}^2 K^4 \alpha (D_\infty (\varphi^{EA}) + \epsilon)}{16S_{EA}^2 \max \{ R, R_\xi \}^2 K^6} \left( \sigma^2 + \frac{8S_{EA}^2 \max \{ R, R_\xi \}^2 K^4 \alpha (D_\infty (\varphi^{EA}) + \epsilon)}{3} \right).
\]

By the estimate (A.19), since \( 0 < \alpha \leq 1 \), and \( 0 < c_{L,N,\mathcal{H}^{EA}} < K^2 \) it is sufficient to show that

\[
D_\infty (\varphi^{EA}) \epsilon + \frac{\epsilon (D_\infty (\varphi^{EA}) + \epsilon)}{6} \leq (D_\infty (\varphi^{EA}) + \epsilon)^2.
\]

This follows from Young’s inequality as \( 2D_\infty (\varphi^{EA}) \epsilon + \epsilon^2 \leq (D_\infty (\varphi^{EA}) + \epsilon)^2 \), and together these results give the desired bound of the proposition.

We can easily derive the desired supremum bound by a covering argument. The estimation of the covering numbers involved will play a critical role in the main theorems and will be done in a dimension dependent way in order to get optimal minimax rates.

**Proposition 16** In the notation of Proposition 15,

\[
P_{\mu^y} \left\{ \sup_{\varphi^{EA} \in \mathcal{H}^{EA}} \frac{D_\infty (\varphi^{EA}) - D_M (\varphi^{EA})}{D_\infty (\varphi^{EA}) + \epsilon} \geq 3\alpha \right\} \leq N \left( \mathcal{H}^{EA}, \frac{\alpha \epsilon}{8S_{EA} \max \{ R, R_\xi \}^2 K^4} \right) e^{-\frac{c_{\mathcal{H}^{EA}}^2 M \epsilon}{32S_{EA} K^4}}
\]

where \( N \left( \mathcal{H}^{EA}, \frac{\alpha \epsilon}{8S_{EA} \max \{ R, R_\xi \}^2 K^4} \right) \) denotes the covering number of \( \mathcal{H}^{EA} \) with radius \( \frac{\alpha \epsilon}{8S_{EA} \max \{ R, R_\xi \}^2 K^4} \).
Proof Let \( \varphi_i^{EA} = \varphi_i^E \oplus \varphi_i^A \in \mathcal{H}^{EA} \), for \( i = 1, \ldots, N \left( \mathcal{H}^{EA}, \frac{\alpha \epsilon}{8S_{EA} \max\{R, R_i\}^2 K^7} \right) \), denote the center of disks \( D_i \) of radius \( \frac{\alpha \epsilon}{8S_{EA} \max\{R, R_i\}^2 K^7} \) covering \( \mathcal{H}^{EA} \). The covering number is finite by the compactness assumption on the hypothesis space. By Lemma 13,

\[
\sup_{\varphi^{EA} \in D_i} \frac{D_\infty(\varphi^{EA}) - D_M(\varphi^{EA})}{D_\infty(\varphi^{EA}) + \epsilon} \geq 3\alpha \quad \Rightarrow \quad \frac{D_\infty(\varphi_i^{EA}) - D_M(\varphi_i^{EA})}{D_\infty(\varphi_i^{EA}) + \epsilon} \geq \alpha.
\]

Now, by Proposition 15, for each \( i \),

\[
P_{\mu^y} \left\{ \sup_{\varphi^{EA} \in \mathcal{H}^{EA}} \frac{D_\infty(\varphi^{EA}) - D_M(\varphi^{EA})}{D_\infty(\varphi^{EA}) + \epsilon} \geq 3\alpha \right\} \leq P_{\mu^y} \left\{ \frac{D_\infty(\varphi_i^{EA}) - D_M(\varphi_i^{EA})}{D_\infty(\varphi_i^{EA}) + \epsilon} \geq \alpha \right\}
\]

\[
\leq e^{-\frac{c_{\mathcal{H}^{EA}} a^2 M \epsilon}{32S_{EA} \max\{R, R_i\}^2 K^6}}.
\]

By definition, \( \mathcal{H}^{EA} \subseteq \bigcup_i D_i \), so that

\[
P_{\mu^y} \left\{ \sup_{\varphi^{EA} \in \mathcal{H}^{EA}} \frac{D_\infty(\varphi^{EA}) - D_M(\varphi^{EA})}{D_\infty(\varphi^{EA}) + \epsilon} \geq 3\alpha \right\}
\]

\[
\leq \sum_i P_{\mu^y} \left\{ \sup_{\varphi^{EA} \in D_i} \frac{D_\infty(\varphi^{EA}) - D_M(\varphi^{EA})}{D_\infty(\varphi^{EA}) + \epsilon} \geq 3\alpha \right\}
\]

\[
\leq N\left( \mathcal{H}^{EA}, \frac{\alpha \epsilon}{8S_{EA} \max\{R, R_i\}^2 K^4} \right) e^{-\frac{c_{\mathcal{H}^{EA}} a^2 M \epsilon}{32S_{EA} \max\{R, R_i\}^2 K^6}}.
\]

\( \square \)

Finally, we state the results for the \( \xi \) variable, the proofs are analogous. The advantage of splitting the theorems will become apparent. Specifically, it allows us to control the covering numbers on the \( EA \) and \( \xi \) hypothesis spaces separately, enabling us to get faster rates than if we viewed the task as estimating all functions simultaneously. This is possible due to the fundamentally decoupled nature of the dynamical system.

Proposition 17 For all \( \epsilon > 0, 0 < \alpha < 1, \varphi^\xi \in \mathcal{H}^\xi \), the following concentration bound holds

\[
P_{\mu^y} \left\{ \sup_{\varphi^\xi \in \mathcal{H}^\xi} \frac{D_\infty(\varphi^\xi) - D_M(\varphi^\xi)}{D_\infty(\varphi^\xi) + \epsilon} \geq \alpha \right\} \leq e^{-\frac{c_{\mathcal{H}^\xi} a^2 M \epsilon}{32S_{\xi} K^4}}.
\]

Proposition 18 In the notation of Proposition 17,

\[
P_{\mu^y} \left\{ \sup_{\varphi^\xi \in \mathcal{H}^\xi} \frac{D_\infty(\varphi^\xi) - D_M(\varphi^\xi)}{D_\infty(\varphi^\xi) + \epsilon} \geq 3\alpha \right\} \leq N\left( \mathcal{H}^\xi, \frac{\alpha \epsilon}{8S_{\xi} R_\xi^2 K^4} \right) e^{-\frac{c_{\mathcal{H}^\xi} a^2 M \epsilon}{32S_{\xi} K^4}},
\]

\( \odot \) Birkhäuser
where \( \mathcal{N} \left( \mathcal{H}, \frac{\alpha \epsilon}{8 S_0 R_2 K^4} \right) \) denotes the covering number of \( \mathcal{H} \) with radius \( \frac{\alpha \epsilon}{8 S_0 R_2 K^4} \).

Now we are ready to present proofs for main concentration theorems.

**Proof of Theorem 2** We start out by setting \( \alpha = \frac{1}{6} \) in Proposition 16, which yields the tightest bound in the argument below. To ease the notation we let \( \hat{\phi}_{L,M}^{E} \mathcal{H}^{E} \oplus \hat{\phi}_{L,M}^{A} \mathcal{H}^{A} \) and similarly for \( \hat{\phi}_{L,\infty}^{E} \mathcal{H}^{E} \). From the Proposition, we have that

\[
\sup_{\phi^{E} \in \mathcal{H}^{E}} \frac{D_{\infty}(\phi^{E}) - D_{M}(\phi^{E})}{D_{\infty}(\phi^{E}) + \epsilon} < \frac{1}{2}
\]

holds true with probability

\[
P \geq 1 - \mathcal{N} \left( \mathcal{H}^{E}, \frac{\epsilon}{48 S_E \max\{R, R_1\}^2 K^4} \right)
\times \exp \left( - \frac{c \mathcal{H}^{E} M \epsilon}{1152 S_{EA}^2 \max\{R, R_1\}^2 K^6} \right).
\]

This immediately implies, by choosing \( \phi^{E} = \hat{\phi}_{L,M}^{E} \mathcal{H}^{E} \), that with probability \( P \)

\[
D_{\infty}(\hat{\phi}_{L,M}^{E} \mathcal{H}^{E}) < 2D_{M}(\hat{\phi}_{L,M}^{E} \mathcal{H}^{E}) + \epsilon.
\]

By definition of \( \hat{\phi}_{L,M}^{E} \mathcal{H}^{E} \) as the minimizer of the empirical error functional \( \mathcal{E}_{M}^{E} \), we see that

\[
D_{M}(\hat{\phi}_{L,M}^{E} \mathcal{H}^{E}) = \mathcal{E}_{M}^{A}(\hat{\phi}_{L,M}^{E} \mathcal{H}^{E}) - \mathcal{E}_{M}^{E}(\hat{\phi}_{L,\infty}^{E} \mathcal{H}^{E}) \leq 0,
\]

and combining this result with equation (A.14) from Proposition 10, we have

\[
c \mathcal{H}^{E} \| \hat{\phi}_{L,M}^{E} \mathcal{H}^{E} - \hat{\phi}_{L,\infty}^{E} \mathcal{H}^{E} \|_{L^{2}(\rho_{T}^{E,L})} \leq D_{\infty}(\hat{\phi}_{L,M}^{E} \mathcal{H}^{E}) < \epsilon, \quad (A.21)
\]

with probability \( P \). With the same probability,

\[
\| \hat{\phi}_{L,M}^{E} \mathcal{H}^{E} - \phi^{E} \|_{L^{2}(\rho_{T}^{E,L})}^{2} \leq 2 \| \hat{\phi}_{L,M}^{E} \mathcal{H}^{E} - \phi^{E} \|_{L^{2}(\rho_{T}^{E,L})}^{2} + 2 \| \hat{\phi}_{L,\infty}^{E} \mathcal{H}^{E} - \phi^{E} \|_{L^{2}(\rho_{T}^{E,L})}^{2}
\]

\[
\leq \frac{2}{c \mathcal{H}^{E}} \left( \epsilon + \inf_{\phi^{E} \in \mathcal{H}^{E}} K^2 \| \phi^{E} - \phi^{E} \|_{L^{2}(\rho_{T}^{E,L})}^{2} \right)
\]

\[
\leq \frac{2}{c \mathcal{H}^{E}} \left( \epsilon + \inf_{\phi^{E} \in \mathcal{H}^{E}} K^2 \max\{R, R_1\}^2 \| \phi^{E} - \phi^{E} \|_{\infty}^{2} \right).
\]
The first inequality follow from the coercivity condition (4.5) and the definition of $\hat{\phi}^E\mathcal{A}_\infty$. The second follows by the definition of the norms. Now for a chosen $0 < \delta < 1$, let

$$1 - \mathcal{N}\left(\mathcal{H}^E\mathcal{A}, \frac{\epsilon}{48S_{EA} \max\{R, R\ddot{x}\}^2 K^4}\right) \times \exp\left(-\frac{c_{\mathcal{H}^E\mathcal{A}} M \epsilon}{1152S_{EA}^2 \max\{R, R\ddot{x}\}^2 K^6}\right) \geq 1 - \delta$$

and solve for $M$. The proof for the part of the system involving $\xi$ is similar. □

**Proof of Theorem 3** To simplify the notation, we use the same conventions as the proof of Theorem 2 and let $D_{M,\infty} = D_{L,\infty}\mathcal{H}^E\mathcal{A}_M$. By definition of the coercivity constant in (4.5), we have the inequality $c_{\bigcup M} \mathcal{H}^E\mathcal{A}_M \leq c_{\mathcal{H}^E\mathcal{A}}$. From an argument analogous to the one used to arrive at Eq. (A.21) in the proof of Theorem 2, we obtain that

$$c_{\bigcup M} \mathcal{H}^E\mathcal{A}_M \|\hat{\phi}^E\mathcal{A}_M - \phi^E\mathcal{A}_M\|^2_{L^2(\rho^E_{T,L})} \leq D_{M,\infty}(\hat{\phi}^E\mathcal{A}_M) + \mathcal{E}^E\mathcal{A}_\infty(\hat{\phi}^E\mathcal{A}_\infty). \quad (A.22)$$

For $\epsilon > 0$, inequality (A.22) implies

$$P_{\mu^y}\{c_{\bigcup M} \mathcal{H}^E\mathcal{A}_M \|\hat{\phi}^E\mathcal{A}_M - \phi^E\mathcal{A}_M\|^2_{L^2(\rho^E_{T,L})} \geq \epsilon\} \leq P_{\mu^y}\{D_{M,\infty}(\hat{\phi}^E\mathcal{A}_M) + \mathcal{E}^E\mathcal{A}_\infty(\hat{\phi}^E\mathcal{A}_\infty) \geq \epsilon\} \leq P_{\mu^y}\{D_{M,\infty}(\hat{\phi}^E\mathcal{A}_M) \geq \frac{\epsilon}{2}\} + P_{\mu^y}\{\mathcal{E}^E\mathcal{A}_\infty(\hat{\phi}^E\mathcal{A}_\infty) \geq \frac{\epsilon}{2}\}.$$

We now bound the two terms in the above expression separately. For the first term, the proof of Theorem 2 shows that

$$P_{\mu^y}\{D_{M,\infty}(\hat{\phi}^E\mathcal{A}_M) \geq \frac{\epsilon}{2}\} \leq \mathcal{N}\left(\mathcal{H}^E\mathcal{A}_M, \frac{\epsilon}{C_1}\right) \exp\left(-\frac{c_{\mathcal{H}^E\mathcal{A}} M \epsilon}{C_2}\right) \leq \mathcal{N}\left(\bigcup M \mathcal{H}^E\mathcal{A}_M, \frac{\epsilon}{C_1}\right) \times \exp\left(-\frac{c_{\bigcup M} \mathcal{H}^E\mathcal{A}_M \epsilon}{C_2}\right)$$

where $C_1 = 96S_{EA}^2 \max\{R, R\ddot{x}\}^2 K^4$, $C_2 = 2304S_{EA}^2 \max\{R, R\ddot{x}\}^2 K^4$, and $\mathcal{N}(\bigcup M \mathcal{H}^E\mathcal{A}_M, \frac{\epsilon}{C_1})$ is finite because of the compactness assumption on $\bigcup M \mathcal{H}^E\mathcal{A}_M$. Summing this bound in $M$,

$$\sum_{M=1}^{\infty} P_{\mu^y}\{D_{M,\infty}(\hat{\phi}^E\mathcal{A}_M) \geq \frac{\epsilon}{2}\} \leq \mathcal{N}\left(\bigcup M \mathcal{H}^E\mathcal{A}_M, \frac{\epsilon}{C_1}\right) \sum_{M=1}^{\infty} \exp\left(-\frac{c_{\bigcup M} \mathcal{H}^E\mathcal{A}_M \epsilon}{C_2}\right) < \infty.$$

For the second term, the bound (A.4) yields

$$\mathcal{E}^E\mathcal{A}_\infty(\hat{\phi}^E\mathcal{A}_\infty) \leq 4K^4 S_{EA} \max\{R, R\ddot{x}\}^2 \inf_{\phi^E\mathcal{A}_M \in \mathcal{H}^E\mathcal{A}_M} \|\phi^E\mathcal{A}_M - \phi^E\mathcal{A}_\infty\|_{\infty} \xrightarrow{M \to \infty} 0.$$
Since $\epsilon$ is fixed, the above formula, together with our assumption on the sequence of hypothesis spaces, implies that $P_\mu^Y \left\{ E_\infty^A(\phi_\infty^A) \geq \frac{\epsilon}{2} \right\} = 0$ for $M$ sufficiently large. So we have $\sum_{M=1}^{\infty} P_\mu^Y \left\{ E_\infty^A(\phi_\infty^A) \geq \frac{\epsilon}{2} \right\} < \infty$. The finiteness of the two sums above implies, by the first Borel–Cantelli Lemma, that

$$
P_\mu^Y \left\{ \limsup_{M \to \infty} \left\{ C \left\| \phi_M^A - \phi^A \right\| L_2(\rho_T^{E.A.L}) \geq \epsilon \right\} \right\} = 0.
$$

Since $\epsilon$ was arbitrary, we have the desired strong consistency of the estimator. An exactly analogous argument gives the result on the part of the system involving $\xi$. □

**Proof of Theorem 4** For part (a), let $\mathcal{H} = \mathcal{K}_S^E \oplus \mathcal{K}_S^A$. Standard results on covering numbers of function spaces (see theorem 2.7.1 of [70]) give us that the covering number of $\mathcal{H}$ satisfies

$$
\mathcal{N}(\mathcal{H}, \epsilon, \| \cdot \|_\infty) \leq C_\mathcal{H} \exp \left( K \sum_{k,k'=1}^{K,K} \left( \frac{1}{\epsilon} \right)^{\gamma_{E,k,k'}^E} + \left( \frac{1}{\epsilon} \right)^{\gamma_{A,k,k'}^A} \right) \leq C_\mathcal{H} \exp \left( 2K^2 \left( \frac{1}{\epsilon} \right) ^{\gamma} \right)
$$

for some absolute constant $C_\mathcal{H}$ depending only on $\mathcal{H}$ and $\gamma$. By assumption on the hypothesis space, we have that $\inf_{\phi^A \in \mathcal{H}} \| \phi^A - \phi^A \|_2 = 0$. From this, the concentration estimate (4.7) together with the covering number bound imply that,

$$
P_\mu^Y \left\{ \left\| \hat{\phi}_{L,M}^A - \phi^A \right\| L_2(\rho_T^{E.A.L}) > \epsilon \right\} \leq \mathcal{N}(\mathcal{H}, C_1 \epsilon, \| \cdot \|_\infty) \exp(-C_2 M \epsilon) \leq C_\mathcal{H} \exp(2K^2(C_1 \epsilon)^{-\gamma} - C_2 M \epsilon) \quad (A.23)
$$

where $C_1 = \frac{c_\mathcal{H}}{48S_{EA} \max\{R,R_1\}^2 K^4}$ and $C_2 = \frac{c_\mathcal{H}}{1152S_{EA} \max\{R,R_1\}^2 K^4}$. Next, define the function

$$
g(\epsilon) := 2K^2(C_1 \epsilon)^{-\gamma} - \frac{C_2 M \epsilon}{2},
$$

which we will minimize to achieve the desired probability bound. By direct calculation, $g(\epsilon) = 0$ if we choose $\epsilon = \epsilon_M = (\frac{C_1}{M})^{\frac{1}{\gamma+1}}$, where $C_3 = \left( \frac{4K^2}{C_2 C_1 \gamma} \right)^{\frac{1}{\gamma+1}}$; moreover the derivative of $g(\epsilon)$ is $\leq 0$ for all $\epsilon \geq \epsilon_M$. Therefore, the bound (A.23) implies

$$
P_\mu^Y \left\{ \left\| \hat{\phi}_{L,M}^A - \phi^A \right\| L_2(\rho_T^{E.A.L}) > \epsilon \right\} \leq \begin{cases} 
\exp \left( \frac{-C_2 M \epsilon}{2} \right), & \epsilon \geq \epsilon_M \\
1, & \epsilon \leq \epsilon_M
\end{cases} \quad (A.24)
$$

Integrating over $\epsilon \in (0, +\infty)$ and using $e^{-x} \leq x + 1$ for all $x \geq 0$, we obtain

$$
\int_0^{\infty} P_\mu^Y \left\{ \left\| \hat{\phi}_{L,M}^A - \phi^A \right\| L_2(\rho_T^{E.A.L}) > \epsilon \right\} d\epsilon \leq \left( \frac{C_4}{M} \right)^{\frac{1}{\gamma+1}} + O \left( \frac{1}{M} \right)
$$
Using coercivity and (4.7), we conclude that

\[
\mathbb{E}_\mu^Y [\| \hat{\phi}^{EA}_{L,M} - \phi^{EA} \|_{L^2(\rho_T^{EA,I})}^2] \leq \frac{C_4}{c^{EA}_{EA}} M^{-\frac{1}{4+1}},
\]

where \( C_4 \) is an absolute constant that only depends on \( K, S_{EA}, R, R_\chi \).

For part (b), we recall (see [25, Proposition 5]) that

\[
\mathcal{N}(B_n, \epsilon, \| \cdot \|_\infty) \leq \left( \frac{4(c_1 + S_{EA})}{\epsilon} \right)^{c_0 K^2 n^\nu}.
\]

Using (4.7), and the approximation assumption, we bound the probability as

\[
P_\mu^Y \left\{ \| \hat{\phi}^{EA}_{L,M} - \phi^{EA} \|_{L^2(\rho_T^{EA,I})}^2 > \epsilon + c_2 n^{-2s} \right\}
\]

\[
= P_\mu^Y \left\{ \| \hat{\phi}^{EA}_{L,M} - \phi^{EA} \|_{L^2(\rho_T^{EA,I})}^2 > t'n^{-2s} + c_2 n^{-2s} \right\}
\]

\[
= P_\mu^Y \left\{ \| \hat{\phi}^{EA}_{L,M} - \phi^{EA} \|_{L^2(\rho_T^{EA,I})}^2 > tn^{-2s} \right\}
\]

\[
\leq \mathcal{N}(B_n, c' tn^{-2s}, \| \cdot \|_\infty) \exp(-c_4 M tn^{-2s})
\]

\[
\leq \left( \frac{c_3}{tn^{-2s}} \right)^{c_0 K^2 n^\nu} \exp(-c_4 M tn^{-2s})
\]

\[
\leq \exp(c_0 K^2 n^\nu \log(c_3) + c_0 K^2 n^\nu |\log(tn^{-2s})| - c_4 M tn^{-2s}),
\]

where \( c_2 = \frac{1}{c_{ln} \mathcal{L}_n} c_1, c_2' = \frac{c_{ln} \mathcal{L}_n}{48(S_{EA} + c_1) \max[R, R_\chi]^2 K^4}, c_3 = \frac{192(S_{EA} + c_1)^2 \max[R, R_\chi]^2 K^4}{c_{EA} \mathcal{L}_n}, c_4 = \frac{1152(S_{EA} + c_1)^2 \max[R, R_\chi]^2 K^4}{c_{ln} \mathcal{L}_n} \) are absolute constants independent of \( M \).

Define

\[
g(n) := c_0 n^\nu K^2 \log(c_3) + c_0 n^\nu K^2 |\log(tn^{-2s})| - \frac{c_4}{2} M tn^{-2s}.
\]

To find the optimal \( n \) in terms of \( M \), we minimize \( g \) in \( n \). By taking a derivative, and solving the corresponding equation, we see that the optimal \( n \) is

\[
n_* = O \left( \left( \frac{M}{\log M} \right)^{\frac{1}{2\nu + 1}} \right),
\]

with a constant depending on \( c_3, c_4, c_2 \) but not on \( M \). We choose \( n_* = \lfloor \left( \frac{M}{\log M} \right)^{\frac{1}{2\nu + 1}} \rfloor \), let \( \epsilon_M = \left( \frac{M}{\log M} \right)^{\frac{2\nu}{2\nu + 1}} \) and

\[
h(\epsilon) := c_0 n_* K^2 \log(c_3) + c_0 n_* K^2 |\log(\epsilon)| - \frac{c_4}{2} M \epsilon.
\]
As above, let $\epsilon = t n_{-2s} = t \epsilon_M$ and consider $h(t \epsilon_M)$. It is easy to see that $\lim_{t \to 0^+} h(t \epsilon_M) = \infty$ and $\lim_{t \to \infty} h(t \epsilon_M) = -\infty$. Together with the continuity of $h$, these facts imply that there exists a constant $c_5$, depending on $K, c_0, c_2, c_3, c_4$, such that $h(c_5 \epsilon_M) = 0$. We further need that $h'(\epsilon) \leq 0$ for all $\epsilon \geq c_5 \epsilon_M$. By taking the derivative of $h$, setting it $\leq 0$, we find that this condition eventually holds by basic calculus. Therefore, if needed, to satisfy the derivative condition, we can enlarge the constant $c_5$ to a constant $c_6$ (independent of $M$) such that $h(\epsilon) \leq 0$ and $h'(\epsilon) \leq 0$ for all $\epsilon \geq c_6 \epsilon_M$. These results imply

$$P_{\mu^X} \left( \| \hat{\Phi}_{L,M,B_{n_s}} - \Phi_{E} \|^2_{L^2(\rho_{E,A,L}^t)} > \epsilon \right) \leq \begin{cases} \exp \left( \frac{-c_5 M \epsilon}{2} \right), & \epsilon \geq c_6 \epsilon_M \\ 1, & \epsilon \leq c_6 \epsilon_M \end{cases}, \quad (A.26)$$

and therefore

$$\int_0^\infty P_{\mu^X} \left( \| \hat{\Phi}_{L,M,B_{n_s}} - \Phi_{E} \|^2_{L^2(\rho_{E,A,L}^t)} > \epsilon \right) d\epsilon \leq C_1 \left( \frac{\log M}{M} \right)^{\frac{2s}{2s+1}},$$

where $C_1$ is a constant depending on $c_0, c_1, s, K, S_{E,A}, R, R_\tilde{x}$.

Now, with $H_{E,M}^{E} = B_{n_s}$ and using (4.7), we have shown the convergence rate

$$\mathbb{E}_{\mu^X} \left[ \| \hat{\Phi}_{L,M,H_{E,M}^{E}} - \Phi_{E} \|^2_{L^2(\rho_{E,A,L}^t)} \right] \leq \frac{c_7}{c_{E}^{E}} \left( \frac{M}{\log M} \right)^{-\frac{2s}{2s+1}},$$

where $c_7$ is an absolute constant that only depends on $s, K, c_0, c_1, S_{E,A}, R, R_\tilde{x}$. □

**Appendix B: Existence, uniqueness and properties of the measures**

In this section, we provide technical details of the analytic properties of the collective system under consideration as well as of the measures that we defined in Sect. 4.1. We emphasize that for the analytic portion of the theory, as we saw with the trajectory prediction result, we view the system (2.2) as coupled (whereas for the learning theory we leverage that they can be decoupled to make the estimation have better performance). We begin by stating that under the assumption that the interaction kernels lie in the corresponding admissible spaces, then the system is well-posed.

**B.1 Well-posedness of second-order heterogeneous systems**

**Proposition 19** Suppose the interaction kernels $\phi^E = (\phi_{kk'}^{E})^{K,K}_{k,k'=1}, \phi^A = (\phi_{kk'}^{A})^{K,K}_{k,k'=1}, \phi^E = (\phi_{kk'}^{E})^{K,K}_{k,k'=1}$ lie in the admissible sets $K_{E,F}, K_{A,F}, K_{E}$ respectively. Where the admissible spaces are defined in (3.11). Then the second-order heterogenous system (2.2) admits a unique global solution in $[0, T]$ for every initial datum $X_0, X_0(0) \in \mathbb{R}^{dN}$, $\mathbb{E}(0) \in \mathbb{R}^N$ and the solution depends continuously on the initial condition.
The proof of Proposition 19 uses Lemma 20 and similar techniques used to prove the well-posedness of the first-order homogeneous system (see Section 6 in [10]) by rewriting the second-order system as a first-order system and then applying standard Caratheodory ODE results.

**Lemma 20** For any $\varphi^E \in \mathcal{K}_{SE}^E$, $\varphi^A \in \mathcal{K}_{SA}^A$, the function

$$F[\varphi^{EA}](x, \dot{x}, s^E, s^A) := \varphi^E(||x||, s^E)x + \varphi^A(||x||, s^A)\dot{x},$$

for $x, \dot{x} \in \mathbb{R}^d$ is Lipschitz continuous on $\mathbb{R}^{2d+p^E+p^A}$ where $p^E, p^A$ are the dimensions of the range of the functions $s^E, s^A$, respectively. Additionally, for any $\varphi^\xi \in \mathcal{K}_{S\xi}^\xi$, the function

$$F[\varphi^\xi](x, s^\xi, \xi) := \varphi^\xi(||x||, s^\xi)\xi$$

is Lipschitz continuous on $\mathbb{R}^{d+1+p^\xi}$, where $p^\xi$ is the dimension of the range of $s^\xi$.

**B.2 Properties of measures**

In this section we state and prove some technical properties of the measures described in Sect. 4.1.

**Lemma 21** Suppose each of the interaction kernels lie in the respective admissible spaces, namely, $\phi^E \in \mathcal{K}_{SE}^E$, $\phi^A \in \mathcal{K}_{SA}^A$, $\phi^\xi \in \mathcal{K}_{S\xi}^\xi$. Then, for each $(k, k')$, the measures $\rho_{E,k,k'}^T$, $\rho_{A,k,k'}^T$, $\rho_{\xi,k,k'}^T$, and $\rho_{E,k,k'}^L$, $\rho_{\xi,k,k'}^L$ defined in Sect. 4.1, are regular Borel probability measures.

The proof is similar to Lemma 1 in [45].

**Proposition 22** Suppose the distribution $\mu_Y$ of the initial condition is compactly supported. Then for each $(k, k')$, the support of the measures $\rho_{E,k,k'}^T$, $\rho_{\xi,k,k'}^L$ (and therefore $\rho_{E,k,k'}^L$, $\rho_{\xi,k,k'}^L$) is also compact.

**Proof** The compact support of the variables $r, \dot{r}, \xi$ and the feature maps follows by the global well-posedness of the system in finite time, together with the Lipschitz assumptions on the non-collective forces. This compact support over a fixed, finite time is what is claimed in Proposition 22.

The main point is that by making reasonable assumptions on the non-collective forces, feature maps, interaction kernels, and the interval of time, together with the assumption that our agents’ initial conditions cannot be arbitrarily far apart, we can derive that the pairwise distance, velocity and $\xi$ will be controlled. Thus, the measures in Sect. 4.1, if given enough trajectories, will be well approximated by the discretized version using the numerical approach described in Sect. 6. Meaning that if we have a reasonable number of trajectories, we can look at the set of pairwise distances, velocities, etc. that these agents explore and bin them to set the support of the interaction kernels.
Explicit values for the constants claimed in the proposition depend on the properties of the non-collective forces, the support and sup-norm of the interaction kernels, the interval $T$, and the number of agents.

Appendix C: Additional performance measures

For measures related to learning the $\xi$-based interaction kernels, we take

$$\delta_{i,i',t}(r, s^\xi, \xi(\mathbf{r}, \mathbf{s}^\xi, \xi))$$

and similarly for $\rho_T^{\xi,L,k,k'}(r, s^\xi, \xi)$, $\rho_T^{\xi,L,M,k,k'}(r, s^\xi, \xi)$. Similarly, $\rho_T^{\xi,k,k'}$ and its time-discretization version, $\rho_T^{\xi,L,k,k'}$, are only used in the theoretical setting, whereas the empirical $\rho_T^{\xi,L,M,k,k'}$ is used in the actual algorithm. We consider direct sums of the measures for the phase variable for ease of notation.

$$\rho_T^{\xi,L} = \bigoplus_{k,k'=1}^{K,K} \rho_T^{\xi,L,k,k'}, \quad \rho_T^{\xi,k,k'} = \bigoplus_{k,k'=1}^{K,K} \rho_T^{\xi,k,k'}, \quad L^2(\rho_T^{\xi,L}) = \bigoplus_{k,k'=1}^{K,K} L^2(\rho_T^{\xi,L,k,k'})$$

Lastly, for the $\xi$-based interaction kernels, i.e., $\hat{\phi}_{kk'}$ versus $\phi_{kk'}^\xi$, we consider the following norm,

$$\|\hat{\phi}_{kk'} - \phi_{kk'}^\xi\|_{L^2(\rho_T^{\xi,k,k'})}^2 = \int_r \int_\xi \int_{s^\xi} (\hat{\phi}_{kk'}(r, \xi, s^\xi) - \phi_{kk'}^\xi(r, \xi, s^\xi))^2 \xi^2 d\rho_T^{\xi,k,k'}(r, \xi, s^\xi).$$

Appendix D: Numerical algorithm

In this section, we will detail the construction of the linear systems to learn $\vec{\alpha}^{EA}$ and $\vec{\alpha}^\xi$.

We start with the procedure of solving for $\vec{\alpha}^{EA}$. First, we build the basis functions for the finite dimensional hypothesis spaces $\mathcal{H}_k^{\xi,k}$, $\mathcal{H}_k^{A,k'}$ using piecewise polynomials or clamped B-splines as the basis functions (see Sect. 5), which altogether are represented as $\mathcal{H}^{EA}$ as in (3.15).
Remark 23 The support of the unknown interaction kernels is not assumed to be known. We build our finite dimensional subspaces, \( \mathcal{H}_{kk'}^{E} \), \( \mathcal{H}_{kk'}^{A} \), based on the empirical observation data. For the support-detection capability of our estimators, see the examples of opinion dynamics in [47, 75].

We utilize the tensor grid of basis functions, i.e., tensor product of basis functions in each dimension of the basis \([ R_{kk'}^{\min, L, M} \mathcal{S}_{kk'}, \mathcal{S}_{kk'}^{E, L, M} ] \) or \([ R_{kk'}^{\min, L, M} \mathcal{S}_{kk'}, \mathcal{S}_{kk'}^{A, L, M} ] \) is the empirical range of \( r \) given by the observation data, similarly for the empirical \( \mathcal{S}_{kk'}^{E, L, M} \) and \( \mathcal{S}_{kk'}^{A, L, M} \) being the range of \( s_{kk'}^{E} \) and \( s_{kk'}^{A} \) given by the observation, respectively. In each dimension\(^3\) of \([ R_{kk'}^{\min, L, M} \mathcal{S}_{kk'}, \mathcal{S}_{kk'}^{E, L, M} ] \) or \([ R_{kk'}^{\min, L, M} \mathcal{S}_{kk'}, \mathcal{S}_{kk'}^{A, L, M} ] \), the basis functions are built as piecewise standard polynomials (or other functions, such as Clamped B-splines, Fourier basis, etc.) uniformly with the number of basis functions being \( n_{kk'}^{E,j} \) or \( n_{kk'}^{A,j} \). Hence \( n_{kk'}^{E} = \prod_{j} n_{kk'}^{E,j} n_{kk'}^{A,j} \) and \( n_{kk'}^{A} = \prod_{j} n_{kk'}^{A,j} n_{kk'}^{E,j} \). Then, we assemble \( \tilde{d}^{(m)} \) as follows,

\[
\tilde{d}^{EA,(m)} = \begin{bmatrix}
\frac{1}{\sqrt{N_{t_1}}} \hat{x}_1(t_1) \\
\vdots \\
\frac{1}{\sqrt{N_{t_L}}} \hat{x}_L(t_L)
\end{bmatrix}.
\]

If \( \hat{x}_i(t_l) \) is not given, a finite difference scheme on \( x_i(t_l) \) or \( \hat{x}_i(t_l) \) is used to approximate \( \hat{x}_i(t_l) \). Next, we build \( \tilde{f}^{(m)} \) as follows,

\[
\tilde{f}^{EA,(m)} = \begin{bmatrix}
\frac{1}{\sqrt{N_{t_1}}} F_{\hat{x}}(x_1(t_1), \hat{x}_1(t_1), \hat{x}_1(t_1)) \\
\vdots \\
\frac{1}{\sqrt{N_{t_L}}} F_{\hat{x}}(x_L(t_L), \hat{x}_L(t_L), \hat{x}_L(t_L))
\end{bmatrix}.
\]

\(^3\) Mixture of basis functions in each dimension is possible, the algorithm does not required the basis functions in each dimension to be of the same kind. We make such assumption for simplicity sake.
Then for the learning matrix, \( \Psi^{E,A,(m)} \in \mathbb{R}^{LNd \times n} \) with \( n = n^E + n^A \). It is a concatenation of two sub-matrix, \( \Psi^{E,(m)} \) and \( \Psi^{A,(m)} \), i.e.,

\[
\Psi^{E,A,(m)} = \begin{bmatrix} \Psi^{E,(m)} & \Psi^{A,(m)} \end{bmatrix}.
\]

For the energy-based learning matrix, \( \Psi^{E,(m)} \), we use a lexicographical order on \((k, k')\) for \( k, k' = 1, \ldots, K \). We define \( n^E_{k,k',\text{prev}} = \sum_{(k_1,k_2) < (k,k')} n^E_{k_1,k_2} \); if \((k, k') = (1, 1)\), we take \( n^E_{1,1,\text{prev}} = 0 \). Then for \( \eta^E_{kk'} = 1, \ldots, n^E_{k,k'} \), \( \Psi^{E,(m)} \) is given as follows,

\[
\Psi^{E,(m)}(li(1:d), \eta^E_{kk'}) = \sum_{i' \in C_{k'}} \frac{1}{\sqrt{N_{k'}}} \psi^x_{k,k',\eta^E_{kk'}}(\|x_{i'}(t_l) - x_i(t_l)\|),
\]

and for \( l = 1, \ldots, L \). Similar process of construction is done for \( \Psi^{A,(m)} \). Then we define,

\[
A^{E,A,(m)} = (\Psi^{E,A,(m)})^T \Psi^{E,A,(m)} \quad \text{and} \quad \bar{b}^{E,A,(m)} = (\Psi^{E,A,(m)})^T (\bar{d}^{E,A,(m)} - \bar{f}^{E,A,(m)}).
\]

And lastly,

\[
A^{E,A}_M = \frac{1}{LM} \sum_{m=1}^M A^{E,A,(m)} \quad \text{and} \quad \bar{b}^{E,A} = \frac{1}{LM} \sum_{m=1}^M \bar{b}^{E,A,(m)}.
\]

Then, \( \bar{\alpha}^{E,A} = \left[(\bar{\alpha}^{E})^T \ (\bar{\alpha}^{A})^T \right]^T \), is obtained by solving

\[
A^{E,A}_M \bar{\alpha}^{E,A} = \bar{b}^{E,A}.
\]

Then, we assemble

\[
\tilde{\phi}^E_{kk'} = \sum_{\eta^E_{kk'} = 1}^{n^E_{kk'}} \alpha^E_{k,k',n^E_{kk'}} \psi^x_{k,k',\eta^E_{kk'}}.
\]

Similar assembly from \( \alpha^{A} \) is done for \( \tilde{\phi}^A_{kk'} \). In the case of using finite difference approximation to approximate the second derivatives of \( x_i \), we end up with

\[
A^{E,A}_M \tilde{\alpha}^{E,A} = \bar{b}^{E,A} + \tilde{\zeta},
\]

where \( \tilde{\zeta} = \mathcal{O}(\frac{T}{L}) \) when a first-order finite difference scheme is used.

Next for \( \tilde{\alpha}^k \), we build the basis functions for each of the finite dimensional spaces \( \mathcal{H}^k_{kk'} \), using piecewise polynomials or clamped B-splines (as in the \( EA \) case, similarly, other many other bases work well in this algorithm). This is an explicit example.
of $\mathcal{H}^\xi$. We utilize the tensor grid of basis functions, i.e., tensor product of basis functions in each dimension of the basis $[R_{kk'}^{\min,L,M}, R_{kk'}^{\max,L,M}] \times S_{kk'}^{\xi,L,M}$, where $[R_{kk'}^{\min,L,M}, R_{kk'}^{\max,L,M}]$ is the empirical range of $r$ given by the observation data, similarly for the empirical $S_{kk'}^{\xi,L,M}$ being the range of $s_{kk'}^{\xi}$ given by the observation. And in each dimension of $[R_{kk'}^{\min,L,M}, R_{kk'}^{\max,L,M}] \times S_{kk'}^{\xi,L,M}$, the basis functions are built as piecewise standard polynomials (or other functions, such as Clamped B-splines, Fourier basis, etc.) uniformly with the number of basis functions being $n_{\xi,j}$. Hence

$$n_{kk'} = \prod_{j=1}^{1+\nu_{k,k'}} n_{\xi,j}.$$ We let

$$\tilde{d}^{\xi,(m)} := \begin{bmatrix} \frac{1}{\sqrt{N_k}} \xi_1(t_1) \\ \vdots \\ \frac{1}{\sqrt{N_k}} \xi_N(t_1) \\ \vdots \\ \frac{1}{\sqrt{N_k}} \xi_1(t_L) \\ \vdots \\ \frac{1}{\sqrt{N_k}} \xi_N(t_L) \end{bmatrix}, \quad \tilde{f}^{\xi,(m)} := \begin{bmatrix} \frac{1}{\sqrt{N_k}} F^{\xi}(x_1(t_1), \dot{x}_1(t_1), \xi_1(t_1)) \\ \vdots \\ \frac{1}{\sqrt{N_k}} F^{\xi}(x_N(t_1), \dot{x}_N(t_1), \xi_N(t_1)) \\ \vdots \\ \frac{1}{\sqrt{N_k}} F^{\xi}(x_1(t_L), \dot{x}_1(t_L), \xi_1(t_L)) \\ \vdots \\ \frac{1}{\sqrt{N_k}} F^{\xi}(x_N(t_L), \dot{x}_N(t_L), \xi_N(t_L)) \end{bmatrix},$$

and

$$\Psi^{\xi,(m)}(li(1 : d), \eta^{\xi,k,k'}) = \sum_{i' \in C_{k,k'}} \psi^{\xi}_{k,k',\eta^{\xi},i'}(\|x_{i'}(t_i) - x_i(t_i)\|, s_{\xi,i'}^{\xi}(t_i) - \xi_i(t_i)), \quad i \in C_k,$$

Finally we define,

$$A^{\xi,(m)} = (\Psi^{\xi,(m)})^T \psi^{\xi,(m)} \quad \text{and} \quad \tilde{b}^{\xi,(m)} = (\Psi^{\xi,(m)})^T (\tilde{d}^{\xi,(m)} - \tilde{f}^{\xi,(m)}).$$

and

$$A^{\xi} = \frac{1}{LM} \sum_{m=1}^{M} A^{\xi,(m)} \quad \text{and} \quad \tilde{b}^{\xi} = \frac{1}{LM} \sum_{m=1}^{M} \tilde{b}^{\xi,(m)}.$$

Thus, $\alpha^{\xi}$ is obtained by solving

$$A^{\xi} \alpha^{\xi} = \tilde{b}^{\xi}.$$
Then, we assemble

\[
\hat{\phi}_{kk'}^{\xi} = \sum_{\xi_{kk'}^1 = 1}^{n_{kk'}^1} \alpha_{k,k',k_{kk'},k_{kk'}}^{\xi} \psi_{k,k',k_{kk'},k_{kk'}}^{\xi}
\]

### Appendix E: Control of trajectory error

In this section, for the convenience of the reader, we gather a few of the technical tools used in the analysis of the system. These are fundamental results necessary for developing the trajectory prediction, the measure support, and the existence and uniqueness results. We also include some of the necessary results on covering numbers of function spaces used for the learning theory.

The first theorem we present is an iterated Grönwall type result that allows us to analyze the trajectory error of the full system \( Y(t) \).

**Theorem 24** Let \( u(t), a(t), \) and \( b(t) \) be nonnegative continuous functions in \( J = [\alpha, \beta] \), and suppose that

\[
u(t) \leq a(t) + b(t) \left[ \int_{\alpha}^{t} k_1(t, t_1) u(t_1) dt_1 + \cdots + \int_{\alpha}^{t} \left( \int_{\alpha}^{t_1} \cdots \left( \int_{\alpha}^{t_{n-1}} k_n(t, t_1, \ldots, t_n) u(t_n) dt_n \right) \cdots dt_1 \right) \right]
\]

for all \( t \in J \), where \( k_i(t, t_1, \ldots, t_i) \) are nonnegative continuous functions in \( J_{i+1}, i = 1, 2, \ldots, n \), which are nondecreasing in \( t \in J \) for all fixed \( (t_1, \ldots, t_i) \in J_i, i = 1, 2, \ldots, n \). Then, for all \( t \in J \)

\[
u(t) \leq a(t) + b(t) \int_{\alpha}^{t} \hat{R}[a](t, s) \exp \left( \int_{s}^{t} \hat{R}[b](t, \tau) d\tau \right) ds
\]

where, for all \( (t, s) \in J_2 \)

\[
\hat{R}[w](t, s) = k_1(t, s) w(s) + \int_{\alpha}^{s} k_2(t, s, t_2) w(t_2) dt_2 + \sum_{i=3}^{n} \int_{\alpha}^{s} \left( \int_{\alpha}^{t_2} \cdots \left( \int_{\alpha}^{t_{n-1}} \int_{\alpha}^{t_{i-1}} k_i(t, s, t_2, \ldots, t_i) w(t_i) dt_i \right) \cdots dt_2 \right)
\]

for each continuous function \( w(t) \) in \( J \).

**Proof** See [17].

Now we are ready to present the results for trajectory prediction.
Proof of Theorem 5  We introduce the function
\[ F[\varphi^{E_A}](x, \dot{x}, s^E, s^A) := \varphi^E(||x||, s^E)x + \varphi^A(||x||, s^A)\dot{x} \]
defined on \( \mathbb{R}^{2d+p^E+p^A} \) for functions \( \varphi^E \in L^\infty([0, R] \times \mathbb{S}^E), \varphi^A \in L^\infty([0, R] \times \mathbb{S}^A) \).
Similarly, let \( F[\varphi^\xi](x, \xi, s^\xi) := \varphi^\xi(||x||, s^\xi)\xi \). By assumption, \( \hat{Y}_0 = Y_0 \) and \( \hat{Y}_0 = \hat{Y}_0 \). For every \( t \in [0, T] \), by the fundamental theorem of calculus and the triangle inequality, we have
\[
\|X_t - \hat{X}_t\|^2_S \leq \sum_{j=1}^{K} \sum_{i \in C_j} \frac{1}{N_j} \left\| \int_0^t \int_{s=0}^u (\dot{x}_i(s) - \dot{\hat{x}}_i(s)) ds du \right\|^2
\]
\[
\leq t^2 \sum_{j=1}^{K} \sum_{i \in C_j} \frac{1}{N_j} \int_0^t \int_{s=0}^u \|\ddot{x}_i(s) - \ddot{\hat{x}}_i(s)\|^2 ds du
\]
\[
= t p \int_0^t \int_{s=0}^u \|\ddot{X}_s - \ddot{\hat{X}}_s - \ddot{\hat{x}}_i(s)\|^2 ds du
\]
\[
= 2T^2 \int_0^t \int_{s=0}^u \|\ddot{X}_s - \ddot{\hat{X}}_s\|^2 ds du + 2T^2 \int_0^t \int_{s=0}^u \|\ddot{x}_i(s) - \ddot{\hat{X}}_s\|^2 ds du
\]
\[
= 2T^2 \int_0^t \int_{s=0}^u \|\ddot{x}_i(s) - \ddot{\hat{x}}_i(s)\|^2 ds du + 2T^2 \int_0^t \int_{s=0}^u \|\ddot{x}_i(s) - \ddot{\hat{x}}_i(s)\|^2 ds du
\]
(E.1)
\[
= 2T^2 \int_0^t \int_{s=0}^u \|\ddot{x}_i(s) - \ddot{\hat{x}}_i(s)\|^2 ds du + 2T^2 \int_0^t \int_{s=0}^u \|\ddot{x}_i(s) - \ddot{\hat{x}}_i(s)\|^2 ds du
\]
(E.2)

First we introduce the convenient notations of
\[
s^E_{ii'} = s^E_{(k_i, k_i')} (x_i, \dot{x}_i, \xi_i, \dot{\xi}_i, \dot{x}_i', \dot{\xi}_i'), \quad \hat{s}^E_{ii'} = \hat{s}^E_{(k_i, k_i')} (\hat{x}_i, \dot{\hat{x}}_i, \hat{\xi}_i, \dot{\hat{\xi}}_i, \dot{\hat{x}}_i', \dot{\hat{\xi}}_i') \]
(E.3)

with analogous formulae for \( s^E_{ii'}, s^A_{ii'}, s^A_{ii'} \).

Above we have introduced the term
\[
I = \left\| F[\varphi^{E_A}](X_s, V_s, \Xi_s) - F[\varphi^{E_A}](\hat{X}_s, \hat{V}_s, \hat{\Xi}_s) \right\|_S^2,
\]
which can be expressed explicitly as
\[
I = \left\| \sum_{j'=1}^{K} \sum_{i' \in C_{j'}} \frac{1}{N_{j'}} (F[\varphi^{E_A}_{jj'}](r_{ii'}(s), \dot{r}_{ii'}(s), s^E_{ii'}(s), s^A_{ii'}(s))) \right\|_S^2,
\]
Using the Lipschitz property of So that we have the bound

\[
21
\]

\[
\| \mathbf{f}^{nc, \dot{x}}(\mathbf{x}, \mathbf{V}, \mathbf{Z}) - \mathbf{f}^{nc, \dot{x}}(\widehat{\mathbf{x}}, \widehat{\mathbf{V}}, \widehat{\mathbf{Z}}) \|^2_S
\]

\[
= \sum_{j=1}^K \sum_{i \in C_j} \frac{1}{N_{ij}} \| F_{ij}^x(\mathbf{x}(s), \dot{\mathbf{x}}(s), \xi_i(s)) - F_{ij}^x(\widehat{\mathbf{x}}(s), \dot{\mathbf{x}}(s), \hat{\xi}(s)) \|^2
\]

\[
\leq \sum_{j=1}^K \sum_{i \in C_j} \frac{1}{N_{ij}} \text{Lip}^2[F_{ij}^x] \left( \| \mathbf{x}(s) - \hat{\mathbf{x}}(s) \|^2 + \| \dot{\mathbf{x}}(s) - \hat{\mathbf{x}}(s) \|^2 + \| \xi_i(s) - \hat{\xi}(s) \|^2 \right)
\]

\[
\leq \max_i \text{Lip}^2[F_{ij}^x] \| \mathbf{Y}_s - \hat{\mathbf{Y}}_s \|^2_{\mathcal{J}}
\]

So that we have the bound

\[
2T^2 \int_{u=0}^t \int_{s=0}^u \| \mathbf{f}^{nc, \dot{x}}(\mathbf{x}, \mathbf{V}, \mathbf{Z}) - \mathbf{f}^{nc, \dot{x}}(\widehat{\mathbf{x}}, \widehat{\mathbf{V}}, \widehat{\mathbf{Z}}) \|^2_S dsdu \quad \text{(E.5)}
\]

\[
\leq 2T^2 \int_{u=0}^t \int_{s=0}^u \max_i \text{Lip}^2[F_{ij}^x] \| \mathbf{Y}_s - \hat{\mathbf{Y}}_s \|^2_{\mathcal{J}} dsdu \quad \text{(E.6)}
\]

Now we break up \( I \) using the triangle inequality and get that \( I \leq I_1 + I_2 \) where

\[
I_1 = \left\| \left( \sum_{j'=1}^K \sum_{i' \in C_{j'}} \frac{1}{N_{j'}} (F[\hat{\varphi}_{jj'}^E](\mathbf{r}_{ii'}(s), \mathbf{r}_{ii'}(s), \mathbf{s}_{ii'}^E(s), \mathbf{s}_{ii'}^A(s))
\]

\[
- F[\hat{\varphi}_{jj'}^E](\mathbf{x}(s) - \mathbf{x}(s), \dot{\mathbf{x}}(s) - \dot{\mathbf{x}}(s), \mathbf{s}_{ij}^E(s), \mathbf{s}_{ij}^A(s))) \right) \right\|_{\mathcal{S}}^2
\]

\[
I_2 = \left\| \left( \sum_{j'=1}^K \sum_{i' \in C_{j'}} \frac{1}{N_{j'}} (F[\hat{\varphi}_{jj'}^E](\mathbf{r}_{ii'}(s), \mathbf{r}_{ii'}(s), \mathbf{s}_{ii'}^E(s), \mathbf{s}_{ii'}^A(s))
\]

\[
- F[\hat{\varphi}_{jj'}^E](\mathbf{r}_{ii'}(s), \mathbf{r}_{ii'}(s), \mathbf{s}_{ii'}^E(s), \mathbf{s}_{ii'}^A(s))) \right) \right\|_{\mathcal{S}}^2
\]

Using the Lipschitz property of \( F[\hat{\varphi}_{jj'}^E] \) we get that, since

\[
I_1 = \sum_{j=1}^K \sum_{i \in C_j} \frac{1}{N_{ij}} \sum_{j'=1}^K \sum_{i' \in C_{j'}} \frac{1}{N_{j'}} (F[\hat{\varphi}_{jj'}^E](\mathbf{r}_{ii'}(s), \mathbf{r}_{ii'}(s), \mathbf{s}_{ii'}^E(s), \mathbf{s}_{ii'}^A(s))
\]
\[ -F[\widehat{\phi}_{jj'}^{EA}](x_i(s) - \hat{x}_i(s), \hat{x}_i(s) - \hat{x}_i(s), s_{ii'}^E(s), s_{ii'}^A(s)) \]

then,

\[
I_1 \leq K \sum_{j=1}^{K} \sum_{i \in C_j} \frac{1}{N_j} \sum_{j'=1}^{K} \sum_{i' \in C_{j'}} \frac{1}{N_{j'}} \left( \text{Lip}[F[\widehat{\phi}_{jj'}^{EA}]](x_{i'}(s) - \hat{x}_{i'}(s), \hat{x}_{i'}(s) - \hat{x}_{i'}(s) + s_{ii'}^E(s) - s_{ii'}^A(s)) \right)^2.
\]

By the assumptions on the feature maps, we have that

\[ \|s_{ii'}^E(s) - s_{ii'}^A(s)\| \leq \text{Lip}[s_{ii'}^E(s)] \| (x_{i'}(s) - \hat{x}_{i'}(s), \hat{x}_{i'}(s) - \hat{x}_{i'}(s), x_{i'}(s) - \hat{x}_{i'}(s), \xi_{i'}(s) - \hat{\xi}_{i'}(s)) \|
\]

Combining these bounds we see that,

\[
I_1 \leq K \sum_{j=1}^{K} \sum_{i \in C_j} \frac{1}{N_j} \sum_{j'=1}^{K} \sum_{i' \in C_{j'}} \frac{1}{N_{j'}} \left( \max_{j,j'} \left( \text{Lip}[F[\widehat{\phi}_{jj'}^{EA}]](\text{Lip}[s_{ij'}^E(s)] + 1) \right)^2 \right) \left( \|x_{i'}(s) - \hat{x}_{i'}(s)\| + \|\hat{x}_{i'}(s) - \hat{\hat{x}}_{i'}(s)\| + \|\xi_{i'}(s) - \hat{\xi}_{i'}(s)\| \right)^2.
\]

Let \( \tilde{S} = \max(S_E, S_A)^2, J = (\max_{j,j'} \text{Lip}[s_{ij'}^E(s), s_{ij'}^A(s)] + 1)^2 \), and then let \( P = \tilde{S}J \) and we get by Young’s inequality that,

\[
I_1 \leq 4KP\|Y_s - \hat{Y}_s\| \tilde{S}^{\frac{1}{2}}, \tag{E.8}
\]

and performing a similar analysis we get that

\[
I_2 \leq 4KP\|Y_s - \hat{Y}_s\| \tilde{S}^{\frac{1}{2}}.
\]

So gathering terms, we can reexpress (E.1) as

\[
\left\| X_t - \hat{X}_t \right\|_{\tilde{S}}^2 \leq 2T^2 \int_{u=0}^{t} \int_{s=0}^{u} \left\| \hat{X}_s - f_{\text{NC}}(X_s, V_s, \Xi_s) - f_{\Phi}^{EA}(X_s, V_s, \Xi_s) \right\|_{\tilde{S}}^2 ds du
\]

\[ + 2T^2(F + 8KP) \int_{u=0}^{t} \int_{s=0}^{u} \left\| Y_s - \hat{Y}_s \right\|_{\tilde{S}}^2 ds du \tag{E.9}
\]

where \( F = \max_i \text{Lip}[F_i^{\hat{\phi}}] \). Performing an analogous analysis on \( \|V_t - \hat{V}_t\|_{\tilde{S}}, \|\Xi_t - \hat{\Xi}_t\|_{\tilde{S}}^2 \), with some additional effort, one can get the following result on the phase variable.
With this notation, we can rewrite the above bound as

\[ \| \Xi_t - \hat{\Xi}_t \|^2 \leq 2T(8QK + F^\xi) \int_{s=0}^{t} \| Y_s - \hat{Y}_s \|_2 ds + 2T \int_{s=0}^{t} \| \hat{\Xi}_s - \mathbf{f}^{\text{nc},\xi}(X_s, V_s, \Xi_s) + \mathbf{f}^{\Phi}(X_s, V_s, \Xi_s) \|^2_2 ds \]

(E.10)

where \( F^\xi = \max_{j} \text{Lip}[F^\xi_j] \) and \( Q = \max(H, S^\xi) \) where \( H = \max_{j,j'} \text{Lip}[s^E_{(j,j')}, s^A_{(j,j')}] \). Similarly, we have that,

\[ \| V_t - \hat{V}_t \|^2 \leq 2T \int_{s=0}^{t} \| \hat{\Xi}_s - \mathbf{f}^{\text{nc},\xi}(X_s, V_s, \Xi_s) + \mathbf{f}^{\Phi^A}(X_s, V_s, \Xi_s) \|^2_2 ds + 2T(F + 8KP) \int_{s=0}^{t} \| Y_s - \hat{Y}_s \|^2_2 ds \]

(E.11)

Gathering the bounds (E.9, E.10, E.11), we have that

\[ \| \hat{Y}_t - Y_t \|^2 \leq 2T(8KP + F + 8QK + F^\xi) \int_{s=0}^{t} \| \hat{Y}_s - Y_s \|^2_2 ds + 2T^2(8KP + F) \int_{u=0}^{t} \int_{s=0}^{u} \| \hat{Y}_s - Y_s \|^2_2 ds du \]

\[ a(t) \left\{ \begin{align*}
+ 2T^2 \int_{u=0}^{t} \int_{s=0}^{u} \| \hat{X}_s - \mathbf{f}^{\text{nc},\xi}(X_s, V_s, \Xi_s) - \mathbf{f}^{\Phi^A}(X_s, V_s, \Xi_s) \|^2_2 ds du \\
+ 2T \int_{s=0}^{t} \| \hat{X}_s - \mathbf{f}^{\text{nc},\xi}(X_s, V_s, \Xi_s) - \mathbf{f}^{\Phi^A}(X_s, V_s, \Xi_s) \|^2_2 ds \\
+ 2T \int_{s=0}^{t} \| \hat{\Xi} - \mathbf{f}^{\text{nc},\xi}(X_s, V_s, \Xi_s) - \mathbf{f}^{\Phi^A}(X_s, V_s, \Xi_s) \|^2_2 ds
\end{align*} \right. \]

where we denote the last three lines by \( a(t) \) and notice that this is a nondecreasing function in \( t \). We also denote \( A_1 = 2T(8KP + F + 8QK + F^\xi) \) and \( B_1 = 2T^2(8KP + F) \). Now use theorem 24, which is in [17] and is originally in Bainov and Simeonov. With this notation, we can rewrite the above bound as

\[ \| \hat{Y}_t - Y_t \|^2 \leq A_1 \int_{s=0}^{t} \| \hat{Y}_s - Y_s \|^2_2 ds + B_1 \int_{u=0}^{t} \int_{s=0}^{u} \| \hat{Y}_s - Y_s \|^2_2 ds du + a(t) \]

(E.12)

And so in the notation of Theorem 24 we have \( u(t) = \| \hat{Y}_t - Y_t \|^2_2 \), \( b(t) = 1 \), \( k_1(t_1, t_2) = A_1 \) and \( k_2(t_1, t_2) = B_1 \), so that for all \( t \) we have

\[ \| \hat{Y}_t - Y_t \|^2 \leq a(t) + \int_{0}^{t} \hat{R}[a](t,s) \exp \left( \int_{s}^{t} \hat{R}[b](t,\tau) d\tau \right) ds \]

and we have the simple bounds

\[ \hat{R}[a](t,s) = a(t) + \int_{0}^{s} B_1 a(t_2) dt_2 \leq a(T) + B_1 T a(T) \]

(E.13)
\[
\begin{align*}
\hat{R}[b](t, \tau) &= A_1 + \int_0^\tau 1dy = A_1 + \tau \\
\end{align*}
\]

(E.14)

So that,

\[
\|\hat{Y}_t - Y_t\|_2^2 \leq a(T) + [a(T) + B_1 T a(T)] \int_{s=0}^{t} \exp \left( \int_{s}^{t} (A_1 + \tau) d\tau \right) ds \\
\leq a(T) + [a(T) + B_1 T a(T)] \int_{s=0}^{T} \exp \left( \int_{0}^{T} A_1 + \tau d\tau \right) ds \\
= a(T) + [a(T) + B_1 T a(T)] T \exp(A_1 T + T^2/2) \\
= a(T)(1 + (1 + B_1 T) T \exp(A_1 T + T^2/2))
\]

So that we can immediately conclude the first assertion of the theorem,

\[
\sup_{t \in [0, T]} \|\hat{Y}_t - Y_t\|_2^2 \leq a(T)(1 + (T + B_1 T^2) \exp(A_1 T + T^2/2))
\]

Lastly, we can use the results of Sect. A to get the key result on the expected supremum error. We take expectation on each of the three terms of \(a(T)\) and normalize them so they are in the form of the results of A.

\[
\frac{1}{T^2} \int_{u=0}^{T} \int_{s=0}^{T} \mathbb{E}_{\mu^y} \|\tilde{X}_s - f^{nc,\xi}(X_s, V_s, \Xi_s) - \tilde{f}^{EA}(X_s, V_s, \Xi_s)\|_S^2 dsdu \leq K^2 \|\tilde{\phi}^{EA} - \phi^{EA}\|_{L^2(\rho^{EA}_T)}^2
\]

We similarly get that,

\[
\frac{1}{T} \int_{s=0}^{T} \mathbb{E}_{\mu^y} \|\tilde{X}_s - f^{nc,\xi}(X_s, V_s, \Xi_s) - \tilde{f}^{\xi}(X_s, V_s, \Xi_s)\|_S^2 ds \leq K^2 \|\tilde{\phi}^{\xi} - \phi^{\xi}\|_{L^2(\rho^{\xi}_T)}^2
\]

(E.15)

and can get an analogous bound for the remaining term of \(a(T)\). These bounds together lead to

\[
a(T) \leq (2T^4 K^2 + 2T^2 K^2) \|\tilde{\phi}^{EA} - \phi^{EA}\|_{L^2(\rho^{EA}_T)}^2 + 2T^2 K^2 \|\tilde{\phi}^{\xi} - \phi^{\xi}\|_{L^2(\rho^{\xi}_T)}^2
\]

which implies the desired result. \(\Box\)

References

1. Abaid, N., Porfiri, M.: Fish in a ring: spatio-temporal pattern formation in one-dimensional animal groups. J. R. Soc. Interface 7, 1441–1453 (2010)
2. Albi, G., Balagué, D., Carrillo, J.A., Brecht, J.V.: Stability analysis of flock and mill rings for second order models in swarming. SIAM J. Appl. Math. 74, 794–818 (2014)
3. Ballerini, M., Cabibbo, N., Candelier, R., Cavagna, A., Cisbani, E., Giardina, I., Lecomte, V., Orlandi, A., Parisi, G., Procaccini, A., Viale, M., Zdravkovic, V.: Interaction ruling animal collective behavior depends on topological rather than metric distance: evidence from a field study. Proc. Natl. Acad. Sci. U.S.A. 105, 1232–1237 (2008)

4. Bellomo, N., Degond, P., Tadmor, E. (eds.): Active Particles, vol. 1. Springer International Publishing AG, Cham (2017)

5. Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M., Walzak, A.M.: Statistical mechanics for natural flocks of birds. Proc. Natl. Acad. Sci. U.S.A. 109, 4786–4791 (2012)

6. Binev, P., Cohen, A., Dahmen, W., DeVore, R., Temlyakov, V.: Universal algorithms for learning theory part I: piecewise constant functions. J. Mach. Learn. Res. 6, 1297–1321 (2005)

7. Bishwal, J.P.N., et al.: Estimation in interacting diffusions: continuous and discrete sampling. Appl. Math. 2, 1154–1158 (2011)

8. Blodel, V., Hendricks, J., Tsitsiklis, J.: On Krause’s multi-agent consensus model with state-dependent connectivity. IEEE Trans. Autom. Control 54, 2586–2597 (2009)

9. Bongard, J., Lipson, H.: Automated reverse engineering of nonlinear dynamical systems. Proc. Natl. Acad. Sci. U.S.A. 104, 9943–9948 (2007)

10. Bongini, M., Fornasier, M., Hansen, M., Maggioni, M.: Inferring interaction rules from observations of evolutive systems I: the variational approach. Math. Models Methods Appl. Sci. 27, 909–951 (2017)

11. Boninsegnia, L., Nüske, F., Clementi, C.: Sparse learning of stochastic dynamical equations. J. Chem. Phys. 148, 241723 (2018)

12. Brunel, N.: Parameter estimation of ODEs via nonparametric estimators. Electron. J. Stat. 2, 1242–1267 (2011)

13. Brunton, S., Kutz, N., Proctor, J.: Data-drive discovery of governing physical laws. SIAM News, 50 (2017)

14. Brunton, S., Proctor, J., Kutz, J.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci. U.S.A. 113, 3932–3937 (2016)

15. Cao, J., Wang, L., Xu, J.: Robust estimation for ordinary differential equation models. Biometrics 67, 1305–1313 (2011)

16. Chen, X.: Maximum likelihood estimation of potential energy in interacting particle systems from single-trajectory data. Electron. Commun. Probab. 26, 1–13 (2021)

17. Cho, Y., Sever, S., Kim, Y.-H.: On some Gronwall type inequalities with iterated integrals. Math. Commun. 12, 63–73 (2007)

18. Chuang, Y., Huang, Y., D’Orsogna, M., Bertozzi, A.: Multi-vehicle flocking: scalability of cooperative control algorithms using pairwise potentials. In: IEEE International Conference on Robotics and Automation, pp. 2292–2299 (2007)

19. Chuang, Y.-L., Chou, T., D’Orsogna, M.R.: Swarming in viscous fluids: three-dimensional patterns in swimmer- and force-induced flows. Phys. Rev. E 93, 1–12 (2016)

20. Chuang, Y.-L., D’Orsogna, M.R., Marthaler, D., Bertozzi, A.L., Chayes, L.S.: State transitions and the continuum limit for a 2D interacting, self-propelled particle system. Physica D Nonlinear Phenom. 232, 33–47 (2007)

21. Couzin, I., Krause, J., Franks, N., Levin, S.: Effective leadership and decision-making in animal groups on the move. Nature 433, 513–516 (2005)

22. Cucker, F., Dong, J.-G.: A general collision-avoiding flocking framework. IEEE Trans. Autom. Control 56, 1124–1129 (2011)

23. Cucker, F., Drozdeki, E.: Flocking in noisy environments. J. Math. Pures Appl. 89, 278–296 (2008)

24. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bull. Am. Math. Soc. 39, 1–49 (2002)

25. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bull. Am. Math. Soc. 39, 1–49 (2002)

26. Cucker, F., Smale, S.: Emergent behavior in flocks. IEEE Trans. Autom. Control 52, 852 (2007)

27. Dahmen, W., DeVore, R., Scherer, K.: Multi-dimensional spline approximation. SIAM J. Numer. Anal. 17, 380–402 (1980)

28. Dattner, I., Khassen, C.: Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. Electron. J. Stat. 9, 1939–1973 (2015)

29. de Boor, C., DeVore, R.: Approximation by Smooth Multivariate Splines. Trans. Am. Math. Soc. 276, 775 (1983)
30. Della Maestra, L., Hoffmann, M.: The Lan property for Mckean–Vlasov models in a mean-field regime, arXiv preprint arXiv:2205.05932 (2022)
31. DeVore, R., Kerkyacharian, G., Picard, D., Temlyakov, V.: Approximation methods for supervised learning. Found. Comput. Math. 6, 3–58 (2006)
32. Feng, J., Ren, Y., Tang, S.: Data-driven discovery of interacting particle systems using gaussian processes, arXiv preprint arXiv:2106.02735 (2021)
33. Genon-Catalot, V., Larédo, C.: Inference for ergodic Mckean–Vlasov stochastic differential equations with polynomial interactions (2022)
34. Gomes, S.N., Stuart, A.M., Wolfram, M.-T.: Parameter estimation for macroscopic pedestrian dynamics models from microscopic data. SIAM J. Appl. Math. 79, 1475–1500 (2019)
35. Grégoire, G., Chaté, H.: Onset of collective and cohesive motion. Phys. Rev. Lett. 92 (2004)
36. Györfi, L., Kohler, M., Krzyzak, A., Walk, H.: A Distribution-Free Theory of Nonparametric Regression. Springer, New York (2002)
37. Han, X., Shen, Z., Wang, W., Di, Z.: Robust reconstruction of complex networks from sparse data. Phys. Rev. Lett. 114, 028701 (2015)
38. Kang, S., Liao, W., Liu, Y.: Ident: identifying differential equations with numerical time evolution, arXiv preprint arXiv:1904.03538 (2019)
39. Kasonga, R.A.: Maximum likelihood theory for large interacting systems. SIAM J. Appl. Math. 50, 865–875 (1990)
40. Katz, Y., Tunstrom, K., Ioannou, C., Huepe, C., Couzin, I.: Inferring the structure and dynamics of interactions in schooling fish. Proc. Natl. Acad. Sci. U.S.A. 108, 18720–18725 (2011)
41. Krause, U.: A discrete nonlinear and non-autonomous model of consensus formation. Commun. Differ. Equ. 2000, 227–236 (2000)
42. Li, Z., Lu, F., Maggioni, M., Tang, S., Zhang, C.: On the identifiability of interaction functions in systems of interacting particles, arXiv preprint arXiv:1912.11965 (2019)
43. Liang, H., Wu, H.: Parameter estimation for differential equation models using a framework of measurement error in regression models. J. Am. Stat. Assoc. 103, 1570–1583 (2008)
44. Long, Z., Lu, Y., Ma, X., Dong, B.: PDE-net: learning PDEs from data, arXiv preprint arXiv:1710.09668 (2017)
45. Lu, F., Maggioni, M., Tang, S.: Learning interaction kernels in heterogeneous systems of agents from multiple trajectories (2019)
46. Lu, F., Maggioni, M., Tang, S.: Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories. Found. Comput. Math., 1–55 (2021)
47. Lu, F., Zhong, M., Tang, S., Maggioni, M.: Nonparametric inference of interaction laws in systems of agents from trajectory data. Proc. Natl. Acad. Sci. U.S.A. 116, 14424–14433 (2019)
48. Lukeman, R., Li, Y., Edelstein-Keshet, L.: Inferring individual rules from collective behavior. Proc. Natl. Acad. Sci. U.S.A. 107, 12576–12580 (2010)
49. Maggioni, M., Miller, J., Zhong, M.: Agent-based learning of celestial dynamics from ephemerides (2020) (in preparation)
50. Messeneger, D.A., Bortz, D.M.: Learning mean-field equations from particle data using wsindy, arXiv preprint arXiv:2101.07821 (2021)
51. Messeneger, D.A., Wheeler, G.E., Liu, X., Bortz, D.M.: Learning anisotropic interaction rules from individual trajectories in a heterogeneous cellular population, arXiv preprint arXiv:2204.14141 (2022)
52. Miao, H., Xia, X., Perelson, A., Wu, H.: On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM Rev. 53, 3–39 (2011)
53. Mostch, S., Tadmor, E.: Heterophilious dynamics enhances consensus. SIAM Rev. 56, 577–621 (2014)
54. O’Keeffe, K., Bettstetter, C.: A review of swarmalators and their potential in bio-inspired computing, p. 85 (2019)
55. O’Keeffe, K.P., Evers, J.H., Kolokolnikov, T.: Ring states in swarmalator systems. Phys. Rev. E 98 (2018)
56. O’Keeffe, K.P., Hong, H., Strogatz, S.H.: Oscillators that sync and swarm. Nat. Commun. 8, 1–12 (2017)
57. Raissi, M.: Deep hidden physics models: deep learning of nonlinear partial differential equations. J. Mach. Learn. Res. 19, 932–955 (2018)
58. Raissi, M., Karniadakis, G.: Hidden physics models: machine learning of nonlinear partial differential equations. J. Comput. Phys. 357, 125–141 (2018)
59. Ramsay, J., Hooker, G., Campbell, D., Cao, J.: Parameter estimation for differential equations: a generalized smoothing approach. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 69, 741–796 (2007)
60. Rudy, S., Brunton, S., Proctor, J., Kutz, N.: Data-driven discovery of partial differential equations. Sci. Adv. 3, e1602614 (2017)
61. Schaeffer, H., Tran, G., Ward, R.: Extracting sparse high-dimensional dynamics from limited data. SIAM J. Appl. Math. 78, 3279–3295 (2018)
62. Schmidt, M., Lipson, H.: Distilling free-form natural laws from experimental data. Science 324, 81–85 (2009)
63. Schumaker, L.: Spline Functions: Basic Theory, 3rd edn. Cambridge University Press, Cambridge (2007)
64. Sharrock, L., Kantas, N., Parpas, P., Pavliotis, G.A.: Parameter estimation for the mckean-vlasov stochastic differential equation, arXiv preprint arXiv:2106.13751 (2021)
65. Shu, R., Tadmor, E.: Anticipation breeds alignment, arXiv preprint arXiv:1905.00633 (2019)
66. Strogatz, S.H.: From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. Physica D 143, 1–20 (2000)
67. Tonstrom, K., Katz, Y., Ioannou, C.C., Huepe, C., Kutz, M.J., Couzin, I.D.: Collective states, multi-stability and transitional behavior in schooling fish. Comput. Biol. 9 (2013)
68. Tran, G., Ward, R.: Exact recovery of chaotic systems from highly corrupted data. Multiscale Model. Simul. 15, 1108–1129 (2017)
69. Tsybakov, A.: Introduction to Nonparametric Estimation, 1st edn. Springer, New York (2008)
70. van der Vaart, A., Wellner, J.: Weak Convergence and Empirical Processes with Applications to Statistics, 1st edn. Springer, New York (1996)
71. Varah, J.: A spline least squares method for numerical parameter estimation in differential equations. SIAM J. Sci. Stat. Comput. 3, 28–46 (1982)
72. Vicsek, T., Czirók, A., Ben-Jacob, E., Cohen, I., Shochet, O.: Novel type of phase transition in a system of self-driven particles. Phys. Rev. Lett. 75, 1226–1229 (1995)
73. Yao, R., Chen, X., Yang, Y,: Mean-field nonparametric estimation of interacting particle systems, arXiv preprint arXiv:2205.07937 (2022)
74. Zhang, S., Lin, G.: Robust data-driven discovery of governing physical laws with error bars. Proc. R. Soc. A Math. Phys. Eng. Sci. 474, 20180305 (2018)
75. Zhong, M., Miller, J., Maggioni, M.: Data-driven discovery of emergent behaviors in collective dynamics. Physica D Nonlinear Phenom., 132542 (2020)