Journal of Machine Learning Research 22 (2021) 1-67 Submitted 10/19; Revised 1/21; Published 1/21

Learning interaction kernels in heterogeneous systems of agents from multiple trajectories

Fei Lu[1,3] FEILU@MATH.JHU.EDU

Mauro Maggioni[1,2,3] MAUROMAGGIONIJHU@ICLOUD.COM

Sui Tang[1,4*] SUITANG@MATH.UCSB.EDU

[1] *Departments of Mathematics,* [2] *Applied Mathematics and Statistics,*
[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*

Editor: Jie Peng

Abstract

Systems of interacting particles, or agents, have wide applications in many disciplines, including Physics, Chemistry, Biology and Economics. These systems are governed by interaction laws, which are often unknown: estimating them from observation data is a fundamental task that can provide meaningful insights and accurate predictions of the behaviour of the agents. In this paper, we consider the inverse problem of learning interaction laws given data from multiple trajectories, in a nonparametric fashion, when the interaction kernels depend on pairwise distances. We establish a condition for learnability of interaction kernels, and construct an estimator based on the minimization of a suitably regularized least squares functional, that is guaranteed to converge, in a suitable L^2 space, at the optimal min-max rate for 1-dimensional nonparametric regression. We propose an efficient learning algorithm to construct such estimator, which can be implemented in parallel for multiple trajectories and is therefore well-suited for the high dimensional, big data regime. Numerical simulations on a variety examples, including opinion dynamics, predator-prey and swarm dynamics and heterogeneous particle dynamics, suggest that the learnability condition is satisfied in models used in practice, and the rate of convergence of our estimator is consistent with the theory. These simulations also suggest that our estimators are robust to noise in the observations, and can produce accurate predictions of trajectories in large time intervals, even when they are learned from observations in short time intervals.

Keywords: Interacting particle systems; inverse problems; Monte Carlo sampling; regularized least squares; nonparametric statistics.

1. Introduction

Systems of interacting particles and agents arise in a wide variety of disciplines including interacting particle systems in Physics (see D'Orsogna et al. (2006); Vicsek et al. (1995);

* Corresponding author

Chuang et al. (2007); Bellomo et al. (2017)), predator-swarm systems in Biology (see Hemelrijk and Hildenbrandt (2011); Toner and Tu (1995)) , and opinions on interacting networks in social science (see Olfati-Saber and Murray (2004); Mostch and Tadmor (2014); Ke et al. (2002)). The interacting system is often modeled by a system of ODEs where the form of the equations is given and the interaction kernel in those equation is constructed based on a priori information, experience, and data. Inference of these interaction kernels, ideally even when little or no a priori information is given, is useful for modeling and predictions, and yet it is a fundamental challenge. In the past, due to the limited amount of data, the estimation of interaction kernels often relied on strong a priori assumptions, which reduced the problem to estimating a small number of scalar parameters indexing a small, given family of possible kernels. The increased collection and availability of data, computational power, and data storage, makes it interesting to develop techniques for the automatic discovery of interaction laws from data, under minimal assumptions on the form of such laws. We consider the following inverse problem: given trajectory data of a particle/agent system, collected from different experiment trials, how to discover the interaction law? We use tools from statistical and machine learning to propose a learning algorithm guided by rigorous analysis to estimate the interaction kernels, and accurately predict the dynamics of the system using the estimated interaction kernels.

Many governing laws of complex dynamical processes are presented in the form of (ordinary or partial) differential equations. The problem of learning differential equations from data, including those arising from systems of interacting agents, has attracted continuous attention of researchers from various disciplines. Pioneering work of learning interaction kernels in system of interacting agents can be found in the work of Lukeman et al. (2010); Katz et al. (2011); Wang et al. (2018), where the true kernels are assumed to lie in the span of a predetermined set of template functions and parametric regression techniques are used to recover the kernels either from real or the synthetic trajectory data. In Bongini et al. (2017), the authors considered learning interaction kernels from data collected from a single trajectory in a nonparametric fashion, and a recovery guarantee is proved in the limit of the number of agents going to infinity. For nonlinear coupled dynamical systems, symbolic regression techniques have been employed to reveal governing laws in various systems from experiment trajectory data sets without a priori knowledge of the underlying dynamics (see Bongard and Lipson (2007); Schmidt and Lipson (2009)). Recently, the problem of learning high dimensional nonlinear differential equations from the synthetic trajectory data, where the dynamics are governed by a few numbers of active terms in a prescribed large dictionary, has received significant attention. Sparse regression approaches include SINDy (Brunton et al. (2016); Rudy et al. (2017); Brunton et al. (2017); Boninsegna et al. (2018)), LASSO (Schaeffer et al. (2013); Han et al. (2015); Kang et al. (2019)), threshold sparse Bayesian regression (Zhang and Lin (2018)), have been shown to enable effective identification of the active terms in the underlying ODEs or PDEs, given the trajectory data. In some cases, recovery guarantees have been established under suitable assumptions on the noise in the observation and randomness of data (Tran and Ward (2017); Schaeffer et al. (2018)). There are also approaches using deep learning techniques to learn ODEs (see Raissi et al. (2018); Rudy et al. (2019)) and PDEs (see Raissi (2018); Raissi and Karniadakis (2018); Long et al. (2018)) from the synthetic trajectory/solution data sets or other types of observations (e.g. of boundary values). In the Statistics community, parameter estimation for

differential equations from trajectory data has been studied, among others, in Varah (1982); Brunel (2008); Liang and Wu (2008); Cao et al. (2011); Ramsay et al. (2007), and references therein. The form of the differential equations is given, and the unknown is the (possibly high-dimensional) parameter $\boldsymbol{\theta}$. Approaches include the trajectory matching method, that chooses $\boldsymbol{\theta}$ so as to maximize agreement with trajectory data; the gradient matching method that seeks $\boldsymbol{\theta}$ to fit the right-hand side of the ODEs to the velocities of the trajectory data; and the parameter cascading method that combines the virtues of these two methods, while avoiding the heavy computational overhead of the trajectory matching method and is applicable to partial/indirect observations which the gradient matching method currently can not handle. A review of the parameter estimation problem in systems of ODEs may be found in Ramsay and Hooker (2018). The identifiability of $\boldsymbol{\theta}$ in general systems of nonlinear ODEs from trajectory data is challenging and a topic of current research. It is often assumed that the parameters are identifiable from the trajectory data, i.e, different $\boldsymbol{\theta}$ would yield different trajectories during the observation period. There is no easy way to check this assumption from data, and characterizations of identifiability exist for some special cases (e.g., see Dattner and Klaassen (2015) for systems of ODEs with a linear dependence on (known functions of) the parameters. We refer the interested reader to Miao et al. (2011) and references therein for a comprehensive survey of this topic.

In this paper, we restrict our attention to learning governing laws in first order particle-/agent-based systems with pairwise interactions, whose magnitudes only depend on pairwise interactions and mutual distances. We consider an N-agent system with K types of agents in the Euclidean space $\mathbb{R}^d$; we denote by $\{C_k\}_{k=1}^K$ the partition of the set of indices $\{1,\dots,N\}$ of the agents corresponding to their type. The agents evolve according to the system of ODEs :

$$\dot{\boldsymbol{x}}_i(t) = \sum_{i'=1}^{N} \frac{1}{N_{\mathcal{k}_{i'}}} \phi_{\mathcal{k}_i \mathcal{k}_{i'}}(||\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)||)(\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)) \,, i = 1, \cdots, N \,. \tag{1.1}$$

where $\dot{\boldsymbol{x}}_i(t) := \frac{d}{dt}\boldsymbol{x}_i(t)$;$\mathcal{k}_i$ is the index of the type of agent i, i.e. $i \in C_{\mathcal{k}_i}$; $N_{\mathcal{k}_{i'}}$ is the number of agents of type $C_{\mathcal{k}_{i'}}$. The *interaction kernel* $\phi_{\mathcal{k}_i \mathcal{k}_{i'}} : \mathbb{R}_+ \to \mathbb{R}$ governs how agents in type $C_{\mathcal{k}_i}$ influence agents in type $C_{\mathcal{k}_{i'}}$; in particular, note that it is indexed by the ordered pair $(\mathcal{k}_i, \mathcal{k}_{i'})$: this is natural, for example in modeling prey-predator interactions. In this system of ODEs, the velocity of each agent is obtained by superimposing the interactions with all the other agents, with each interaction being in the direction to the other agent, weighted by the interaction kernel evaluated at the distance to the other agent. We will let $r_{ii'} := \|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|$, $\boldsymbol{r}_{ii'} := \boldsymbol{x}_{i'} - \boldsymbol{x}_i$. We assume that the interaction kernels $\phi_{\mathcal{k}_i \mathcal{k}_{i'}}$ are the only unknown factors in (1.1); in particular, the sets C_k's are known. The notation used is summarized in Table 1.

We let $\boldsymbol{X}(t) := (\boldsymbol{x}_i(t))_{i=1}^N$ in the state space $\mathbb{R}^{dN}$ be the vector describing the state of the system; we let $\boldsymbol{\phi} := (\phi_{kk'})_{k,k'=1}^{K,K}$ denote the interaction kernels, and $\mathbf{f}_{\boldsymbol{\phi}}(\boldsymbol{X}) \in \mathbb{R}^{dN}$ be the vectorization of the right hand side of (1.1), which can then be rewritten in the form

$$\dot{\boldsymbol{X}}(t) = \mathbf{f}_{\boldsymbol{\phi}}(\boldsymbol{X}(t)). \tag{1.2}$$

The observational data $\boldsymbol{X}_{\mathrm{traj},M,\mu_0} := \{\boldsymbol{X}^{(m)}(t_l), \dot{\boldsymbol{X}}^{(m)}(t_l)\}_{l,m=1,1}^{L,M}$ consist of the positions and velocities of all agents, observed at time $0 = t_1 < \cdots < t_L = T$, along multiple

Variable	Definition
$\boldsymbol{x}_i(t) \in \mathbb{R}^d$	state vector (position, opinion, etc.) of agent i at time t
$\|\cdot\|$	Euclidean norm in $\mathbb{R}^d$
$\boldsymbol{r}_{ii'}(t), \boldsymbol{r}_{ii''}(t) \in \mathbb{R}^d$	$\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)$, $\boldsymbol{x}_{i''}(t) - \boldsymbol{x}_i(t)$
$r_{ii'}(t), r_{ii''}(t) \in \mathbb{R}^+$	$r_{ii'}(t) = \|\boldsymbol{r}_{ii'}(t)\|, r_{ii''}(t) = \|\boldsymbol{r}_{ii'}(t)\|$
N	number of agents
K	number of types
N_k	number of agents in type k
k_i	type of agent i
C_k	the set of indices of the agents of type k
$\phi_{kk'}$	interaction kernel for the influence of agents of type k on agents of type k'

Table 1: Notation for first-order models

trajectories, indexed by m, started from different initial conditions $\{\boldsymbol{X}^{(m)}(0)\}_{m=1}^{M}$, which we assumed to be i.i.d. samples from a probability measure μ_0 on the state space $\mathbb{R}^{dN}$. For simplicity of notation, we consider the case where the times t_l are equi-spaced, but minor and straightforward modifications yield the general case.

The goal is to estimate the interaction kernels $(\phi_{kk'})_{k,k'=1,1}^{K,K}$ from $\boldsymbol{X}_{\mathrm{traj},M,\mu_0}$ and predict the dynamics given a new initial condition drawn from μ_0. The case of $K = 1$ and $M = 1$ has been considered in Bongini et al. (2017) in the mean field regime ($N \to \infty$); here we consider the case when the number of agents N, as well as the number L of observations per trajectory are fixed, and the number of observed trajectories $M \to \infty$, in the more general case of agents of multiple types, for some instances of which the mean-field theory is fraught with difficulties. No ergodicity requirement on the dynamical system is made, and the time horizon T in which the observations are made if fixed and may be small.

Inspired by the work in Bongini et al. (2017), in our recent work we introduced a risk functional Lu et al. (2019b) that exploits the structure of the system (1.1), and minimize it over a hypothesis function class $\boldsymbol{\mathcal{H}}_M$, to obtain estimators of the true kernels $\boldsymbol{\phi}$:

$$\begin{aligned} \boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi}) :&= \frac{1}{ML} \sum_{m=1,l=1}^{M,L} \left\| \dot{\boldsymbol{X}}^{(m)}(t_l) - \mathbf{f}_{\boldsymbol{\varphi}}(\boldsymbol{X}^{(m)}(t_l)) \right\|_{\mathcal{S}}^2 , \\ \widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M} :&= \operatorname*{arg\,min}_{\boldsymbol{\varphi} \in \boldsymbol{\mathcal{H}}_M} \boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi}) \end{aligned} \tag{1.3}$$

where $\boldsymbol{\varphi} = (\varphi_{kk'})_{k,k'=1}^{K,K} \in \boldsymbol{\mathcal{H}}_M$, $\mathbf{f}_{\boldsymbol{\varphi}}$ denotes the right hand side of (1.2) with the interaction kernels $\boldsymbol{\varphi}$, and $\|\cdot\|_{\mathcal{S}}$ is chosen to equalize the contributions across types accounting for possibly different N_k's:

$$\|\boldsymbol{X}\|_{\mathcal{S}}^2 = \sum_{i=1}^{N} \frac{1}{N_{k_i}} \|\boldsymbol{x}_i\|^2 . \tag{1.4}$$

The weights play a role in balancing the risk functional across different types of agents, and it is particularly important in cases where types have dramatically different cardinalities. For example, in a predator-swarm system with a large number of preys but only a small number of predators, the unweighted error functional would pay most attention to

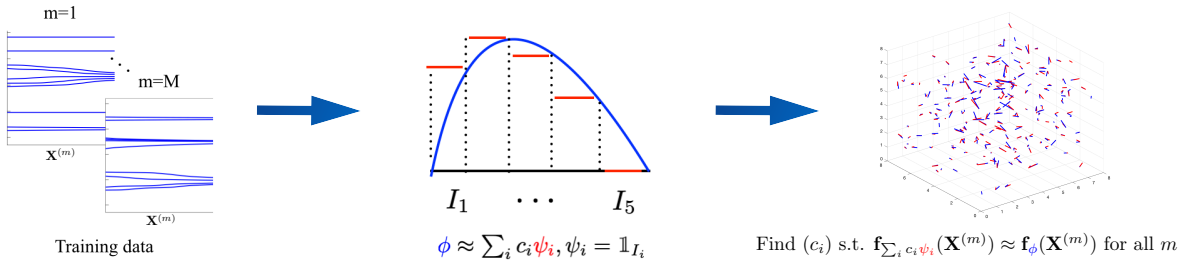

Figure 1: Overview of the learning approach for homogeneous systems. Left: training data. Middle: $\{\psi_i\}$: indicator functions. Right: the blue lines represent the velocity field generated by the true kernel ϕ. The red lines represent the velocity field generated by the estimator.

learning the interaction kernels corresponding to the preys, mostly disregarding the learning of the interaction kernels corresponding to predators, leading to inferior performance in the prediction of the trajectories of the system.

Notation	Definition
M	number of trajectories
L	number of time instances in $[0,T]$
μ_0	initial conditions sampled from μ_0
$\boldsymbol{X}_{\text{traj},M,\mu_0}$	empirical observations
$\|\boldsymbol{X}\|_{\mathcal{S}}$	$\sum_{i=1}^{N} \frac{1}{N_{k_i}} \|\boldsymbol{x}_i\|^2$
$\boldsymbol{\mathcal{E}}_M(\cdot)$	empirical error functional
$\boldsymbol{\phi} = (\phi_{kk'}), k, k' = 1, \cdots, K$	true interaction kernels
$\mathcal{H}_{kk'}$	the hypothesis spaces for $\phi_{kk'}$
$\{\psi_{kk',p}\}_{p=1}^{n_{kk'}}$	basis for $\mathcal{H}_{kk'}$
$\boldsymbol{\mathcal{H}} = \oplus_{kk'}\mathcal{H}_{kk'}, k, k' = 1, \cdots, K$	the hypothesis spaces for $\boldsymbol{\phi}$
$\boldsymbol{\varphi} = (\varphi_{kk'}), k, k' = 1, \cdots, K$	$\boldsymbol{\varphi} \in \boldsymbol{\mathcal{H}}$ with $\varphi_{kk'} \in \mathcal{H}_{kk'}$
$\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}}$	$\text{argmin}_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}}\mathcal{E}_M(\boldsymbol{\varphi})$

Table 2: Notation used in learning approach

A key question addressed in this paper can be informally summarized as:

(Q) For which choice of the hypothesis spaces $\{\boldsymbol{\mathcal{H}}_M\}$ does $\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M} \to \boldsymbol{\phi}$ for $M \to \infty$? For such choices, in which norm does the convergence hold, and what is the rate of convergence?

Our learning problem is closely related to a classical nonparametric regression problem considered by the statistics and machine learning community: given samples $\{(\boldsymbol{z}_i, g(\boldsymbol{z}_i) + \boldsymbol{\epsilon}_i)\}_{i=1}^{M}$ with the $(\boldsymbol{z}_i, \boldsymbol{\epsilon}_i)$'s drawn i.i.d from an unknown joint distribution $\boldsymbol{\rho}$ defined on the sample space, and the noise term satisfies $\mathbb{E}[\boldsymbol{\epsilon}_i] = 0$, the goal is to learn an unknown function $g : \mathbb{R}^D \to \mathbb{R}$ with prior assumptions on its regularity (e.g, g is s-Hölder). A common approach (see Cucker and Smale (2002); Györfi et al. (2002)) is to choose an hypothesis class $\mathcal{F}_M$ depending on the sample size and the regularity of g, and then define $\widehat{g}_{M,\mathcal{F}_M}$ as

the minimizer of the empirical risk functional

$$\widehat{g}_{M,\mathcal{F}_M} := \arg\min_{f\in\mathcal{F}_M} \frac{1}{M}\sum_{i=1}^{M}(g(\boldsymbol{z}_i)+\boldsymbol{\epsilon}_i - f(\boldsymbol{z}_i))^2.$$

If we let $\boldsymbol{z}=\boldsymbol{X}, \boldsymbol{\epsilon}=0$ and $g=\mathbf{f}_{\boldsymbol{\phi}}(\boldsymbol{X})$ in the regression setting above, our trajectory data is of the type needed for regression. However, our data are correlated in time due to the underlying dynamics. Even if we ignore the the lack of independence (e.g., consider $L=1$), the application of the existing regression approaches to learning $\mathbf{f}_{\boldsymbol{\phi}}(\boldsymbol{X}):\mathbb{R}^{dN}\to\mathbb{R}^{dN}$ with noisy observations would lead at best to the optimal min-max rate $O(M^{-\frac{1}{dN}})$, showing the effect of the curse of dimensionality of the state space. This significantly restricts their usability as soon as, say, $dN\geq 16$. While sparsity can provably help ameliorate the curse of dimensionality by the use of well-chosen dictionaries, and so could nonlinear constructions such as neural networks, in practice choices of dictionaries and architectures may not be obvious, and perhaps in some cases even possible (see e.g. empirical results of applying SINDy and simple neural networks in Appendix 7).

We proceed in a different direction: we take the structure of the system of equations (1.1) into account, as well as its symmetries (over permutations of agents of each type, and over translations) and move our regression target to the interaction kernels $\boldsymbol{\phi}=(\phi_{kk'})_{k,k'=1,1}^{K,K}$, to take advantage of the fact that each interaction kernel $\phi_{kk'}$ is a function of one variable only, being defined on $\mathbb{R}^+$. This becomes an inverse problem. The observed variables for each interaction kernel $\phi_{kk'}$ are the pairwise distances $\{r_{ii'}\}_{i\in C_k, i'\in C_{k'}}$, and even in the case of $L=1$, their samples are correlated (e.g. across indices, $r_{ii'}$ and $r_{ii''}$, as well as in time). For $L>1$ the nonlinear forward map of the dynamical system creates complicated dependences. This is in contrast with the i.i.d assumption on the samples of observation variables in a classical regression setting. Furthermore, the values of $\boldsymbol{\phi}$ at the observed variables are not measured directly, but rather linear combinations thereof (the r.h.s. of the system of ODEs) are observed. Constraining upon such observations leads to a system of equations, that is typically underdetermined given the trajectory data (see analysis in section 2.1). This may cause severe ill-posedness in our inverse problem.

Finally, we remark that in this work we consider the regime where either the velocity is observed, or $\frac{T}{L}$ is sufficiently small. In our numerical section 3.1, we consistently use finite differences to obtain accurate estimates of the velocity data from position data (with the velocities being unobserved). While the learning theory framework is valid as long as we have accurate velocity data, in the theory part, for simplicity, we assume the velocity data are observed. In Lu et al. (2020) stochastic interacting particle systems are considered, for which the velocity is unobserved (in fact, it does not exist in the classical sense), and the error due to the approximation of velocities is analyzed.

1.1 Contribution and Main results

The main focus of this work is to provide a theoretical foundation for the nonparametric learning of the interaction kernel in agent-based systems from multiple trajectories. This work is built on the recent works Bongini et al. (2017); Lu et al. (2019b), where the problem was introduced, in various regimes of observations. The numerical experiments in Lu et al.

(2019b) demonstrated the effectiveness of the proposed approach on a wide variety of agent-based systems, with initial theoretical results for the special case $K = 1$ (i.e., homogeneous systems). The new contributions of this work are: (i) the theoretical framework is generalized, with full details, to cover the more general and widely used *heterogeneous* systems (with K different types of agents), including a new analysis of learnability and consistency for multiple kernels. Our results confirm (at least, as far as upper bounds go) that the problem of learning multi-type interaction kernels requires stronger assumptions (in terms of a multi-type coercivity condition) and larger sample size; (ii) we add numerical validation of the learning theory on three representative heterogeneous systems that are used in practice, that provide empirical evidence of the learnability of kernels, of the consistency of the estimator, of the near-optimal convergence rate of our estimators, and of the decay rate of trajectories prediction errors; (iii) we also test the robustness of the learning approach with respect to multiple type of noise in the observations

We leverage classical nonparametric regression techniques (e.g. Fan and Gijbels (1996); Cucker and Smale (2002); Binev et al. (2005); Györfi et al. (2002)), by using the coercivity condition to ensure well-conditioned learnability of the interaction kernels, and by introducing a dynamics-adapted probability measure $\boldsymbol{\rho}_T^L$ on the pairwise distance space. We use $\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)$ as the function space for learning; the performance of the estimator $\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}$ is evaluated by studying its convergence in probability and in expectation as the number of observed trajectories M increases, by providing bounds on

$$P_{\mu_0}\{\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}(r)r - \boldsymbol{\phi}(r)r\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \geq \epsilon\} \quad \text{and} \quad \mathbb{E}_{\mu_0}[\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}(r)r - \boldsymbol{\phi}(r)r\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)}], \quad (1.5)$$

where both the probability and expectation are taken with respect to μ_0, the distribution of the initial conditions of the observed trajectories. Under the coercivity condition, the estimators $\{\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}\}_{M=1}^{\infty}$ obtained in (1.3) are strongly consistent and converge at an optimal min-max rate to the true kernels $\boldsymbol{\phi}$ in terms of M, as if we were in the in-principle-easier (both in terms of dimension and observed quantities) 1-dimensional nonparametric regression setting with noisy observations. We therefore avoid the curse of dimensionality of the state space. Furthermore, in the case of $L = 1$ and that μ_0 is satisfies a suitable exchangeability condition and is Gaussian, we prove that the coercivity condition holds, and show that even the constants in the error bound can be independent of N, making the bounds essentially dimension-free. Numerical simulations suggest that the coercivity condition holds for even larger classes of interaction kernels and initial distributions, and for different values of L as long as $\boldsymbol{\rho}_T^L$ is not degenerate (see Li et al. (2021) for a recent investigation of the coercivity condition).

We exhibit an efficient algorithm to compute the estimators based on the regularized least-squares problem (1.3), and demonstrate the learnability of interaction kernels on various systems, including opinion dynamics, predator-swarm dynamics and heterogeneous particle dynamics. Our theory results holds for rather general hypothesis function spaces, with a wide variety of choices. In our numerical section we shall use local basis functions consisting of piece-wise polynomials due to their simplicity and ease of efficient computation. The numerical results are consistent with the convergence rate from the learning theory and demonstrate its applicability. In particular, the convergence rate has no dependency on the dimension of the state space of the system, and therefore avoids the curse of dimensionality and makes these estimators well-suited for the high dimensional data regime. The numerical

results also suggest that our estimators are robust to noise and predict the true dynamics faithfully, in particular, the collective behaviour of agents, in a large time interval, even though they are learned from trajectory data collected in a very short time interval.

1.2 Discussion and future work

The regime we considered in this manuscript is that the data is collected from M independent short trajectories, and the convergence rate of our estimators is with respect to M. In such a regime, we require neither that the system to be ergodic nor the trajectory to be long, nor the number of observations along a trajectory to be large. It is different from regime where the data is collected from a long trajectory of an ergodic system. There is of course a natural connection between these two regimes, though, when the system is ergodic. In that case we may view the long trajectory as many short trajectories starting from initial conditions sampled from the invariant measure, immediately obtaining similar results using our framework. We need the proper initial distributions such that the coercivity condition holds true and therefore the inverse problem is well-posed. In fact, many types of distributions, particularly the common distributions such as Gaussian or uniform distribution, satisfy the coercivity condition, and therefore ensure the identifiability of the interaction kernel. Our regime is of interest in many applications either when the model is not ergodic, or when the invariant measure is unavailable, or when observation naturally come in the form of multiple, relatively short trajectories.

Another important issue in our learning problem is the effective sample size (ESS) of the data, particularly the ESS with respect to the regression measure ρ_T^L. We have explored this issue in Lu et al. (2019b), as well as in the case of stochastic interacting particle systems in Lu et al. (2020). For non-ergodic systems (which are the focus of this manuscript), our results in Lu et al. (2019b) (see e.g. Fig.S16) suggest that the ESS does not necessarily increase in L because the $L^2(\rho_T^L)$ error does not always decrease in ML. For ergodic stochastic systems, we obtain in Lu et al. (2020) optimal convergence rate in ML, as suggested by the ergodic theory that the ESS is proportional to ML.

The learning theory and algorithm developed in this paper could be extended to an even larger variety of agent-based systems, including second-order systems Miller et al. (2020), stochastic interacting agent systems Lu et al. (2020), and discrete-time systems with non-smooth interaction kernels; these cases require different analyses, and will be explored in separate works.

1.3 Notation

Throughout this paper, we use bold letters to denote the vectors or vector-valued functions. Let $\mathcal{K}$ be a compact (or precompact) set of Euclidean space; Lebesgue measure will be assumed unless otherwise specified. We define the following function spaces

- $L^\infty(\mathcal{K})$: the space of bounded scalar valued functions on $\mathcal{K}$ with the infinity norm

$$\|g\|_\infty = \operatorname*{ess\,sup}_{x\in\mathcal{K}} |g(x)|;$$

- $\boldsymbol{L}^\infty(\mathcal{K}) := \bigoplus_{k,k'=1,1}^{K,K} L^\infty(\mathcal{K})$ with $\|\boldsymbol{f}\|_\infty = \max_{k,k'} \|f_{kk'}\|_\infty, \forall \boldsymbol{f} \in \boldsymbol{L}^\infty(\mathcal{K})$;

- $C(\mathcal{K})$: the closed subspace of $L^\infty(\mathcal{K})$ consisting of continuous functions;
- $C_c(\mathcal{K})$: the set of functions in $C(\mathcal{K})$ with compact support;
- $C^{k,\alpha}(\mathcal{K})$ for $k \in \mathbb{N}, 0 < \alpha \leq 1$: the space of k times continuously differentiable functions whose k-th derivative is Hölder continuous of order α. In the special case of $k = 0$ and $\alpha = 1$, $g \in C^{0,1}(\mathcal{K})$ is called Lipschitz function space, and denote by $\mathrm{Lip}(\mathcal{K})$. The Lipschitz seminorm of $g \in \mathrm{Lip}(\mathcal{K})$ is defined as $\mathrm{Lip}[g] := \sup_{x \neq y} \frac{|g(x)-g(y)|}{\|x-y\|}$.

We use $||\cdot||_\infty$ as the default norm to define the compactness of sets in $\boldsymbol{L}^\infty(\mathcal{K})$ and its subspaces. The prior on the interaction kernels $\boldsymbol{\phi}$ is that $\boldsymbol{\phi}$ belong to a compact set of $\mathbf{Lip}(\mathcal{K}) := \bigoplus_{k,k'=1,1}^{K,K} \mathrm{Lip}(\mathcal{K})$. The regularity condition is presented either by specifying a compact set in a function class (e.g, $C^{k,\alpha}(\mathcal{K})$) or by quantifying the rate of approximation by a chosen sequence of linear spaces. We will restrict the estimators to a function space that we call the *hypothesis space*; $\boldsymbol{\mathcal{H}}$ will be a subset of $\boldsymbol{L}^\infty(\mathcal{K})$, where we show that the minimizer of (1.3) exists. We will focus on the compact (finite- or infinite-dimensional) subset of $\boldsymbol{L}^\infty(\mathcal{K})$ in the theoretical analysis, however in the numerical implementation we will use finite-dimensional linear spaces. While these linear subspaces are not compact, it is shown that the minimizers over the whole space behave essentially in the same way as the minimizers over compact sets of linear subspaces (e.g., see Theorem 11.3 in Györfi et al. (2002)). We shall therefore assume the compactness of the hypothesis space in the theoretical analysis, following the spirit of Cucker and Smale (2002); Binev et al. (2005); DeVore et al. (2004).

1.4 Relevant background on interacting agent systems

The interacting agent systems considered in this paper follow the equations of motion (1.1). These equations may be derived from physical laws for particle dynamics in gradient form: we can think of each agent as a particle and for type-k agents, there is an associated potential energy function depending only on the pairwise distance between agents:

$$\mathcal{U}_k(\boldsymbol{X}(t)) := \sum_{i \in C_k} \left(\sum_{i' \in C_k} \frac{1}{2N_k} \Phi_{kk}(r_{ii'}(t)) + \sum_{i' \notin C_k} \frac{1}{N_{k_{i'}}} \Phi_{k k_{i'}}(r_{ii'}(t)) \right) \tag{1.6}$$

with $\phi_{k k_{i'}}(r) = \Phi'_{k k_{i'}}(r)/r$. The evolution of agents in each type is driven by the minimization of this potential energy function. This relationship justifies the appearance of the functions $\{\phi_{kk'}(r)r\}_{k,k'=1}^{K,K}$ in (1.5) and in our main results.

In the special case of $K = 1$, the system is said to be *homogeneous*. It is one of the simplest models of interacting agents, yet it can yield complex, emergent dynamics (Kolokolnikov et al. (2013)). A prototypical example is opinion dynamics in social sciences, where the interaction kernel could be an increasing or decreasing positive function, modeling, respectively, heterophilious and homophilous opinion interactions (Mostch and Tadmor (2014)), or particle systems in Physics where all particles are identical (e.g. a monoatomic metal). In the case of $K \geq 2$, the system is said to be *heterogeneous*. Prototypical examples include interacting particle systems with different particle types (e.g. composite materials with multiple atomic types) and predator-prey systems, where the interaction kernels

may be negative for small r, inducing the repulsion, and positive for large r, inducing the attraction.

There has been a line of research fitting real data in systems of form (1.1) across various disciplines. We refer readers the application in chemistry using Lennard Jones potential to Cisneros et al. (2016), application in the exploratory data analysis for animal movement to Lukeman et al. (2010); Brillinger et al. (2011, 2012). When $K = 1$ and the interaction kernel is an indicator function, the corresponding system is the well-known Hegselmann-Krause Model (also called flocking model in some literatures of computational social science),we refer readers to De et al. (2014); Abebe et al. (2018) for details. Recently, the systems have also been applied to learn the celestial dynamics from Jet Propulsion Lab's data Zhong et al. (2020b,a), or learning the cell dynamics from microscopy videos (personal communication)

We consider $\boldsymbol{\phi} \in \boldsymbol{L}^\infty([0, R])$ with the radius R representing the maximum range of interaction between agents. We further assume that $\boldsymbol{\phi}$ lies in the *admissible set*

$$\boldsymbol{\mathcal{K}}_{R,S} := \{\boldsymbol{\varphi} = (\varphi_{kk'})_{k,k'=1}^{K,K} : \ \varphi_{kk'} \in C_c^{0,1}([0, R]), \|\varphi_{kk'}\|_\infty + \mathrm{Lip}[\varphi_{kk'}] \leq S\} \tag{1.7}$$

for some $S > 0$. For $\boldsymbol{\phi} \in \boldsymbol{\mathcal{K}}_{R,S}$, the system (1.1) is well-posed for any given initial condition (i.e. there is a unique solution, that can be extended to all times) and it is expected to converge for $t \to \infty$ to configurations of points whose mutual distances are close to local minimizers of the potential energy function in (1.6), corresponding to steady states of evolution. We refer to Kolokolnikov et al. (2013); Chen (2014) for the qualitative analysis of this type of systems.

1.5 Outline and organization

The remainder of the paper is organized as follows. In section 2, we present the learning theory that establishes a theoretical framework for analyzing the performance of the proposed learning algorithm. We then discuss the numerical implementation of the learning algorithm in section 3 and 4.1, and its performance in various numerical examples in section 4. Section 5 presents some theoretical results for the coercivity condition, a key condition for achieving the optimal convergence rate of interaction kernels. Finally, we present the proof of the main Theorems in the Appendix.

2. Learning theory

2.1 Measures and function spaces adapted to the dynamics

To measure the accuracy of the estimators, we introduce a probability measure, dependent on the distribution of the initial condition μ_0 and the underlying dynamical system, and then define the function space for learning. We start with a heuristic argument. In the case $K = 1$, the interaction kernel ϕ depends only on one variable, but it is observed through a collection of non-independent linear measurements with values $\dot{\boldsymbol{x}}_i$, the l.h.s. of (1.1), at locations $r_{ii'} := \|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|$, with coefficients $\boldsymbol{r}_{ii'} := \boldsymbol{x}_{i'} - \boldsymbol{x}_i$. One could attempt to recover $\{\phi(r_{ii'})\}_{i,i'}$ from the equations of $\dot{\boldsymbol{x}}_i$'s by solving the corresponding linear system. Unfortunately, this linear system is usually underdetermined as dN (number of known quantities) $\leq \frac{N(N-1)}{2}$ (number of unknowns) and in general one will not be able to recover the values of ϕ at locations $\{r_{ii'}\}_{i,i'}$. We take a different route, to leverage observations

through time: we note that the pairwise distances $\{r_{ii'}\}_{i,i'}$ are "equally" important in a homogeneous system, and introduce a probability density ρ_T^L on $\mathbb{R}_+$

$$\rho_T^L(r) := \frac{1}{\binom{N}{2}L} \sum_{l=1}^{L} \mathbb{E}_{\mu_0} \sum_{i,i'=1, i<i'}^{N} \delta_{r_{ii'}(t_l)}(r)\,, \tag{2.1}$$

where the expectation in (2.1) is with respect to the distribution μ_0 of the initial condition. By the law of large numbers, this density is the limit, as $M \to \infty$, of the empirical measure of pairwise distance

$$\rho_T^{L,M}(r) := \frac{1}{\binom{N}{2}LM} \sum_{l,m=1}^{L,M} \sum_{i,i'=1, i<i'}^{N} \delta_{r_{ii'}^{(m)}(t_l)}(r)\,. \tag{2.2}$$

The measure ρ_T^L is intrinsic to the dynamical system and independent of the observations. It can be thought of as an "occupancy" measure, in the sense that for any interval $I \subset \mathbb{R}_+$, $\rho_T^L(I)$ is the probability of seeing a pair of agents at a distance between them equal to a value in I, averaged over the observation time. It measures how much regions of $\mathbb{R}_+$ on average (over the observed times and with respect to the distribution μ_0 of the initial conditions) are explored by the dynamical system. Highly explored regions are where the learning process ought to be successful, since these are the areas with enough samples from the dynamics to enable the reconstruction of the interaction kernel. Therefore, a natural metric to measure the regression error is the mean square distance in $L^2(\rho_T^L)$: for an estimator $\widehat{\phi}_{M,\mathcal{H}}$, we let

$$\mathrm{dist}(\widehat{\phi}_{M,\mathcal{H}}, \phi) = \|\widehat{\phi}_{M,\mathcal{H}}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T^L)} = \left(\int_{r=0}^{\infty} \left|\widehat{\phi}_{M,\mathcal{H}}(r)r - \phi(r)r\right|^2 \rho_T^L(dr) \right)^{\frac{1}{2}}. \tag{2.3}$$

If trajectories were observed continuously in time, we could consider

$$\rho_T(r) = \frac{1}{\binom{N}{2}T} \int_{t=0}^{T} \mathbb{E}_{\mu_0} \sum_{i,i'=1, i<i'}^{N} \delta_{r_{ii'}(t)}(r)\, dt\,. \tag{2.4}$$

The natural generalizations of ρ_T^L and ρ_T, defined in (2.1) and (2.4), to the heterogeneous case, for each $k, k' = 1, \ldots, K$, are the probability measures on $\mathbb{R}_+$ (in discrete and continuous time respectively)

$$\rho_T^{L,kk'}(r) = \frac{1}{LN_{kk'}} \sum_{l=1}^{L} \mathbb{E}_{\mu_0} \sum_{\substack{i \in C_k, i' \in C_{k'} \\ i \neq i'}} \delta_{r_{ii'}(t_l)}(r) \tag{2.5}$$

$$\rho_T^{kk'}(r) = \frac{1}{TN_{kk'}} \int_{t=0}^{T} \mathbb{E}_{\mu_0} \sum_{\substack{i \in C_k, i' \in C_{k'} \\ i \neq i'}} \delta_{r_{ii'}(t)}(r) dt, \tag{2.6}$$

where $N_{kk'} = N_k N_{k'}$ when $k \neq k'$ and $N_{kk'} = \binom{N_k}{2}$ when $k = k'$. The error of an estimator $\widehat{\phi}_{kk'}$ will be measured by $||\widehat{\phi}_{kk'}(\cdot)\cdot - \phi_{kk'}(\cdot)\cdot||_{L^2(\rho_T^{L,kk'})}$ as in (2.3). For simplicity of notation,

we write

$$\boldsymbol{\rho}_T^L = \bigoplus_{k,k'=1,1}^{K,K} \rho_T^{L,kk'}, \boldsymbol{\rho}_T = \bigoplus_{k,k'=1,1}^{K,K} \rho_T^{kk'}, \quad \boldsymbol{L}^2(\boldsymbol{\rho}_T^L) = \bigoplus_{k,k'=1,1}^{K,K} L^2(\rho_T^{L,kk'}), \tag{2.7}$$

with $\|\boldsymbol{\varphi}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} = \sum_{kk'} \|\boldsymbol{\varphi}_{kk'}\|^2_{\rho_T^{L,kk'}}$ for any $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\boldsymbol{\rho}_T^L)$.

Well-posedness and properties of measures

The probability measures $\{\rho_T^{L,kk'}\}_{k,k'=1,1}^{K,K}$ and their continuous counterpart are well-defined, thanks to the following lemma:

Lemma 1 *Suppose $\boldsymbol{\phi} \in \boldsymbol{\mathcal{K}}_{R,S}$ (see the admissible set defined in*(1.7)*). Then for each (k,k'), the measures $\rho_T^{L,kk'}$ and $\rho_T^{kk'}$, defined in* (2.5) *and* (2.6)*, are regular Borel probability measures on $\mathbb{R}_+$. They are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$ provided that μ_0 is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{dN}$.*

We emphasize that the measures $\rho_T^{L,kk'}$ and $\rho_T^{kk'}$ are both averaged-in-time measures of the pairwise distances, and they have the same properties. The only difference is that they correspond to discrete-time and continuous-time observations, respectively. In the following, we analyze only the discrete-time observation case using $\rho_T^{L,kk'}$, and all the arguments can be extended directly to the continuous-time observation case using $\rho_T^{kk'}$.

The measures $\rho_T^{kk'}$ and $\rho_T^{L,kk'}$ are compactly supported provided that μ_0 is:

Proposition 2 *Suppose the distribution μ_0 of the initial condition is compactly supported. Then there exists $R_0 > 0$, such that for each (k,k'), the support of the measure $\rho_T^{kk'}$ (and therefore $\rho_T^{L,kk'}$), is contained in $[0,R_0]$ with $R_0 = 2C_0 + 2K\|\boldsymbol{\phi}\|_\infty RT$ where C_0 only depends on supp(μ_0).*

The proofs of these results are postponed to sec.6.2.

2.2 Learnability: a coercivity condition

A fundamental question is the well-posedness of the inverse problem of learning the interaction kernels. Since the least square estimator always exists for compact sets in $\boldsymbol{L}^\infty([0,R])$, learnability is equivalent to the convergence of the estimators to the true interaction kernels as the sample size increases (i.e. $M \to \infty$) and as the compact sets contain better and better approximations to the true kernels $\boldsymbol{\phi}$. To ensure such a convergence, one would naturally wish: (i) that the true kernel $\boldsymbol{\phi}$ is the unique minimizer of the expectation of the error functional (by the law of large numbers)

$$\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}) \coloneqq \lim_{M\to\infty} \boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi}) = \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_{\mu_0}\left[\|\dot{\boldsymbol{X}}(t_l) - \mathbf{f}_{\boldsymbol{\varphi}}(\boldsymbol{X}(t_l))\|^2_{\mathcal{S}}\right]; \tag{2.8}$$

(ii) that the error of the estimator, say $\widehat{\boldsymbol{\phi}}$, is small once $\boldsymbol{\mathcal{E}}_\infty(\widehat{\boldsymbol{\phi}})$ is small since $\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\phi}) = 0$.

Note that $\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi})$ is a quadratic functional of $\boldsymbol{\varphi}-\boldsymbol{\phi}$; by Jensen's inequality, we have

$$\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}) < K^2 \left\|\boldsymbol{\varphi}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\right\|^2_{L^2(\rho_T^L)} .$$

If we bound $\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi})$ from below by $\left\|\boldsymbol{\varphi}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\right\|^2_{L^2(\boldsymbol{\rho}_T^L)}$, we can conclude (i) and (ii) above. This suggests the following coercivity condition:

Definition 3 (Coercivity condition for first-order systems) *Consider the dynamical system defined in* (1.1) *at time instants* $0 = t_1 < t_2 < \cdots < t_L = T$*, with random initial condition distributed according to the probability measure* μ_0 *on* $\mathbb{R}^{dN}$*. We say that it satisfies the coercivity condition on a hypothesis space* $\boldsymbol{\mathcal{H}}$ *with a constant* $c_{L,N,\boldsymbol{\mathcal{H}}}$ *if*

$$c_{L,N,\boldsymbol{\mathcal{H}}} := \inf_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}\setminus\{\mathbf{0}\}} \frac{\frac{1}{L}\sum_{l=1}^L \mathbb{E}_{\mu_0}\left[\left\|\mathbf{f}_{\boldsymbol{\varphi}}(\boldsymbol{X}(t_l))\right\|_S^2\right]}{\left\|\boldsymbol{\varphi}(\cdot)\cdot\right\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)}} > 0\,. \tag{2.9}$$

A similar definition holds for continuous observations on the time interval $[0,T]$*, upon replacing the sum over observations with an integral over* $[0,T]$*.*

The coercivity condition plays a key role in the learning of the kernel. It ensures learnability by ensuring the uniqueness of minimizer of the expectation of the error functional, and by guaranteeing convergence of estimator. To see this, apply the coercivity inequality to $\boldsymbol{\varphi}-\boldsymbol{\phi}$ and suppose $\boldsymbol{\varphi}-\boldsymbol{\phi}$ lies in $\boldsymbol{\mathcal{H}}$, we obtain

$$c_{L,N,\boldsymbol{\mathcal{H}}} \left\|\boldsymbol{\varphi}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\right\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \le \boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}). \tag{2.10}$$

From the facts that $\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}) \ge 0$ for any $\boldsymbol{\varphi}$ and that $\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\phi}) = 0$, we conclude that the true kernel $\boldsymbol{\phi}$ is the unique minimizer of the $\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi})$. Furthermore, the coercivity condition enables us to control the error of the estimator by the discrepancy between the error functional and its expectation (see Proposition 18), therefore guaranteeing convergence of the estimator. Finally, the coercivity constant controls the condition number of the inverse problem, guaranteeing numerical stability. We study the consistency and rate of convergence in the next section.

2.3 Consistency and rate of convergence

We start from a concentration estimate, in which the coercivity condition plays a fundamental role.

Theorem 4 *Suppose that* $\boldsymbol{\phi} \in \boldsymbol{\mathcal{K}}_{R,S}$*. Let* $\boldsymbol{\mathcal{H}}_M \subset \boldsymbol{L}^\infty([0,R])$ *be a compact (with respect to the* ∞*-norm) convex set bounded above by* $S_0 \ge S$*. Assume that the coercivity condition* (2.9) *holds true on* $\boldsymbol{\mathcal{H}}_M$*. Then for any* $\epsilon > 0$*, the estimate*

$$c_{L,N,\boldsymbol{\mathcal{H}}_M} \|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \le 2 \inf_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}_M} \|\boldsymbol{\varphi}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} + 2\epsilon \tag{2.11}$$

holds true with probability at least $1-\delta$*, provided that* $M \ge \frac{1152 S_0^2 R^2 K^4}{c_{L,N,\boldsymbol{\mathcal{H}}_M}\epsilon}\big(\log(\mathcal{N}(\boldsymbol{\mathcal{H}}_M, \frac{\epsilon}{48 S_0 R^2 K^4})) + \log(\frac{1}{\delta})\big)$*, where* $\mathcal{N}(\boldsymbol{\mathcal{H}}_M, \frac{\epsilon}{48 S_0 R^2 K^4})$ *is the covering number of* $\boldsymbol{\mathcal{H}}_M$ *with respect to the* ∞*-norm.*

If we choose a family of compact convex hypothesis space $\boldsymbol{\mathcal{H}}_M$ that contain better and better approximations to the true interaction kernels $\boldsymbol{\phi}$; then the concentration estimate (2.11) yields the following consistency result:

Theorem 5 (Consistency of estimators) *Suppose that $\{\boldsymbol{\mathcal{H}}_M\}_{M=1}^{\infty} \subset \boldsymbol{L}^\infty([0,R])$ is a family of compact convex subsets such that*

$$\inf_{\boldsymbol{f}\in\boldsymbol{\mathcal{H}}_M} \|\boldsymbol{f}-\boldsymbol{\phi}\|_\infty \xrightarrow{M\to\infty} 0$$

Suppose $\cup_M \boldsymbol{\mathcal{H}}_M$ is compact in $\boldsymbol{L}^\infty([0,R])$ and the coercivity condition holds true on $\cup_M \boldsymbol{\mathcal{H}}_M$. Then

$$\lim_{M\to\infty} \|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} = 0 \text{ with probability one.}$$

Given data collected from M trajectories, we would like to choose the best $\boldsymbol{\mathcal{H}}_M$ to maximize the accuracy of the estimator. Theorem 4 highlights two competing issues. On one hand, we would like the hypothesis space $\boldsymbol{\mathcal{H}}_M$ to be large so that the bias $\inf_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}_M} \|\boldsymbol{\varphi}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)}$ (or $\inf_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}} \|\boldsymbol{\varphi}-\boldsymbol{\phi}\|^2_\infty$) is small. On the other hand, we would like $\boldsymbol{\mathcal{H}}_M$ to be small so that the covering number $\mathcal{N}(\boldsymbol{\mathcal{H}}_M, \frac{\epsilon}{48S_0R^2K^3})$ is small. This is the classical bias-variance trade-off in statistical estimation. As is standard in nonparametric regression, the rate of convergence depends on the regularity condition of the true interaction kernels and the hypothesis space, as is demonstrated in the following proposition. We show that the optimal min-max rate of convergence for 1-dimensional regression with noisy observations is achieved by choosing suitable hypothesis spaces, with dimension dependent on M:

Theorem 6 *Let $\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}$ be a minimizer of the empirical error functional defined in (1.3) over the hypothesis space $\boldsymbol{\mathcal{H}}_M$.*
(a) If we choose $\boldsymbol{\mathcal{H}}_M \equiv \boldsymbol{\mathcal{K}}_{R,S}$, assume that the coercivity condition holds true on $\boldsymbol{\mathcal{K}}_{R,S}$, then there exists a constant $C = C(K,S,R)$ such that

$$\mathbb{E}_{\mu_0}[\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)}] \le \frac{C}{c_{L,N,\boldsymbol{\mathcal{H}}_M}} M^{-\frac{1}{4}}.$$

(b) Assume that $\{\boldsymbol{\mathcal{L}}_n\}_{n=1}^{\infty}$ is a sequence of linear subspaces of $\boldsymbol{L}^\infty([0,R])$, such that

$$dim(\boldsymbol{\mathcal{L}}_n) \le c_0K^2n\,, \quad \inf_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{L}}_n} \|\boldsymbol{\varphi}-\boldsymbol{\phi}\|_\infty \le c_1 n^{-s} \tag{2.12}$$

for some constants $c_0, c_1 > 0, s \ge 1$. Such a sequence of linear spaces exists, for example, when $\boldsymbol{\phi} \in C^{k,\alpha}$ with $s = k+\alpha$, it is approximated by $\boldsymbol{\mathcal{L}}_n$: piecewise polynomials of degree at least $\lfloor s-1 \rfloor$, defined on n uniform subintervals of $[0,R]$. Suppose the coercivity condition holds true on the set $\boldsymbol{\mathcal{L}} := \cup_n \boldsymbol{\mathcal{L}}_n$. Define $\boldsymbol{\mathcal{B}}_n$ to be the central ball of $\boldsymbol{\mathcal{L}}_n$ with the radius (c_1+S). Then by choosing $\boldsymbol{\mathcal{H}}_M = \boldsymbol{\mathcal{B}}_{n(M)}$, with $n(M) \asymp (\frac{M}{\log M})^{\frac{1}{2s+1}}$, there exists a constant $C = C(K,S,R,c_0,c_1)$ such that

$$\mathbb{E}_{\mu_0}[\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)}] \le \frac{C}{c_{L,N,\boldsymbol{\mathcal{L}}}} \left(\frac{\log M}{M}\right)^{\frac{s}{2s+1}}. \tag{2.13}$$

We remark that C increases (polynomially) with K, the number of agent types, consistently with the expectation that the multi-type estimation problem is harder than the single-type problem. We do not expect, however, the dependency of C on K to be sharp. A tighter bound would take into account, for example, how similar the interaction kernels between different types are. This is an interesting direction of future research.

Proof For part (i), denote $\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{K}}_{R,S}$, and recall that for $\epsilon > 0$, the covering number of $\boldsymbol{\mathcal{H}}$ (with respect to the ∞-norm) satisfies

$$\mathcal{N}(\boldsymbol{\mathcal{H}}, \epsilon) \leq e^{C_1 K^2 \epsilon^{-1}}$$

where C_1 is an absolute constant (see e.g. (Cucker and Smale, 2002, Proposition 6)), and $\inf_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}} \|\boldsymbol{\varphi} - \boldsymbol{\phi}\|_\infty^2 = 0$. Then estimate (2.11) gives

$$\begin{aligned} P_{\mu_0}\{\|\widehat{\boldsymbol{\phi}}_{L,M,\boldsymbol{\mathcal{H}}}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} > \epsilon\} &\leq \mathcal{N}(\boldsymbol{\mathcal{H}}, \frac{\epsilon^2 c_{L,N,\boldsymbol{\mathcal{H}}}}{48SR^2K^4}) e^{\frac{-c_{L,N,\boldsymbol{\mathcal{H}}}^2 M\epsilon^2}{1152S^2K^4}} \\ &\leq e^{\frac{48SR^2K^6C_1}{c_{L,N,\boldsymbol{\mathcal{H}}}}\epsilon^{-2} - \frac{c_{L,N,\boldsymbol{\mathcal{H}}}^2 M\epsilon^2}{1152S^2R^2K^6}}. \end{aligned} \tag{2.14}$$

Define $g(\epsilon) := \frac{48SR^2K^6C_1}{c_{L,N,\boldsymbol{\mathcal{H}}}}\epsilon^{-2} - \frac{c_{L,N,\boldsymbol{\mathcal{H}}}^2 M\epsilon^2}{2304S^2R^2K^4}$, which is a decreasing function of ϵ. By direct calculation, $g(\epsilon) = 0$ if $\epsilon = \epsilon_M = (\frac{C_2}{M})^{\frac{1}{4}}$, where $C_2 = \frac{11092S^3K^{10}C_1}{c_{L,N,\boldsymbol{\mathcal{H}}}^3}$. Thus, we obtain

$$P_{\mu_0}\{\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} > \epsilon\} \leq \begin{cases} e^{\frac{-c_{L,N,\boldsymbol{\mathcal{H}}}^2 M\epsilon^2}{2304S^2K^4}}, & \epsilon \geq \epsilon_M \\ 1, & \epsilon \leq \epsilon_M \end{cases}$$

Integrating over $\epsilon \in (0, +\infty)$ gives

$$\mathbb{E}_{\mu_0}[\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)}] \leq \frac{C_3}{c_{L,N,\boldsymbol{\mathcal{H}}}} M^{-\frac{1}{4}},$$

where $C_3 = C(K, S, R)$ is an absolute constant only depends on K, S and R.

For part (ii), recall that for $\epsilon > 0$, the covering number of $\boldsymbol{\mathcal{B}}_n$ by ϵ-balls is bounded above by $(4(c_1 + S)/\epsilon)^{c_0 K^2 n}$(see (Cucker and Smale, 2002, Proposition 5)). From estimate (2.11), we obtain

$$\begin{aligned} P_{\mu_0}\{\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{B}}_n}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \geq \epsilon + c_2 n^{-s}\} &\leq (\frac{c_3}{\epsilon^2})^{c_0K^2n} e^{-c_4 M\epsilon^2} \\ &= e^{c_0K^2n\log(c_3) + 2c_0K^2n|\log(\epsilon)| - c_4M\epsilon^2}, \end{aligned} \tag{2.15}$$

where $c_2 = \sqrt{\frac{1}{c_{L,N,\cup_n\boldsymbol{\mathcal{L}}_n}}}c_1$, $c_3 = \frac{192(S+c_1)^2R^2K^4}{c_{L,N,\cup_n\boldsymbol{\mathcal{L}}_n}}$, and $c_4 = \frac{c_{L,N,\cup_n\boldsymbol{\mathcal{L}}_n}^2}{1152(S+c_1)^2R^2K^6}$ are absolute constants independent of M. Define

$$g(\epsilon) := c_0 nK^2 \log(c_3) + 2c_0 nK^2 |\log(\epsilon)| - \frac{c_4}{2} M\epsilon^2.$$

Set $n_* = (\frac{M}{\log M})^{\frac{1}{2s+1}}$, and consider $g(c\epsilon_M)$ with $\epsilon_M = (\frac{\log M}{M})^{\frac{s}{2s+1}} = n_*^{-s}$ as a function of c. By calculation, $g(c\epsilon_M)$ is a decreasing function of c. We have $\lim_{c\to 0^+} g(c\epsilon_M) = \infty$ and

$\lim_{c\to\infty} g(c\epsilon_M) = -\infty$. Therefore, there exists a constant c_5 depending only on K, c_0, c_3, c_4 such that $g(c_5\epsilon_M) = 0$. This gives

$$P_{\mu_0}\{\|\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{B}}_{n_*}}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} > \epsilon\} \le \begin{cases} e^{\frac{-c_4}{2}M\epsilon^2}, & \epsilon \ge c_5\epsilon_M \\ 1, & \epsilon \le c_5\epsilon_M \end{cases}.$$

Therefore, with $\boldsymbol{\mathcal{H}}_M = \boldsymbol{\mathcal{B}}_{n_*}$,

$$\mathbb{E}_{\mu_0}[\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)}] \le \frac{c_6}{c_{L,N,\cup_n\boldsymbol{\mathcal{H}}_n}}\left(\frac{\log M}{M}\right)^{\frac{s}{2s+1}},$$

where c_6 is an absolute constant only depending on K, S, c_0, c_1. ■

The convergence rate $\frac{s}{2s+1}$ coincides with the convergence rate for 1-dimensional regression, where one can observe directly noisy values of the target function at sample points drawn i.i.d from $\boldsymbol{\rho}_T^L$, for the set of functions satisfying the approximation property (2.12). It is the optimal min-max rate for functions $C^{k,\alpha}$ with $s = k + \alpha$. Obtaining this optimal rate is satisfactory, because we do not observe the values $\{\phi_{kk'}(\|\boldsymbol{x}_{i'}^{(m)}(t_l) - \boldsymbol{x}_i^{(m)}(t_l)\|)\}_{l=1,i,i'=1,\ m=1,k,k'=1}^{L,N,N,M,K,K}$ from the observations of the trajectories of the states. The only randomness is in the M samples, via the random initial condition. It is perhaps a shortcoming of our result that there is no dependence on L nor N in our upper bounds, especially since numerical examples in Lu et al. (2019b) suggest that the error does decrease with L. In the case of $K = 1$ and N large, the results in Bongini et al. (2017) suggest rates no better than $N^{-\frac{1}{d}}$, i.e. they are cursed by the dimensionality of the space in which the agents move, albeit recent work by some of the authors of Bongini et al. (2017) suggest better results, with rates similar to ours but in the case of $N \to +\infty$ may be possible (personal communication).

2.4 Accuracy of trajectory prediction

Once an estimator $\widehat{\boldsymbol{\phi}}_{L,M,\boldsymbol{\mathcal{H}}}$ is obtained, a natural question is the accuracy of trajectory prediction based on the estimated kernel. The next proposition 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 $\boldsymbol{L}^2(\boldsymbol{\rho}_T)$ error of the estimator. This further validates the effectiveness of our error functional and $\boldsymbol{L}^2(\boldsymbol{\rho}_T)$-metric to assess the quality of the estimator.

Proposition 7 *Suppose $\widehat{\boldsymbol{\phi}} \in \boldsymbol{\mathcal{K}}_{R,S}$. Denote by $\widehat{\boldsymbol{X}}(t)$ and $\boldsymbol{X}(t)$ the solutions of the systems with kernels $\widehat{\boldsymbol{\phi}} = (\widehat{\phi}_{kk'})_{k,k'=1}^{K,K}$ and $\boldsymbol{\phi}$ respectively, starting from the same initial condition. Then for each trajectory we have*

$$\sup_{t\in[0,T]}\|\widehat{\boldsymbol{X}}(t) - \boldsymbol{X}(t)\|_{\mathcal{S}}^2 \le 2T\exp(8T^2K^2S^2)\int_0^T \|\dot{\boldsymbol{X}}(s) - \mathbf{f}_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{X}(s))\|_{\mathcal{S}}^2 dt\,,$$

and on average with respect to the initial distribution μ_0

$$\mathbb{E}_{\mu_0}[\sup_{t\in[0,T]}\|\widehat{\boldsymbol{X}}(t) - \boldsymbol{X}(t)\|_{\mathcal{S}}^2] \le 2T^2K^2\exp(8T^2K^2S^2)\|\widehat{\boldsymbol{\phi}}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T)}^2,$$

where the measure $\boldsymbol{\rho}_T$ is defined by (2.4).

Proof Recall that $\boldsymbol{r}_{ii'} := \boldsymbol{x}_{i'} - \boldsymbol{x}_i$ and $\widehat{\boldsymbol{r}}_{ii'} := \widehat{\boldsymbol{x}}_{i'} - \widehat{\boldsymbol{x}}_i$. To simplify the notation, we introduce the function $F_{[\varphi]}(\boldsymbol{z}) := \varphi(\|\boldsymbol{z}\|)\boldsymbol{z}$, defined on $\mathbb{R}^d$ for $\varphi \in L^\infty([0,R])$. Since $\widehat{\boldsymbol{\phi}} := (\widehat{\phi}_{kk'})_{k,k'=1,1}^{K,K} \in \boldsymbol{\mathcal{K}}_{R,S}$, we obtain $\mathrm{Lip}[F_{[\widehat{\phi}_{kk'}]}] \le S$ for each pair (k,k'). For every $t \in [0,T]$, we have

$$\begin{aligned}\|\boldsymbol{X}(t) - \widehat{\boldsymbol{X}}(t)\|_{\mathcal{S}}^2 &= \sum_{j=1}^K \sum_{i\in C_j} \frac{1}{N_j} \left\| \int_0^t (\dot{\boldsymbol{x}}_i(s) - \dot{\widehat{\boldsymbol{x}}}_i(s)) ds \right\|^2 \le t \sum_{j=1}^K \sum_{i\in C_j} \frac{1}{N_j} \int_0^t \left\| \dot{\boldsymbol{x}}_i(s) - \dot{\widehat{\boldsymbol{x}}}_i(s) \right\|^2 ds \\ &\le 2T \int_0^t \left\| \dot{\boldsymbol{X}}(s) - \mathbf{f}_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{X}(s)) \right\|_{\mathcal{S}}^2 ds + 2T \int_0^s I(s) ds,\end{aligned}$$

where

$$I(s) = \left\| \sum_{j'=1}^K \sum_{i'\in C_{j'}} \frac{1}{N_{j'}} \left(F_{\widehat{\phi}_{jj'}}(\boldsymbol{r}_{ii'}(s)) - F_{\widehat{\phi}_{jj'}}(\widehat{\boldsymbol{r}}_{ii'}(s)) \right) \right\|_{\mathcal{S}}^2.$$

By the triangle inequality, we have $I(s) \le I_1(s) + I_2(s)$, where

$$I_1(s) = \left\| \sum_{j'=1}^K \sum_{i'\in C_{j'}} \frac{1}{N_{j'}} \left(F_{\widehat{\phi}_{jj'}}(\boldsymbol{r}_{ii'}(s)) - F_{\widehat{\phi}_{jj'}}(\boldsymbol{x}_i(s) - \hat{\boldsymbol{x}}_{i'}(s)) \right) \right\|_{\mathcal{S}}^2,$$

$$I_2(s) = \left\| \sum_{j'=1}^K \sum_{i'\in C_{j'}} \frac{1}{N_{j'}} \left(F_{\widehat{\phi}_{jj'}}(\widehat{\boldsymbol{r}}_{ii'}(s)) - F_{\widehat{\phi}_{jj'}}(\boldsymbol{x}_i(s) - \hat{\boldsymbol{x}}_{i'}(s)) \right) \right\|_{\mathcal{S}}^2.$$

Estimating by Jensen or Hölder inequalities, we obtain

$$\begin{aligned} I_1(s) &\le \sum_{j=1}^K \sum_{i\in C_j} \frac{1}{N_j} \left| \sum_{j'=1}^K \sum_{i'\in C_{j'}} \frac{1}{N_{j'}} \mathrm{Lip}[F_{\widehat{\phi}_{jj'}}] \|\boldsymbol{x}_{i'}(s) - \hat{\boldsymbol{x}}_{i'}(s)\| \right|^2 \\ &\le K \sum_{j=1}^K \sum_{i\in C_j} \frac{1}{N_j} \sum_{j'=1}^K \sum_{i'\in C_j'} \frac{\mathrm{Lip}^2[F_{\widehat{\phi}_{jj'}}]}{N_{j'}} \|\boldsymbol{x}_{i'}(s) - \hat{\boldsymbol{x}}_{i'}(s)\|^2 \\ &\le K (\sum_{j=1}^K \max_{j'} \mathrm{Lip}^2[F_{\widehat{\phi}_{jj'}}]) \|\boldsymbol{X}(s) - \hat{\boldsymbol{X}}(s)\|_{\mathcal{S}}^2 \\ &\le K^2 S^2 \|\boldsymbol{X}(s) - \hat{\boldsymbol{X}}(s)\|_{\mathcal{S}}^2. \end{aligned}$$

Similarly,

$$I_2(s) \le K \max_j (\sum_{j'=1}^K \mathrm{Lip}^2[F_{\widehat{\phi}_{jj'}}]) \|\boldsymbol{X}(s) - \hat{\boldsymbol{X}}(s)\|_{\mathcal{S}}^2 \le K^2 S^2 \|\boldsymbol{X}(s) - \hat{\boldsymbol{X}}(s)\|_{\mathcal{S}}^2.$$

Combining above inequalities with Gronwall's inequality yields the first inequality in the proposition. The second inequality follows by combining the above with Proposition 6.3, which implies

$$\frac{1}{T} \int_0^T \mathbb{E} \left\| \dot{\boldsymbol{X}}(s) - \mathbf{f}_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{X}(s)) \right\|_{\mathcal{S}}^2 ds < K^2 \left\| \widehat{\boldsymbol{\phi}}(\cdot) \cdot - \boldsymbol{\phi}(\cdot) \cdot \right\|_{L^2(\boldsymbol{\rho}_T)}^2.$$

■

3. Algorithm

Recall that our goal is to learn the interaction kernels $\boldsymbol{\phi}$ from the observational data

$$\{\boldsymbol{x}_i^{(m)}(t_l), \dot{\boldsymbol{x}}_i^{(m)}(t_l)\}_{i=1,m=1,l=1}^{N,M,L},$$

consisting of the positions and velocities of agents observed at equidistant time instances $0 = t_1 < t_2 < \cdots < t_L = T$ with M i.i.d initial conditions drawn from a probability measure μ_0 on $\mathbb{R}^{dN}$. Our estimator $\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}}$ is obtained by minimizing the empirical error functional

$$\boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi}) = \frac{1}{LM} \sum_{l=1,m=1,i=1}^{L,M,N} \frac{1}{N_{\mathsf{k}_i}} \left\| \dot{\boldsymbol{x}}_i^{(m)}(t_l) - \sum_{i'=1}^{N} \frac{1}{N_{\mathsf{k}_{i'}}} \varphi_{\mathsf{k}_i \mathsf{k}_{i'}}(r_{ii'}^{(m)}(t_l)) \boldsymbol{r}_{ii'}^{(m)}(t_l) \right\|^2, \tag{3.1}$$

over all possible $\boldsymbol{\varphi} = \{\varphi_{kk'}\}_{k,k'=1}^{K}$ in a suitable hypothesis space $\boldsymbol{\mathcal{H}} = \bigoplus_{k,k'=1,1}^{K,K} \mathcal{H}_{kk'}$. In section 2, we analyzed the performance of estimators over compact convex subsets of $\boldsymbol{L}^\infty([0,R])$. However, to compute these estimators numerically, one has to solve a constrained quadratic minimization problem, which is computationally demanding. Fortunately, as in the standard nonparametric setting references, such a costly constrained optimization is unnecessary, and one can simply compute the minimizer by least-squares over the linear finite-dimensional hypothesis spaces because one can prove by standard truncation arguments that the learning theory is still applicable to the truncation of these estimators obtained by the unconstrained optimization (see chapter 11 in Györfi et al. (2002)). In the following, we solve the minimization problem by choosing a suitable set of basis functions for $\boldsymbol{\varphi}$ and compute the minimizer by regularized (i.e. constrained to $\boldsymbol{\varphi}$) least squares in a fashion that is amenable to efficient parallel implementation.

3.1 Numerical implementation

3.1.1 Choice of the hypothesis spaces and their basis

We use local basis functions to capture local features of the interaction kernels, such as the sharp jumps: each hypothesis space $\mathcal{H}_{kk'}$ is an $n_{kk'}$-dimensional space spanned by $\{\psi_{kk',p}\}_{p=1}^{n_{kk'}}$, a set of piecewise polynomial functions of degree s, with s being the order of local differentiability of the true kernel. The dimension $n_{kk'}$ is chosen to be a scalar multiple of the optimal dimension $n_* = (\frac{M}{\log M})^{\frac{1}{2s+1}}$ of the hypothesis space, as in Theorem 6. For simplicity, we set these piecewise polynomials to be supported on a uniform partition of the interval $[0,R]$, where the radius R is the largest observed pairwise distance.

3.1.2 Velocity data of agents

When only the position data are available, the velocity data may be approximated numerically. In our numerical experiments, $\dot{\boldsymbol{x}}_i^{(m)}(t_l)$ is approximated by backward differences:

$$\dot{\boldsymbol{x}}_i^{(m)}(t_l) \approx \Delta \boldsymbol{x}_i^m(t_l) = \frac{\boldsymbol{x}_i^{(m)}(t_{l+1}) - \boldsymbol{x}_i^{(m)}(t_l)}{t_{l+1} - t_l}, \quad \text{for } 1 \le l \le L.$$

The error of the backward difference approximation is of order $O(T/L)$, leading to a comparable bias in the estimator, as we shall see in 3.2. Hereafter we assume that T/L is sufficiently small so that the error is negligible relative to the statistical error.

3.1.3 The numerical implementation

With these basis functions, denoting $\varphi_{kk'}(r) = \sum_{p=1}^{n_{kk'}} a_{kk',p}\psi_{kk',p}(r) \in \mathcal{H}_{kk'}$ for some constant coefficients $(a_{kk',p})_{p=1}^{n_{kk'}}$, we can rewrite the error functional in (3.1) as

$$\boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi}) = \frac{1}{LM}\sum_{l=1,m=1,i=1}^{L,M,N}\frac{1}{N_{k_i}}\left\|\sum_{i'=1}^{N}\frac{1}{N_{k_{i'}}}\Big(\sum_{p=1}^{n_{k_ik_{i'}}} a_{k_ik_{i'},p}\psi_{k_ik_{i'},p}\Big)(r_{ii'}^{(m)}(t_l))\boldsymbol{r}_{i,i'}^{(m)}(t_l) - \dot{\boldsymbol{x}}_i^{(m)}(t_l)\right\|^2,$$

which is a quadratic functional with respect to the coefficient vector $\vec{a}$ of $\boldsymbol{\varphi}$:

$$\frac{1}{LM}\sum_{m=1}^{M}\|\Psi_{\boldsymbol{\mathcal{H}}}^{(m)}\vec{a} - \vec{d}_L^{(m)}\|_2^2.$$

Here the vectors $\vec{a}$ and $\vec{d}_L^{(m)}$ are

$$\vec{a} = \begin{pmatrix} a_{11,1} \\ \vdots \\ a_{11,n_{11}} \\ \vdots \\ a_{KK,1} \\ \vdots \\ a_{KK,n_{KK}} \end{pmatrix} \in \mathbb{R}^{\sum_{k,k'=1}^{K} n_{kk'}}, \quad \vec{d}_L^{(m)} = \begin{pmatrix} (1/N_{k_1})^{1/2}\dot{\boldsymbol{x}}_1^{(m)}(t_1) \\ \vdots \\ (1/N_{k_N})^{1/2}\dot{\boldsymbol{x}}_N^{(m)}(t_1) \\ \vdots \\ (1/N_{k_1})^{1/2}\dot{\boldsymbol{x}}_1^{(m)}(t_L) \\ \vdots \\ (1/N_{k_N})^{1/2}\dot{\boldsymbol{x}}_N^{(m)}(t_L) \end{pmatrix} \in \mathbb{R}^{LNd},$$

and the learning matrix $\Psi_{\boldsymbol{\mathcal{H}}}^{(m)} \in \mathbb{R}^{LNd\times\sum_{k,k'=1}^{K} n_{kk'}}$ is defined as follows: partition the columns into K^2 regions with each region indexed by the pair (k,k'), with $k,k' = 1,\cdots,K$; the usual lexicographic partial ordering is placed on these pairs, namely $(k_1,k_1') < (k_2,k_2')$ iff $k_1 < k_2$ or $k_1 = k_2$ and $k_1' < k_2'$; then in the region of the columns of $\Psi_{\boldsymbol{\mathcal{H}}}^{(m)}$ corresponding to (k,k'), the entries of the learning matrix are

$$\Psi_{\boldsymbol{\mathcal{H}}}^{(m)}(li, \tilde{n}_{kk'} + p) = \sqrt{\frac{1}{N_{k_i}}}\sum_{i'\in C_{k'}}\frac{1}{N_{k'}}\psi_{kk',p}(r_{ii'}^{(m)}(t_l))\boldsymbol{r}_{ii'}^{(m)}(t_l) \in \mathbb{R}^d,$$

for $i \in C_k$ and $1 \le l \le L$, and $\tilde{n}_{kk'} = \sum_{(k_1,k_1')<(k,k')} n_{k_1k_1'}$.

Then we solve the least squares problem $\arg\min_{\vec{a}} \frac{1}{LM}\sum_{m=1}^{M}\|\Psi_{\boldsymbol{\mathcal{H}}}^{(m)}\vec{a} - \vec{d}_L^{(m)}\|_2^2$ by the normal equation

$$\underbrace{\frac{1}{M}\sum_{m=1}^{M}A_{\boldsymbol{\mathcal{H}}}^{(m)}}_{A_{M,\boldsymbol{\mathcal{H}}}}\vec{a} = \frac{1}{M}\sum_{m=1}^{M}b_{\boldsymbol{\mathcal{H}}}^{(m)}, \tag{3.2}$$

where the trajectory-wise regression matrices are

$$A_{\boldsymbol{\mathcal{H}}}^{(m)} := \frac{1}{LN}(\Psi_{\boldsymbol{\mathcal{H}}}^{(m)})^T\Psi_{\boldsymbol{\mathcal{H}}}^{(m)}, \quad b_{\boldsymbol{\mathcal{H}}}^{(m)} := \frac{1}{LN}(\Psi_{\boldsymbol{\mathcal{H}}}^{(m)})^T\vec{d}_{\boldsymbol{\mathcal{H}}}^{(m)}.$$

Note that the matrices $A_{\boldsymbol{\mathcal{H}}}^{(m)}$ and $b_{\boldsymbol{\mathcal{H}}}^{(m)}$ for different trajectories may be computed in parallel. The size of the matrices $A_{\boldsymbol{\mathcal{H}}}^{(m)}$ is $(\sum_{kk'} n_{kk'}) \times (\sum_{kk'} n_{kk'})$, and there is no need to read and store all the data at once, thereby dramatically reducing memory usage. These are the main reasons why we solve the normal equations instead of the linear system directly associated to the least squares problem; the disadvantage of this approach is that the condition number of $A_{\boldsymbol{\mathcal{H}}}^{(m)}$ is the square of the condition number of the matrix of the linear system, and in situations where the latter is large, passing to the normal equations is not advised. We summarize the learning algorithm in the following table.

Algorithm 1 Learning interaction kernels from observations

Input: Data $\{\boldsymbol{X}^{(m)}(t_l)\}_{l=1,m=1}^{L+1,M}$ and the interval $[0, R]$[1].

1: Use the backward finite difference method to compute the velocity data (if not given) to obtain $\{\boldsymbol{X}^{(m)}(t_l), \dot{\boldsymbol{X}}^{(m)}(t_l)\}_{l=1,m=1}^{L,M}$

2: For each pair (k, k'), partition the interval $[0, R]$ into $\frac{n_{kk'}}{s+1}$ uniform sub-intervals and construct the piecewise polynomial functions of degree s, $\{\psi_{kk',p}\}_{p=1}^{n_{kk'}}$, over the partition.

3: Assemble (in parallel) the normal equation as in (3.2)

$$A_{M,\boldsymbol{\mathcal{H}}}\vec{a} = \vec{b}_{M,\boldsymbol{\mathcal{H}}}$$

4: Solve for $\vec{a}$

$$\vec{a} = A_{M,\boldsymbol{\mathcal{H}}}^{\dagger}\vec{b}_{M,\mathcal{H}}, \tag{3.3}$$

where $A_{M,\boldsymbol{\mathcal{H}}}^{\dagger}$ is the pseudo-inverse of $A_{M,\boldsymbol{\mathcal{H}}} = \Psi_{M,\boldsymbol{\mathcal{H}}}^{\top}\Psi_{M,\boldsymbol{\mathcal{H}}}$ and $\vec{b}_{M,\boldsymbol{\mathcal{H}}} = \Psi_{M,\boldsymbol{\mathcal{H}}}^{\top}\vec{d}_M$.

Output: The estimator $\widehat{\phi} = (\sum_{p=1}^{n_{kk'}} a_{kk',p}\psi_{kk',p})_{k,k'=1}^{K}$.

The total computational cost of constructing estimators, given P CPU's, is $O(ML\frac{N^2d}{P}n^2 + n^3)$. This becomes $O((L\frac{N^2d}{P} + C)M^{1+\frac{2}{2s+1}})$ when n is chosen optimally according to our Theorem and ϕ is at least Lipschitz (corresponding to the index of regularity $s \geq 1$ in the theorem.

3.2 Well-conditioning from coercivity

We could choose different bases of $\boldsymbol{\mathcal{H}}$ to construct the regression matrix $A_{M,\boldsymbol{\mathcal{H}}}$ in (3.2); although the minimizer in $\boldsymbol{\mathcal{H}}$ is of course independent of the choice of basis, the condition number of $A_{L,M,\boldsymbol{\mathcal{H}}}$ does depend on the choice of basis, affecting the numerical performance. The question is, how do we choose the basis functions so that the matrix $A_{L,M,\boldsymbol{\mathcal{H}}}$ is well-conditioned? We show that if the basis is orthonormal in $\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)$, the coercivity constant

[1] R is assumed know; if not, we could estimate it using $R_{\max,M} := \max_{i,i',l,m} \|\boldsymbol{x}_i^{(m)}(t_l) - \boldsymbol{x}_{i'}^{(m)}(t_l)\|$.

provides a lower bound on the smallest singular value of $A_{L,M,\boldsymbol{\mathcal{H}}}$, therefore providing control on the condition number of $A_{L,M,\boldsymbol{\mathcal{H}}}$.

To simplify the notation, we introduce a bilinear functional $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ on $\boldsymbol{\mathcal{H}} \times \boldsymbol{\mathcal{H}}$

$$\langle\!\langle \boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 \rangle\!\rangle := \frac{1}{L}\sum_{l,i=1}^{L,N} \frac{1}{N_{\mathcal{k}_i}} \mathbb{E}_{\mu_0}\left[\left\langle \sum_{i'=1}^{N} \frac{1}{N_{\mathcal{k}_{i'}}} \varphi_{1,\mathcal{k}_i\mathcal{k}_{i'}}(r_{ii'}(t))\boldsymbol{r}_{ii'}(t), \sum_{i'=1}^{N} \frac{1}{N_{\mathcal{k}_{i'}}} \varphi_{2,\mathcal{k}_i\mathcal{k}_{i'}}(r_{ii'}(t))\boldsymbol{r}_{ii'}(t)\right\rangle\right] \tag{3.4}$$

for any $\boldsymbol{\varphi}_1 = (\varphi_{1,kk'})_{k,k'}$, and $\boldsymbol{\varphi}_2 = (\varphi_{2,kk'})_{k,k'} \in \boldsymbol{\mathcal{H}}$. For each pair (k, k') with $1 \le k, k' \le K$, let $\{\psi_{kk',1}, \cdots, \psi_{kk',n_{kk'}}\}$ be a basis of $\mathcal{H}_{kk'} \subset L^\infty([0, R])$ such that

$$\langle \psi_{kk',p}(\cdot)\cdot, \psi_{kk',p'}(\cdot)\cdot \rangle_{L^2(\rho_T^{L,kk'})} = \delta_{p,p'}, \|\psi_{kk',p}\|_\infty \le S_0. \tag{3.5}$$

For each $\psi_{kk',n_{kk'}} \in \mathcal{H}_{kk'}$, we denote by $\boldsymbol{\psi}_{kk',n_{kk'}}$ its canonical embedding in $\boldsymbol{\mathcal{H}}$. Adopting the usual lexicographic partial ordering on pairs (k, k'), we reorder the basis $\{\boldsymbol{\psi}_{kk',1}, \cdots, \boldsymbol{\psi}_{kk',n_{kk'}}\}$ to be $\{\boldsymbol{\psi}_{1+\tilde{n}_{kk'}}, \cdots, \boldsymbol{\psi}_{n_{kk'}+\tilde{n}_{kk'}}\}$, where $\tilde{n}_{kk'} = \sum_{(k_1,k_1')<(k,k')} n_{k_1k_1'}$. Set $n = \sum_{k,k'} n_{kk'} = dim(\boldsymbol{\mathcal{H}})$; then for any function $\boldsymbol{\varphi} \in \boldsymbol{\mathcal{H}}$, we can write $\boldsymbol{\varphi} = \sum_{p=1}^n a_p \boldsymbol{\psi}_p$. We have:

Proposition 8 *Define* $A_{\infty,\boldsymbol{\mathcal{H}}} = \left(\langle\!\langle \boldsymbol{\psi}_p, \boldsymbol{\psi}_{p'} \rangle\!\rangle\right)_{p,p'} \in \mathbb{R}^{n\times n}$. *Then the coercivity constant of* $\boldsymbol{\mathcal{H}} = \text{span}\{\boldsymbol{\psi}_p\}_{p=1}^n$ *is the smallest singular value of* $A_{\infty,\boldsymbol{\mathcal{H}}}$, *i.e.*

$$\sigma_{\min}(A_{\infty,\boldsymbol{\mathcal{H}}}) = c_{L,N,\boldsymbol{\mathcal{H}}}$$

with $c_{L,N,\boldsymbol{\mathcal{H}}}$ *defined in* (2.9). *Moreover, for large* M, *the smallest singular value of* $A_{M,\boldsymbol{\mathcal{H}}}$

$$\sigma_{\min}(A_{M,\boldsymbol{\mathcal{H}}}) \ge 0.9 c_{L,N,\boldsymbol{\mathcal{H}}}$$

with probability at least $1 - 2n\exp(-\frac{c_{L,N,\boldsymbol{\mathcal{H}}}^2 M}{200n^2c_1^2 + \frac{10c_{L,N,\boldsymbol{\mathcal{H}}}c_1}{3}n})$ *with* $c_1 = K^4R^2S_0^2 + 1$.

Proof Due to properties (3.5), the set of functions $\{\boldsymbol{\varphi}_p\}_{p=1}^n \subset \boldsymbol{\mathcal{H}}$ is orthonormal in the sense $\langle \boldsymbol{\psi}_p(\cdot)\cdot, \boldsymbol{\psi}_{p'}(\cdot)\cdot \rangle_{L^2(\boldsymbol{\rho}_T^L)} = \delta_{pp'}$. Then for any $\boldsymbol{\varphi} = \sum_{p=1}^n a_p\boldsymbol{\varphi}_p \in \boldsymbol{\mathcal{H}}$,

$$\begin{aligned}\vec{a}^T A_{\infty,\boldsymbol{\mathcal{H}}}\vec{a} = \langle\!\langle \sum_{p=1}^n a_p\boldsymbol{\psi}_p, \sum_{p=1}^n a_p\boldsymbol{\psi}_p \rangle\!\rangle =& \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}_{\mu_0}\left[\left\|\mathbf{f}_{\sum_{p=1}^n a_p\boldsymbol{\psi}_p}(\boldsymbol{X}(t_l))\right\|_{\mathcal{S}}^2\right] \\ \ge \sigma_{\min}(A_{M,\boldsymbol{\mathcal{H}}})\|\vec{a}\|^2 = \sigma_{\min}(A_{M,\boldsymbol{\mathcal{H}}})&\left\|\boldsymbol{\varphi}(\cdot)\cdot\right\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)}^2.\end{aligned}$$

Thus, the coercivity constant $c_{L,N,\boldsymbol{\mathcal{H}}}$ is $\sigma_{\min}(A_{M,\boldsymbol{\mathcal{H}}})$, since this lower bound is in fact realized by the eigenvector corresponding to the singular value $\sigma_{\min}(A_{M,\boldsymbol{\mathcal{H}}})$.

From the definition of $A_{M,\boldsymbol{\mathcal{H}}}$ in (3.2) and the fact that $A_{\infty,\boldsymbol{\mathcal{H}}} = \mathbb{E}[A_{\boldsymbol{\mathcal{H}}}^{(m)}]$, we have $\lim_{M\to\infty} A_{M,\boldsymbol{\mathcal{H}}} = A_{\infty,\boldsymbol{\mathcal{H}}}$ by the Law of Large Numbers. Using the matrix Bernstein inequality (Theorem 6.1.1 in Tropp (2015)) to control the smallest singular value of $A_{M,\boldsymbol{\mathcal{H}}}$, we obtain that $\sigma_{min}(A_{M,\boldsymbol{\mathcal{H}}})$ is bounded below by $0.9c_{L,N,\boldsymbol{\mathcal{H}}}$ with the desired probability. ■

Proposition 8 also implies that the $O(T/L)$ error in the finite difference approximations leads to a bias of order $O(T/L)$ in the estimator (with high probability). This can be

derived from equation (3.2): the error in the finite difference approximation leads to a bias of order $O(T/L)$ on the vector $b^{(m)}_{\boldsymbol{\mathcal{H}}}$, which is whence passed to the estimator linearly, as $A^{-1}_{M,\boldsymbol{\mathcal{H}}}\frac{1}{M}\sum_{m=1}^{M} b^{(m)}_{\boldsymbol{\mathcal{H}}}$. With high probability, the bias is of the same order as the finite difference error since the smallest singular value of the regression matrix $A_{M,\boldsymbol{\mathcal{H}}}$ is bounded below by the coercivity constant.

From Proposition 8 we see that, for each hypothesis space $\boldsymbol{\mathcal{H}}_{kk'}$, it is important to choose a basis that is well-conditioned in $\boldsymbol{L}^2(\boldsymbol{\rho}^L_T)$, instead of in $\boldsymbol{L}^\infty([0,R])$, for otherwise the matrix $A_{\infty,\boldsymbol{\mathcal{H}}}$ in the normal equations may be ill-conditioned or even singular. This issue can deteriorate in practice when the unknown ρ^L_T is replaced by the empirical measure $\boldsymbol{\rho}^{L,M}_T$. It is therefore advisable to either use piecewise polynomials on a partition of the support of $\boldsymbol{\rho}^{L,M}_T$ or use the pseudo-inverse to avoid the artificial singularity.

4. Numerical examples

We report in this section the learning results of three widely used examples of first-order interacting agent systems: opinion dynamics from social sciences, predator-swarm dynamics from Biology and a heterogeneous particle system inspired by particle Physics. Numerical results demonstrate that our learning algorithm can produce an accurate estimation of the true interaction kernels from observations made in a very short time, and can predict the dynamics, and even collective behaviour of agents, in a larger time interval. We also demonstrate numerically that as the number of observed trajectories M increases, the errors in the estimation of the interaction kernel and in the predicted trajectories decay at rates agreeing with the theory in Section 2. The theoretical results along with the numerical validation shows that our estimators are statistically optimal and computationally efficient.

4.1 Numerical setup

We begin by specifying in detail the setup for the numerical simulations in the examples that follow. We use a large number $M_{\rho^L_T}$ of independent trajectories, not to be used elsewhere, to obtain an accurate approximation of the unknown probability measure ρ^L_T in (2.1). In what follows, to keep the notation from becoming cumbersome, we denote by ρ^L_T this empirical approximation. We run the dynamics over the observation time $[t_1, t_L]$ with M different initial conditions (drawn from the dynamics-specific probability measure μ_0), and the observations consist of the state vector, with no velocity information, at L equidistant time samples in the time interval $[t_1, t_L]$. All ODE systems are evolved using ode15s in MATLAB® with a relative tolerance at 10^{-5} and absolute tolerance at 10^{-6}.

In the numerical experiments, we shall use piecewise constant or piecewise linear functions to estimate the interaction kernels and then use the estimators to predict the dynamics. In order to reduce the stiffness of the differential equations with estimated interaction kernels, we choose a fine grid to linearly interpolate the estimator (and exterpolate it with a constant). This results in Lipschitz continuous estimators.

We report the relative $\boldsymbol{L}^2(\rho^L_T)$ error of our estimators. In the spirit of Theorem 7, we also report the error on trajectories $\boldsymbol{X}(t)$ and $\widehat{\boldsymbol{X}}(t)$ generated by the system with the true interaction kernel and with the learned interaction kernel, respectively, on both the training time interval $[t_1, t_L]$ and on a future time interval $[t_L, t_f]$, with both the same initial

conditions as those used for training, and on new initial conditions (sampled according to μ_0), where the max-in-time trajectory prediction error over time interval $[T_0, T_1]$ is defined as

$$\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}\right\|_{\mathrm{TM}([T_0,T_1])} = \sup_{t\in[T_0,T_1]} \left\|\boldsymbol{X}(t) - \hat{\boldsymbol{X}}(t)\right\|_{\mathcal{S}}. \tag{4.1}$$

The trajectory error will be estimated using M trajectories (we report the mean of the error). We run a total of 10 independent learning trials and compute the mean of the corresponding estimators, their errors, and the trajectory errors just discussed.

Finally, for each example, we also consider the case of noisy observations of the positions. With noise added in the position, the finite difference method used to estimate the velocities will amplify the noise: the error in the estimated velocity will scale as $\frac{\mathrm{std(noise)}}{t_{l+1}-t_l}$ (Wagner et al. (2015)). This issue is treated in the topic of numerical differentiation of noisy data and several approaches have been developed, include the total variation regularization approach (Chartrand (2011)) used in Brunton et al. (2016); Zhang and Lin (2018), high order finite difference method used in Tran and Ward (2017), global and local smoothing techniques (see Knowles and Renka (2014); Wagner et al. (2015); Cleveland and Devlin (1988)) used in Wu and Xiu (2019); Kang et al. (2019), but no technique robust enough to work across a wide variety of examples seems to exist. We have tried to use these techniques in our examples: a combination of position data smoothing techniques and total variation regularization approach worked well in the opinion dynamics but no technique worked well in the Predator-Swarm Dynamics and particle dynamics, likely due to the large Lipschitz constant of the response functions in these two systems. We leave the development of robust techniques, and their theoretical analysis, of this important problem to future work. Here we investigate the effect of noise in learning interaction kernels from an empirical perspective: we consider the simpler setting where we assume the velocities are observed, but both position and velocities are corrupted by noise. In the case of additive noise, the observations are

$$\{(\boldsymbol{X}^{(m)}(t_l) + \eta_{1,l,m}, \dot{\boldsymbol{X}}^{(m)}(t_l)) + \eta_{2,l,m}\}_{l=1,m=1}^{L,M},$$

while in the case of multiplicative noise they are

$$\{(\boldsymbol{X}^{(m)}(t_l) \cdot (1+\eta_{1,l,m}), \dot{\boldsymbol{X}}^{(m)}(t_l)) \cdot (1+\eta_{2,l,m})\}_{l=1,m=1}^{L,M},$$

where in both cases $\eta_{1,l,m}$ and $\eta_{2,l,m}$ are i.i.d. samples from Unif.$([-\sigma, \sigma])$.

For each example, 6 plots display and summarized our numerical results:

- in the first plot, we compare the estimated interaction kernels (after smoothing) to the true interaction kernel(s) , for different values of M, with mean and standard deviation computed over a number of learning trials. In the background we compare $\boldsymbol{\rho}_T^L$ (computed on $M_{\boldsymbol{\rho}_T^L}$ trajectories, as described above) and $\boldsymbol{\rho}_T^{L,M}$ (generated from the observed data consisting of M trajectories).

- The second plot compares the true trajectories (evolved using the true interaction law(s)) and predicted trajectories (evolved using the learned interaction law(s) from a small number M of trajectories) over two different set of initial conditions – one taken from the training data, and one new, randomly generated from μ_0. It also includes the comparison between the true trajectories and the trajectories generated

with the learned interaction kernels, but for a different system with the number of agents $N_{\text{new}} = 4N$, over one set of randomly chosen initial conditions.

- The third plot displays the convergence rate of mean trajectory error with respect to M, both on the training time interval and the future time interval, in which we also compare them with the convergence rate of estimated interaction kernels (those used to produce the predicted trajectories).
- The fourth plot displays the coercivity constant over the hypothesis spaces used in the experiments (see Algorithm 2) and the convergence rate of interaction kernels with and without the observation of true velocities. To validate the applicability of the main Theorems, we report the relative $\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)$ errors of the piecewise polynomial estimators (without smoothing), just as in the main Theorem 6.
- The fifth plot compares the estimators learned from the noisy observations with the true kernels, and their performance in trajectory prediction.
- The last plot shows the convergence rate of our estimators and their smoothed ones when the observations are contaminated by noises.

Algorithm 2 Estimation of the coercivity constant over the hypothesis space $\boldsymbol{\mathcal{H}}$

Input: A set of basis functions $\{\boldsymbol{\varphi}_p\}_{p=1}^n \subset \boldsymbol{\mathcal{H}}$

1: Generate the position data $\{\boldsymbol{X}^{(m)}(t_l)\}_{l=1,m=1}^{L,M}$ with $M = 10^5$.

2: Use the data in step 1 to compute an empirical measure $\tilde{\boldsymbol{\rho}}_T^L$.

3: Apply the Gram-Schmidt process on the $\{\boldsymbol{\varphi}_p\}_{p=1}^n$ to get a new set of basis functions $\{\tilde{\boldsymbol{\varphi}}_p\}_{p=1}^n$ that satisfy

$$\langle \tilde{\boldsymbol{\varphi}}_p(\cdot)\cdot, \tilde{\boldsymbol{\varphi}}_{p'}(\cdot)\cdot \rangle_{\boldsymbol{L}^2(\tilde{\boldsymbol{\rho}}_T^L)} = \delta_{pp'}$$

4: Use the data in step 1 and $\{\tilde{\boldsymbol{\varphi}}_p\}_{p=1}^n$ to assemble the matrix $A_{M,\boldsymbol{\mathcal{H}}}$ (see equation 3.2) and compute its minimal eigenvalue.

Output: $\sigma_{min}(A_{M,\boldsymbol{\mathcal{H}}})$.

4.2 Opinion dynamics

One successful application of first order systems is opinion dynamics in social sciences (see Krause (2000); Blodel et al. (2009); Mostch and Tadmor (2014); Brugna and Toscani (2015); Couzin et al. (2005) and references therein). The interaction function ϕ models how the opinions of pairs of people influence each other. We consider a homogeneous case with interaction kernel defined as

$$\phi(r) = \begin{cases} 0.4, & 0 \le r < \frac{1}{\sqrt{2}} - 0.05, \\ -0.3\cos(10\pi(r - \frac{1}{\sqrt{2}} + 0.05)) + 0.7, & \frac{1}{\sqrt{2}} - 0.05 \le r < \frac{1}{\sqrt{2}} + 0.05, \\ 1, & \frac{1}{\sqrt{2}} + 0.05 \le r < 0.95, \\ 0.5\cos(10\pi(r - 0.95)) + 0.5, & 0.95 \le r < 1.05 \\ 0, & 1.05 \le r \end{cases}$$

This kernel ϕ is compactly supported and Lipschitz continuous with Lipschitz constant 5π. It models heterophilious opinion interactions (see Mostch and Tadmor (2014)) in a homogeneous group of people: each agent is more influenced by its further neighbours than by its closest neighbours. It is shown in Mostch and Tadmor (2014) that heterophilious dynamics enhances consensus: the opinions of agents merge into clusters, with the number of clusters significantly smaller than the number of agents, perhaps contradicting the intuition that would suggest that the tendency to bond more with those who are different rather than with those who are similar would break connections and prevent clusters of consensus.

Suppose the prior information is that ϕ is Lipschitz and compactly supported on $[0, 10]$ (so $R = 10$). Let $\mathcal{H}_n$ be the function space consisting of piecewise constant functions on uniform partitions of [0,10] with n intervals. It is well-known in approximation theory (see the survey DeVore and Lucier (1992)) that $\inf_{\varphi\in\mathcal{H}_n} \|\varphi - \phi\|_\infty \leq \mathrm{Lip}[\phi]n^{-1}$, therefore the conditions in Theorem 6 are satisfied with $s = 1$. Our theory suggests that a choice of dimension n proportional to $(\frac{M}{\log M})^{\frac{1}{3}}$ will yield an optimal convergence rate $M^{-\frac{1}{3}}$ up to a logarithmic factor. We choose $n = 60(\frac{M}{\log M})^{\frac{1}{3}}$. Table 3 summarizes the system and learning parameters.

d	N	$M_{\rho_T^L}$	L	$[t_1; t_L; t_f]$	μ_0	$\deg(\psi)$	n
1	10	10^5	51	$[0; 0.5; 20]$	$\mathcal{U}([0, 8])$	0	$60(\frac{M}{\log M})^{\frac{1}{3}}$

Table 3: (OD) Parameters for the system

Figure 2 shows that as the number of trajectories increases, we obtain more faithful approximations to the true interaction kernel, including near the locations with large derivatives and the support of ϕ. The estimators also perform well near 0, notwithstanding that information of $\phi(0)$ is lost due to the structure of the equations, that have terms of the form $\phi(0)\vec{0} = \vec{0}$.

We then use the learned interaction kernels $\hat{\phi}$ to predict the dynamics, and summarize the results in Figure 3. Even with $M = 16$, our estimator produces very accurate approximations of the true trajectories. Figure 12 displays the accuracy of trajectory predictions. As M increases, the mean trajectory prediction errors decay at the same rate as the convergence rate of the interaction kernel, not only in the training time interval $[0, 0.5]$ (consistently with Theorem 7), but also in the future time interval $[0.5, 20]$ (suggesting the bounds in Theorem 7 may be sometimes overly pessimistic).

To verify the learnability of the interaction kernel, we estimate the coercivity constant on the hypothesis spaces used in the experiments: we partition $[0, 10]$ into n uniform subintervals and choose a set of basis functions consisting of the indicator functions, and then use Algorithm 2. We display the estimated coercivity constant of $\mathcal{H}_n$ for different values of n in Figure 5(a). These numerical results suggest that the coercivity constant, over $L^2([0, 10], \rho_T^L)$, is around 0.08, close to the conjectured lower bound 0.09 based on Theorem 9. We impute this small difference to the finite sample approximation.

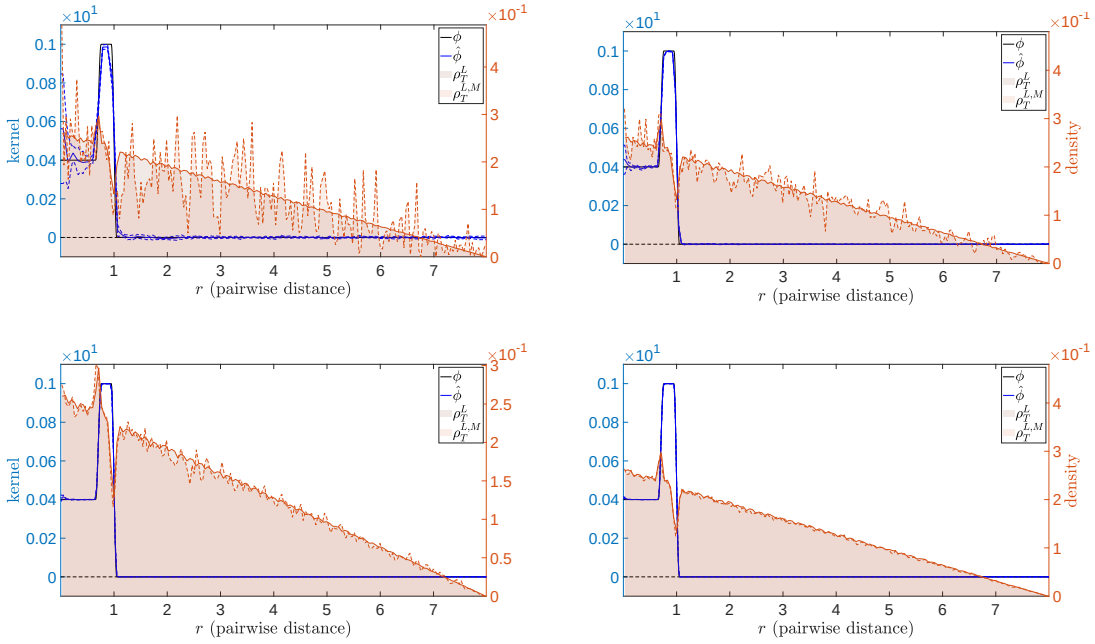

Figure 2: (Opinion Dynamics) Comparison between true and estimated interaction kernels with different values of M, together with histograms (shaded regions) for ρ_T^L and $\rho_T^{L,M}$. In black: the true interaction kernel. In blue: the piecewise constant estimator smoothed by the linear interpolation. From left-top to right-bottom: learning from $M = 2^4, 2^7, 2^{10}, 2^{13}$ trajectories. The standard deviation bars on the estimated interaction kernels become smaller and barely visible. The estimators converge to the true interaction kernel, as also indicated by the relative errors: $(1.5 \pm 0.08) \cdot 10^{-1}$, $(6 \pm 0.5) \cdot 10^{-2}$, $(2.5 \pm 0.03) \cdot 10^{-2}$ and $(8.9 \pm 0.1) \cdot 10^{-3}$.

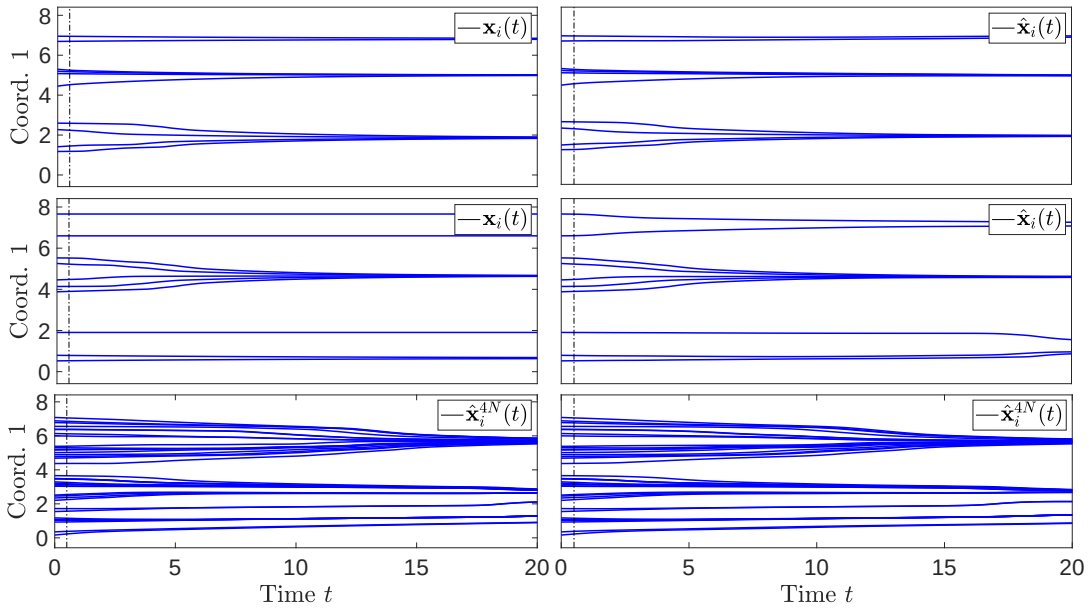

Figure 3: (Opinion Dynamics) $\boldsymbol{X}(t)$ (Left column) and $\hat{\boldsymbol{X}}(t)$ (Right column) obtained with the true kernel ϕ and the estimated interaction kernel $\hat{\phi}$ from $M = 16$ trajectories, for an initial condition in the training data (Top row) and an initial condition randomly chosen (Middle row). The black dashed vertical line at $t = 0.5$ divides the "training" interval $[0, 0.5]$ from the "prediction" interval [0.5, 20]. Bottom row: $\boldsymbol{X}(t)$ and $\hat{\boldsymbol{X}}(t)$ obtained with ϕ and $\hat{\phi}$, for dynamics with larger $N_{new} = 4N$, over one set of initial conditions. We achieve small error in all cases, in particular predicting the number and location of clusters for large time. The mean of max-in-time trajectory errors over 10 learning trials can be found in Figure 12.

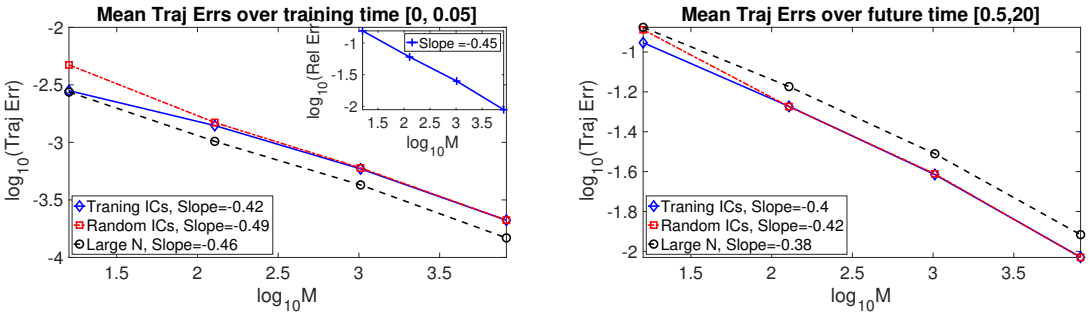

Figure 4: (Opinion Dynamics) Mean errors in trajectory prediction over 10 learning trails using estimated interaction kernels obtained with different values of M: for initial conditions in the training set (Training ICs), randomly drawn from μ_0 (Random ICs), and for a system with $4N$ agents (Large N). Left: Errors over the training time interval [0,0.5]. Right: Errors over the prediction time interval [0.5,20]. Right upper corner inside the left figure: the convergence rate of the relative $L^2(\rho_T^L)$ errors of smoothed estimated interaction kernels. The mean trajectory errors decay at a rate close to the convergence rate of interaction kernels, in agreement with Theorem 7.

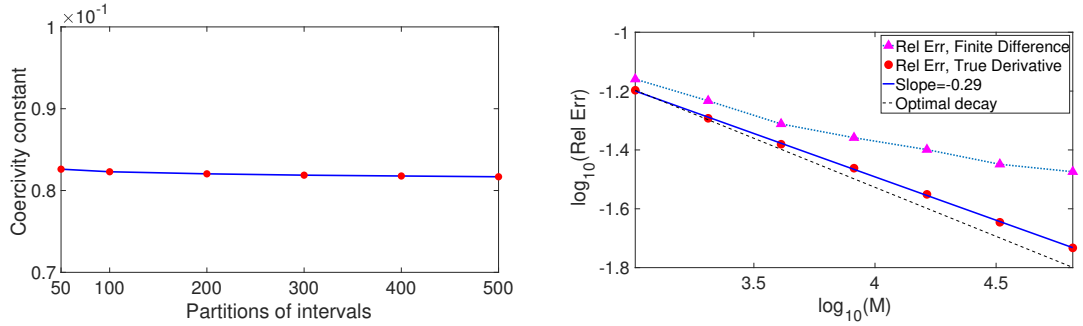

Figure 5: (Opinion Dynamics). Left: The approximate coercivity constant on $\mathcal{H}_n$ that consists of piecewise constant functions over n uniform partitions of $[0, 10]$, obtained from data of 10^5 trajectories. Right: Given true velocities, the convergence rate of the estimators is 0.29, close to the theoretical optimal min-max rate $1/3$ (shown in the black dot line). Otherwise, for unobserved velocities, the curve of the learning error flattens due to the approximation error of velocities by the finite difference method.

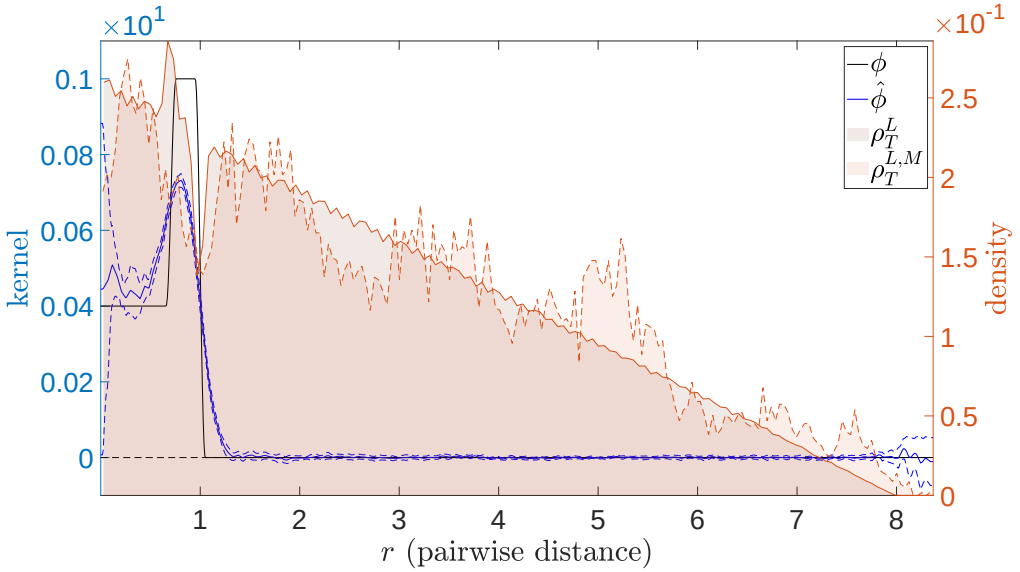

Figure 6: (Opinion Dynamics) Interaction kernel learned with Unif.($[-\sigma, \sigma]$) additive noise, for $\sigma = 0.15$, in the observed positions *and observed velocities*; here $M = 128$, with all the other parameters as in Table 3.

Figure 5(b) shows that the convergence rate of the interaction kernel is $M^{-0.3}$, close to the theoretical optimal rate $M^{-1/3}$ in Theorem 6 up to a logarithmic factor. An interesting phenomenon is that smoothed learned interaction kernel exhibits a convergence rate of $M^{-0.45}$ (see upper-right corners of plots in Figure 12). We explain this phenomenon as follows: the gridded interpolation smoothing techniques make our piecewise constant estimators match well with the true kernel, which is almost piecewise constant, and given the lack of noise, it succeeds in reducing the error in the estimator and yielding an almost parametric convergence rate.

4.3 Predator-Swarm dynamics

There has been a growing literature on modelling interactions between animals of multiple types for the study of animal motion, see Escobedo et al. (2014); Parrish and Edelstein-Keshet (1999); Cohn and Kumar (2009); Nowak (2006); Fryxell et al. (2007). We consider a first-order Predator-Swarm system, modelling interactions between a group of preys and a single predator. The prey-prey interactions have both short-range repulsion to prevent collisions, and long-range attraction to keep the preys in a flock. The preys attract the predator and the predator repels the preys. Since there is only one predator, there are no predator-predator interactions. The intensity of interactions between the single predator and group of preys can be tuned with parameters, determining dynamics with various interesting patterns (from confusing the predator with fast preys, to chasing, to catching up to one prey). We use the set C_1 for the set of preys, and the set C_2 for the single predator.

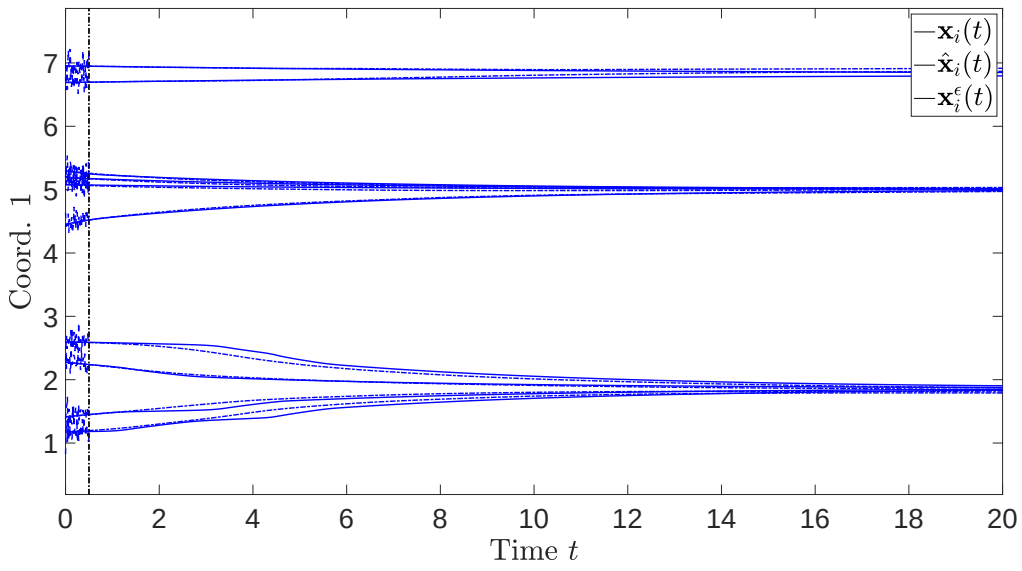

Figure 7: (Opinion Dynamics) One of the observed trajectories before and after being perturbed by the additive noise. The solid lines represent the true trajectory; the dashed semi-transparent lines represent the noisy trajectory used as training data (together with noisy observations of the velocity); the dash dotted lines are the predicted trajectory learned from the noisy trajectory.

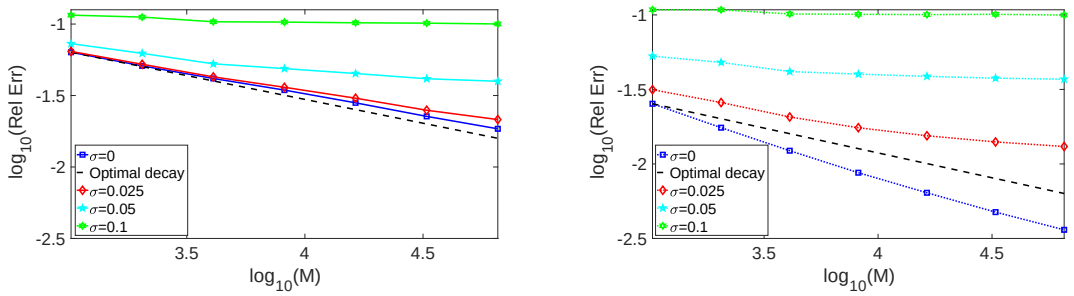

Figure 8: (Opinion Dynamics). The convergence rates of estimators with different levels of additive noise drawn from Unif.($[-\sigma, \sigma]$). The noise causes a flattening of the error, with large noise making improvements in the error negligible as the number of observations increases. Left: Rates for estimators without smoothing. Right: Rates for smoothed estimators.

We consider the interaction kernels

$$\phi_{1,1}(r) = 1 - r^{-2}, \quad \phi_{1,2}(r) = -2r^{-2}, \quad \phi_{2,1}(r) = 3.5r^{-3}, \quad \phi_{2,2}(r) \equiv 0.$$

Since interaction kernels are all singular at $r = 0$, we truncate them at r_{trunc} by connecting it with an exponential function of the form $a\exp(-br)$ so that it has continuous derivative on $\mathbb{R}^+$. The truncation parameters are summarized in Table 4.

kernels	r_{trunc}
$\phi_{1,1}$	0.4
$\phi_{1,2}$	1
$\phi_{2,1}$	0.4
$\phi_{2,2}$	0

Table 4: (PS) Truncation parameters for the Prey-Predator kernels

d	N_1	N_2	$M_{\rho_T^L}$	L	$[t_1; t_L; t_f]$	μ_0	$\deg(\psi_{kk'})$	$n_{kk'}$
2	9	1	10^5	100	[0; 1; 20]	Preys: Unif. disk [0,0.5] Predators: Unif. ring [0.8,1]	1	$100(\frac{M}{\log M})^{\frac{1}{5}}$

Table 5: (PS) System and learning Parameters for the Predator-Swarming system

In the numerical experiments, the initial positions of the preys are sampled from the uniform distribution on the disk with radius 0.5, and the initial position of the predator is sampled from the uniform distribution in the ring with radii between 0.8 and 1. The dynamics mimics the following real situation: preys gather and scatter in a small area; the predator approaches the preys gradually and begins to chase the preys within a small distance; although the predator is able to catch up with the swarm as a whole, the individual prey is able to escape by "confusing" the predator: the preys form a ring with the predator at the centre. Finally, they form a flocking behaviour, i.e., they all run in the same direction.

In this example, we assume that the prior information is that each interaction kernel $\boldsymbol{\phi}_{kk'}$ is in the 2-Hölder space, i.e., its derivative is Lipschitz. Note that the true interaction kernels are not compactly supported. However, our theory is still applicable to this case: due to the compact support of μ_0 and decay of $\boldsymbol{\phi}$ at ∞, Grownwall's inequality implies that, for a sufficiently large R (depending only on $\text{supp}(\mu_0)$, $\|\boldsymbol{\phi}\|_\infty$ and T), $\boldsymbol{\phi}$ and $\boldsymbol{\phi}1_{[0,R]}$ would produce the same dynamics on $[0, T]$ for any initial conditions sampled from μ_0, but now $\boldsymbol{\phi} = \boldsymbol{\phi}1_{[0,R]}$ is in the function space $\mathcal{K}_{R,S}$. Therefore, we can still assume that $\boldsymbol{\phi}$ is compactly supported. Here, we choose $R = 10$ and $\mathcal{H}_{n_{kk'}}$ to be the function space that consists of piecewise linear functions on the uniform partition of [0,10] with n intervals. It is well-known in approximation theory (e.g. DeVore and Lucier (1992)) that $\inf_{\varphi\in\mathcal{H}_n} \|\varphi - \phi_{kk'}1_{[0,10]}\|_\infty \leq \text{Lip}[\phi'_{kk'}]n^{-2}$. Therefore the conditions in Theorem 6 are satisfied with $s = 2$. Our theory suggests that any choice of dimension n that is proportional to $(\frac{M}{\log M})^{1/5}$ yields an optimal convergence rate $M^{-\frac{2}{5}}$, up to a logarithmic factor. We choose $n = 100(\frac{M}{\log M})^{1/5}$ here. The system and learning parameters for Predator-Swarm dynamics are summarized in Table 5.

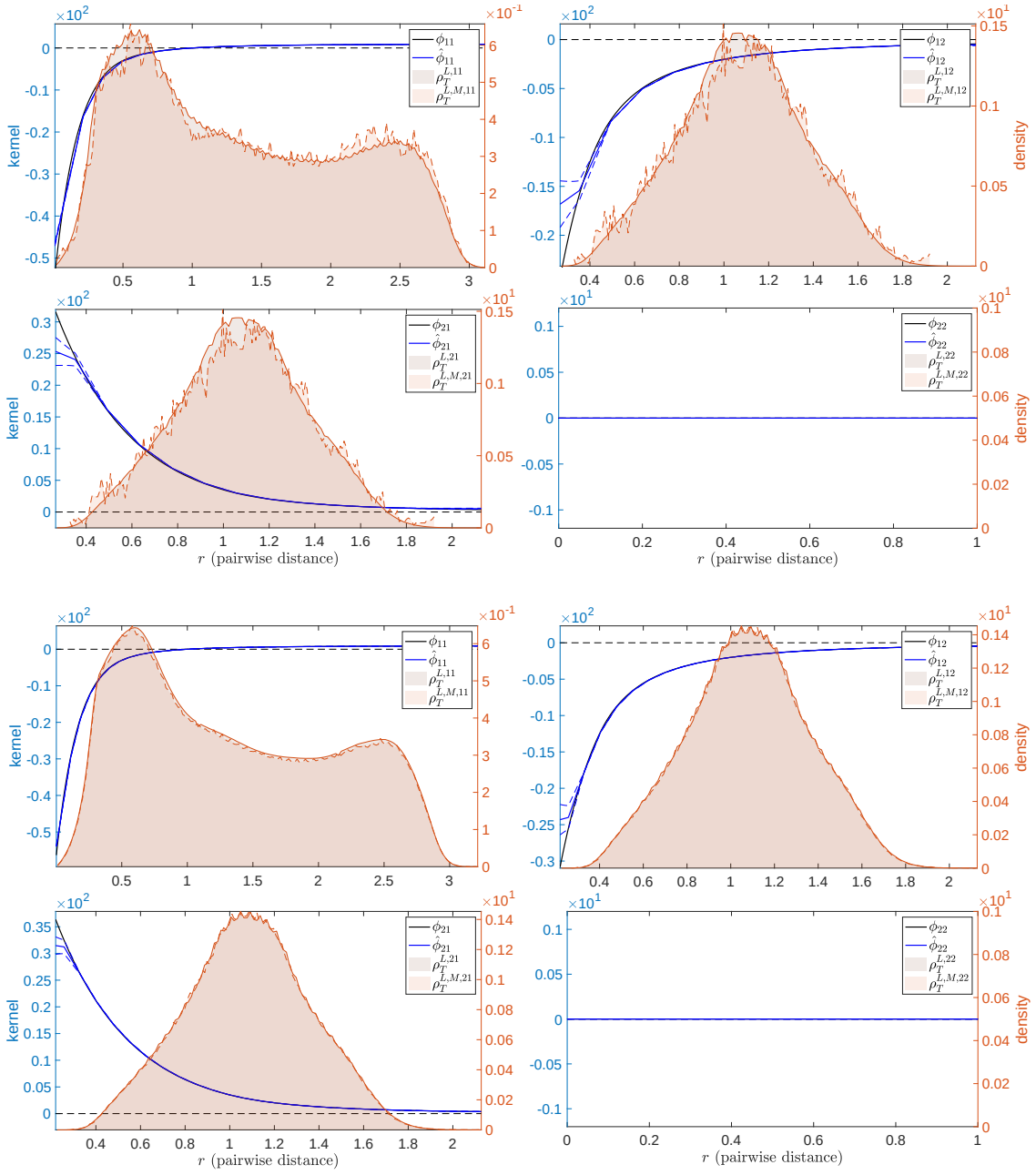

Figure 9: (Predator-Swarm Dynamics) Comparison between true and estimated interaction kernels with $M = 16$ (Top) and 1024 (Bottom). In black: the true interaction kernels. In blue: the learned interaction kernels using piecewise linear functions. When M increases from 16 to 1024, the standard deviation bars on the estimated interaction kernels become smaller and less visible. The relative errors in $\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)$ for the interaction kernels are $(9 \pm 2) \cdot 10^{-3}$ and $(2.5 \pm 0.05) \cdot 10^{-3}$.

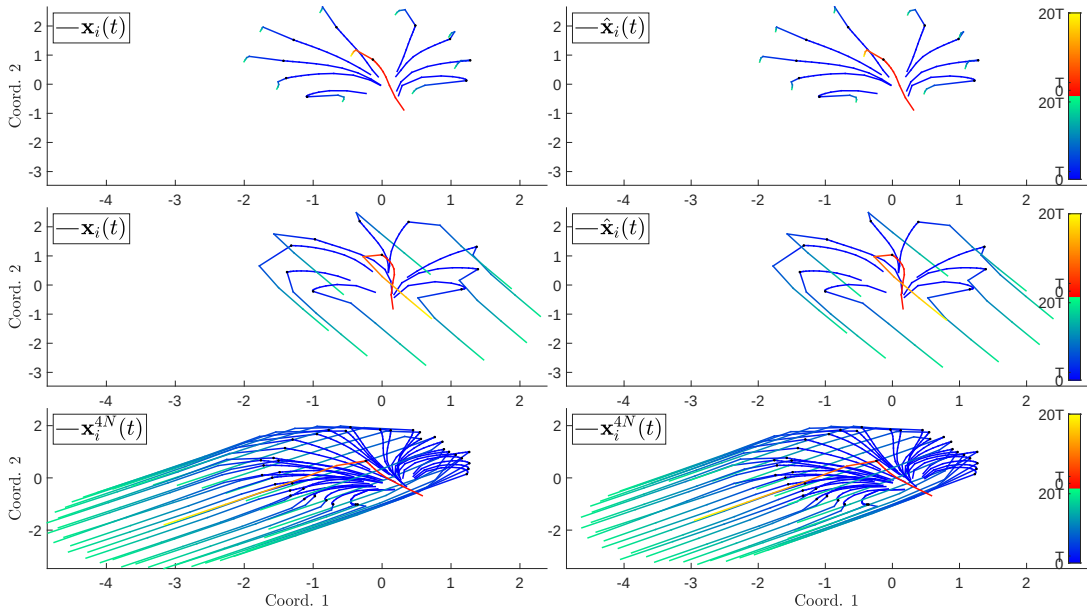

Figure 10: (PS) $\boldsymbol{X}(t)$ and $\hat{\boldsymbol{X}}(t)$ obtained with $\boldsymbol{\phi}$ and $\hat{\boldsymbol{\phi}}$ learned from $M = 16$ trajectories respectively: for an initial condition in the training data (Top) and an initial condition randomly chosen (Middle). The black dot at $t = 1$ divides the "training" interval $[0, 1]$ from the "prediction" interval $[1, 20]$. Bottom: $\boldsymbol{X}(t)$ and $\hat{\boldsymbol{X}}(t)$ obtained with ϕ and $\hat{\phi}$ learned from $M = 16$ trajectories respectively, for dynamics with larger $N_{new} = 4N$, over a set of initial conditions. We achieve small errors in all cases, in particular we predict successfully the flocking time and direction. The means of trajectory errors can be found in Figure 11.

Figure 9 indicates that the estimators match the true interaction kernels extremely well except for a small bias at locations near 0. We impute this error near 0 to two reasons: (i) the strong short-range repulsion between agents force the pairwise distances to stay bounded away from $r = 0$, yielding a $\boldsymbol{\rho}_T^L$ that is nearly singular near 0, so that there are only a few samples to learn the interaction kernels near 0. We see that as M increases, the error near 0 is getting smaller, and we expect it to converge to 0. (ii) Information of $\boldsymbol{\phi}(0)$ is lost due to the structure of the equations, as we mentioned earlier in the previous example, which may cause the error in the finite difference approximation of velocities to affect the reconstruction of values near 0.

Figure 10 shows that with a rather small M, the learned interaction kernels not only produce an accurate approximation of the transient behaviour of the agents over the training time interval $[t_1, t_L]$, but also of the flocking behaviour over the large time interval $[t_L, t_f]$ including the time of formation and the direction of a flocking, which is perhaps beyond expectations.

Figure 11(a) shows that the mean trajectory errors over 10 learning trials decay with M at a rate 0.32 on the training time interval $[0, 1]$, matching the convergence rate of smoothed kernels, even in the case of a new system with $4N$ agents. This agrees with Theorem 7 on

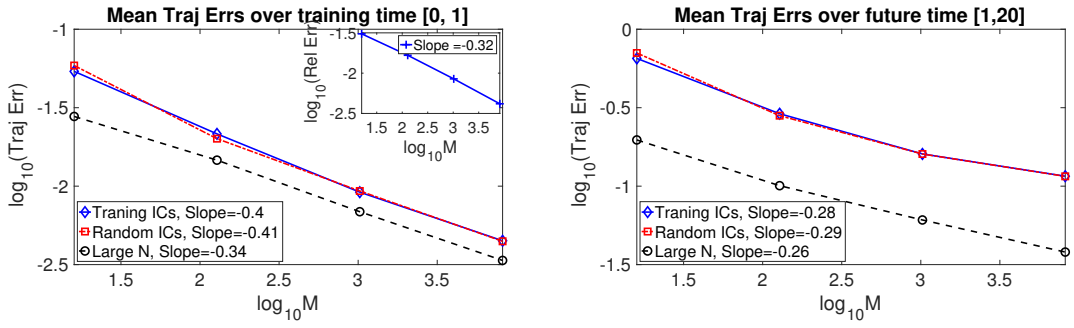

Figure 11: (Predator-Swarm Dynamics) Mean errors in trajectory prediction over 10 learning trials using estimated interaction kernels obtained with different values of M: for initial conditions in the training set (Training ICs), randomly drawn from μ_0 (Random ICs), and for a system with $4N$ agents (Large N). Left: Errors over the training time interval [0,1]. Right: Errors over the future time interval [1,20]. Upper right corner of left figure: the convergence rate of the smoothed learned interaction kernels. The decay rate of the mean trajectory prediction errors over the training time is faster than that of interaction kernels. On the prediction time interval, we still achieve good accuracy for trajectories, with a rate a bit slower than that of interaction kernels.

the convergence rate of trajectories over the training time. For the prediction time interval $[1, 20]$, our learned interaction kernels also produced very accurate approximations of true trajectories in all cases, see Figure 11(b).

Next, we study the learnability of the estimated interaction kernels in this system. As demonstrated by Proposition 8, the coercivity constant is the minimal eigenvalue of $A_{\infty,\boldsymbol{\mathcal{H}}}$, which in our cases is blocked diagonal: one block for learning prey-prey and prey-predator interactions from velocities of preys, and the other block for learning predator-prey interaction from velocities of the predator. We display the minimal eigenvalues for each block in Figure 12(a). We see that the minimal eigenvalue of the prey-prey block matrix stays around $2 \cdot 10^{-2}$ and the predator-predator matrix stays around $0.7 \cdot 10^{-2}$ as partitions get finer. We therefore conjecture that the coercivity constant over $\boldsymbol{L}^2([0, 10], \boldsymbol{\rho}_T^L)$ is about $0.7 \cdot 10^{-2}$.

When true velocities are observed, we obtain a convergence rate for $\|\hat{\boldsymbol{\phi}}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|_{L^2(\boldsymbol{\rho}_T^L)}$ around $M^{-0.35}$ $(\log M/M)^{-0.4})$ (see Figure 12(b)), which matches our theoretical results and is close to the optimal min-max rate $M^{-2/5}$ for regression with noisy observations up to a logarithmic factor. If the velocities were not observed, the convergence rate would be affected. In the right upper corner of Figure 11(a), we see that the convergence rate of the smoothed estimators is around $M^{-0.32}$ if we use the finite difference method to estimate the unobserved velocities, leading to an error in the velocities of size $O(10^{-2})$.

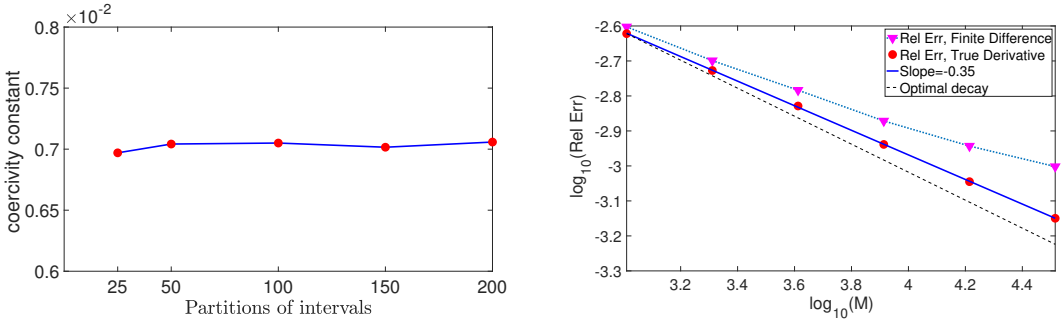

Figure 12: (PS). Left: The coercivity constant on $\mathcal{H}_n$ consisting of piecewise linear functions over n-uniform partitions of the support of $\boldsymbol{\rho}_T^L$, computed from data consisting of $M = 10^5$ trajectories.Right: the relative $L^2(\boldsymbol{\rho}_T^L)$ errors decay at a rate about $(M)^{-0.35}$, close to theoretical optimal min-max rate.

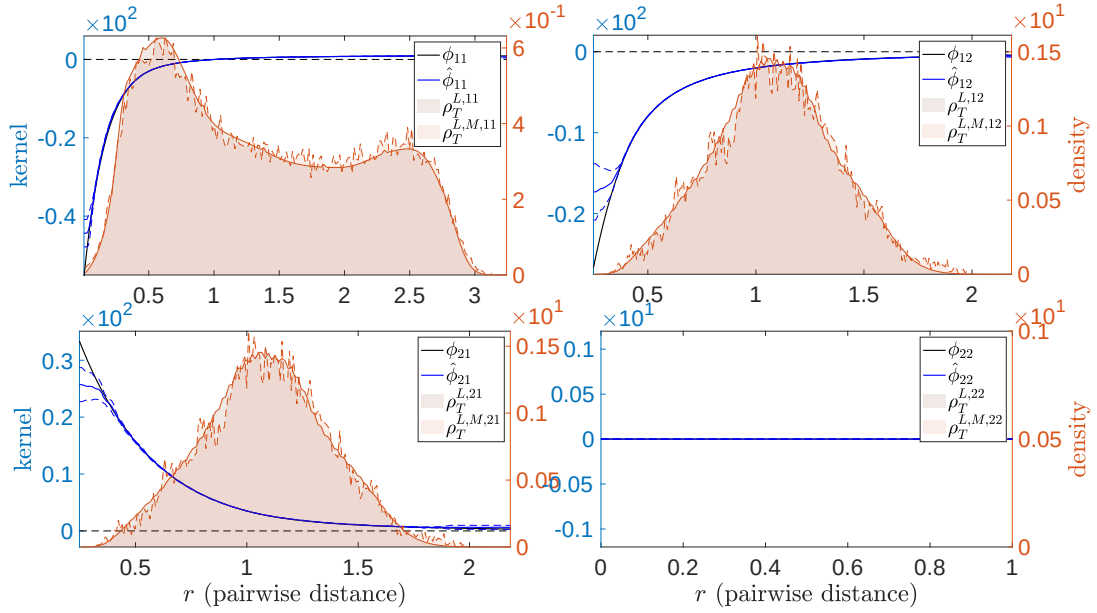

Figure 13: (Predator-Swarm Dynamics) Interaction kernels learned with Unif.($[-\sigma, \sigma]$) multiplicative noise, for $\sigma = 0.1$, in the observed positions *and observed velocities*; here $M = 16$, with all the other parameters as in Table 5.

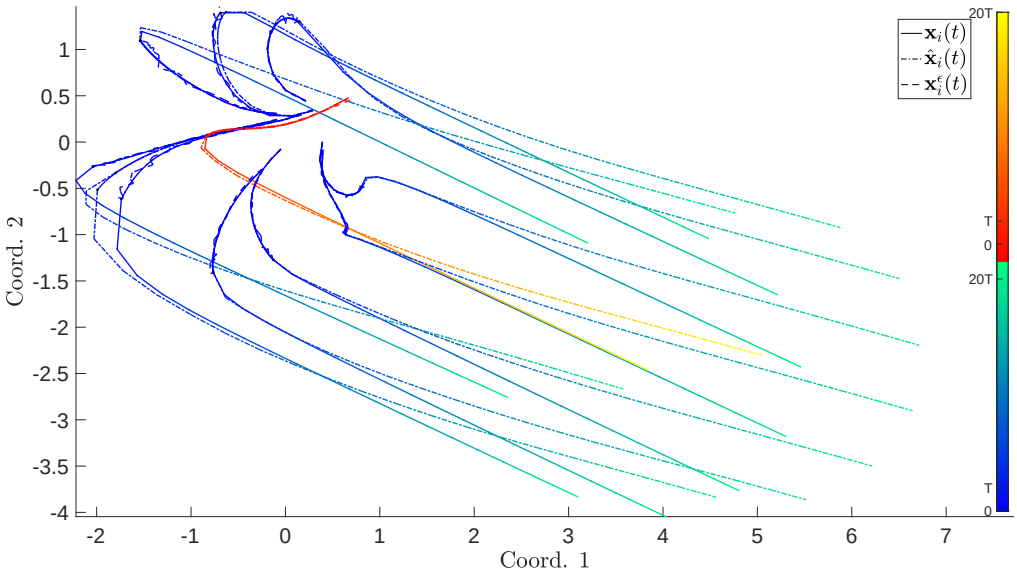

Figure 14: (Predator-Swarm Dynamics) One of the observed trajectories before and after being perturbed by the multiplicative noise drawn from Unif.($[-\sigma, \sigma]$) with $\sigma = 0.1$. The solid lines represent the true trajectory; the dashed semi-transparent lines represent the noisy trajectory used as training data (together with noisy observations of the velocity); the dash dotted lines are the predicted trajectory learned from the noisy trajectory.

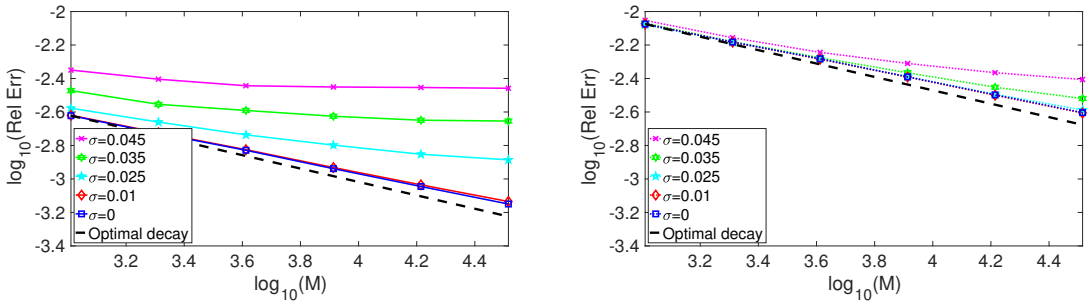

Figure 15: (Predator-Swarm Dynamics). The convergence rates of estimators with different levels of multiplicative noise drawn from Unif.($[-\sigma, \sigma]$). The noise make the learning curve tend to be flat as its level increases. Left: Rates for estimators without smoothing. Right: Rates for smoothed estimators.

4.4 Heterogeneous particle dynamics

We consider here another representative heterogeneous agent dynamics: a particle system with two types of particles (denoted by α and β) governed by Lennard-Jones type potentials (a popular choice for example to model atom-atom interactions in molecular dynamics and materials sciences). The general expression of the Lennard-Jones type potential is

$$\Phi(r) = \frac{p\epsilon}{(p-q)}\left[\frac{q}{p}\left(\frac{r_m}{r}\right)^p - \left(\frac{r_m}{r}\right)^q\right]$$

where ϵ is the depth of the potential well, r is the distance between the particles, and r_m is the distance at which the potential reaches its minimum. At r_m, the potential function has the value $-\epsilon$. The r^{-p} term, which is the repulsive term, describes Pauli repulsion at short ranges due to overlapping electron orbitals, and the r^{-q} term, which is the attractive long-range term, describes attraction at long ranges (modeling van der Waals forces, or dispersion forces). Note that the corresponding Lennard-Jones type kernel $\phi(r) = \frac{\Phi'(r)}{r}$ is singular at $r = 0$: we truncate it at r_{trunc} by connecting it with an exponential function of the form $a\exp(-br^{12})$ so that it is continuous with a continuous derivative on $\mathbb{R}^+$.

In our notation for heterogeneous systems, the set C_1 corresponds to α-particles, and C_2 corresponds to β-particles. The intensity of interaction(s) between particles can be tuned with parameters, determining different types of mixture of particles. In the numerical simulations, we consider interaction kernels $\phi_{1,1}$, $\phi_{1,2}$, $\phi_{2,1}$ and $\phi_{2,2}$ with parameters summarized in Table 6.

kernels	p	q	ϵ	r_m	r_{trunc}
$\phi_{1,1}$	4	1	10	0.8	0.68
$\phi_{1,2}$	8	2	1.5	0.5	0.4
$\phi_{2,1}$	8	2	1.5	0.5	0.4
$\phi_{2,2}$	5	2	5	1	0.8

Table 6: (LJ) Parameters for the Lennard Jones kernels

d	N_1	N_2	$M_{\rho_T^L}$	L	$[t_1; t_L; t_f]$	μ_0	$\deg(\psi_{kk'})$	$n_{kk'}$
2	5	5	10^5	100	$[0; 0.05; 2]$	$\mathcal{N}(\mathbf{0}, I_{20\times 20})$	1	$300(\frac{M}{\log M})^{\frac{1}{5}}$

Table 7: (LJ) Parameters for the system

In the experiments, the particles are drawn i.i.d from standard Gaussian distribution in $\mathbb{R}^2$. In this system, the particle-particle interactions are all short-range repulsions and long-range attractions. The short-range repulsion force prevents the particles to collide and long-range attractions keep the particles in the flock. Both the α-particles and β-particles form the crystal-like clusters. Moreover, the α-β potential function is the same with the β-α potential function. Both of them have the smallest r_m and relative large attractive force (see Table 6) when an α-particle is far from a β-particle. As a result, the attraction force between two different types of particles will force them to mix homogeneously.

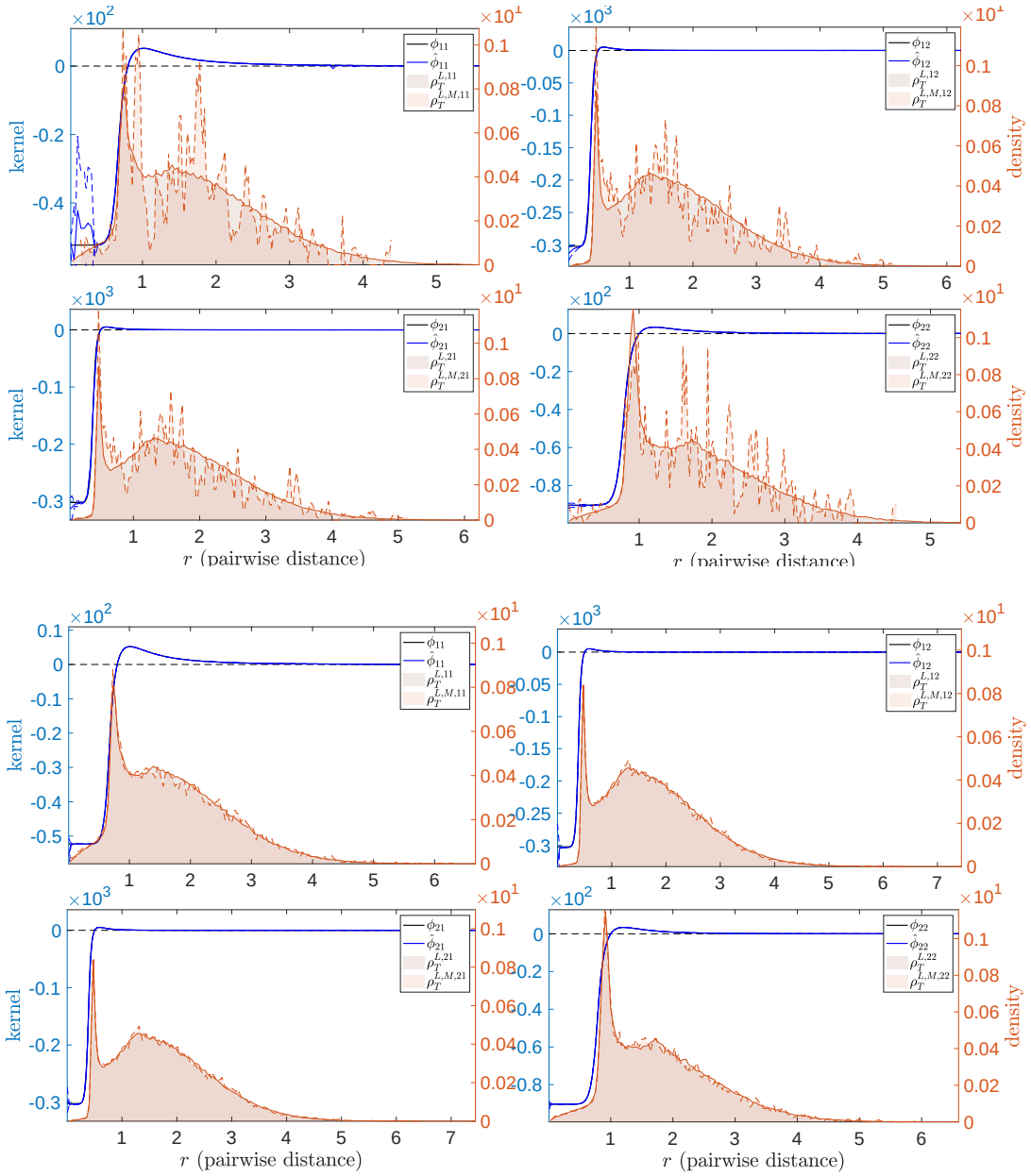

Figure 16: (Lennard Jones Dynamics) Comparison between true and estimated interaction kernels with $M = 16$ (Top) and $M = 512$ (Bottom). In black: the true interaction kernels. In blue: the learned interaction kernels using piecewise linear functions. The learned interaction kernels with $M = 16$ match the true kernels very well except at the region near 0. For a larger $M = 512$, the learned interaction kernels match more faithfully to the true interaction kernels at locations near 0; the standard deviation bars on the estimated interaction kernels become smaller and less visible. The relative errors in $\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)$ norm for the kernels are $(4 \pm 2) \cdot 10^{-2}$ and $(1.2 \pm 0.003.5) \cdot 10^{-2}$.

Since the system evolves to equilibrium configurations very quickly, we observe the dynamics up to a time t_L which is a fraction of the equilibrium time. Since the particles only explore a bounded region due to the large-range attraction, $\boldsymbol{\rho}_T^L$ is essentially supported on a bounded region (see the histogram background of Figure 16), on which the interaction kernels are in the 2-Hölder space. Therefore, our learning theory is still applicable in this case. Similar to the learning in Predator-Swarm dynamics, the estimator of each $\phi_{kk'}$ belongs to a piecewise linear function space over a uniform partition of n intervals. The observation and learning parameters are summarized in Table 7. As reported in Figure 16, with rather small M, the learned interaction kernels $\widehat{\boldsymbol{\phi}}$ approximate the true interaction kernels $\boldsymbol{\phi}$ very well in the regions with large $\boldsymbol{\rho}_T^L$, i.e., regions with an abundance of observed values of pairwise distances to reconstruct the interaction kernels. The reasons for the error near 0 are similar to those for Predator-Swarm dynamics.

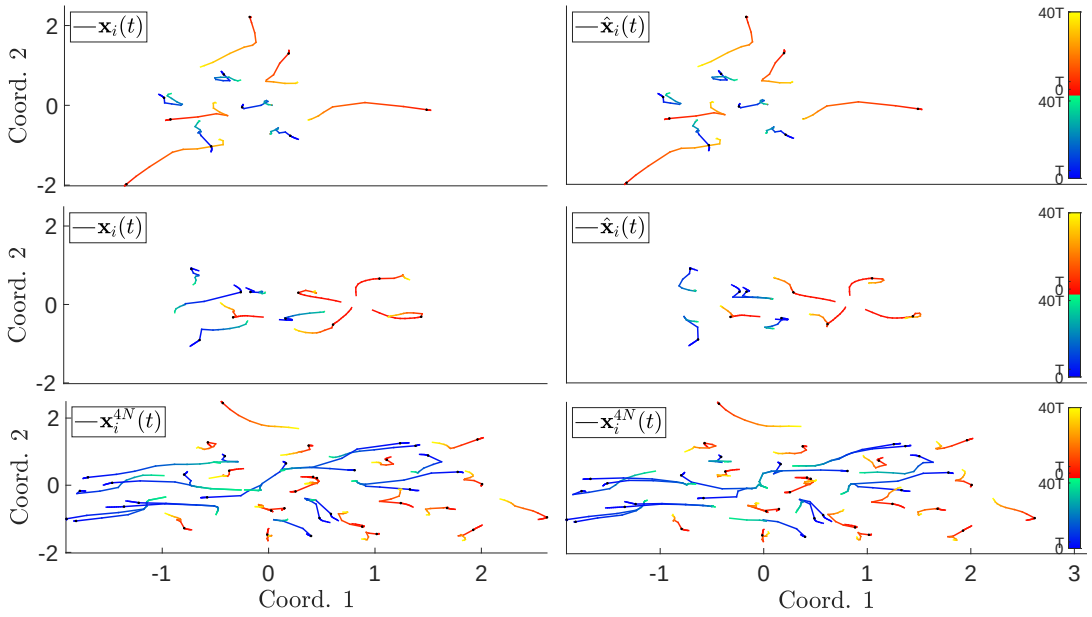

Figure 17: (Lennard Jones Dynamics) $\boldsymbol{X}(t)$ and $\hat{\boldsymbol{X}}(t)$ obtained with $\boldsymbol{\phi}$ and $\widehat{\boldsymbol{\phi}}$ learned with $M = 16$ for an initial condition in the training data (Top row) and a new initial condition random drawn from μ_0 (Middle row). The black dot at $t = 0.05$ divides the training time interval $[0, 0.05]$ from the prediction time interval $[0.05, 2]$. Bottom row: $\boldsymbol{X}(t)$ and $\hat{\boldsymbol{X}}(t)$ obtained with ϕ and $\hat{\phi}$ learned from $M = 16$ trajectories respectively, for dynamics with larger $N_{new} = 4N$, over a set of initial conditions. We achieve small errors, in average, in particular predicting the time and shape of particle aggregation. The means of trajectory prediction errors are in Figure 18.

Figure 17 shows that the learned interaction kernels not only produce an accurate approximation of transient behaviour of particles on the training time interval $[t_1, t_L]$, but also the aggregation of particles over a much larger prediction time interval $[t_L, t_f]$.

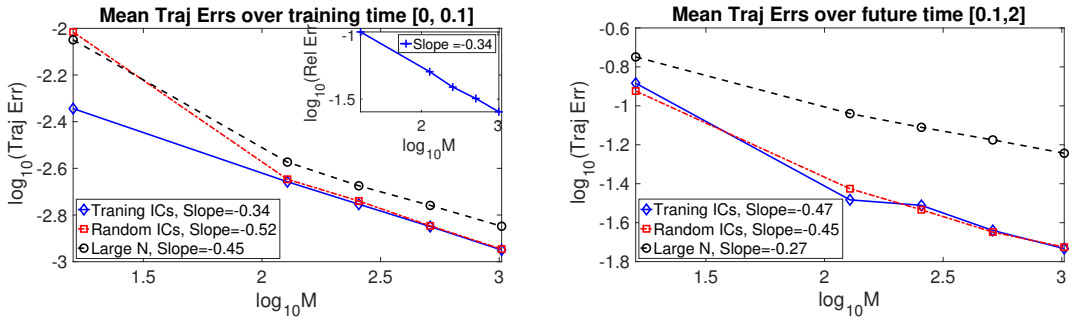

Figure 18: (Lennard Jones Dynamics) Mean errors in trajectory prediction over 10 learning trails using estimated interaction kernels obtained with different values of M: for initial conditions in the training set (Training ICs), randomly drawn from μ_0 (Random ICs), and for a system with $4N$ agents (Large N). Left: Errors over the training time interval [0, 0.05]. Right: Errors over the future time interval [0.05, 2]. Upper right corner of left figure: the convergence rate of kernels (used to predict dynamics). The learning curves of trajectory prediction errors over the training time interval are almost the same with those of interaction kernels. On the prediction time interval, we still achieve a good accuracy of trajectory prediction and just a slightly slower rate for predicting trajectories in all cases.

We summarize the mean trajectory prediction errors over 10 learning trials in Figure 18. It shows that the convergence rate of the trajectory errors over the training time interval $[0, 0.05]$ is the same with the convergence rate of the kernels. For the prediction time interval $[0.05, 2]$, our learned interaction kernels still produced very accurate approximations of true trajectories in all cases, as demonstrated in Figure 18(b).

We then compute the minimal eigenvalue of $A_{\infty,\boldsymbol{\mathcal{H}}}$, inspired by Proposition 8. In this case, $A_{\infty,\boldsymbol{\mathcal{H}}}$ is block diagonal: one block for learning $\phi_{1,1}$ and $\phi_{1,2}$ from velocities of α-particles, and the other block for learning $\phi_{2,1}$ and $\phi_{2,2}$ from velocities of the β-particles. We display the minimal eigenvalues for each block in Figure 19(a). We see that the minimal eigenvalue of the type 1 block matrix stay around $8.7 \cdot 10^{-2}$ and the type 2 block matrix stay around $8.9 \cdot 10^{-2}$ as the partitions get finer, suggesting a positive coercivity constant over $\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)$.

We choose the partition number $n_{kk'}$ for learning $\phi_{kk'}$ according to Theorem 6. Given true velocities, we obtain a convergence rate for $\|\hat{\boldsymbol{\phi}}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|_{L^2(\boldsymbol{\rho}_T^L)}$ around $M^{-0.35}$ (see right column of Figure 19), which is close to the theoretical optimal min-max rate $M^{-2/5}$. The error in finite difference approximation of the velocities affects the convergence rate, as is demonstrated in the right upper corner of Figure 18(a), where the learning curves tends to flatten.

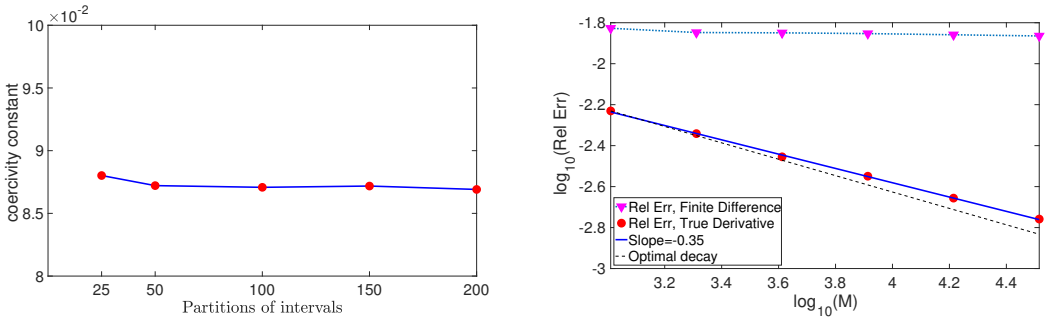

Figure 19: Left: The coercivity constant on $\mathcal{H}_n$ consisting of piecewise linear functions over n-uniform partitions of the support of $\boldsymbol{\rho}_T^L$, computed from data consisting of $M = 10^5$ trajectories. Right: the relative $L^2(\boldsymbol{\rho}_T^L)$ errors decay at a rate $M^{-0.35}$, close to theoretical optimal min-max rate $M^{-0.4}$ up to a logarithmic factor (shown in the black dot line) as in Theorem 6.

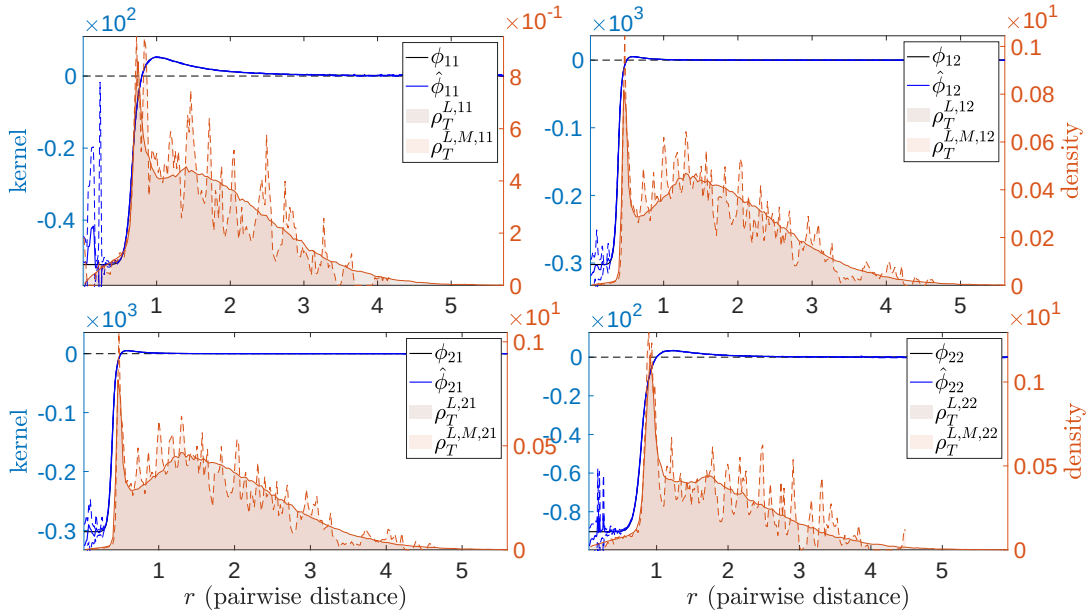

Figure 20: (Lennard Jones Dynamics) Interaction kernels learned with Unif.($[-\sigma, \sigma]$) additive noise, for $\sigma = 0.02$, in the observed positions *and observed velocities*; here $M = 16$, with all the other parameters as in Table 7.

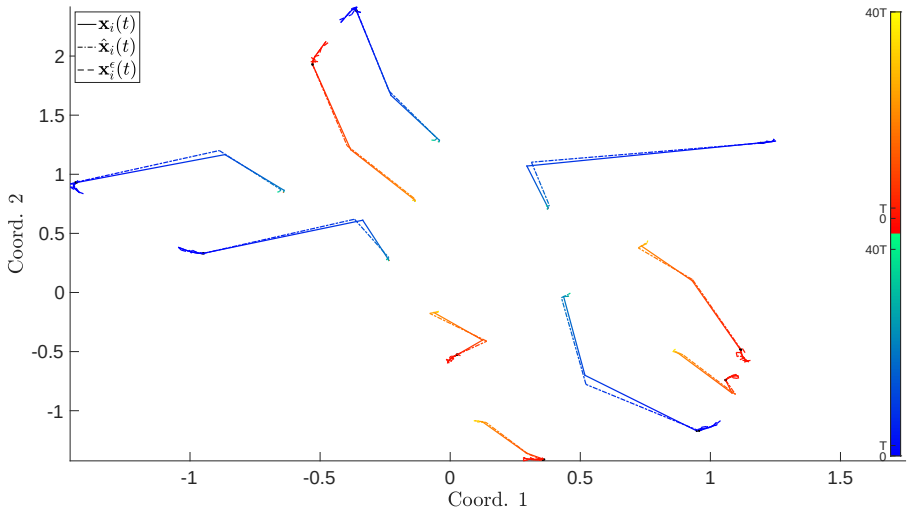

Figure 21: (Lennard Jones Dynamics) One of the observed trajectories before and after being perturbed by the additive noise drawn from Unif.($[-\sigma, \sigma]$) with $\sigma = 0.02$. The solid lines represent the true trajectory; the dashed semi-transparent lines represent the noisy trajectory used as training data (together with noisy observations of the velocity); the dash dotted lines are the predicted trajectory learned from the noisy trajectory.

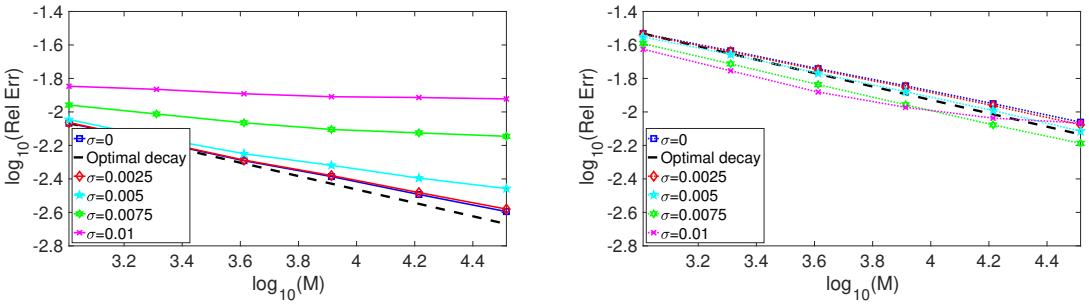

Figure 22: (Lennard Jones Dynamics). The convergence rates of estimators with different levels of additive noise drawn from Unif.($[-\sigma, \sigma]$). Left: Rates for estimators without smoothing. Right: Rates for smoothed estimators.

4.5 Summary of the numerical experiments

- Short time observations don't prevent us from learning the interaction kernels. The randomness of initial conditions enables the agents to explore large regions of state space, and in the space of pairwise distance in a short time. The estimated interaction kernels approximate well in the regions where $\boldsymbol{\rho}_T^L$ is large, i.e. regions with an abundance pairwise distances to reconstruct the interaction kernels. As the number of trajectories increases, we obtain more faithful approximations of the true interaction kernels, agreeing with the consistency Theorem 5. We also see that our estimators, even learned from a small amount of data sampled from a short period of transient dynamics, not only can predict the dynamics on the training time interval $[t_1, t_L]$ but also produce accurate predictions of large time behaviour of the system.

- The decay rate of the mean trajectory prediction errors over the training time interval $[t_1, t_L]$ in terms of M is the same with the convergence rate of the estimated interaction kernels (that are used to predict dynamics), agreeing with Theorem 7.

- The coercivity condition holds on hypothesis spaces that consist of piecewise polynomials, for different kernel functions, and various initial distributions including Gaussian and uniform distributions, and for different L. The convergence rates of the kernels match closely the rate we derived in Theorem 6, which are optimal up to a logarithmic factor.

- Our estimators are robust to the observational noise up to a certain level and still produced rather accurate predictions of the true dynamics. The convergence rate of interaction kernels tends to be flat as the noise level increases, showing our estimators do not overcome the problem exhibited by other estimators of ODEs and PDEs, and do not denoise and recover the true interaction kernel even asymptotically. When the noise is significant, the accuracy of the estimators did not improve as the number of observed trajectories increased.

5. Discussions on the coercivity condition

The coercivity condition on the hypothesis space $\boldsymbol{\mathcal{H}}$, quantized by the constant $c_{L,N,\boldsymbol{\mathcal{H}}}$, plays a vital role in establishing the optimal convergence rate, as is demonstrated in Theorem 4 and Theorem 6. Proposition 8 provides a way to compute the coercivity constant on a finite dimensional $\boldsymbol{\mathcal{H}}$ numerically: it is the minimal eigenvalue of the matrix that yields the estimator by choosing an orthonormal basis of $\boldsymbol{L}^2([0,R], \boldsymbol{\rho}_T^L)$. We have performed extensive numerical experiments to test the coercivity condition for different dynamical systems. Numerical results suggest that the coercivity condition holds true on rather general hypothesis space for a large class of kernel functions, and for various initial distributions including Gaussian and uniform distributions, and for different values of L as long as $\boldsymbol{\rho}_T^L$ is not degenerate.

In the following, we prove the coercivity condition on general compact sets of $\boldsymbol{L}^2([0,R], \boldsymbol{\rho}_T^L)$ under suitable hypotheses, and even independently of N. This implies that the finite sample bounds we achieved in Theorem 6 can be dimension free, suggesting that the coercivity condition may be a fundamental property of the system, even in the mean field limit.

5.1 Homogeneous systems

In a homogeneous system, it is natural to assume the distribution μ_0 of initial conditions is exchangeable (i.e., the distribution is invariant under permutation of components). We prove the coercivity condition for exchangeable Gaussian distributions in the case of $L = 1$. We show that $c_{L,N,\mathcal{H}}$ can be bounded below by a positive constant that is independent of N for any compact set $\mathcal{H}$ in $L^2([0,R],\rho_T^L)$. A key element is connect the coercivity with the strict positiveness of an integral operator, which follows from a Müntz-Szász-type Theorem. We refer to Li et al. (2021) for a further study on the coercivity condition for stochastic homogeneous systems.

Theorem 9 *Consider the system at time $t_1 = 0$ with initial distribution $\boldsymbol{\mu}_0$ being exchangeable Gaussian with $\mathrm{cov}(\boldsymbol{x}_i(t_1)) - \mathrm{cov}(\boldsymbol{x}_i(t_1), \boldsymbol{x}_j(t_1)) = \lambda I_d$ for a constant $\lambda > 0$. Then the coercivity condition holds true on $\mathcal{H} = L^2([0,R],\rho_T^1)$ with the coercivity constant $c_{L,N,\mathcal{H}} = \frac{N-1}{N^2}$. If the hypothesis space $\mathcal{H}$ is a compact subset of $L^2([0,R],\rho_T^1)$, then we have $c_{L,N,\mathcal{H}} = \frac{N-1}{N^2} + \frac{(N-1)(N-2)}{N^2} c_{\mathcal{H}}$ for its coercivity constant, where $c_{\mathcal{H}} > 0$ is independent of N.*

Proof With $L = 1$, the right hand side of the coercivity inequality (2.9) is

$$\begin{aligned}\mathbb{E}_{\mu_0}[\|\mathbf{f}_\varphi(\boldsymbol{X}(t_1))\|_{\mathcal{S}}^2] &= \frac{1}{N^3}\sum_{i=1}^N \mathbb{E}_{\mu_0}\left[\left(\sum_{j=k=1}^N + \sum_{j\neq k=1}^N\right)\varphi(\|\boldsymbol{x}_{ji}(t_1)\|)\varphi(\|\boldsymbol{x}_{ki}(t_1)\|)\langle \boldsymbol{x}_{ji}(t_1), \boldsymbol{x}_{ki}(t_1)\rangle\right] \\ &= \frac{N-1}{N^2}\|\varphi(\cdot)\cdot\|_{L^2([0,R],\rho_T^1)}^2 + \mathcal{R}, \end{aligned} \tag{5.1}$$

where $\mathcal{R} = \frac{1}{N^3}\sum_{i=1}^N \sum_{j\neq k, j\neq i, k\neq i} C_{ijk}$ with

$$C_{ijk} := \mathbb{E}\Big[\varphi(\|\boldsymbol{x}_{ji}(t_1)\|)\varphi(\|\boldsymbol{x}_{ki}(t_1)\|)\langle \boldsymbol{x}_{ji}(t_1), \boldsymbol{x}_{ki}(t_1)\rangle\Big],$$

Because of exchangeability, we have

$$C_{ijk} = \mathbb{E}\Big[\varphi(\|Y-X\|)\varphi(\|Z-X\|)\,\langle Y-X, Z-X\rangle\Big]$$

for all (i,j,k), where X, Y, Z are exchangeable Gaussian random variables with $\mathrm{cov}(X) - \mathrm{cov}(X,Y) = \lambda I_d$. By Lemma 10 below, $C_{ijk} \geq c_{\mathcal{H}}\|\varphi(\cdot)\cdot\|_{L^2([0,R],\rho_T^1)}^2$. Therefore, $\mathcal{R} \geq \frac{(N-1)(N-2)}{N^2} c_{\mathcal{H}}\|\varphi(\cdot)\cdot\|_{L^2([0,R],\rho_T^1)}^2$, and $\langle\!\langle \varphi, \varphi\rangle\!\rangle \geq (\frac{N-1}{N^2} + \frac{(N-1)(N-2)}{N^2} c_{\mathcal{H}})\|\varphi(\cdot)\cdot\|_{L^2([0,R],\rho_T^1)}^2$. ■

The following lemma is a key element in the above proof of the coercivity condition.

Lemma 10 *Let X, Y, Z be exchangeable Gaussian random vectors in $\mathbb{R}^d$ with $\mathrm{cov}(X) - \mathrm{cov}(X,Y) = \lambda I_d$ for a constant $\lambda > 0$. Let ρ_T^1 be a probability measure over $\mathbb{R}^+$ with density function $C_\lambda^{-1} r^{d-1} e^{-\frac{1}{4\lambda}r^2}$ where $C_\lambda = \frac{1}{2}(4\lambda)^{\frac{d}{2}}\Gamma(\frac{d}{2})$. Then,*

$$\mathbb{E}\left[\varphi(|X-Y|)\varphi(|X-Z|)\langle X-Y, X-Z\rangle\right] \geq c_{\mathcal{X}}\|\varphi(\cdot)\cdot\|_{L^2(\rho_T^1)}^2 \tag{5.2}$$

for all $\varphi(\cdot)\cdot \in \mathcal{X}$, with $c_{\mathcal{X}} > 0$ if $\mathcal{X}$ is compact and $c_{\mathcal{X}} = 0$ if $\mathcal{X} = L^2([0,R],\rho_T^1)$.

Proof Let $(U,V) = (X-Y, X-Z)$. Note that the covariance matrix of (U,V) is $\lambda \begin{pmatrix} 2I_d & I_d \\ I_d & 2I_d \end{pmatrix}$. Then

$$\begin{aligned}\mathbb{E}\left[\varphi(|X-Y|)\varphi(|X-Z|)\langle X-Y, X-Z\rangle\right] =& \ \mathbb{E}\left[\varphi(|U|)\varphi(|V|)\langle U,V\rangle\right] \\ =& \int_0^\infty \int_0^\infty \varphi(r) r \varphi(s) s G(r,s) d\rho_T^1(r) d\rho_T^1(s), \quad (5.3)\end{aligned}$$

where the integral kernel $G : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ is

$$G(r,s) = e^{-\frac{1}{12\lambda}(r^2+s^2)} \begin{cases} \frac{1}{2} C_d \left(e^{c_\lambda rs} - e^{-c_\lambda rs}\right), & \text{if } d = 1; \\ C_d \int_{S_1}\int_{S_1} \langle \xi, \eta\rangle e^{c_\lambda rs\langle \xi,\eta\rangle} \frac{d\xi d\eta}{|S_1|^2}, & \text{if } d \geq 2 \end{cases}$$

with $C_d = \left(\frac{\sqrt{3}}{2}\right)^{-d}$, $c_\lambda = \frac{1}{3\lambda}$ and with $|S_1| = 2\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}$ being the surface area of the unit sphere. Define

$$G_R(r,s) = \begin{cases} G(r,s), & 0 \leq r,s \leq R; \\ 0, & \text{otherwise.} \end{cases} \quad (5.4)$$

Note that $G_R(r,s) \in L^2([0,R]\times[0,R], \rho_T^1 \times \rho_T^1)$ is real and symmetric, so its associated integral operator $Q_R : L^2([0,R],\rho_T^1) \to L^2([0,R],\rho_T^1)$

$$Q_R g(r) = \int_0^R G_R(r,s) g(s) d\rho_T^1(s), \quad r \in [0,R] \quad (5.5)$$

is symmetric and compact. Following from (5.3), for any $\varphi(\cdot)\cdot \in L^2([0,R],\rho_T^1)$

$$\mathbb{E}\left[\varphi(|X-Y|)\varphi(|X-Z|)\langle X-Y, X-Z\rangle\right] = \langle Q_R\varphi(\cdot)\cdot, \varphi(\cdot)\cdot\rangle_{L^2([0,R],\rho_T^1)}. \quad (5.6)$$

To show the existence of $c_{\mathcal{X}} \geq 0$ in (5.2), we show that Q_R is strictly positive. We first show that $\langle Q_R g, g\rangle_{L^2([0,R],\rho_T^1)} \geq 0$ for any $g \in L^2([0,R],\rho_T^1)$. When $d = 1$, we have from Taylor expansion that

$$G(r,s) = e^{-\frac{1}{12\lambda}(r^2+s^2)} \sum_{k=0}^\infty \frac{1}{k!} a_k (rs)^k \quad (5.7)$$

with $a_k = \frac{1}{2} C_d c_\lambda^k \left(1 - (-1)^k\right)$. When $d \geq 2$, using the fact that

$$\langle \xi,\eta\rangle e^{c_\lambda rs\langle\xi,\eta\rangle} = \frac{1}{rs}\frac{\partial}{\partial c_\lambda} e^{crs\langle\xi,\eta\rangle} = \sum_{k=1}^\infty \frac{1}{(k-1)!} c_\lambda^{k-1} (rs)^{k-1} \langle\xi,\eta\rangle^k,$$

and the fact that

$$b_{k+1} = \int_{S_1}\int_{S_1} \langle\xi,\eta\rangle^k \frac{d\xi d\eta}{|S_1|^2} \begin{cases} = 0, & \text{for even } k; \\ \in (0,1), & \text{for odd } k, \end{cases}$$

we obtain again (5.7) with $a_k = C_d c_\lambda^k b_{k+1} \geq 0$. Note that for either $d = 1$ or $d \geq 2$, we have $a_k > 0$ when k is odd and $a_k = 0$ when k is even. Therefore, for any $g \in L^2([0,R], \rho_T^1)$ we have

$$\langle Q_R g, g \rangle_{L^2([0,R],\rho_T^1)} = \sum_{k=1, k \text{ odd}}^{\infty} \frac{a_k}{k!} \left(\int_0^R g(r) r^k e^{-\frac{1}{12\lambda} r^2} d\rho_T^1 \right)^2 \geq 0.$$

Next we show $\langle Q_R g, g \rangle_{L^2([0,R],\rho_T^1)} = 0$ implies $g = 0$. Suppose $\langle Q_R g, g \rangle_{L^2([0,R],\rho_T^1)} = 0$: this means that

$$g(r) r e^{-\frac{1}{12\lambda} r^2} \perp \text{span}\{1, r^2, r^4, r^6, \cdots\} \subset L^2([0,R], \rho_T^1).$$

But $\text{span}\{1, r^2, r^4, r^6, \cdots\}$ is a dense set in $L^2([0,R], \rho_T^1)$ by Müntz-Szász Theorem (Borwein and Erdélyi, 1997, Theorem 6.5), therefore $g(r) r e^{-\frac{1}{12\lambda} r^2} = 0$ and hence $g = 0$. This proves that Q_R is strictly positive, which implies that Q_R only has positive eigenvalues with 0 as an accumulation point of eigenvalues. Therefore, for $\varphi(\cdot)\cdot$ in the compact set $\mathcal{X}$ of $L^2([0,R], \rho_T^1)$, we have

$$\langle Q_R \varphi(\cdot)\cdot, \varphi(\cdot)\cdot \rangle_{L^2([0,R],\rho_T^1)} \geq c_{\mathcal{X}} \| \varphi(\cdot) \cdot \|^2_{L^2([0,R],\rho_T^1)},$$

where $c_{\mathcal{X}} > 0$ only depends on $\mathcal{X}$. ■

The following lemma shows that for any $R > 0$, the norm of the operator Q_R is strictly less than 1.

Lemma 11 *The compact operator $Q_R : L^2([0,R], \rho_T^1) \to L^2([0,R], \rho_T^1)$ defined in (5.5) satisfies $\|Q_R\| < 1$.*

Proof Note that $\|Q_R\| \leq 1$ follows directly from the Cauchy-Schwarz inequality:

$$\begin{aligned} \langle Q_R(\varphi), \varphi \rangle_{L^2([0,R],\rho_T^1)} &= \mathbb{E}\left[\varphi(|U|)\varphi(|V|) \langle \frac{U}{|U|}, \frac{V}{|V|} \rangle \right] \leq \mathbb{E}\left[\varphi(|U|)\varphi(|V|) \right] \\ &\leq \sqrt{\mathbb{E}\left[\varphi^2(|V|)\right]} \sqrt{\mathbb{E}\left[\varphi^2(|U|)\right]} = \|\varphi\|^2_{L^2([0,R],\rho_T^1)}, \end{aligned}$$

for any $\varphi \in L^2([0,R], \rho_T^1)$. Suppose $\|Q_R\| = 1$. Since Q_R is compact, then there exists an eigenfunction:

$$Q_R g(r) = g(r),\ r \in [0,R].$$

Now define $\hat{g}(r) = \begin{cases} g(r), & 0 \leq r \leq R; \\ 0, & R < r \leq 2R \end{cases}$. Note that, from its definition in (5.4), $G_{2R}(r,s) = G_R(r,s)$ for all $(r,s) \in [0,R] \times [0,R]$; therefore for $r \in [0,R]$,

$$Q_{2R}\hat{g}(r) = \int_0^{2R} \hat{g}(s) G_{2R}(r,s) d\rho_T^1(s) = \int_0^R g(s) G_{2R}(r,s) d\rho_T^1(s) = Q_R g(r).$$

Therefore, using the fact that $\|Q_{2R}\| \leq 1$, we have

$$\begin{aligned} \langle Q_{2R}\hat{g}, Q_{2R}\hat{g} \rangle_{L^2([0,2R],\rho_T^1)} &= \langle Q_R g, Q_R g \rangle_{L^2([0,R],\rho_T^1)} + \langle Q_{2R}\hat{g}, Q_{2R}\hat{g} \rangle_{L^2([R,2R],\rho_T^1)} \\ &\leq \|\hat{g}\|^2_{L^2([0,2R],\rho_T^1)} = \|g\|^2_{L^2([0,R],\rho_T^1)}. \end{aligned}$$

This means that $Q_{2R}\hat{g} = 0$ a.e. on $[R, 2R]$. However, we now show that $Q_{2R}\hat{g}(r)$ is real analytic over $(0, 2R)$, which leads to a contradiction. To see the analyticity of $Q_{2R}\hat{g}(r)$, we use the power series representation of the kernel $G_{2R}(r, s)$ defined in (5.7), to see that for $s \in [0, 2R]$

$$(Q_{2R}\hat{g})(s) = \int_0^R G_{2R}(r,s)g(r)d\rho_T^1(r) = e^{-\frac{1}{12\lambda}s^2}\int_0^R \sum_{k=1,\text{odd}}^{\infty} \frac{1}{k!}a_k r^k s^k g(r) e^{-\frac{1}{12\lambda}r^2} d\rho_T^1(r)$$
$$= C_d e^{-\frac{1}{12\lambda}s^2} \sum_{k=1,\text{odd}}^{\infty} \frac{1}{k!}\frac{b_{k+1}}{3^k}c_k s^k,$$

where $c_k = \int_0^\infty r^k e^{-\frac{1}{12\lambda}r^2} g(r) d\rho_T^1(r)$ and $b_{k+1} \in (0, 1)$. Then $|c_k| < \|\varphi\|_{L^2(\rho_T^1)}\|r^k\|_{L^2(\rho_T^1)}$. By computation, $\|r^k\|_{L^2(\rho_T^1)} = \sqrt{3^k\Gamma(k + \frac{d-1}{2})/\Gamma(\frac{d}{2})}$. According to Stirling's formula, $\Gamma(z + 1) \sim \sqrt{2\pi z}(z/e)^z$ for positive z and $k! \sim (k/e)^k\sqrt{2\pi k}$; applying the root test, we conclude that the convergence radius for the above series on the right is infinity. Therefore, it is a real analytic function over $(0, 2R)$. ■

Theorem 9 shows a particular case in which the coercivity constant $c_{L,N,\mathcal{H}}$ is positive uniformly in N, and therefore coercivity is a property also of the system in the limit as $N \to \infty$, satisfying the mean-field equations. The coercivity condition has been further discussed in Li et al. (2021), where it is proved for a class of linear and nonlinear stochastic systems of interacting agents.

5.2 Heterogeneous systems

Intuitively, learning interaction kernels in heterogeneous systems seems more difficult than in homogeneous systems, as the observed velocities are the superposition of multiple interaction kernels evaluated at different locations. However, the numerical experiments in subsection 4.1 demonstrate the efficiency of our algorithm in learning multiple interaction kernels simultaneously from their superpositions. In this section, we generalize the arguments of Theorem 9 to heterogeneous systems. In particular, this requires considering the coercivity condition on the function space of multiple interaction kernels. For simplicity of notation, we consider a system with $K = 2$ types of agents, with $C_1 = \{1, \ldots, N\}$ and $C_2 = \{N + 1, \ldots, 2N\}$.

Theorem 12 *Consider the system at time $t_1 = 0$ with initial distribution $\boldsymbol{\mu}_0$ being exchangeable Gaussian with $\text{cov}(\boldsymbol{x}_i(t_1)) - \text{cov}(\boldsymbol{x}_i(t_1), \boldsymbol{x}_j(t_1)) = \lambda I_d$ for a constant $\lambda > 0$. Then the coercivity condition holds true on a hypothesis space $\boldsymbol{\mathcal{H}}$ that is compact in $\boldsymbol{L}^2([0, R], \boldsymbol{\rho}_T^1)$, with constant $c_{L,N,\boldsymbol{\mathcal{H}}} > \frac{(1-c'_{\boldsymbol{\mathcal{H}}})(N-1)}{N^2}$, where $c'_{\boldsymbol{\mathcal{H}}} < 1$.*

Proof Since the initial distribution is exchangeable Gaussian and $L = 1$, we have $\rho_T^{1,kk'} \equiv \rho_T^1$, a probability measure over $\mathbb{R}^+$ with density function $C_\lambda r^{d-1}e^{-\frac{1}{3\lambda}r^2}$ where $C_\lambda = \frac{1}{2}(3\lambda)^{\frac{d}{2}}\Gamma(\frac{d}{2})$.

For $\boldsymbol{\varphi} = (\varphi_{kk'})_{k=1,k'=1}^{2,2} \in \boldsymbol{\mathcal{H}}$,

$$\begin{aligned}\mathbb{E}_{\mu_0}\big\|\mathbf{f}_{\boldsymbol{\varphi}}(\boldsymbol{X}(0))\big\|_{\mathcal{S}}^2 = &\frac{N-1}{N^2}\|\varphi_{11}(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)} + \frac{1}{N}\|\varphi_{12}(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)} + \frac{N-1}{N}\mathcal{R}_1 \\ &+ \frac{1}{N}\|\varphi_{21}(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)} + \frac{N-1}{N^2}\|\varphi_{22}(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)} + \frac{N-1}{N}\mathcal{R}_2,\end{aligned} \tag{5.8}$$

where

$$\mathcal{R}_1 = \frac{N-2}{N}\langle Q\varphi_{11}(\cdot)\cdot, \varphi_{11}(\cdot)\cdot\rangle_{L^2([0,R],\rho_T^1)} + 2\langle Q\varphi_{11}(\cdot)\cdot, \varphi_{12}(\cdot)\cdot\rangle_{L^2([0,R],\rho_T^1)} + \langle Q\varphi_{12}(\cdot)\cdot, \varphi_{12}(\cdot)\cdot\rangle_{L^2([0,R],\rho_T^1)},$$
$$\mathcal{R}_2 = \frac{N-2}{N}\langle Q\varphi_{22}(\cdot)\cdot, \varphi_{22}(\cdot)\cdot\rangle_{L^2([0,R],\rho_T^1)} + 2\langle Q\varphi_{22}(\cdot)\cdot, \varphi_{21}(\cdot)\cdot\rangle_{L^2([0,R],\rho_T^1)} + \langle Q\varphi_{21}(\cdot)\cdot, \varphi_{21}(\cdot)\cdot\rangle_{L^2([0,R],\rho_T^1)},$$

and Q is the integral operator defined in (5.5). Suppose $\boldsymbol{\mathcal{H}} = \bigoplus_{k,k'=1,1}^{2,2}\mathcal{H}_{k,k'}$ with $\mathcal{H}_{k,k'} \subset L^2([0,R],\rho_T^1)$. Let

$$c'_{\boldsymbol{\mathcal{H}}} = \max_{k,k',f\in\mathcal{H}_{k,k'}} \frac{\langle Qf(\cdot)\cdot, f(\cdot)\cdot\rangle_{L^2([0,R],\rho_T^1)}}{\langle f(\cdot)\cdot, f(\cdot)\cdot\rangle_{L^2([0,R]\rho_T^1)}};$$

by Lemma 11, we know $c'_{\boldsymbol{\mathcal{H}}} = \max_{kk'} c'_{\mathcal{H}_{kk'}} < 1$, then we have

$$\|\varphi_{kk'}(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)} \geq \langle Q\varphi_{kk'}(\cdot)\cdot, \varphi_{kk'}(\cdot)\cdot\rangle_{L^2([0,R],\rho_T^1)} + (1-c'_{\boldsymbol{\mathcal{H}}})\|\varphi_{kk'}(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)} \tag{5.9}$$

Combining (5.9) with (5.8) yields

$$\begin{aligned}\mathbb{E}_{\mu_0}\big\|\mathbf{f}_{\boldsymbol{\varphi}}(\boldsymbol{X}(0))\big\|_{\mathcal{S}}^2 \geq &\frac{(1-c'_{\boldsymbol{\mathcal{H}}})(N-1)}{N^2}\left(\|\varphi_{11}(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)} + \|\varphi_{22}(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)}\right) \\ &+ \frac{(1-c'_{\boldsymbol{\mathcal{H}}})}{N}\left(\|\varphi_{12}(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)} + \|\varphi_{21}(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)})\right) \\ &+ \frac{N-1}{N}\left(\langle Q\psi_1,\psi_1\rangle_{L^2([0,R],\rho_T^1)} + \langle Q\psi_2,\psi_2\rangle_{L^2([0,R],\rho_T^1)}\right)\end{aligned} \tag{5.10}$$

where

$$\psi_1 = \sqrt{\frac{N-1}{N}}\varphi_{11}(\cdot)\cdot + \sqrt{\frac{N}{N-1}}\varphi_{12}(\cdot)\cdot \quad \text{and} \quad \psi_2 = \sqrt{\frac{N-1}{N}}\varphi_{22}(\cdot)\cdot + \sqrt{\frac{N}{N-1}}\varphi_{21}(\cdot)\cdot. \tag{5.11}$$

Note that $Q : L^2([0,R],\rho_T^1) \to L^2([0,R],\rho_T^1)$ defined on (5.5) is a compact strictly positive operator. Therefore (5.10) yields that

$$\mathbb{E}_{\mu_0}\big\|\mathbf{f}_{\boldsymbol{\varphi}}(\boldsymbol{X}(0))\big\|_{\mathcal{S}}^2 > \frac{(1-c'_{\boldsymbol{\mathcal{H}}})(N-1)}{N^2}\|\boldsymbol{\varphi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2([0,R],\boldsymbol{\rho}_T^1)}.$$

■

We remark that the inequality (5.10) indicates that $c_{L,N,\boldsymbol{\mathcal{H}}}$ could be independent of N if

$$\langle Q\psi_1,\psi_1\rangle_{L^2([0,R],\rho_T^1)} + \langle Q\psi_2,\psi_2\rangle_{L^2([0,R],\rho_T^1)} \approx \|\boldsymbol{\varphi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2([0,R],\boldsymbol{\rho}_T^1)} \tag{5.12}$$

where the functions ψ_1 and ψ_2 are defined in (5.11). This would be implied by: for $k = 1, 2$, the hypothesis spaces $\mathcal{H}_{k,1}$ and $\mathcal{H}_{k,2}$ have a positive angle as subsets of $L^2([0, R], \rho_T^1)$, so that for $f \in \mathcal{H}_{k,1}$ and $g \in \mathcal{H}_{k,2}$

$$\|f(\cdot)\cdot + g(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)} \geq c'(\|f(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)} + \|g(\cdot)\cdot\|^2_{L^2([0,R],\rho_T^1)}). \tag{5.13}$$

For example, if the supports of true interaction kernels $\phi_{k,1}$ and $\phi_{k,2}$ are disjoint for $k = 1, 2$ and this information is available a priori, then we could choose $\mathcal{H}_{k,1}$ and $\mathcal{H}_{k,2}$ consisting of appropriate functions with disjoint supports. In this case, (5.13) is true with $c' = 1$. Using arguments as in the case of homogeneous systems, one can then show the coercivity constant is positive and independent of N.

Acknowledgments

We would like to thank anonymous reviewers for their constructive comments, which led to significant improvements to the paper. FL and MM are grateful for partial support from NSF-1913243, FL for NSF-1821211; MM is partially supported by NSF-1837991, 1546392, 2031985 and AFOSR-FA9550-17-1-0280/-0288; ST for the AMS Simons travel grant and a start-up fund sponsored by University of California Santa Barbara.

6. Appendix: proofs

In this section, we provide technical details of our main results. For reader's convenience and the sake of a self-contained presentation, we first show that the first order heterogeneous systems (1.1) are well-posed provided the interaction kernels are in the admissible space.

6.1 Well posedness of first order heterogeneous systems

Proposition 13 *Suppose the kernels $\boldsymbol{\phi} = (\phi_{kk'})_{k,k'=1}^K$ lie in the admissible set $\boldsymbol{\mathcal{K}}_{R,S}$, i.e., $\phi_{kk'} \in \mathcal{K}_{R,S}$. Then the first order heterogenous system (1.1) admits a unique global solution in $[0, T]$ for every initial datum $\boldsymbol{X}(0) \in \mathbb{R}^{dN}$ and the solution depends continuously on the initial condition.*

The proof of the Proposition 13 uses Lemma 14 and the same techniques for proving the well-posedness of the homogeneous system (see Section 6 in Bongini et al. (2017)).

Lemma 14 *For any $\varphi \in \mathcal{K}_{R,S}$, the function*

$$F_{[\varphi]}(\boldsymbol{x}) = \varphi(\|\boldsymbol{x}\|)\boldsymbol{x} \quad , \boldsymbol{x} \in \mathbb{R}^d$$

is Lipschitz continuous on $\mathbb{R}^d$.

6.2 Proofs of properties of measures

Proof [Proof of Lemma 1]For each (k, k') with $1 \leq k, k' \leq K$, t and each Borel set $A \subset \mathbb{R}_+$, define

$$\varrho^{kk'}(t, A) = \frac{1}{LN_{kk'}}\mathbb{E}_{\mu_0} \sum_{\substack{i\in C_k, i'\in C_{k'} \\ i\neq i'}} \mathbb{1}_A(r_{ii'}(t)),$$

where $\mathbb{1}_A$ is the indicator function of the set A. Clearly, the measure $\varrho^{kk'}(t,\cdot)$ is the average of the probability distributions of the pairwise distances between type k agents and type k' agents, and therefore it is a probability distribution, and so is $\rho_T^{L,kk'} = \frac{1}{L}\sum_{l=1}^{L}\varrho^{kk'}(t_l,\cdot)$.

To show that $\rho_T^{kk'}(\cdot) = \frac{1}{T}\int_0^T \varrho^{kk'}(t,\cdot)dt$ is well-defined and is a probability measure, it suffices to show that the mapping $t \in [0,T] \to \varrho^{kk'}(t,A)$ is lower semi-continuous for every open set $A \subset \mathbb{R}_+$, and is upper semi-continuous for any compact set A. Fix $t \in [0,T]$. Due to the continuity of the solution to ODE system (see Proposition 13), we have $\left\|\boldsymbol{X}(t_n) - \boldsymbol{X}(t)\right\| \to 0$ if $t_n \to t$, therefore $r_{ii'}(t_n)$ converges to $r_{ii'}(t)$ for each pair (i,i') with $1 \le i,i' \le N$. Since the indicator function of an open set is lower semi-continuous, whereas the indicator function of a closed set is upper semi-continuous, the conclusion follows from the Portmanteau Lemma.

To prove the absolute continuity of $\rho_T^{L,kk'}$ and $\rho_T^{kk'}$ with respect to the Lebesgue measure, let $A \subset \mathbb{R}_+$ with Lebesgue measure zero. Let $P_{ii'}(\boldsymbol{X}) = r_{ii'} : \mathbb{R}^{dN} \to \mathbb{R}_+$ for $i,i' = 1,\dots,N$, and denote $F_t : \mathbb{R}^{dN} \to \mathbb{R}^{dN}$ the forward map such that $\boldsymbol{X}_t = F_t(\boldsymbol{X}_0)$. Observe that $B_{ii'} = P_{ii'}^{-1}(A)$ in a Lebesgue null set in $\mathbb{R}^{dN}$ for each (i,i'), and that the forward map F_t of the dynamical system is continuous, we have

$$\mathbb{E}_{\mu_0}[\mathbb{1}_A(r_{ii'}(t))] = \mu_0\left(F_t^{-1}(P_{ij}^{-1}(A))\right) = 0$$

for each t and each pair (i,j). As a consequence,

$$\rho_T^{L,kk'}(A) = \frac{1}{LN_{kk'}}\sum_{l=1}^{L}\sum_{\substack{i\in C_k, i'\in C_{k'}\\ i\neq i'}} \mathbb{E}_{\mu_0}[\mathbb{1}_A(r_{ii'}(t_l))] = 0$$

and similarly $\rho_T^{kk'}(A) = 0$ by Fubini's Theorem. The previous analysis also implies that, for any Borel set A,

$$\rho_T^{L,kk'}(A) = \sup\{\rho_T^{L,kk'}(K), K \subset A, K \text{ compact}\} = \inf\{\rho_T^{L,kk'}(O), A \subset O, O \text{ open}\}.$$

Therefore, the measure $\rho_T^{L,kk'}$ is a regular measure on $\mathbb{R}_+$. ■

Proof [Proof of Proposition 2] By integration of (1.1) we obtain

$$\|\boldsymbol{x}_i(t)\| \le \|\boldsymbol{x}_i(0)\| + \int_0^t \sum_{i'=1}^{N}\frac{1}{N_{k_{i'}}}|\phi_{k_i k_{i'}}(r_{ii'}(s))|r_{ii'}(s)ds \le \|\boldsymbol{x}_i(0)\| + K\|\boldsymbol{\phi}\|_\infty Rt.$$

Using the fact that μ_0 is compactly supported, we obtain

$$\max_i \|\boldsymbol{x}_i(t)\| \le C_0 + K\|\boldsymbol{\phi}\|_\infty Rt.$$

for some constant C_0 depending only on the size of the support of μ_0. Therefore,

$$\max_{ii'} r_{ii'}(t) \le 2\max_i \|\boldsymbol{x}_i(t)\| \le R_0\,, \quad 0 \le t \le T$$

where $R_0 = 2C_0 + 2K\|\boldsymbol{\phi}\|_\infty RT$. The conclusion follows.

■

6.3 Proofs of Convergence of Estimators

Throughout this section, we assume that

Assumption 15 *$\boldsymbol{\mathcal{H}}$ is a compact convex subset of $\boldsymbol{L}^\infty([0,R])$ and is bounded above by $S_0 \geq S$.*

It is easy to see that $\boldsymbol{\mathcal{H}}$ can be naturally embedded as a compact set of $\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)$. Assumption 15 ensures the existence of minimizers to the error functional $\boldsymbol{\mathcal{E}}_M$ defined in (1.3). We shall first estimate discrepancy of this functional, prove the uniqueness of their minimizers, and then establish uniform estimates on the defect between $\boldsymbol{\mathcal{E}}_M$ and $\boldsymbol{\mathcal{E}}_\infty$.

For $t \in [0,T]$ and $\boldsymbol{\varphi} \in \boldsymbol{\mathcal{H}}$, we introduce the random variable

$$\boldsymbol{\mathcal{E}}_{\boldsymbol{X}(t)}(\boldsymbol{\varphi}) := \left\| \dot{\boldsymbol{X}}(t) - \mathbf{f}_{\boldsymbol{\varphi}}(\boldsymbol{X}(t)) \right\|_{\mathcal{S}}^2, \tag{6.1}$$

where $\|\cdot\|_{\mathcal{S}}$ is defined in (1.4). From its definition, we have $\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}) = \frac{1}{L}\sum_{l=1}^L \mathbb{E}_{\mu_0}[\boldsymbol{\mathcal{E}}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi})]$.

Continuity of the error functionals over $\boldsymbol{\mathcal{H}}$

Proposition 16 *For $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 \in \boldsymbol{\mathcal{H}}$, we have*

$$|\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}_1) - \boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}_2)| \leq K^2 \|\boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \|2\boldsymbol{\phi}(\cdot)\cdot - \boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \tag{6.2}$$

$$|\boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi}_1) - \boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi}_2)| \leq K^4 \|\boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|_\infty \|2\boldsymbol{\phi}(\cdot)\cdot - \boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|_\infty \tag{6.3}$$

Proof Let $\boldsymbol{\varphi}_1 = (\varphi_{kk',1})$ and $\boldsymbol{\varphi}_2 = (\varphi_{kk',2})$. In the following, we use k, k' and k'' to index agent types; recall that C_k is the index set of agents of type k. Using Jensen's inequality,

$$\begin{aligned}
&|\mathcal{E}_{\boldsymbol{X}(t)}(\boldsymbol{\varphi}_1) - \mathcal{E}_{\boldsymbol{X}(t)}(\boldsymbol{\varphi}_2)| \\
&= \left| \sum_{k=1}^K \frac{1}{N_j} \sum_{i\in C_k} \Big\langle \sum_{k'=1}^K \frac{1}{N_{j'}} \sum_{i'\in C_{k'}} (\varphi_{kk',1} - \varphi_{kk',2})(r_{ii'})\boldsymbol{r}_{ii'}, \sum_{k''=1}^K \frac{1}{N_{k''}} \sum_{i'\in C_{k''}} (2\phi_{kk'} - \varphi_{kk',1} - \varphi_{kk',2})(r_{ii'})\boldsymbol{r}_{ii'} \Big\rangle \right| \\
&\leq \sum_{k=1}^K \sum_{k'=1}^K \sum_{k''=1}^K \frac{1}{N_k} \sum_{i\in C_k} \| \frac{1}{N_{k'}} \sum_{i'\in C_{k'}} (\varphi_{kk',1} - \varphi_{kk',2})(r_{ii'})\boldsymbol{r}_{ii'} \| \| \frac{1}{N_{k''}} \sum_{i'\in C_{k''}} (2\phi_{kk''} - \varphi_{kk'',1} - \widehat{\phi}_{kk'',2})(r_{ii'})\boldsymbol{r}_{ii'} \| \\
&< \sum_{j=1}^K \sum_{k'=1}^K \sum_{k''=1}^K \sqrt{\frac{1}{N_k N_{k'}} \sum_{i\in C_k, i'\in C_{k'}} (\varphi_{kk',2} - \varphi_{kk',1})^2 (r_{ii'}) r_{ii'}^2} \times \\
&\qquad \sqrt{\frac{1}{N_k N_{k''}} \sum_{i\in C_k, i'\in C_{k''}} (2\phi_{kk''} - \varphi_{kk'',1} - \phi_{kk'',2})^2 (r_{ii'}) r_{ii'}^2} \\
&< \sum_{k=1}^K \sum_{k'=1}^K \sum_{k''=1}^K \|\varphi_{kk',2}(\cdot)\cdot - \varphi_{kk',1}(\cdot)\cdot\|_{L^2(\hat{\rho}_T^{t,kk'})} \|2\phi_{kk''}(\cdot)\cdot - \varphi_{kk'',1}(\cdot)\cdot - \phi_{kk'',2}(\cdot)\cdot\|_{L^2(\hat{\rho}_T^{t,kk''})} \\
&\leq K^2 \|\boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|_{\boldsymbol{L}^2(\hat{\boldsymbol{\rho}}_T^t)} \|2\boldsymbol{\phi}(\cdot)\cdot - \boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|_{\boldsymbol{L}^2(\hat{\boldsymbol{\rho}}_T^t)},
\end{aligned} \tag{6.4}$$

where

$$\widehat{\rho}_T^{t,kk'} = \frac{1}{LN_{kk'}} \sum_{l=1}^{L} \sum_{\substack{i\in C_k, i'\in C_{k'} \\ i\neq i'}} \delta_{r_{ii'}(t)}(r)dt, \text{ and } \widehat{\boldsymbol{\rho}}_T^t = \bigoplus_{k,k'=1,1}^{K,K} \widehat{\rho}_T^{t,kk'}.$$

Therefore

$$\begin{aligned}
&\big|\frac{1}{L}\sum_{l=1}^{L}\mathcal{E}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi}_1) - \frac{1}{L}\sum_{l=1}^{L}\mathcal{E}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi}_2)\big| \le \frac{1}{L}\sum_{l=1}^{L}\big|\mathcal{E}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi}_2) - \mathcal{E}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi}_1)\big| \\
&< \frac{K^2}{L}\sum_{l=1}^{L}\|\boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|_{\boldsymbol{L}^2(\widehat{\boldsymbol{\rho}}_T^{t_l})}\|2\boldsymbol{\phi}(\cdot)\cdot - \boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|_{\boldsymbol{L}^2(\widehat{\boldsymbol{\rho}}_T^{t_l})} \\
&\le K^2\sqrt{\frac{1}{L}\sum_{l=1}^{L}\|\boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\widehat{\boldsymbol{\rho}}_T^{t_l})}}\sqrt{\frac{1}{L}\sum_{l=1}^{L}\|2\boldsymbol{\phi}(\cdot)\cdot - \boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\widehat{\boldsymbol{\rho}}_T^{t_l})}} \\
&= K^2\|\boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|_{\boldsymbol{L}^2(\widehat{\boldsymbol{\rho}}_T^{L})}\|2\boldsymbol{\phi}(\cdot)\cdot - \boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|_{\boldsymbol{L}^2(\widehat{\boldsymbol{\rho}}_T^{L})} \qquad (6.5)\\
&\le K^4\|\boldsymbol{\varphi}_1(\cdot)\cdot - \boldsymbol{\varphi}_2(\cdot)\cdot\|_\infty\|2\boldsymbol{\phi}(\cdot)\cdot - \boldsymbol{\varphi}_1(\cdot) - \boldsymbol{\varphi}_2(\cdot)\cdot\|_\infty \qquad (6.6)\\
&\le K^4R^2\|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_\infty\|2\boldsymbol{\phi} - \boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_\infty \qquad (6.7)
\end{aligned}$$

Taking expectation with respect to μ_0 on both sides and using (6.5) yields the first inequality. Since

$$|\boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi}_1) - \boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi}_2)| \le \frac{1}{M}\sum_{m=1}^{M}\big|\frac{1}{L}\sum_{l=1}^{L}\mathcal{E}_{\boldsymbol{X}^{(m)}(t_l)}(\boldsymbol{\varphi}_1) - \frac{1}{L}\sum_{l=1}^{L}\mathcal{E}_{\boldsymbol{X}^{(m)}(t_l)}(\boldsymbol{\varphi}_2)\big|,$$

the second inequality in proposition follows by applying (6.6).

■

The following lemma can be immediately deduced using (6.2), (6.3), and (6.7).

Lemma 17 *For all $\boldsymbol{\varphi} \in \boldsymbol{\mathcal{H}}$, we define the defect function*

$$L_M(\boldsymbol{\varphi}) = \boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi}). \qquad (6.8)$$

Then for $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 \in \boldsymbol{\mathcal{H}}$, the estimate

$$|L_M(\boldsymbol{\varphi}_1) - L_M(\boldsymbol{\varphi}_2)| \le 2K^4R^2\|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_\infty\|\boldsymbol{\varphi}_1 + \boldsymbol{\varphi}_2 - 2\boldsymbol{\phi}\|_\infty$$

holds true surely.

Uniqueness of minimizers over compact convex space

Recall the bilinear functional $\langle\!\langle\cdot,\cdot\rangle\!\rangle$ introduced in (3.4)

$$\langle\!\langle\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2\rangle\!\rangle := \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}_{\mu_0}\Big[\,\langle\mathbf{f}_{\boldsymbol{\varphi}_1}(\boldsymbol{X}(t_l)), \mathbf{f}_{\boldsymbol{\varphi}_2}(\boldsymbol{X}(t_l))\rangle_{\mathcal{S}}\,\Big],$$

for any $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 \in \boldsymbol{\mathcal{H}}$. Then the coercivity condition (2.9) can be rephrased as: for all $\boldsymbol{\varphi} \in \boldsymbol{\mathcal{H}}$

$$c_{L,N,\boldsymbol{\mathcal{H}}} \|\boldsymbol{\varphi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \leq \langle\!\langle \boldsymbol{\varphi}, \boldsymbol{\varphi} \rangle\!\rangle$$

Proposition 18 *Let*

$$\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}} := \arg\min_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}} \boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi});$$

then for all $\boldsymbol{\varphi} \in \mathcal{H}$,

$$\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_\infty(\widehat{\boldsymbol{\phi}}_{\infty,\mathcal{H}}) \geq c_{L,N,\boldsymbol{\mathcal{H}}} \|\boldsymbol{\varphi}(\cdot)\cdot - \widehat{\boldsymbol{\phi}}_{\infty,\mathcal{H}}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)}. \tag{6.9}$$

As a consequence, the minimizer of $\boldsymbol{\mathcal{E}}_\infty$ over $\boldsymbol{\mathcal{H}}$ is unique in $\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)$.

Proof For $\boldsymbol{\varphi} \in \boldsymbol{\mathcal{H}}$, we have

$$\begin{aligned}
\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_\infty(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}) &= \langle\!\langle \boldsymbol{\varphi} - \boldsymbol{\phi}, \boldsymbol{\varphi} - \boldsymbol{\phi} \rangle\!\rangle - \langle\!\langle \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}} - \boldsymbol{\phi}, \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}} - \boldsymbol{\phi} \rangle\!\rangle \\
&= \langle\!\langle \boldsymbol{\varphi} - \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}, \boldsymbol{\varphi} + \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}} - 2\boldsymbol{\phi} \rangle\!\rangle \\
&= \langle\!\langle \boldsymbol{\varphi} - \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}, \boldsymbol{\varphi} - \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}} \rangle\!\rangle + 2\langle\!\langle \boldsymbol{\varphi} - \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}, \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}} - \boldsymbol{\phi} \rangle\!\rangle.
\end{aligned}$$

Note that by the coercivity condition, $\langle\!\langle \boldsymbol{\varphi} - \widehat{\boldsymbol{\phi}}_{\infty,\mathcal{H}}, \boldsymbol{\varphi} - \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}} \rangle\!\rangle \geq c_{L,N,\boldsymbol{\mathcal{H}}} \|\boldsymbol{\varphi}(\cdot)\cdot - \widehat{\boldsymbol{\varphi}}_{\infty,\boldsymbol{\mathcal{H}}}(\cdot)\cdot\|^2_{L^2(\boldsymbol{\rho}_T^L)}$. Therefore, to prove (6.9), it suffices to show that $\langle\!\langle \boldsymbol{\varphi} - \widehat{\boldsymbol{\phi}}_{L,\infty,\boldsymbol{\mathcal{H}}}, \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}} - \boldsymbol{\phi} \rangle\!\rangle \geq 0$. To see this, the convexity of $\boldsymbol{\mathcal{H}}$ implies $t\boldsymbol{\varphi} + (1-t)\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}} \in \boldsymbol{\mathcal{H}}$, $\forall t \in [0,1]$. For $t \in (0,1]$, we have

$$\boldsymbol{\mathcal{E}}_\infty(t\boldsymbol{\varphi} + (1-t)\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}) - \boldsymbol{\mathcal{E}}_\infty(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}) \geq 0$$

since $\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}$ is a minimizer in $\boldsymbol{\mathcal{H}}$, therefore

$$\begin{aligned}
&t\langle\!\langle \boldsymbol{\varphi} - \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}, t\boldsymbol{\varphi} + (2-t)\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}} - 2\boldsymbol{\phi} \rangle\!\rangle \geq 0 \\
\Leftrightarrow\ &\langle\!\langle \boldsymbol{\phi} - \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}, t\boldsymbol{\phi} + (2-t)\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\phi}} - 2\boldsymbol{\phi} \rangle\!\rangle \geq 0.
\end{aligned}$$

Since the bilinear functional $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ is continuous over $\boldsymbol{\mathcal{H}} \times \boldsymbol{\mathcal{H}}$ (see Proposition 6.3), letting $t \to 0^+$, by a continuity argument we have $\langle\!\langle \boldsymbol{\varphi} - \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}, 2\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}} - 2\boldsymbol{\phi} \rangle\!\rangle \geq 0$.

■

Uniform estimates on defect functions

Lemma 19 *Denote the minimizer of $\boldsymbol{\mathcal{E}}_\infty(\cdot)$ over $\boldsymbol{\mathcal{H}}$ by*

$$\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}} = \arg\min_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}} \boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}). \tag{6.10}$$

For any $\boldsymbol{\varphi} \in \boldsymbol{\mathcal{H}}$, define

$$\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}) := \boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_\infty(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}), \tag{6.11}$$

$$\mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}) := \boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_M(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}). \tag{6.12}$$

For all $\epsilon > 0$ and $0 < \alpha < 1$, if $\boldsymbol{\varphi}_1 \in \boldsymbol{\mathcal{H}}$ satisfies

$$\frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_1) - \mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_1)}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_1) + \epsilon} < \alpha\,,$$

then for all $\boldsymbol{\varphi}_2 \in \boldsymbol{\mathcal{H}}$ such that $\|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_\infty \leq \frac{\alpha\epsilon}{8S_0R^2K^4}$ we have

$$\frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) - \mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2)}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) + \epsilon} < 3\alpha.$$

Proof For $\boldsymbol{\varphi} \in \boldsymbol{\mathcal{H}}$, recall the definition (6.8) of the defect function $L_M(\boldsymbol{\varphi}) = \boldsymbol{\mathcal{E}}_{L,\infty}(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi})$. We have

$$\begin{aligned}\frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) - \mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2)}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) + \epsilon} &= \frac{\boldsymbol{\mathcal{E}}_\infty(\boldsymbol{\varphi}_2) - \boldsymbol{\mathcal{E}}_\infty(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}) - (\boldsymbol{\mathcal{E}}_M(\boldsymbol{\varphi}_2) - \boldsymbol{\mathcal{E}}_M(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}))}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) + \epsilon} \\ &= \frac{L_M(\boldsymbol{\varphi}_2) - L_M(\boldsymbol{\varphi}_1)}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) + \epsilon} + \frac{L_M(\boldsymbol{\varphi}_1) - L_M(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}})}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) + \epsilon}\end{aligned}$$

By Lemma 17, we have

$$L_M(\boldsymbol{\varphi}_2) - L_M(\boldsymbol{\varphi}_1) \leq 8S_0R^2K^4\|\boldsymbol{\varphi}_2 - \boldsymbol{\varphi}_1\|_\infty \leq \alpha\epsilon\,.$$

Notice that $\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) \geq 0$ and therefore,

$$\frac{L_M(\boldsymbol{\varphi}_1) - L_M(\boldsymbol{\varphi}_2)}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) + \epsilon} \leq \alpha\,.$$

For the second term, we use Proposition 6.3 and the fact $\alpha < 1$ to obtain

$$\mathcal{E}_\infty(\boldsymbol{\varphi}_1) - \mathcal{E}_\infty(\boldsymbol{\varphi}_2) < 4S_0R^2K^4\|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_\infty < \epsilon.$$

This implies that

$$\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_1) - \mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) = \mathcal{E}_\infty(\boldsymbol{\varphi}_1) - \mathcal{E}_\infty(\boldsymbol{\varphi}_2) < \epsilon \leq \epsilon + \mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2),$$

which implies that

$$\frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_1) + \epsilon}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) + \epsilon} \leq 2.$$

But then

$$\frac{L_M(\boldsymbol{\varphi}_1) - L_M(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}})}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) + \epsilon} = \frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_1) - \mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_1)}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_1) + \epsilon}\frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_1) + \epsilon}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_2) + \epsilon} < 2\alpha,$$

and the conclusion follows by summing the last two estimates. ■

Proposition 20 *For all $\epsilon > 0$ and $0 < \alpha < 1$, we have*

$$P_{\mu_0}\left\{ \sup_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}} \frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}) - \mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}) + \epsilon} \geq 3\alpha \right\} \leq \mathcal{N}\left(\boldsymbol{\mathcal{H}}, \frac{\alpha\epsilon}{8S_0R^2K^4}\right) e^{-\frac{c_{L,N,\boldsymbol{\mathcal{H}}}\alpha^2 M\epsilon}{32S_0K^4}}$$

where the covering number $\mathcal{N}(\boldsymbol{\mathcal{H}}, \delta)$ denotes the minimal number of balls in $\boldsymbol{\mathcal{H}}$, with respect to the ∞-norm, with radius δ covering $\boldsymbol{\mathcal{H}}$.

Proof For all $\epsilon > 0$ and $0 < \alpha < 1$, we first show that for any $\boldsymbol{\varphi} \in \boldsymbol{\mathcal{H}}$,

$$P_{\mu_0}\left\{ \frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}) - \mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}) + \epsilon} \geq \alpha \right\} \leq e^{\frac{-c_{L,N,\boldsymbol{\mathcal{H}}}\alpha^2 M\epsilon}{32S_0^2K^4}}.$$

We consider the random variable $\frac{1}{L}\sum_{l=1}^{L}\big(\boldsymbol{\mathcal{E}}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_{\boldsymbol{X}(t_l)}(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}})\big)$, and let σ^2 be its variance. From (6.2) and the coercivity condition (2.9), we obtain that

$$\begin{aligned}
\sigma^2 &\leq \mathbb{E}\big[\big|\frac{1}{L}\sum_{l=1}^{L}\big(\boldsymbol{\mathcal{E}}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_{\boldsymbol{X}(t_l)}(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}})\big)\big|^2\big] \\
&\leq \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\big[\big|\boldsymbol{\mathcal{E}}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_{\boldsymbol{X}(t_l)}(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}})\big|^2\big] \\
&\leq K^4\|\boldsymbol{\varphi}(\cdot)\cdot - \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)}\|\boldsymbol{\varphi}(\cdot)\cdot + \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}(\cdot)\cdot - 2\boldsymbol{\phi}(\cdot)\cdot\|^2_{\infty} \\
&\leq \frac{K^4}{c_{L,N,\boldsymbol{\mathcal{H}}}}(\boldsymbol{\mathcal{E}}_{\infty}(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_{\infty}(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}))\|\boldsymbol{\varphi}(\cdot)\cdot + \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}(\cdot)\cdot - 2\boldsymbol{\phi}(\cdot)\cdot\|^2_{\infty} \\
&\leq \frac{(2S_0+2S)^2R^2K^4}{c_{L,N,\boldsymbol{\mathcal{H}}}}(\boldsymbol{\mathcal{E}}_{\infty}(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_{\infty}(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}})) \\
&\leq \frac{16S_0^2R^2K^4}{c_{L,N,\boldsymbol{\mathcal{H}}}}\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}).
\end{aligned} \tag{6.13}$$

We also have that almost surely

$$|\frac{1}{L}\sum_{l=1}^{L}\big(\boldsymbol{\mathcal{E}}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_{\boldsymbol{X}(t_l)}(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}})\big)| \leq 8S_0^2R^2K^4.$$

Applying the one-sided Bernstein's inequality to the random variable

$$\frac{1}{L}\sum_{l=1}^{L}\big(\boldsymbol{\mathcal{E}}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi}) - \boldsymbol{\mathcal{E}}_{\boldsymbol{X}(t_l)}(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}})\big),$$

we obtain that

$$P_{\mu_0}\big\{\frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}) - \mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}) + \epsilon} \geq \alpha\big\} \leq e^{-\frac{\alpha^2(\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon)^2M}{2(\sigma^2+\frac{8S_0^2R^2K^4\alpha(\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon)}{3})}}.$$

Now we estimate

$$\frac{c_{L,N,\boldsymbol{\mathcal{H}}}\epsilon}{32S_0^2R^2K^6} \leq \frac{(\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon)^2}{2(\sigma^2+\frac{8S_0^2R^2K^4\alpha(\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon)}{3})}\,, \text{i.e.}$$
$$\frac{c_{L,N,\boldsymbol{\mathcal{H}}}\epsilon}{16S_0^2R^2K^6}(\sigma^2+\frac{8S_0^2R^2K^4\alpha(\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon)}{3}) \leq (\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon)^2.$$

By the estimate (6.13), since $0<\alpha\leq 1$ and $0<c_L<K^2$ it suffices to show

$$\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})\epsilon+\frac{\epsilon(\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon)}{6} \leq (\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon)^2,$$

which follows from $2\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})\epsilon+\epsilon^2 \leq (\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon)^2$.

Given $\epsilon>0$, consider $\boldsymbol{\varphi}_j$ such that the disks D_j centered at $\boldsymbol{\varphi}_j$ and with radius $\frac{\alpha\epsilon}{8S_0R^2K^4}$ cover $\boldsymbol{\mathcal{H}}$ for $j=1,\cdots,\mathcal{N}(\boldsymbol{\mathcal{H}},\frac{\alpha\epsilon}{8S_0R^2K^4})$. By Lemma 19, we have

$$\sup_{\boldsymbol{\varphi}\in D_j}\frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})-\mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon} \geq 3\alpha \Rightarrow \frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_j)-\mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_j)}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_j)+\epsilon} \geq \alpha.$$

We conclude that, for each j,

$$P_{\mu_0}\left\{\sup_{\boldsymbol{\varphi}\in D_j}\frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})-\mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon} \geq 3\alpha\right\} \leq P_{\mu_0}\left\{\frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_j)-\mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_j)}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}_j)+\epsilon} \geq \alpha\right\} \leq e^{-\frac{c_{L,N,\boldsymbol{\mathcal{H}}}\alpha^2M\epsilon}{32S_0^2R^2K^6}}.$$

Since $\boldsymbol{\mathcal{H}}=\cup_j D_j$,

$$\begin{aligned}P_{\mu_0}\left\{\sup_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}}\frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})-\mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon} \geq 3\alpha\right\} &\leq \sum_j P_{\mu_0}\left\{\sup_{\boldsymbol{\varphi}\in D_j}\frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})-\mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon} \geq 3\alpha\right\}\\ &\leq \mathcal{N}(\boldsymbol{\mathcal{H}},\frac{\alpha\epsilon}{8S_0R^2K^4})e^{-\frac{c_{L,N,\boldsymbol{\mathcal{H}}}\alpha^2M\epsilon}{32S_0^2R^2K^6}}.\end{aligned}$$

■

Proof [**of Theorem** 4] Put $\alpha=\frac{1}{6}$ in Proposition 20. We know that, with probability at least

$$1-\mathcal{N}(\boldsymbol{\mathcal{H}},\frac{\epsilon}{48S_0R^2K^4})e^{-\frac{c_{L,N,\boldsymbol{\mathcal{H}}}M\epsilon}{1152S_0^2R^2K^6}},$$

we have

$$\sup_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}}\frac{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})-\mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})}{\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\epsilon} < \frac{1}{2},$$

and therefore, for all $\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}$,

$$\frac{1}{2}\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi}) < \mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\boldsymbol{\varphi})+\frac{1}{2}\epsilon.$$

Taking $\boldsymbol{\varphi}=\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}}$, we have

$$\mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}}) < 2\mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}})+\epsilon\,.$$

But $\mathcal{D}_{M,\boldsymbol{\mathcal{H}}}(\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}}) = \mathcal{E}_M(\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}}) - \mathcal{E}_M(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}) \leq 0$ and hence by Proposition 18 we have

$$c_{L,N,\boldsymbol{\mathcal{H}}}\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}}(\cdot)\cdot - \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \leq \mathcal{D}_{\infty,\boldsymbol{\mathcal{H}}}(\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}}) < \epsilon.$$

Therefore,

$$\begin{aligned}
\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} &\leq 2\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}}(\cdot)\cdot - \widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}(\cdot)\cdot\|^2_{L^2(\boldsymbol{\rho}_T^L)} + 2\|\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \\
&\leq \frac{2}{c_{L,N,\boldsymbol{\mathcal{H}}}}(\epsilon + \inf_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}}\|\boldsymbol{\varphi}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)}) \\
&\leq \frac{2}{c_{L,N,\boldsymbol{\mathcal{H}}}}(\epsilon + \inf_{\boldsymbol{\varphi}\in\boldsymbol{\mathcal{H}}} R^2\|\boldsymbol{\varphi} - \boldsymbol{\phi}\|^2_\infty)
\end{aligned}$$

where the last two inequalities follows from the coercivity condition and by the definition of $\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}}$. Given $0 < \delta < 1$, we let

$$1 - \mathcal{N}(\boldsymbol{\mathcal{H}}, \frac{\epsilon}{48S_0R^2K^4})e^{-\frac{c_{L,N,\boldsymbol{\mathcal{H}}}M\epsilon}{1152S_0^2R^2K^6}} \geq 1 - \delta$$

and the conclusion follows. ■

Proof [**of Theorem** 5] According to the definition of the coercivity constant, we have $c_{L,N,\boldsymbol{\mathcal{H}}_M} \geq c_{L,N,\cup_M\boldsymbol{\mathcal{H}}_M}$. Then by the proof of Theorem 4, we obtain

$$c_{L,N,\cup_M\boldsymbol{\mathcal{H}}_M}\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \leq \boldsymbol{\mathcal{D}}_{\infty,\boldsymbol{\mathcal{H}}_M}(\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}) + \boldsymbol{\mathcal{E}}_\infty(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}_M}) \tag{6.14}$$

For any $\epsilon > 0$, the inequality (6.14) implies that

$$\begin{aligned}
&P_{\mu_0}\{c_{L,N,\cup_M\boldsymbol{\mathcal{H}}_M}\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \geq \epsilon\} \\
&\quad\leq P_{\mu_0}\{\boldsymbol{\mathcal{D}}_{\infty,\boldsymbol{\mathcal{H}}_M}(\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}) + \boldsymbol{\mathcal{E}}_\infty(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}_M}) \geq \epsilon\} \\
&\quad\leq P_{\mu_0}\{\boldsymbol{\mathcal{D}}_{\infty,\boldsymbol{\mathcal{H}}_M}(\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}) \geq \frac{\epsilon}{2}\} + P_{\mu_0}\{\boldsymbol{\mathcal{E}}_\infty(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}_M}) \geq \frac{\epsilon}{2}\}.
\end{aligned}$$

From the proof of Theorem 4, we obtain that

$$\begin{aligned}
P_{\mu_0}\{\boldsymbol{\mathcal{D}}_{\infty,\boldsymbol{\mathcal{H}}_M}(\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}) \geq \epsilon\} &\leq \mathcal{N}(\boldsymbol{\mathcal{H}}_M, \frac{\epsilon}{48S^2R^2K^4})e^{-\frac{c_{L,N,\boldsymbol{\mathcal{H}}_M}M\epsilon}{1152S^2R^2K^6}} \\
&\leq \mathcal{N}(\cup_M\boldsymbol{\mathcal{H}}_M, \frac{\epsilon}{48S^2R^2K^4})e^{-\frac{c_{L,N,\cup_M\boldsymbol{\mathcal{H}}_M}M\epsilon}{1152S^2R^2K^6}} \\
&\leq C(\cup_M\boldsymbol{\mathcal{H}}_M, \epsilon)e^{-\frac{c_{L,N,\cup_M\boldsymbol{\mathcal{H}}_M}M\epsilon}{1152S^2R^2K^6}},
\end{aligned}$$

where C_1 is an absolute constant independent of M and $C(\cup_M\boldsymbol{\mathcal{H}}_M, \epsilon)$ is a finite positive constant due to the compactness of $\cup_M\boldsymbol{\mathcal{H}}_M$. Therefore,

$$\sum_{M=1}^\infty P_{\mu_0}\{\boldsymbol{\mathcal{D}}_{\infty,\boldsymbol{\mathcal{H}}_M}(\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}) \geq \epsilon\} \leq \sum_{M=1}^\infty C(\cup_M\boldsymbol{\mathcal{H}}_M, \epsilon)e^{-\frac{c_{L,N,}M\epsilon}{1152S^2R^2K^6}} < \infty.$$

On the other hand, the estimate (6.2) yields that

$$\boldsymbol{\mathcal{E}}_\infty(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}_M}) \leq 4K^4SR^2 \inf_{\boldsymbol{f}\in\boldsymbol{\mathcal{H}}_M} \|\boldsymbol{f}-\boldsymbol{\phi}\|_\infty \xrightarrow{M\to\infty} 0.$$

Therefore, $P_{\mu_0}\{\boldsymbol{\mathcal{E}}_\infty(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}_M}) \geq \epsilon\} = 0$ when M is large enough. So we have

$$\sum_{M=1}^{\infty} P_{\mu_0}\{\boldsymbol{\mathcal{E}}_\infty(\widehat{\boldsymbol{\phi}}_{\infty,\boldsymbol{\mathcal{H}}_M}) \geq \epsilon\} < \infty$$

By the Borel-Cantelli Lemma, we have

$$P_{\mu_0}\big\{\limsup_{M\to\infty}\{c_{L,N,\cup_M\boldsymbol{\mathcal{H}}_M}\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \geq \epsilon\}\big\} = 0,$$

which is equivalent to

$$P_{\mu_0}\big\{\liminf_{M\to\infty}\{c_{L,N,\cup_M\boldsymbol{\mathcal{H}}_M}\|\widehat{\boldsymbol{\phi}}_{M,\boldsymbol{\mathcal{H}}_M}(\cdot)\cdot - \boldsymbol{\phi}(\cdot)\cdot\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)} \leq \epsilon\}\big\} = 1.$$

The conclusion follows. ■

7. Appendix: Empirical comparison with SINDy and neural network

Our learning approach exploits the structure of the vector field $\mathbf{f}_\phi : \mathbb{R}^{dN} \to \mathbb{R}^{dN}$ in our dynamical system, first learns an estimator ϕ and therefore obtains an estimator of $\mathbf{f}_\phi$. In this section, we used two other approaches to learn $\mathbf{f}_\phi$ directly from trajectory data: the first approach is SINDy (Brunton et al. (2016)), which aims to represent or approximate each row of $\mathbf{f}_\phi$ as a linear combination of only a small number of elements in a (typically large) dictionary; the second one is regular neural networks. While there are some theoretical results for SINDy that, under suitable assumptions on the dynamics and sampling of observations, guarantee a sample complexity fundamentally dependent of the sparsity level of the r.h.s. $\mathbf{f}_\phi$ in terms of the dictionary (with additional log factors in the dictionary size), it is not clear which dictionary and which sparsity levels are achievable for interacting particle systems of the type considered here, both in general (without additional information) nor even in the simple examples we consider. In fact, it seems that the best sparsity achievable could be no better than $O(N^2)$, unless the special form of the $\mathbf{f}_\phi$ is used. This is of course even worse than the dimension dN of the state space, as soon as $N \geq d$. SINDy also provides no computational advantages, albeit for very large systems it may be possible to achieve better computational costs with the use of randomized algorithms. We are no aware of theoretical results neither on the statistical performance nor the computational cost of the simple neural networks we tried, and we draw no conclusions from our, admittedly very limited, experiments with neural networks. In general, given the optimal statistical performance of our approach, and its close-to-optimal computational cost, these experimental comparisons are mostly for completeness.

We consider a first order homogeneous system of the form

$$\dot{\boldsymbol{X}}(t) = \mathbf{f}_\phi(\boldsymbol{X}(t)), [\mathbf{f}_\phi(\boldsymbol{X}(t))]_i = \frac{1}{N}\sum_{i'=1}^{N}\phi(\|\boldsymbol{x}_i(t)-\boldsymbol{x}_{i'}(t)\|)(\boldsymbol{x}_{i'}(t)-\boldsymbol{x}_i(t)),$$

with: (i) the constant kernel $\phi(r) \equiv 1$, and (ii) the cosine kernel $\phi(r) = \cos(\frac{\pi r}{2}), 0 < r < 1$; other parameters for the experimental setup can be found in Table 8. These interaction kernels have been chosen so that the function $\mathbf{f}_\phi$ to be learned is very regular. We apply SINDy with a reasonably large dictionary consisting of monomials up to order 2, sines and cosines of frequencies $\{k\}_{k=1}^{10}$ (Code available in Brunton et al. (2016)); for neural networks we consider a three-layer FNN (Feed-Forward Neural network) with $[25, 25, 10]$ hidden units trained using Matlab©2019b Neural Network Function Fitting toolbox. Below, we summarize the results:

- for the constant kernel in (1), all three approaches perform well in learning $\mathbf{f}_\phi$ (see Table 9) and trajectory prediction (see Table 10). In this case, the vector field $\mathbf{f}_\phi$ is linear and lies in the span of dictionary used in SINDy. However, FNN uses much longer time in learning than other two approaches;

- for the cosine kernel in (2), $\mathbf{f}_\phi$ does not lie in the span of the dictionary used in SINDy. Although the estimators of SINDy and FNN have small errors in terms of function fitting and trajectory prediction on the training data set, both estimators perform poorly outside of the training data, see Table 11 and Table 12, as well as Figure 23 and Figure 24 for predicted trajectories.

We draw the following conclusion:

- SINDy estimators' performance, as is well-known, critically depends on the design and use a dictionary of functions such that the true vector field has a (near-) sparse representation in such dictionary. One could consider increasing the size of the dictionary used in SINDy, however, the number of basis functions increases exponentially in the ambient dimension. In 10 dimensions or more (like in our numerical examples) one cannot even construct such bases on a standard workstation. The only possible solution would be to carefully design dictionaries for these right-hand sides, and perhaps these dictionaries would even need to be specific to the underlying interaction kernel, therefore requiring significant additional information. This is specifically what we are trying to avoid in this work, but of course we do recognize that there are situations where such additional information is very much available and could (should!) be used (e.g. certain physical systems where physical laws are strongly suspected to be in a certain form). Alternatively, scalable randomized algorithms and adaptive approximation procedures may enable one to scale to larger dimensions.

- "Regular" neural networks fail on learning $\mathbf{f}_\phi$ with good generalization, as they also do not incorporate underlying physics of the system, e.g., translation invariance, and invariance under permutations of agents. As a consequence, they do not generalize to states in regions of state space far from regions in the training data, even if they are translationally or permutationally near to states in the training set. However, it is certainly possible that a neural network with a suitable architecture (perhaps also incorporating time, or a graph structure as in graph neural networks), combined with a suitably initialized algorithm with a suitable set of parameters will converge to a satisfactory estimator for a given system, or maybe even for large families of systems, such as those considered in this paper.

State dim	d	N	μ_0	Training time	Future time	$\deg(\psi)$	n
10	1	10	$\mathcal{U}([0,8])$	[0,1]	[1,20]	0	$60(\frac{M}{\log M})^{\frac{1}{3}}$

Table 8: Parameters for the system and our proposed algorithm.

	Our algorithm		SINDy		FNN[25,25,10]	
$M = 100, L = 100$	[0,1]	[1,20]	[0,1]	[1,20]	[0,1]	[1,20]
Training ICs	$7.2 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	$1.5 \cdot 10^{-9}$	$5.0 \cdot 10^{-3}$	$2.7 \cdot 10^{-4}$	$5.2 \cdot 10^{-3}$
Test ICs	$7.5 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	$5.0 \cdot 10^{-3}$	$5.0 \cdot 10^{-3}$	$7.8 \cdot 10^{-3}$	$5.3 \cdot 10^{-3}$
Running time for learning	$4.6 \cdot 10$ seconds [2]		$1.9 \cdot 10^{-1}$ seconds		$1.9 \cdot 10^{3}$ seconds	

Table 9: Constant kernel. Relative empirical mean squared error of $\frac{\|\mathbf{f}_\phi - \hat{\mathbf{f}}_\phi\|_{L^2}}{\|\mathbf{f}_\phi\|_{L^2}}$ on trajectory data sets

	Our algorithm		SINDy		FNN[25,25,10]	
$M = 100, L = 100$	[0,1]	[1,20]	[0,1]	[1,20]	[0,1]	[1,20]
$\text{Mean}_{\text{Training ICs}}$	$4.0 \cdot 10^{-3}$	$4.0 \cdot 10^{-3}$	$4.0 \cdot 10^{-3}$	$4.0 \cdot 10^{-3}$	$4.0 \cdot 10^{-3}$	$4.0 \cdot 10^{-3}$
$\text{Mean}_{\text{Test ICs}}$	$4.0 \cdot 10^{-3}$	$4.0 \cdot 10^{-3}$	$4.0 \cdot 10^{-3}$	$4.0 \cdot 10^{-3}$	$4.5 \cdot 10^{-3}$	$4.5 \cdot 10^{-3}$
Running time	$8.9 \cdot 10$ seconds		4.4 seconds		$9.0 \cdot 10$ seconds	

Table 10: Constant kernel. Empirical mean of Max-in-time trajectory prediction error

	Our algorithm		SINDy		FNN([25 25 10])	
$M = 200, L = 100$	[0,1]	[1,20]	[0,1]	[1,20]	[0,1]	[1,20]
Training ICs	$6.0 \cdot 10^{-2}$	$4.6 \cdot 10^{-2}$	$8.1 \cdot 10^{-2}$	3.3	$5.6 \cdot 10^{-1}$	3.8
Test ICs	$6.3 \cdot 10^{-2}$	$4.8 \cdot 10^{-2}$	3.3	4.8	6.0	9.7
Running time for learning	$1.2 \cdot 10$ seconds [2]		$2.8 \cdot 10$ seconds		$1.8 \cdot 10^{3}$ seconds	

Table 11: Cosine kernel. Relative empirical mean squared error of $\frac{\|\mathbf{f}_\phi - \hat{\mathbf{f}}_\phi\|_{L^2}}{\|\mathbf{f}_\phi\|_{L^2}}$ on trajectory data sets

[1] The time is calculated by running the software code available in Lu et al. (2019a). It contains time for computing all parameters needed in the package and therefore more than the actual fitting time.

[2] The same package used as in footnote [1].

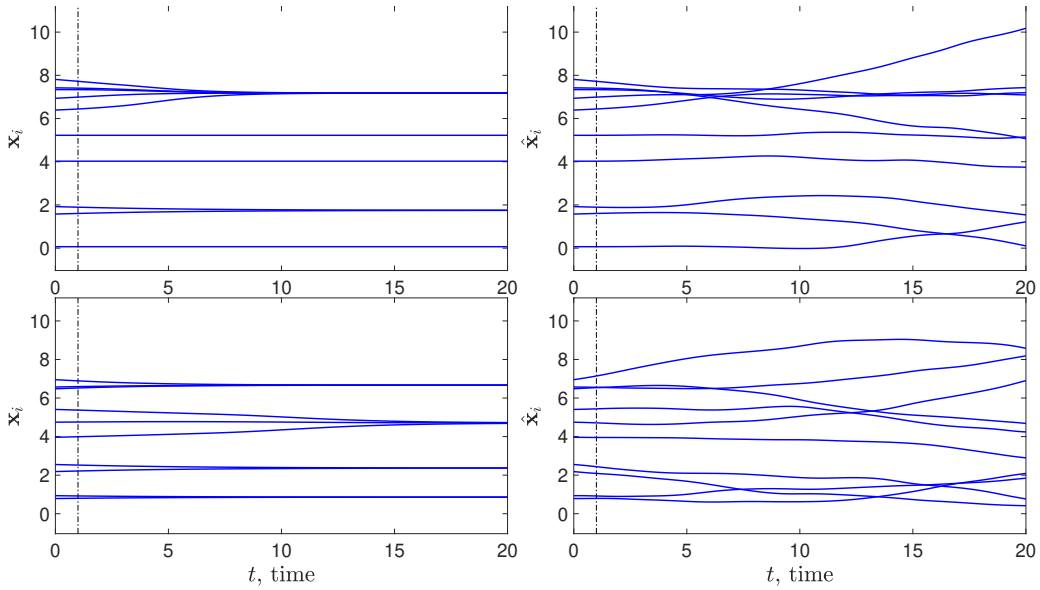

Figure 23: **Cosine kernel: trajectory prediction using SINDy.** $\boldsymbol{X}(t)$ (Left column) and $\hat{\boldsymbol{X}}(t)$ (Right column) obtained with the true kernel ϕ and the SINDy estimator $\hat{\mathbf{f}}_\phi$ from $M = 200$ trajectories, for an initial condition in the training data (Top row) and an initial condition randomly chosen (bottom row). The black dashed vertical line at $t = 1$ divides the "training" interval $[0, 1]$ from the "prediction" interval $[1, 20]$. The mean of max-in-time trajectory prediction errors over 200 experiments can be found in Table 12.

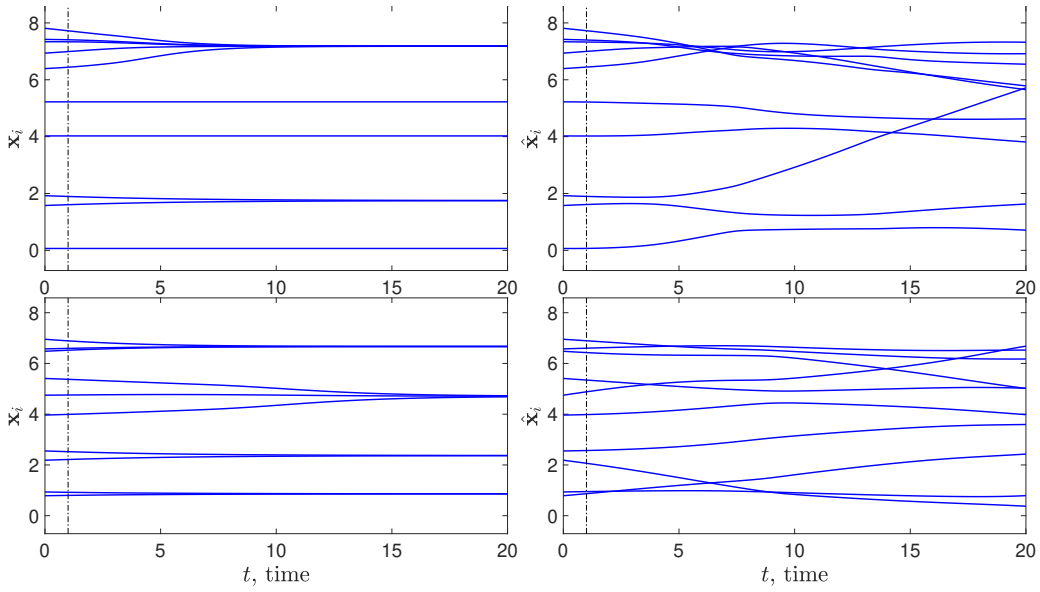

Figure 24: **Cosine kernel: trajectory prediction using FNN.** $\boldsymbol{X}(t)$ (Left column) and $\hat{\boldsymbol{X}}(t)$ (Right column) obtained with the true kernel ϕ and the FNN estimator $\hat{\mathbf{f}}_\phi$ from $M = 200$ trajectories, for an initial condition in the training data (Top row) and an initial condition randomly chosen (bottom row). The black dashed vertical line at $t = 1$ divides the "training" interval $[0, 1]$ from the "prediction" interval $[1, 20]$. The mean of max-in-time trajectory prediction errors over 200 experiments can be found in Table 12.

	Our algorithm		SINDy		FNN[25,25,10]	
$M = 200, L = 100$	[0,1]	[1,20]	[0,1]	[1,20]	[0,1]	[1,20]
$\text{Mean}_{\text{Training ICs}}$	$4.4 \cdot 10^{-4}$	$1.1 \cdot 10^{-2}$	$1.2 \cdot 10^{-3}$	$9.0 \cdot 10^{-1}$	$7.8 \cdot 10^{-3}$	$9.6 \cdot 10^{-1}$
$\text{Mean}_{\text{Test ICs}}$	$5.0 \cdot 10^{-4}$	$1.2 \cdot 10^{-2}$	$1.0 \cdot 10^{-1}$	1.7	$1.0 \cdot 10^{-1}$	1.5
Running time	$3.1 \cdot 10$ seconds		$1.3 \cdot 10$ seconds		$4.0 \cdot 10^2$ seconds	

Table 12: Cosine kernel. Empirical mean of Max-in-time trajectory prediction error

References

R. Abebe, J. Kleinberg, D. Parkes, and C. Tsourakakis. Opinion dynamics with varying susceptibility to persuasion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1089–1098, 2018.

N. Bellomo, P. Degond, and E. Tadmor, editors. *Active Particles, Volume 1.* Springer International Publishing AG, Switerland, 2017.

P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov. Universal algorithms for learning theory part i: piecewise constant functions. *Journal of Machine Learning Research*, 6(Sep):1297–1321, 2005.

V. Blodel, J. Hendricks, and J. Tsitsiklis. On Krause's multi-agent consensus model with state-dependent connectivity. *Automatic Control, IEEE Transactions on*, 54(11):2586 – 2597, 2009.

J. Bongard and H. Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 104 (24):9943–9948, 2007.

M. Bongini, M. Fornasier, M. Hansen, and M. Maggioni. Inferring interaction rules from observations of evolutive systems I: The variational approach. *Mathematical Models and Methods in Applied Sciences*, 27(05):909–951, 2017.

L. Boninsegna, F. Nüske, and C. Clementi. Sparse learning of stochastic dynamical equations. *The Journal of Chemical Physics*, 148(24):241723, 2018.

P. Borwein and T. Erdélyi. Generalizations of müntz's theorem via a remez-type inequality for müntz spaces. *Journal of the American Mathematical Society*, 10(2):327–349, 1997.

D. Brillinger, H. Preisler, and M. Wisdom. Modelling particles moving in a potential field with pairwise interactions and an application. *Brazilian Journal of Probability and Statistics*, 25(3):421–436, 2011.

D. Brillinger, H. Preisler, A. Ager, and J. Kie. The use of potential functions in modelling animal movement. In *Selected Works of David Brillinger*, pages 385–409. Springer, 2012.

C. Brugna and G. Toscani. Kinetic models of opinion formation in the presence of personal conviction. *Physical Review E*, 92(5):052818, 2015.

N. Brunel. Parameter estimation of ODEs via nonparametric estimators. *Electronic Journal of Statistics*, 2:1242–1267, 2008.

S. Brunton, J. Proctor, and J. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 113(15):3932–3937, 2016.

S. Brunton, N. Kutz, and J. Proctor. Data-drive discovery of governing physical laws. *SIAM News*, 50(1), 2017.

J. Cao, L. Wang, and J. Xu. Robust estimation for ordinary differential equation models. *Biometrics*, 67(4):1305–1313, 2011.

R. Chartrand. Numerical differentiation of noisy, nonsmooth data. *ISRN Applied Mathematics*, 2011, 2011.

X. Chen. *Multi-agent systems with reciprocal interaction laws*. PhD thesis, Harvard University, 2014.

Y. Chuang, M. D'Orsogna, D. Marthaler, A. Bertozzi, and L. Chayes. State transition and the continuum limit for the 2D interacting, self-propelled particle system. *Physica D*, 232:33 – 47, 2007.

G. Cisneros, K. Wikfeldt, L. Ojamäe, J. Lu, Y. Xu, H. Torabifard, A. Bartok, G. Csanyi, V. Molinero, and F. Paesani. Modeling molecular interactions in water: From pairwise to many-body potential energy functions. *Chemical reviews*, 116(13):7501–7528, 2016.

W. Cleveland and S. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

H. Cohn and A. Kumar. Algorithmic design of self-assembling structures. *Proceedings of the National Academy of Sciences of the United States of America*, 106:9570 – 9575, 2009.

I. Couzin, J. Krause, N. Franks, and S. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433(7025):513 – 516, 2005.

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

I. Dattner and C. Klaassen. Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electronic Journal of Statistics*, 9 (2):1939–1973, 2015.

A. De, S. Bhattacharya, P. Bhattacharya, N. Ganguly, and S. Chakrabarti. Learning a linear influence model from transient opinion dynamics. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 401–410, 2014.

R. DeVore and B. Lucier. Wavelets. *Acta Numerica*, 1:1–56, 1992.

R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov. Mathematical methods for supervised learning. *IMI Preprints*, 22:1–51, 2004.

M. D'Orsogna, Y. Chuang, A. Bertozzi, and L. Chayes. Self-propelled particles with soft-core interactions: patterns, stability, and collapse. *Physical Review Letter*, 96:104 – 302, 2006.

R. Escobedo, C. Muro, L. Spector, and R. P. Coppinger. Group size, individual role differentiation and effectiveness of cooperation in a homogeneous group of hunters. *Journal of the Royal Society Interface*, 11:20140204, 2014.

J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications*. 1996.

J. Fryxell, A. Mosser, A. Sinclair, and C. Packer. Group formation stabilizes predator-prey dynamics. *Nature*, 449:1041 – 1043, 2007.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer, New York, 2002.

X. Han, Z. Shen, W. Wang, and Z. Di. Robust reconstruction of complex networks from sparse data. *Physical Review Letters*, 114(2):028701, 2015.

C. Hemelrijk and H. Hildenbrandt. Some causes of the variable shape of flocks of birds. *PloS One*, 6(8):e22479, 2011.

S. Kang, W. Liao, and Y. Liu. Ident: Identifying differential equations with numerical time evolution. *arXiv preprint arXiv:1904.03538*, 2019.

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

J. Ke, J. Minett, C. Au, and W. Wang. Self-organization and selection in the emergence of vocabulary. *Complexity*, 7:41 – 54, 2002.

I. Knowles and R. Renka. Methods for numerical differentiation of noisy data. *Electronic Journal of Differential Equations*, 21:235–246, 2014.

T. Kolokolnikov, J. Carrillo, A. Bertozzi, R. Fetecau, and M. Lewis. Emergent behaviour in multi-particle systems with non-local interactions. *Physica D*, 260:1–4, 2013.

U. Krause. A discrete nonlinear and non-autonomous model of consensus formation. *Communications in difference equations*, pages 227 – 236, 2000.

Z. Li, F. Lu, M. Maggioni, S. Tang, and C. Zhang. On the identifiability of interaction functions in systems of interacting particles. *Stochastic Processes and their Applications*, 132:135 – 163, 2021.

H. Liang and H. Wu. Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103(484):1570–1583, 2008.

Z. Long, Y. Lu, X. Ma, and B. Dong. PDE-net: Learning PDEs from data. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3208–3216, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

F. Lu, M. Zhong, S. Tang, and M. Maggioni. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proceedings of the National Academy of Sciences of the Unite States of America*, 116(29):14424–14433, 2019a.

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

F. Lu, M. Maggioni, and S. Tang. Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories. *arXiv preprint arXiv:2007.15174*, 2020.

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

H. Miao, X. Xia, A. Perelson, and H. Wu. On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Review*, 53(1):3–39, 2011.

J. Miller, S. Tang, M. Zhong, and M. Maggioni. Learning theory for inferring interaction kernels in second-order interacting agent systems. *arXiv preprint arXiv:2010.03729*, 2020.

S. Mostch and E. Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56 (4):577 – 621, 2014.

M. Nowak. Five rules for the evolution of cooperation. *Science*, 314:1560 – 1563, 2006.

R. Olfati-Saber and R. Murray. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control*, 49(9):1520–1533, 2004.

J. Parrish and L. Edelstein-Keshet. Complexiy, pattern, and evolutionary trade-offs in animal aggregation. *Science*, 284:99 – 101, 1999.

M. Raissi. Deep hidden physics models: Deep learning of nonlinear partial differential equations. *The Journal of Machine Learning Research*, 19(1):932–955, 2018.

M. Raissi and G. Karniadakis. Hidden physics models: Machine learning of nonlinear partial differential equations. *Journal of Computational Physics*, 357:125–141, 2018.

M. Raissi, P. Perdikaris, and G. Karniadakis. Multistep neural networks for data-driven discovery of nonlinear dynamical systems. *arXiv preprint arXiv:1801.01236*, 2018.

J. Ramsay and G. Hooker. Dynamic data analysis: Modeling data with differential equations. *Springer Series in Statistics*, 2018.

J. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, 2007.

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

S. Rudy, S. Brunton, J. Proctor, and N. Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.

H. Schaeffer, R. Caflisch, C. Hauck, and S. Osher. Sparse dynamics for partial differential equations. *Proceedings of the National Academy of Sciences of the United States of America*, 110(17):6634–6639, 2013.

H. Schaeffer, G. Tran, and R. Ward. Extracting sparse high-dimensional dynamics from limited data. *SIAM Journal on Applied Mathematics*, 78(6):3279–3295, 2018.

M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

J. Toner and Y. Tu. Long-range order in a two-dimensional dynamical xy model: how birds fly together. *Physical Review Letters*, 75(23):4326, 1995.

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

J. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

J. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, 3(1):28–46, 1982.

T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel Type of Phase Transition in a System of Self-Driven Particles. *Physical Review Letters*, 75(6):1226 – 1229, 1995.

J. Wagner, P. Mazurek, and R. Morawski. Regularised differentiation of measurement data. In *Proc. XXI IMEKO World Congress" Measurement in Research and Industry*, pages 1–6, 2015.

S. Wang, E. Herzog, W. Kiss, I. Schwartz, G. Bloch, M. Sebek, D. Granados-Fuentes, L. Wang, and J. Li. Inferring dynamic topology for decoding spatiotemporal structures in complex heterogeneous networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115(37):9300–9305, 2018.

K. Wu and D. Xiu. Numerical aspects for approximating governing equations using data. *Journal of Computational Physics*, 384:200–221, 2019.

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

M. Zhong, J. Miller, and M. Maggioni. Data-driven modeling of celestial dynamics from Jet Propulsion Laboratory's development ephemeris. *In preparation*, 2020a.

M. Zhong, J. Miller, and M. Maggioni. Data-driven discovery of emergent behaviors in collective dynamics. *Physica D: Nonlinear Phenomena*, page 132542, 2020b.