Nonparametric inference of interaction laws in systems of agents from trajectory data

Fei Lu[a,b,c], Ming Zhong[b], Sui Tang[a], and Mauro Maggioni[a,b,c,d,1]

[a]Department of Mathematics, Johns Hopkins University, Baltimore, MD 21218; [b]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218; [c]Institute for Data Intensive Engineering and Science, Johns Hopkins University, Baltimore, MD 21218; and [d]Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, MD 21218

Inferring the laws of interaction in agent-based systems from observational data is a fundamental challenge in a wide variety of disciplines. We propose a nonparametric statistical learning approach for distance-based interactions, with no reference or assumption on their analytical form, given data consisting of sampled trajectories of interacting agents. We demonstrate the effectiveness of our estimators both by providing theoretical guarantees that avoid the curse of dimensionality and by testing them on a variety of prototypical systems used in various disciplines. These systems include homogeneous and heterogeneous agent systems, ranging from particle systems in fundamental physics to agent-based systems that model opinion dynamics under the social influence, prey–predator dynamics, flocking and swarming, and phototaxis in cell dynamics.

data-driven modeling | dynamical systems | agent-based systems

1. Introduction

Systems of interacting agents arise in a wide variety of disciplines, including Physics, Biology, Ecology, Neurobiology, Social Sciences, and Economics (e.g., refs. 1–4 and references therein). Agents may represent particles, atoms, cells, animals, neurons, people, rational agents, opinions, etc. The understanding of agent interactions at the appropriate scale in these systems is as fundamental a problem as the understanding of interaction laws of particles in Physics.

How can laws of interaction between agents be discovered? In Physics, vast knowledge and intuition exist to formulate hypotheses about the form of interactions, inspiring careful experiments and accurate measurements, that together lead to the inference of interaction laws. This is a classical area of research, dating back to at least Gauss, Lagrange, and Laplace (5), that plays a fundamental role in many disciplines. In the context of interacting agents at the scale of complex organisms, there are fewer controlled experiments possible and few "canonical" choices for modeling the interactions. Different types and models of interactions have been proposed in different scientific fields and fit to experimental data, which in turn may suggest new modeling approaches, in a model–data validation loop. Often, the form of governing interaction laws is chosen a priori, within perhaps a small parametric family, and the aim is often to reproduce only qualitatively, and not quantitatively, some of the macroscopic features of the observed dynamics, such as the formation of certain patterns.

Our work fits at the boundary between statistical/machine learning and dynamical systems, where equations are estimated from observed trajectory data, and inference takes into account assumptions about the form of the equations governing the dynamics. Since the past decade, the rapidly increasing acquisition of data, due to decreasing costs of sensors and measurements, has made the learning of large and complex systems possible, and there has been an increasing interest in inference techniques that are model-agnostic and scalable to high-dimensional systems and large datasets.

We establish statistically sound, dynamically accurate, computationally efficient techniques* for inferring these interaction laws from trajectory data. We propose a nonparametric approach for learning interaction laws in particle and agent systems, based on observations of trajectories of the states (e.g., position, opinion, etc.) of the systems, on the assumption that the interaction kernel depends on pairwise distances only, unlike recent efforts that either require feature libraries or parametric forms for such interactions (6–10), or aim at identifying only the type of interaction from a small set of possible types (11–13). We consider a least-squares (LS) estimator, classical in the area of inverse problems (dating back to Legendre and Gauss), suitably regularized and tuned to the learning of the interaction kernel in agent-based systems.

The unknown is the interaction kernel, a function of pairwise distances between agents of the systems. While the values of this function are not observed, in contrast to the standard regression problems, we are able to show that our estimator converges at an optimal rate as if we were in the 1D regression setting. In particular, the learning rate has no dependency on the dimension of the state space of the system, therefore avoiding any curse of dimensionality, and making these estimators well-suited for the modern high-dimensional data regime. It may be easily extended to a variety of complex systems; here, we consider first- and second-order models, with single and multiple types of agents, and with interactions with simple environments. We demonstrate with examples that the theoretical guarantees on the performance of the estimator make it suitable for testing hypotheses on underlying models of interactions,

assisting an investigator in choosing among different possible (nonparametric) models.

Finally, our estimator is constructed with algorithms that are computationally efficient (with complexity $O(LN^2M)$ when the interaction kernel is Lipschitz; *SI Appendix*, section 2F) and may be implemented in a streaming fashion: It is, therefore, well-suited for large datasets.

2. Learning Interaction Kernels

We start with a model that is used in a wide variety of interacting agent systems [e.g., physical particles or influence propagation in a population (14, 15)]: Consider $N > 1$ agents $\{x_i\}_{i=1}^N$ in $\mathbb{R}^d$, evolving according to the system of ordinary differential equations (ODEs)

$$\dot{x}_i(t) = \frac{1}{N} \sum_{i'=1}^N \phi(\|x_{i'}(t) - x_i(t)\|)(x_{i'}(t) - x_i(t)), \quad [1]$$

where $\dot{x}_i(t) = \frac{d}{dt}x_i(t)$; $\|\cdot\|$ is the Euclidean norm, and $\phi: \mathbb{R}_+ \to \mathbb{R}$ is the interaction kernel. In other words, every agent's velocity is obtained by superimposing the interactions with all of the other agents, each weighted in a way dependent on the distance to the interacting agent. In a prototypical example—e.g., arising in particle systems (Section 2B) and flocking systems—the interaction kernel may be negative for small distances, inducing repulsion, and attractive for large distances. Let $X := (x_i)_{i=1}^N \in \mathbb{R}^{dN}$ be the state vector for all of the agents, $r_{ii'}(t) := x_{i'}(t) - x_i(t)$ and $r_{ii'}(t) := \|r_{ii'}(t)\|$. The evolution Eq. **1** is the gradient flow for the potential energy $\mathcal{U}(X(t)) := \frac{1}{2N}\sum_{i \neq i'} \Phi(r_{ii'}(t))$, with $\phi(\cdot) = \Phi'(\cdot)/\cdot$. The function $\phi(\cdot)\cdot$ reappears naturally below, the fundamental reason being its relationship with $\mathcal{U}$ and Φ. Our observations are positions along trajectories: $X_{tr} := \{X^m(t_l)\}_{l=1,m=1}^{L,M}$, with $0 = t_1 < \ldots < t_L = T$ being the times at which observations occur, and m indexing M different trajectories. Velocities $\dot{X}^m(t_l)$ are approximated by finite differences. The M initial conditions (ICs) $X_0^m := X^m(0)$ are drawn independently at random from a probability measure μ_0 on $\mathbb{R}^{dN}$.

Our goal is to infer, in a nonparametric fashion, the interaction kernel ϕ, by constructing an estimator $\hat{\phi}$ from training data. A fundamental statistical problem that involves estimating a function is regression: Given samples $(z_i, g(z_i))_{i=1}^n$, with the z_i's independent and identically distributed (i.i.d.) samples from an (unknown) measure ρ_Z in $\mathbb{R}^D$, and g a suitably regular (say, Hölder s) unknown function $\mathbb{R}^D \to \mathbb{R}$, one constructs an estimator $\hat{g}$ such that $\|\hat{g} - g\|_{L^2(\rho_Z)} \lesssim n^{-\frac{s}{2s+D}}$, with high probability (over the z_i's). This rate is optimal in a minimax sense (16), and its dramatic degradation with D is a manifestation of the curse of dimensionality. Upon rewriting Eq. **1** as $\dot{X} = \mathbf{f}_\phi(X)$, our observations (with either approximated or directly observed velocities) resemble those needed for regression if we thought of $Z = X$ as a random variable, and $g = \mathbf{f}_\phi$. However, our observations are not i.i.d. samples of X with respect to any probability measure, the lack of independence being the most glaring aspect. If we nevertheless pursued this line of thought, we would be hit with the curse of dimensionality in trying to learn the target function $g = \mathbf{f}_\phi$ on the state space $\mathbb{R}^{dN}$, leading to a rate $n^{-O(1/dN)}$ for regression. This renders this approach useless in practice as soon as, say, $dN \geq 20$. A direct application of existing approaches (e.g., refs. 6–8), developed for low-dimensional systems, go in this direction, These works would try to ameliorate this curse of dimensionality by requiring $\mathbf{f}_\phi$ to be well-approximated by a linear combination of a small number of functions in a known large dictionary. While such dictionaries may be known for specific problems, they are usually not given in the case of complex, agent-based systems. Finally, such dictionaries typically grow

dramatically in size with the dimension (here, dN), and existing guarantees that avoid the curse of dimensionality require further, strong assumptions on the measurements or the dynamics.

We proceed in a different direction, aiming for the flexibility of a nonparametric model while exploiting the structure of the system in Eq. **1**. The target function ϕ depends on just one variable (pairwise distance), but it is observed through a collection of nonindependent linear measurements (the left-hand side of Eq. **1**), at locations $r_{ii'}^m(t_l) = \|x_{i'}^m(t_l) - x_i^m(t_l)\|$, with coefficients $r_{ii'}^m(t_l) = x_{i'}^m(t_l) - x_i^m(t_l)$, as in the right-hand side of Eq. **1**. When the t_l's are equidistant in time, we consider an estimator minimizing the empirical error functional

$$\mathcal{E}_{L,M}(\varphi) := \frac{1}{LMN} \sum_{l,m,i=1}^{L,M,N} \left\| \dot{x}_i^m(t_l) - \mathbf{f}_\varphi(x^m(t_l))_i \right\|^2, \quad [2]$$

$$\hat{\phi} = \hat{\phi}_{L,M,\mathcal{H}} := \arg\min_{\varphi \in \mathcal{H}} \mathcal{E}_{L,M}(\varphi), \quad [3]$$

where $\mathcal{H}$ is a hypothesis space of functions $\mathbb{R}_+ \to \mathbb{R}$, of dimension n (we will choose n dependent on M). We introduce a natural probability measure ρ_T on $\mathbb{R}_+$ adapted to the dynamics: It can be thought of as an "occupancy" measure, in the sense that for any interval I, $\rho_T(I)$ is the probability (over the random ICs distributed according to μ_0) of seeing a pair of agents with a distance between them being a value in I, averaged over the time interval $[0, T]$; see Eq. **4** for a formal definition.

We measure the performance of $\hat{\phi}$ in terms of the error $\|\hat{\phi}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T)}$. Theorem (Thm.) 3.3, our main result, will bound this error by $\tilde{O}(M^{-s/(2s+1)})$ if ϕ is Hölder s: This is the optimal exponent for learning ϕ if we were in the (more favorable) 1D regression setting! We therefore completely avoid the curse of dimensionality. In fact, we show under some rather general assumptions that not only the rate, but even the constants in the bound are independent of N, making the bounds essentially dimension-free. It is crucial that ρ_T has wide support in order for the error to be informative. When the system is ergodic, we expect ρ_T to have a large support for large T, as the system explores its ergodic distribution. However, many deterministic systems of interest may reach a stationary state (as in the cases of the Lennard–Jones or opinion dynamics, to be considered momentarily), in which case ρ_T becomes highly concentrated on a finite set for large T: In these cases, it may be more relevant to consider T small compared with the relaxation time.

We are also interested in whether trajectories $X(t)$ of the true system are well-approximated by trajectories $\hat{X}(t)$ of the system governed by the interaction kernel $\hat{\phi}$, on both the "training" time interval $[0, T]$ and after time T. Proposition (Prop.) 3.4 below bounds $\sup_{t \in [0, T']} \|\hat{X}(t) - X(t)\|$ in terms of $\|\hat{\phi}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T)}$, at least for T' not too large; this further validates the use of $L^2(\rho_T)$. We will report on this distance for both $T' = T$ and $T' > T$ ("prediction" regime).

Finally, while the error $\|\hat{\phi}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T)}$ is unknown in practice (since ϕ is unknown), our results give guarantees on its size, which in turn imply guarantees on accuracy of trajectory predictions. Proxies for the error on trajectories, for example, by holding out portions of trajectories during the training phase, may be derived from data. These measures of error may be used to test and validate different models of the dynamics: Too large an error with one model may invalidate it and suggest that a different one (e.g., second vs. first order or multiple vs. single agent types) should be used (Section 5).

A. Different Sampling Regimes and Randomness. The total number of observations is (number of ICs)× (number of temporal

observations in $[0, T]) = M \times L$, each in $\mathbb{R}^{dN}$. We will consider several regimes:

Many short time trajectories. T is small, L is small (e.g., $L = 1$), and M is large (many ICs sampled from μ_0);

Single large time trajectory. T is large (even comparable to the relaxation time of the system if applicable), L is large, and $M = 1$ (or very small);

Intermediate time scale. T, L and M are all not small, but none is very large, corresponding to multiple "medium"-length trajectories, with several different ICs.

Randomness is injected via the ICs, and in our main results in Section 3, the sample size will be M. If the system is ergodic, the regimes above are partially related to each other, at least when the ICs are sampled from the ergodic distribution μ_{erg}. Indeed, at times much larger than the mixing time T_{mix}, the state of the system becomes indistinguishable from a random sample of μ_{erg}, and we may interpret the subsequent part of the trajectory as a new trajectory with that IC. The M observed trajectories of length $T \gg T_{\mathrm{mix}}$ are then equivalent to $M \times T/T_{\mathrm{mix}}$ trajectories of length T_{mix}, to which our results apply. In regimes when M is very small or μ_0 is very concentrated, there is little randomness: The problem is close to a fixed-design inverse problem, which is solvable if the dynamics produces different-enough pairwise distances.

B. Example: Interacting Particles with the Lennard–Jones Potential. We illustrate the learning procedure on a particle system with $N = 7$ particles in $\mathbb{R}^2$, interacting according to Eq. **1** with $\phi(r) = \Phi'_{LJ}(r)/r$, where $\Phi_{LJ}(r) := 4\epsilon \left((\sigma/r)^{12} - (\sigma/r)^6 \right)$ is the Lennard–Jones potential, consisting of a strong near-field repulsion and a long-range attraction. The system converges quickly to equilibrium configurations, which often consist of ordered, crystal-like structures. This example is challenging for various reasons: the Interaction kernel is unbounded, has unbounded support, and equilibrium is reached quickly, reducing the amount of information in trajectories. *SI Appendix*, section 3B contains a detailed description of the experiments. Fig. 1 demonstrates that the estimators approximate the true kernel well in different sampling regimes and that the trajectories of the true system are well-approximated by those of the learned system both in the "training" interval ($[t_0, T]$) and in the "prediction" interval ($[T, 50 T]$ and $[T, 2 T]$ respectively for the two regimes). We also show, as a simple example of transfer learning, that we can use the interaction kernel learned on the system with N particles to accurately predict trajectories of a system with $4N$ particles.

The rate of decay of the estimation error is close to the optimal rate in Thm. 3.3 (Fig. 2); this is a consequence of two factors: the use of an empirical approximation to ρ_T^L and
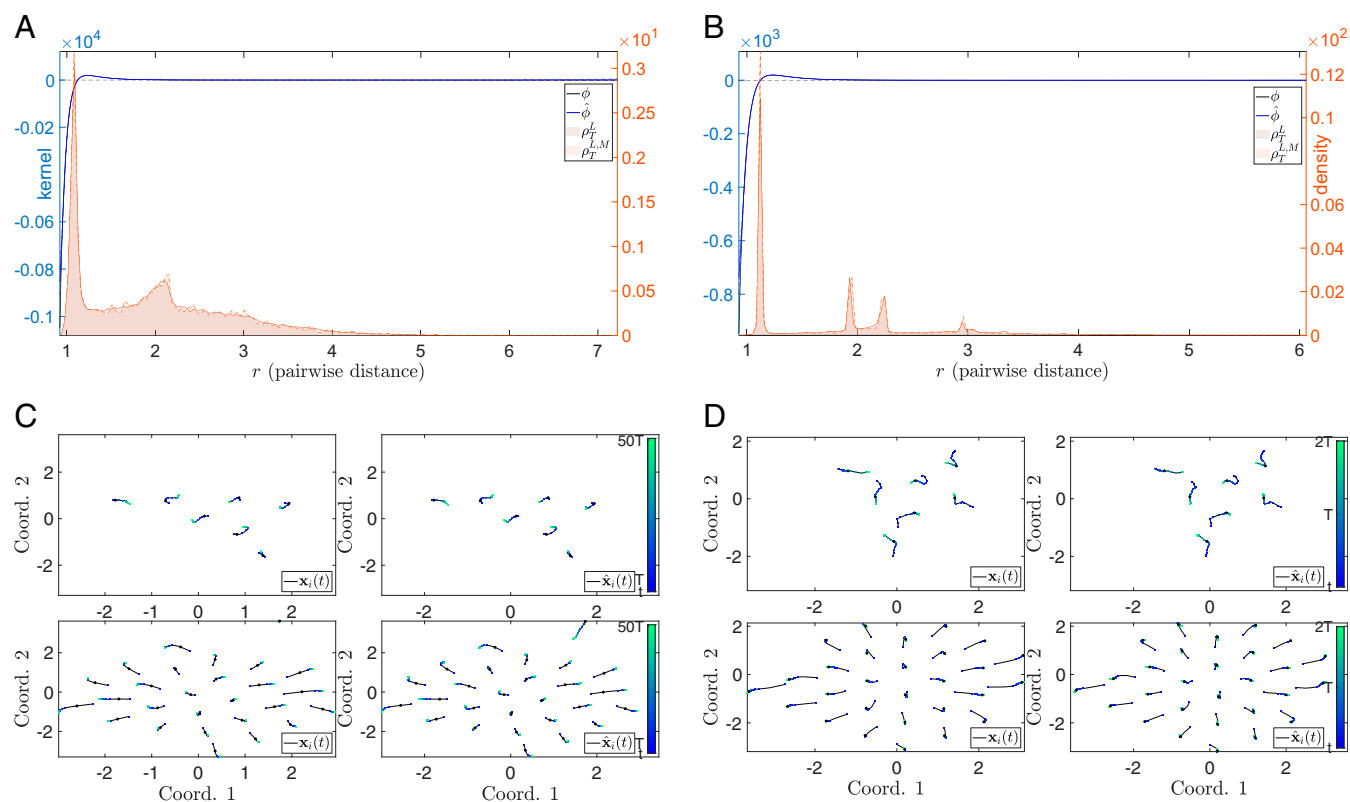
Fig. 1. Interaction kernel estimation and trajectory prediction for the Lennard–Jones system. (*A* and *B*) Estimators $\hat{\phi}$ (in blue) of the true interaction kernel ϕ (in black) in two sampling regimes: many short-time trajectories (*A*) and a few large-time trajectories (*B*). The proposed nonparametric estimators perform extremely well—the means and SDs of the relative $L^2(\rho_T^L)$ errors are $6.6 \cdot 10^{-2} \pm 5.0 \cdot 10^{-3}$ and $7.2 \cdot 10^{-2} \pm 1.0 \cdot 10^{-2}$, respectively, over 10 independent learning runs. The SD (dashed) lines on the estimated kernel are so small to be barely visible. In both cases, we superimpose histograms of ρ_T^L (estimated from a large number of trajectories, outside of training data) and $\rho_T^{L,M}$ (estimated from the M training data trajectories; *SI Appendix*, Eq. 5). The estimators belong to a hypothesis space $\mathcal{H}_n$ of piecewise linear functions with equidistant knots and yield accurate estimators in $L^2(\rho_T^L)$. Note that we observe the dynamics starting from a suitable $t_0 > 0$, due to the singularity of Lennard–Jones kernel at $r = 0$. See *SI Appendix*, section 3B for details about the setup and results. (*C* and *D*) The true and predicted trajectories for the N-particle system (*Upper*) and a $4N$-particle system (*Lower*) with interaction kernels learned on the N-particle system, for randomly sampled ICs. *C* and *D* show true and predicted trajectories for systems with interaction kernels learned in *A* and *B*, respectively. The blue-to-green color gradient indicates the movement of particles in time (see color scales on the side). We achieve small errors in predicting the trajectories in all cases, even when we transfer the interaction kernel learned on an N-particle system to predict trajectories of a system with $4N$ particles. Coord., coordinates.
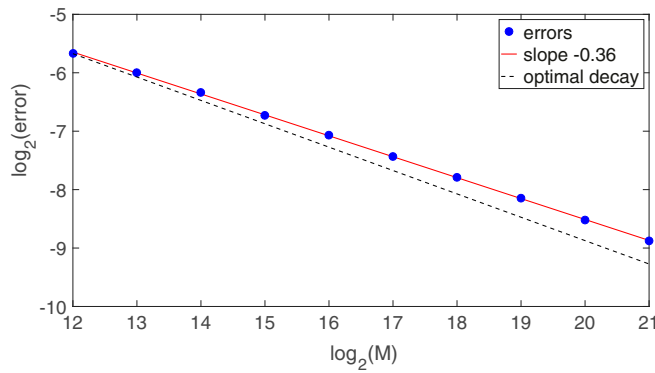
Fig. 2. Learning rate in M for the Lennard–Jones system. The estimation error in $L^2(\rho_T^L)$ decays at rate 0.36, close to the optimal rate 0.4 for admissible kernels; Thm. 3.3.

the blowup at 0 of Φ_{LJ}, which is not an admissible kernel as in Thm. 3.3 (see *SI Appendix*, Section 3B for a detailed discussion).

Fig. 3 shows the behavior of the error of the estimators as both L and M are increased. It indicates that a single long trajectory may not contain enough "information" to learn the kernel, at least for deterministic systems approaching a steady state. It also shows the behavior predicted by Thm. 3.3—namely, for each fixed L the error decreases as M increases.

3. Learning Theory

We introduce an error functional based on the structure of the dynamical system $\dot{X} = \mathbf{f}_\phi(X)$, whose minimizer will be our estimator of the interaction kernel ϕ. We consider kernels in the admissible set $\mathcal{K}_{R,S} := \{\phi \in C^1(\mathbb{R}_+): \text{supp}(\phi) \subset [0, R], \sup_{r \in [0,R]} |\phi(r)| + |\phi'(r)| \leq S\}$, for some $R, S > 0$. The boundedness of ϕ and ϕ' ensures the global well-posedness of the system in Eq. **1**. The restriction $\text{supp}(\phi) \subset [0, R]$ models the finite range of interaction between agents, and it may be relaxed to $\phi \in W^{1,\infty}(\mathbb{R}_+)$ with a suitable decay.

A. Probability Measures Adapted to the Dynamics. To measure the quality of the estimator of the interaction kernel ϕ, we introduce two probability measures on $\mathbb{R}_+$, the space of pairwise distances $r_{ii'}^m(t_l) = \|\mathbf{x}_{i'}^m(t_l) - \mathbf{x}_i^m(t_l)\|$. We consider the expectation of the empirical measure of pairwise distances, for continuous and discrete time observations, respectively:

$$\rho_T(r) := \frac{1}{\binom{N}{2}T} \int_{t=0}^{T} \mathbb{E}_{\mathbf{X}_0 \sim \mu_0} \left[\sum_{i,i'=1, i<i'}^{N} \delta_{r_{ii'}(t)}(r) \, dt \right], \quad \textbf{[4]}$$

$$\rho_T^L(r) := \frac{1}{\binom{N}{2}L} \sum_{l=1}^{L} \mathbb{E}_{\mathbf{X}_0 \sim \mu_0} \left[\sum_{i,i'=1, i<i'}^{N} \delta_{r_{ii'}(t_l)}(r) \right]. \quad \textbf{[5]}$$

The expectations are over the ICs, with distribution μ_0. The measure ρ_T is intrinsic to the dynamical system, dependent on μ_0 and the time scale T, and independent of the observation data. ρ_T^L depends also on the sampling scheme $\{t_l\}_{l=1}^{L}$ in time. Both are Borel probability measures on $\mathbb{R}_+$ (*SI Appendix*, Lemma 1.1), measuring how much regions of $\mathbb{R}_+$ on average (over the observed times and ICs) are explored by the system. Highly explored regions are where the learning process ought to be more accurate, as they are populated by more "samples" of pairwise distances. We will measure the estimation error of our estimators in $L^2(\rho_T)$ or $L^2(\rho_T^L)$.

We report here on the analysis in the discrete-time observation case, most relevant in practice, with ρ_T^L; the arguments, however, also apply to continuous-time observations, with ρ_T.

B. Learnability: The Coercivity Condition. A fundamental question is the learnability of the kernel, i.e., the convergence of the estimator $\hat{\phi}_{L,M,\mathcal{H}}$ defined in Eq. **3** to the true kernel ϕ as the sample size increases (i.e., $M \to \infty$) and $\mathcal{H}$ increases in a suitable way. The following condition, similar to the one introduced in ref. 17 for studying the mean field limit ($N \to \infty$), ensures learnability and well-posedness of the estimation.

Definition 3.1. The dynamical system in Eq. **1**, with IC sampled from μ_0 on $\mathbb{R}^{dN}$, satisfies the **coercivity condition** on a set $\mathcal{H}$ if there exists a constant $c_{L,N,\mathcal{H}} > 0$ such that for all $\varphi \in \mathcal{H}$ with $\varphi(\cdot) \cdot \in L^2(\rho_T^L)$,

$$c_{L,N,\mathcal{H}} \|\varphi(\cdot) \cdot\|_{L^2(\rho_T^L)}^2 \leq \frac{1}{NL} \sum_{l,i=1}^{L,N} \mathbb{E} \left\| \frac{1}{N} \sum_{i'=1}^{N} \varphi(r_{ii'}(t_l)) \mathbf{r}_{ii'}(t_l) \right\|^2.$$

$$\textbf{[6]}$$

The coercivity condition ensures learnability, by implying the uniqueness of minimizer of $\mathcal{E}_{L,\infty}(\varphi) := \mathbb{E}[\mathcal{E}_{L,M}(\varphi)]$ and, eventually, the convergence of estimators through a control of the error of the estimator in $L^2(\rho_T^L)$ (*SI Appendix*, Thm. 1.2 and Prop. 1.3). Thm. 3.1 proves that the coercivity condition holds under suitable hypotheses, even independently of N; numerical tests suggest that it holds generically over larger classes of interaction kernels and distributions of ICs, for large L, and as long as ρ_T^L is not degenerate (*SI Appendix*, Fig. S6). Finally, $c_{L,N,\mathcal{H}}$ also controls the condition number of the matrix in the LS problem yielding the estimator (see *SI Appendix*, Prop. 2.1 for details).

We prove that coercivity holds when μ_0 is exchangeable (i.e., the distribution is invariant under permutation of components), Gaussian, and $L = 1$. Numerical tests (*SI Appendix*, Fig. S6) suggest that the coercivity condition holds true for a larger class of interaction kernels, for various initial distributions including Gaussian and uniform distributions, and for large L, as long as ρ_T^L is not degenerate. We conjecture that the coercivity condition holds true in much greater generality (but not always!), leaving a detailed investigation to future work.

Theorem 3.1. *Suppose $L = 1, N > 1$ and assume that the distribution of $\mathbf{X}(t_1) = (\mathbf{x}_1(t_1), \dots, \mathbf{x}_N(t_1))$ is exchangeable Gaussian with $\text{cov}(X_i) - \text{cov}(X_i, \mathbf{x}_{i'}) = \lambda I_d$ for a constant $\lambda > 0$. Then, the*
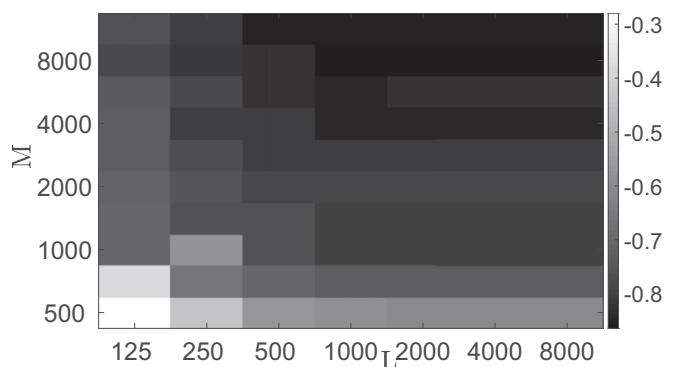
Fig. 3. The relative error of the estimated kernel as a function of M, L for the Lennard–Jones system. The relative error, in $\log_{10}$ scale, of $\hat{\phi}$ decreases both in L and M, in fact, roughly in the product ML, at least when M and L are not too small. $M = 1$ does not seem to suffice, no matter how large L is, due to the limited amount of "information" contained in a single trajectory.

coercivity condition holds true with $c_{L,N,\mathcal{H}} = \frac{(N-1)(N-2)}{N^2} c_{\mathcal{H}} + \frac{N-1}{N^2}$, where $c_{\mathcal{H}}$ is independent of N, is positive for any compact $\mathcal{H} \subset L^2(\rho_T^L)$, and is zero for $\mathcal{H} = L^2(\rho_T^L)$.

In this setting, the analysis of the coercivity constant $c_{L,N,\mathcal{H}}$ is based on the exchangeability of the initial distribution of the agents and relates coercivity to a positive integral kernel:

Lemma 3.2. *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 $\lambda > 0$. Suppose $L = 1$. Then, there is a positive definite integral kernel $\mathcal{K}(r,s) : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ such that for any $g \in L^2(\rho_T^L)$*

$$\mathbb{E}\left[g(|X-Y|)g(|X-Z|)\langle X-Y, X-Z \rangle \right]$$
$$= \iint g(r)r g(s)s \mathcal{K}(r,s)\, dr ds,$$

where $\rho_T^L(r) \propto r^{d-1} e^{-r^2/3}$, since $L = 1$. Therefore, there exists $c_{\mathcal{H}} \geq 0$, depending only on $\mathcal{H} \subset L^2(\rho_T^L)$, such that for $g \in \mathcal{H}$

$$\iint g(r)r g(s)s \mathcal{K}(r,s)\, dr ds \geq c_{\mathcal{H}} \|g(\cdot)\cdot\|^2_{L^2(\rho_T^L)},$$

and $c_{\mathcal{H}} > 0$ if $\mathcal{H}$ is compact in $L^2(\rho_T^L)$.

We conclude that under the assumptions of Thm. 3.1, if $\mathcal{H}$ is compact, then $c_{L,N,\mathcal{H}}$ is bounded below uniformly in N.

C. Optimal Rates of Convergence. The classical bias–variance trade-off in statistical estimation guides the selection of a hypothesis space $\mathcal{H}$, whose dimension will depend on M, the number of observed trajectories. On the one hand, $\mathcal{H}$ should be large so that the bias (distance between the true kernel ϕ and $\mathcal{H}$) is small; on the other hand, $\mathcal{H}$ should be small so that variance of the estimator is small. In the extreme case where $\mathcal{H} = \mathcal{K}_{R,S}$, the bias is 0, the variance of the estimator dominates, and we obtain the bound $\mathbb{E}[\|\widehat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T^L)}] \leq CM^{-1/4}$ (*SI Appendix*, Prop. 1.5). In fact, significantly better rates may be achieved for regular ϕ's:

Theorem 3.3. *Assume that $\phi \in \mathcal{K}_{R,S}$. Let $\{\mathcal{H}_n\}_n$ be a sequence of subspaces of $L^\infty([0,R])$, with $\dim(\mathcal{H}_n) \leq c_0 n$ and $\inf_{\varphi \in \mathcal{H}_n} \|\varphi - \phi\|_{L^\infty([0,R])} \leq c_1 n^{-s}$, for some constants $c_0, c_1, s > 0$. Assume that the coercivity condition holds on $\mathcal{H} := \overline{\cup_{n=1}^\infty \mathcal{H}_n}$. Such a sequence exists, for example, if ϕ is s-Hölder regular, and can be chosen so that $\mathcal{H}$ is compact in $L^2(\rho_T^L)$. Choose $n_* = (M/\log M)^{1/(2s+1)}$. Then, there exists a constant $C = C(c_0, c_1, R, S)$ such that*

$$\mathbb{E}\left[\|\widehat{\phi}_{L,M,\mathcal{H}_{n_*}}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T^L)} \right] \leq \frac{C}{c_{L,N,\mathcal{H}}} \left(\frac{\log M}{M} \right)^{\frac{s}{2s+1}}. \quad \text{[7]}$$

The rate [i.e., the exponent $s/(2s+1)$] we achieve is *optimal*: It coincides with the minimax rate in the classical regression setting, where one can observe directly noisy values of an s-Hölder regression function at the sample points. We obtain this optimal rate, even if we do not observe the values $\{\phi(r_{ii'}^m(t_l))\}_{l,i,i',m}$, but a "mixture" of them in the observed trajectory data. Many choices of $\{\mathcal{H}_n\}$ are consistent with the requirements in the theorem, e.g., splines on increasingly finer grids, or band-limited functions with increasing frequency limits. These choices affect the constants in Eq. **7**, the computational complexity of computing $\widehat{\phi}_{L,M,\mathcal{H}_{n_*}}$, but not the rate in M. While the rate is independent of the dimension dN of the state space, the constant may depend on d and N via $c_{L,N,\mathcal{H}}$. However, we expect that under rather general conditions, beyond those in Thm. 3.1, $c_{L,N,\mathcal{H}}$ is, in fact, lower-bounded independently of N for any compact subset $\mathcal{H}$ of $L^2(\rho_T^L)$ and is a fundamental property of the mean field limit ($N \to \infty$) of the system.

One shortcoming of our result is that the rate is not a function of the total number of observations, which is $O(LN^2M)$ (we have $LN^2/2$ pairwise distances for each of the M trajectories), but only of M, the number of *random* samples. Numerical experiments (see Fig. 3 and similar experiments for the other systems, reported in *SI Appendix*) suggest that the estimator improves as L increases, at least to a point, limited by the "information" in a single trajectory. Comparing to ref. 17, where the mean field limit $N \to \infty$, $M = 1$, is studied, we see the rates in ref. 17 seem no better than $N^{-1/d}$, i.e., they are cursed by dimension. So are sparsity-based inference techniques such as those in refs. 6–8, 11, and 18, which also require a good dictionary of template functions, are not non-parametric (at least in the form therein presented), and lack performance guarantees, except in some cases under stringent assumptions.

Our work here may be compared with the classical parameter estimation problem for the ODE models (19–22), where one is interested in estimating the vector parameter $\boldsymbol{\theta}$ in the ODE model $\dot{X} = f(X(t), t, \boldsymbol{\theta})$ from the observation of a single noisy trajectory. Our error functional, in spirit, is the same with the gradient-matching method (also called the two-stage method) used in the parameter-estimation problems (23–27). A challenging problem is the identifiability of $\boldsymbol{\theta}$. We refer the reader (28) for the statistical analysis and (29) (and references therein) for a comprehensive survey of this topic. However, the problem and approach we considered here are different from the
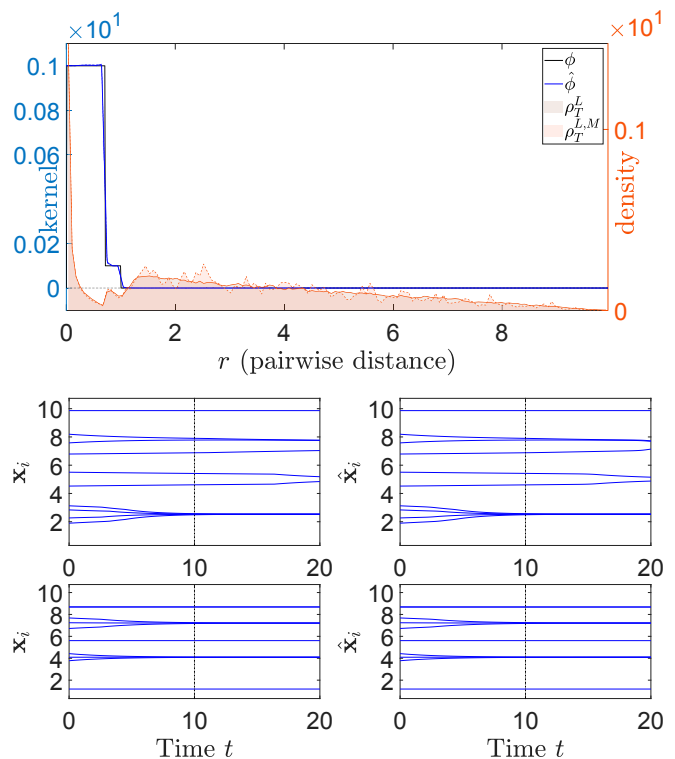
Fig. 4. Opinion dynamics. (*Upper*) Comparison between true and estimated interaction kernel, together with histograms for ρ_T^L and $\rho_T^{L,M}$. The mean and SD of the relative error for the interaction kernel are $1.6 \cdot 10^{-1} \pm 2.3 \cdot 10^{-3}$ over 10 independent learning runs. The SD lines (in dashed lines) on the estimated kernel are so small to be barely visible. (*Lower*) Trajectories $X(t)$ and $\widehat{X}(t)$ obtained with ϕ and $\hat{\phi}$, respectively, for an IC in the training data (top row) and an IC randomly chosen (bottom row). The black dashed vertical line at $t = T$ divides the "training" interval $[0, T]$ from the "prediction" interval $[T, T_f]$ (which in this case, $T_f = 2T$). We achieve small errors in all cases, in particular predicting number and location of clusters for large time.
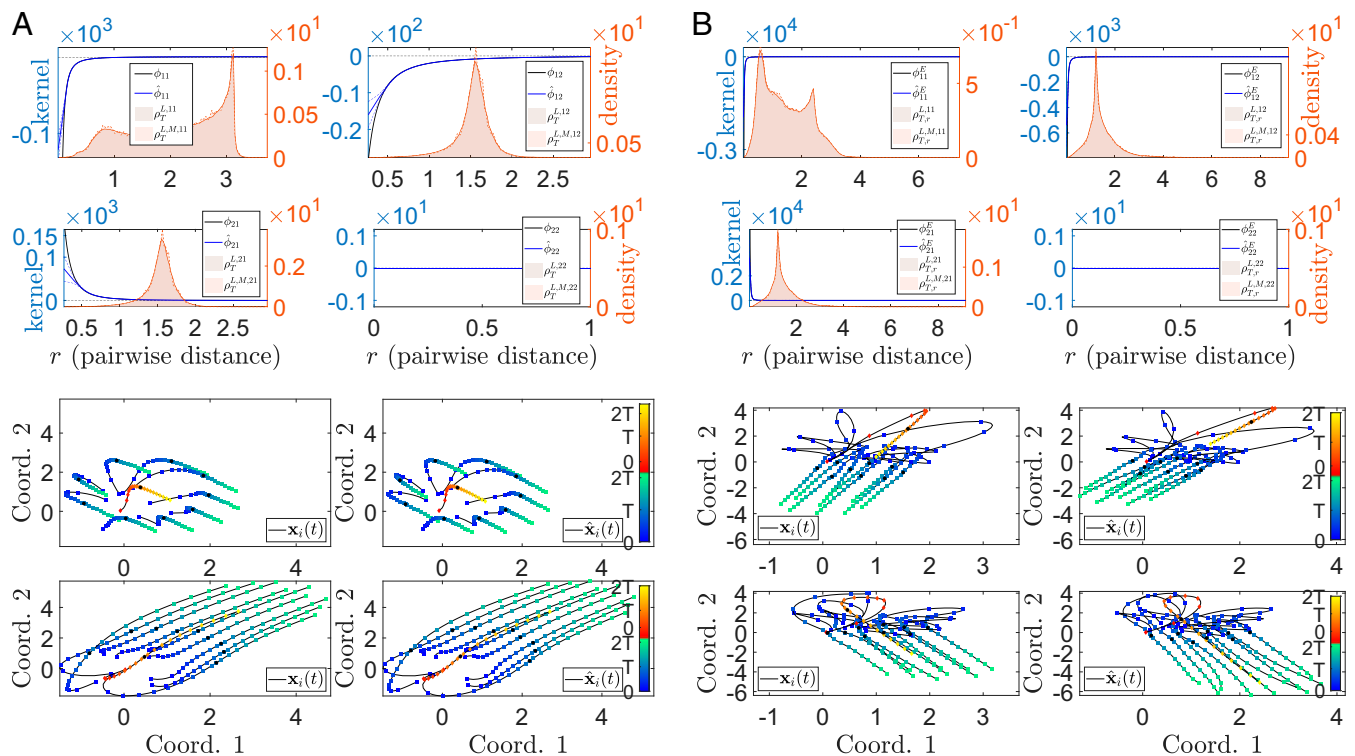
Fig. 5. Estimation of interaction kernels and trajectory prediction for predator–swarm first- and second-order systems. Results for the first-order (*A*) and second-order (*B*) predator–swarm systems, as described in Sections 4 and 5, are shown. For each system (corresponding to each column), *Upper* represents $\phi_{k,k'}$ and $\widehat{\phi}_{k,k'}$, superimposed with the histograms of ρ_T^L (estimated from a large number of trajectories, outside of training data) and ρ_T^{LM} (estimated from the *M* training data trajectories; *SI Appendix*, Eq. 5). *Lower* shows trajectories $X(t)$ and $\widehat{X}(t)$ of the corresponding (original and estimated) systems, evolved from the same ICs as the training data (third row) and newly sampled ICs (fourth row), over both the training time interval $[0, T]$ and in the future $[T, 2T]$ (color bars; the black dots in the trajectories correspond to $t = T$). For trajectories generated by the predator–swarm system, red-to-yellow lines indicate the movement of predators, whereas the blue-to-green lines indicate the movement of prey. The color gradients indicate time; see the color scales on the side of the plots. The estimators $\widehat{\phi}_{k,k'}$ perform extremely well: with negligible differences in the regions with large ρ_T^L and with possibly larger errors in regions with small ρ_T^L (where the SDs over 10 independent learning runs become visible). The $L^2(\rho_T^L)$ errors of the estimators are reported numerically in *SI Appendix*, section 3. Note that they are truncated to a constant while preserving continuity, when there are no samples (e.g., r near 0 or r very large). The measure ρ_T^L is quite smooth but can have interesting features; ρ_T^{LM} is typically a noisy version of ρ_T^L. The trajectories of the estimated system are typically good approximations to those of the original system, on both ICs in the training data and newly sampled ICs. The error of the estimated trajectories increases with time, as expected, albeit it still typically excellent also in the "prediction" time interval $[T, 2T]$, showing that the bounds in Prop. 3.4, while sharp in general, may be overly pessimistic in some practical cases. Some slightly larger errors are present in some trajectories, e.g., when prey and predators get much closer to each other than they did in the training data. Coord., coordinates.

parameter-estimation problem in several aspects. First of all, our state variable X enters into the domain of the ϕ (via its "projection" onto pairwise distance), while the parameter vector θ is decoupled from the state variable X. Moreover, our estimator is nonparametric—i.e., the goal is to estimate a function ϕ (a vector infinite dimensions) instead of a finite-dimensional vector θ of parameters. Finally, we establish identifiability conditions for ϕ from the perspective that the observations are i.i.d. trajectories with random ICs, in contrast with the identifiability of θ from observations along a fixed single trajectory with i.i.d. noise. We would like to mention the different, but related, problem of inferring potentials from ground states and unstable modes (for example, ref. 30), as well as recent results on existence and properties of ground states for systems with nonlocal interactions (31).

D. Trajectory-Based Performance Measures. It is important not only that $\widehat{\phi}$ is close to ϕ, but also that the dynamics of the system governed by $\widehat{\phi}$ approximate well the original dynamics. The error in prediction may be bounded trajectory-wise by a continuous-time version of the error functional and bounded in average by the $L^2(\rho_T)$ error of the estimated kernel (further evidence of the usefulness of ρ_T):

Proposition 3.4. *Assume* $\widehat{\phi}(\|\cdot\|)\cdot \in \mathrm{Lip}(\mathbb{R}^d)$, *with Lipschitz constant* C_{Lip}. *Let* $\widehat{X}(t)$ *and* $X(t)$ *be the solutions of systems with kernels* $\widehat{\phi}$ *and* ϕ, *respectively, started from the same IC. Then, for each trajectory*

$$\sup_{t \in [0, T]} \|\widehat{X}(t) - X(t)\|^2 \leq 2\, T e^{8\, T^2\, C_{\mathrm{Lip}}^2} \int_0^T \|\dot{X}(t) - \mathbf{f}_{\widehat{\phi}}(X(t))\|^2\, dt,$$

*and on average with respect to the distribution* μ_0 *of ICs:*

$$\mathbb{E}_{\mu_0}\left[\sup_{t \in [0, T]} \|\widehat{X}(t) - X(t)\|\right] \leq C\sqrt{N}\|\hat{\phi}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T)},$$

where the measure ρ_T *is defined in Eq. 4 and* $C = C(T, C_{\mathrm{Lip}})$.

4. Extensions: Heterogeneous Agent Systems, First and Second Order

The method proposed extends naturally to a large variety of interacting agent systems arising in a multitude of applications (4), including systems with multiple types of agents, driven by second-order equations, and including interactions with an environment. For detailed discussions of related topics on self-organized dynamics, we refer the readers to refs. 3 and 32–35 and the recent surveys (36, 37).
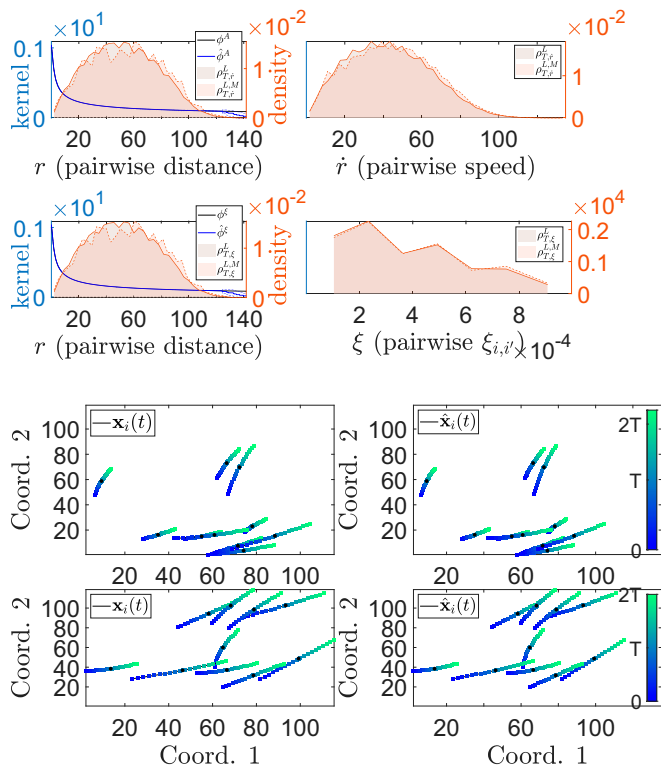
Fig. 6. Estimation of interaction kernels (*Upper*) and trajectory prediction (*Lower*) for the Phototaxis system. Results for the Phototaxis systems, as described in Sections 4 and 5, are shown. (*Upper*) *Left* represents ϕ^A vs. $\hat{\phi}^A$ (top row), and ϕ^ξ vs. $\hat{\phi}^\xi$ (bottom row), superimposed with the histograms of $\rho^L_{T,r}$ and, respectively, $\rho^{L,M}_{T,r}$. *Right* shows the comparison of the marginal distributions, $\rho^L_{T,r}$ vs. $\rho^{L,M}_{T,r}$ and $\rho^L_{T,\xi}$ vs. $\rho^{L,M}_{T,\xi}$. (*Lower*) *Left* represents the trajectories generated from true interaction kernels, whereas *Right* shows the trajectories generated by the estimated kernels, generated from training IC data (top row) and from a new random IC data (bottom row). In this system, the interaction kernels ϕ^A and ϕ^ξ are the same; the corresponding estimators $\hat{\phi}^A$ and $\hat{\phi}^\xi$ are both learned accurately, but note that they are being learned from two different sets of data, $(r,\dot{r})$ and (r,ξ), respectively. In both cases, data are scarce or missing for large values r, leading to estimators tapering to 0 faster than the true interaction kernels. However, despite the undesired tail-end behavior of our estimators, the estimators perform extremely well in regenerating the trajectories. See *SI Appendix*, section 3 for more details. Coord., coordinate.

A. First-Order Heterogeneous Agents Systems. Let the agents be divided into K disjoint sets $\{C_k\}_{k=1}^K$ ("types"), with different interaction kernels for each ordered pair of types:

$$\dot{x}_i(t) = \sum_{i'=1}^N \frac{1}{N_{k_{i'}}} \phi_{k_i k_{i'}}(r_{ii'}(t)) r_{ii'}(t), \quad \textbf{[8]}$$

where k_i is the index of the type of agent i—i.e., $i \in C_{k_i}$; $N_{k_{i'}}$ is the number of agents in type $C_{k_{i'}}$; $r_{ii'} = x_{i'} - x_i$ and $r_{ii'} = \|r_{ii'}\|$; $\phi_{kk'} : \mathbb{R}_+ \to \mathbb{R}$ is the interaction kernel governing how agents in type $C_{k'}$ influence agents in type C_k. As usual we let $X := (x_i)_{i=1}^N \in \mathbb{R}^{dN}$ be the vector describing the state of the system. We assume that the interaction kernels $\phi_{k_i k_{i'}}$'s are the only unknown factors in the model; in particular, we know the sets C_k's (i.e., the type of each agent is known). The goal is to infer the interaction kernels $\phi_{kk'}$ from observations $\{X^m(t_l)\}_{l,m=1}^{L,M}$ with $0 = t_1 < \ldots < t_l = T$ and with the ICs $X^m(0) = X_0^m$ randomly sampled from μ_0.

Let $\mathbf{f}_\phi(X^m) \in \mathbb{R}^{dN}$ be the vectorization of the right hand sides of Eq. **8**, and $\phi = (\phi_{kk'})_{k,k'=1}^K$. Dropping from the notation of

quantities that are assumed known, we rewrite the equations for the dynamics in Eq. **8** as $\dot{X}^m = \mathbf{f}_\phi(X^m)$. We use an error functional similar to Eq. **2**, with a weighted norm, to define the estimators:

$$\hat{\phi} := \underset{\varphi \in \mathcal{H}}{\text{argmin}} \frac{1}{ML} \sum_{m=1,l=1}^{M,L} \left\| \dot{X}^m(t_l) - \mathbf{f}_\varphi(x^m(t_l)) \right\|_{\mathcal{S}}^2, \quad \textbf{[9]}$$

where $\varphi = (\varphi_{kk'})_{k,k'=1}^K$, $\hat{\phi} = (\hat{\phi}_{kk'})_{k,k'=1}^K$ and $\|X\|_{\mathcal{S}}^2 := \sum_{i=1}^N \frac{1}{N_{k_i}} \|x_i\|^2$. The weighted norm $\|\cdot\|_{\mathcal{S}}^2$ is introduced so that, when different types of agents have significantly different cardinalities (e.g., a large number of preys vs. a single predator), the error functional will take into suitable consideration the least numerous type. Otherwise, only the interaction kernel of the most numerous type of agents would be accurately learned. Other more general weighting strategies may be considered, with minimal changes to the algorithm.

The generalization of ρ_T^L in Eq. **5** (similarly for ρ_T) to the heterogeneous-agent case is the family, indexed by ordered pairs $\{(k, k')\}_{k,k' \in \{1,\ldots,K\}}$, of probability measures on $\mathbb{R}_+$

$$\rho_T^{L,kk'}(r) = \frac{1}{LN_{kk'}} \sum_{l=1}^L \mathbb{E}_{X_0 \sim \mu_0} \sum_{i \in C_k, i' \in C_{k'}, i \neq i'} \delta_{r_{ii'}(t_l)}(r), \quad \textbf{[10]}$$

where $N_{kk'} = N_k N_{k'}$ when $k \neq k'$ and $N_{kk'} = \binom{N_k}{2}$ when $k = k'$ (for $N_k > 1$, otherwise there is no interaction kernel to learn). The error of an estimator, $\hat{\phi}_{kk'}$, will be measured by $\left\| \hat{\phi}_{kk'}(\cdot)\cdot - \phi_{kk'}(\cdot)\cdot \right\|_{L^2(\rho_T^{L,kk'})}$.

While this case requires learning multiple interaction kernels, it turns out that the learning theory developed for the single-type agent systems can be generalized, and the estimator in Eq. **9** still achieves optimal rates of convergence, and a similar control on the error of predicted trajectories can be obtained.

B. Second-Order Heterogeneous Agent Systems. Here, we focus on a broad family of second-order multitype agent systems (not included, even when rewritten as first-order systems, in the family discussed above). We consider systems with K types of agents:

$$\begin{cases} m_i \ddot{x}_i = F_i^v(\dot{x}_i, \xi_i) + \sum_{i'=1}^N \frac{1}{N_{k_{i'}}} \left(\phi_{k_i k_{i'}}^E(r_{ii'}) r_{ii'} + \phi_{k_i k_{i'}}^A(r_{ii'}) \dot{r}_{ii'} \right) \\ \dot{\xi}_i = F_i^\xi(\xi_i) + \sum_{i'=1}^N \frac{1}{N_{k_{i'}}} \phi_{k_i k_{i'}}^\xi(r_{ii'}) \xi_{ii'}, \end{cases} \quad \textbf{[11]}$$

for $i = 1, \ldots, N$. Here $k_i \in \{1, \ldots, K\}$ is the type of agent i, $\xi_i \in \mathbb{R}$ is a variable modeling the agent's response to the environment (e.g., food/light source), $\xi_{ii'} = \xi_{i'} - \xi_i$, and m_i, N_k, mass of agent i and number of agents of type k; F_i^v, F_i^ξ, noncollective influences on $\dot{x}_i$ and ξ_i; and $\phi_{kk'}^E$, $\phi_{kk'}^A$, $\phi_{kk'}^\xi$, energy-, alignment-, and ξ-type interaction kernels.

Note that here each agent is influenced by a weighted sum of different influences over agents of different types, leading to a rich family of models (including but not limited to prey–predator, leader–follower, and cars–pedestrian models). Using vector notation, let $\mathbf{f}_{\phi^E}(X^m)$ and $\mathbf{f}_{\phi^A}(X^m, \dot{X}^m) \in \mathbb{R}^{dN}$ be the collection of the energy and alignment induced interaction terms respectively, and $\mathcal{F}^v(\dot{X}^m, \Xi^m)_i = F_i^v(\dot{x}_i, \xi_i)$ (similar setup for $\mathcal{F}^\xi(\Xi^m)$ and $\mathbf{f}_{\phi^\xi}(X^m, \Xi^m)$) we can rewrite the equations as:

$$\begin{cases} \ddot{X}^m = \mathcal{F}^v(\dot{X}^m, \Xi^m) + \mathbf{f}_{\phi^E}(X^m) + \mathbf{f}_{\phi^A}(X^m, \dot{X}^m) \\ \dot{\Xi}^m = \mathcal{F}^\xi(\Xi^m) + \mathbf{f}_{\phi^\xi}(X^m, \Xi^m), \end{cases} \quad \textbf{[12]}$$

Table 1. Model selection: First- vs. second-order

System	Learned as first order	Learned as second order
First-order system	**0.01 ± 0.002**	1.6 ± 1.1
Second-order system	1.7 ± 0.3	**0.2 ± 0.06**

The table shows the mean and SD of the errors of estimated trajectories, over $M = 250$ train-test runs, with random ICs in each case. Small errors, consistent with our theory that the errors are on a scale of $M^{-2/5}$, indicate a correct model. The order is correctly identified in each case (highlighted in bold).

where $\phi^E = \{\phi^E_{kk'}\}$, $\phi^A = \{\phi^A_{kk'}\}$ and $\phi^\xi = \{\phi^\xi_{kk'}\}$, with $k, k' = 1, \dots, K$. We assume that the interaction kernels are the only unknowns in the model, to be estimated from the observations $\{X^m(t_l), \dot{X}^m(t_l), \Xi^m(t_l)\}_{l,m=1}^{L,M}$, with M ICs $X_0^m := X^m(0)$, $\dot{X}_0^m := \dot{X}^m(0)$, and $\Xi_0^m := \Xi^m(0)$ sampled independently from μ_0^X, $\mu_0^{\dot{X}}$, and μ_0^Ξ, respectively. With $\ddot{X}^m(t_l)$ approximated by finite difference, we construct estimators similar to those in Eq. **2**:

$$(\widehat{\phi}^E, \widehat{\phi}^A) := \underset{\varphi^E, \varphi^A \in \mathcal{H}^\nu}{\operatorname{argmin}} \frac{1}{ML} \sum_{m,l=1}^{M,L} \left\| \ddot{X}^m(t_l) - \mathcal{F}^\nu(\dot{X}^m(t_l), \Xi^m(t_l)) \right.$$
$$\left. - \mathbf{f}_{\varphi^E}(X^m(t_l)) - \mathbf{f}_{\varphi^A}(X^m(t_l), \dot{X}^m(t_l)) \right\|_{\mathcal{S}}^2, \quad [13]$$

and the interactions acting on the auxiliary variable ξ_i can be obtained separately as

$$\widehat{\phi}^\xi := \underset{\phi^\xi \in \mathcal{H}^\xi}{\arg\min} \frac{1}{ML} \sum_{m=1,l=2}^{M,L} \|\dot{\Xi}_l^m - \mathcal{F}^\xi(\Xi_l^m) - \mathbf{f}_{\phi^\xi}(X_l^m, \Xi_l^m)\|_{\mathcal{S}}^2,$$

where $\dot{\Xi}_l^m = \dot{X}^m(t_l)$, $X_l^m = X^m(t_l)$, $\Xi_l^m = \Xi^m(t_l)$, $\widehat{\phi}^\xi = \{\hat{\phi}^\xi_{kk'}\}_{k,k'=1}^K$, and the state space norm $\|\cdot\|_{\mathbb{S}}$ is defined similarly to the first-order case. Here, we are using a vectorized notation for $\varphi^E, \varphi^A, \mathcal{H}^\nu$ (a suitable product hypothesis space). To measure performance, for each pair (k, k'), we define a probability measure on $\mathbb{R}_+ \times \mathbb{R}_+$

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

and another probability measure on $\mathbb{R}_+ \times \mathbb{R}_+$,

$$\rho_{T,r,\xi}^{L,kk'}(r, \xi) = \frac{1}{LN_{kk'}} \sum_{l=1}^L \mathbb{E} \sum_{i \in C_k, i' \in C_{k'}, i \neq i'} \delta_{r_{ii'}(t_l), \xi_{ii'}(t_l)}(r, \xi),$$

where the expectation is with respect to ICs distributed according to $\mu_0^X \times \mu_0^{\dot{X}} \times \mu_0^\Xi$, and we let $\dot{r} = \|\dot{r}\|$ (with abuse of notation), $\xi_{ii'}(t) = |\xi_{i'}(t) - \xi_i(t)|$, $N_{kk'} = N_k N_{k'}$ if $k \neq k'$ and $N_{kk'} = \binom{N_k}{2}$ if $k = k'$ (and $N_k > 1$, as there is no kernel to learn if $N_k = 1$). Let $\rho_{T,r}^{kk'}$ be the marginal of $\rho_T^{kk'}$ with respect to r. We will measure the errors for $\hat{\phi}^E_{kk'}(r)r$, $\hat{\phi}^A_{kk'}(r)\dot{r}$, and $\widehat{\phi}^\xi_{kk'}(r)\xi$ in $L^2(\rho_{T,r}^{kk'})$, $L^2(\rho_T^{kk'})$, and $L^2(\rho_{T,r,\xi}^{kk'})$, respectively.

The algorithm to construct the estimator in Eq. **13** generalizes that for the first-order single-type agent systems, and involves a LS problem with a structured matrix with K^2 vertical bands indexed by (k, k'), accommodating the estimators for the interaction kernels. Note that such an LS problem takes into account, as it should, the dependencies in learning the various interaction kernels, all at once.

We note that while of course the second-order system may be written as a first-order system in the variables x_i and $v_i = \dot{x}_i$; even when $F_i^\nu \equiv 0$ and $\phi^A_{k_i, k_{i'}} \equiv 0$, the resulting equations for (x_i, v_i) are different from those governing the first-order systems considered above in Eq. **8**.

5. Examples

We consider the learning of interaction kernels and the prediction of trajectories for three canonical categories of examples of self-organized dynamics (see *SI Appendix,* section 3 for details).

Opinion Dynamics These are first-order ODE systems with a single type of agent, with bounded, discontinuous, compactly supported, and attraction-only interaction kernels. They model how the opinions of people influence each other and how consensus is formed based on different kinds of influence functions (refs. 14, 15, and 38 and references therein).

Predator–Swarm System We consider a first-order system with a single predator and a swarm of prey, with the interaction kernels (prey–prey, predator–prey, and prey–predator) similar to Lennard–Jones kernels (with appropriate signs to model attractions and repulsions). Different chasing patterns arise depending on the relative interaction strength of predator–prey vs. prey–predator interactions. We also consider a second-order
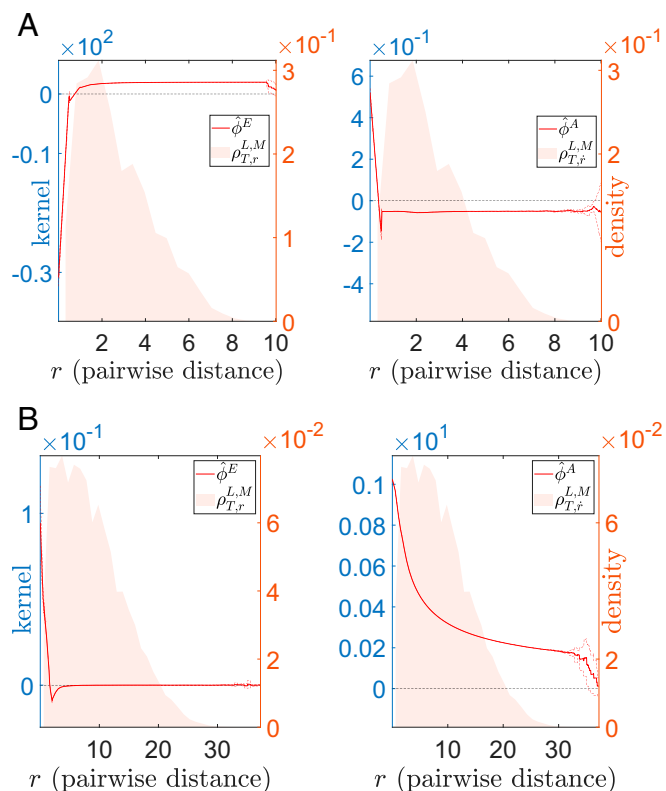
Fig. 7. Model selection: energy-based vs. alignment-based. The estimated interaction kernels for an energy-based model (*A*) and an alignment-based model (*B*). For each model, we compute two estimators: an energy-based interaction kernel $\hat{\phi}^E$ and an alignment-based interaction kernel $\hat{\phi}^A$. Our estimators correctly identify the type of model in each case: The $L^2(\rho_{T,r}^L)$ norm of $\hat{\phi}^E$ is significantly larger than that of $\hat{\phi}^A$ (means and SDs: **18.8 ± 0.4** vs. 6.5 ± 0.3) for the energy-based model, and the $L^2(\rho_{T,r}^L)$ norm of $\hat{\phi}^A$ is larger than that of $\hat{\phi}^E$ (means and SDs: **27.6 ± 0.7** vs. $2.4 \cdot 10^{-2} \pm 0.1$) for the alignment-based model. Note that the *y* axes are on very different scales.

predator–swarm system, with the collective interaction acting on accelerations, leading to even richer dynamics and chasing patterns (e.g., refs. 39–41).

Phototaxis This is a second-order ODE system with a single type of agents interacting in an environment, modeling phototactic bacteria moving toward a far-away fixed light source. The response of the bacteria to the light source is represented in the auxiliary variable ξ_i as the excitation level for each bacteria i (e.g., refs. 42–44). Another example which we do not pursue here is the Vicsek model (45), which fits perfectly in our model upon choosing $\xi_i = \theta_i$ (θ_i: moving direction of agent i).

In our experiments, we report the measure $\rho_T^{L,M}$ estimated from the training data, our estimator, and similarly in the case of noisy observations; we measure performance in terms of (relative) $L^2(\rho_T^L)$ error of the kernel estimators and of distance between true trajectories $X(t)$ and estimated trajectories $\hat{X}(t)$, on both the "training" interval $[0, T]$ (where observations were given) and in the future $[T, 2T]$ (predictions). See Prop. 3.4, where the bounds may be overly pessimistic, especially for systems tending to stable configurations. Our estimator performs extremely well in all these examples: The interaction kernels are accurately estimated, and the trajectories are accurately predicted. We refer the reader to Fig. 4 for the results of the opinion dynamics, Fig. 5 for the results of the predator–swarm dynamics, Fig. 6 for the results of the phototaxis, and *SI Appendix*, section 3 for further details on the setup for the experiments and a comprehensive report of all of the results, as well as a detailed description of the final algorithm and its computation complexity in *SI Appendix*, section 2.

Model Selection and Transfer Learning. We also consider the use of our method for model selection, where the theoretical guarantees on learning the interaction kernels and on predicting trajectories are used to decide between different models for the dynamics. We consider two examples of model selection, to test whether: (*i*) a second-order system is driven by energy-based or alignment-based interactions; or (*ii*) a heterogeneous agent system is driven by first- or second-order ODEs. For each of them, we construct two estimators assuming either case and then select models according to the performance of the estimators in predicting trajectories. See Table 1 and Fig. 7 for results and discussions and *SI Appendix*, section 3E for details.

As a simple example of transfer learning, we use the interaction kernel learned on a system with N agents to accurately predict trajectories of the same type of system but with more agents ($4N$ in our simulations); the interaction kernel acts as a sort of "latent variable" that seamlessly enables transfer across such related systems. In *SI Appendix*, section 3, we report the corresponding results, for all of the systems considered (see, however, Fig. 1 for the Lennard–Jones system).

Noisy Observations. Our estimators appear robust under observation noise, namely, if the observed positions and derivatives are corrupted by noise. Fig. 8 demonstrates the kernel estimation and trajectory prediction for the first-order predator–swarm system when only noisy observations are available. Similar results (reported in *SI Appendix*, section 3) are obtained in all of the other systems considered.

Choice of the Basis of the Hypothesis Space. Our learning approach is robust to the choice of hypothesis space $\mathcal{H}$, as long as the coercivity condition is satisfied by $\mathcal{H}$ (or the sequence $\mathcal{H}_n$). Additionally, different well-conditioned bases may be used in $\mathcal{H}$ to compute the projection onto $\mathcal{H}$, implying, together with the coercivity condition, a control of the condition number of the LS problem (*SI Appendix*, Prop. 2.1). To demonstrate this numerically, we compare the B-splines linear basis with the piecewise polynomial basis on the 1st-order predator–swarm system, with results shown in *SI Appendix*, Fig. S8.
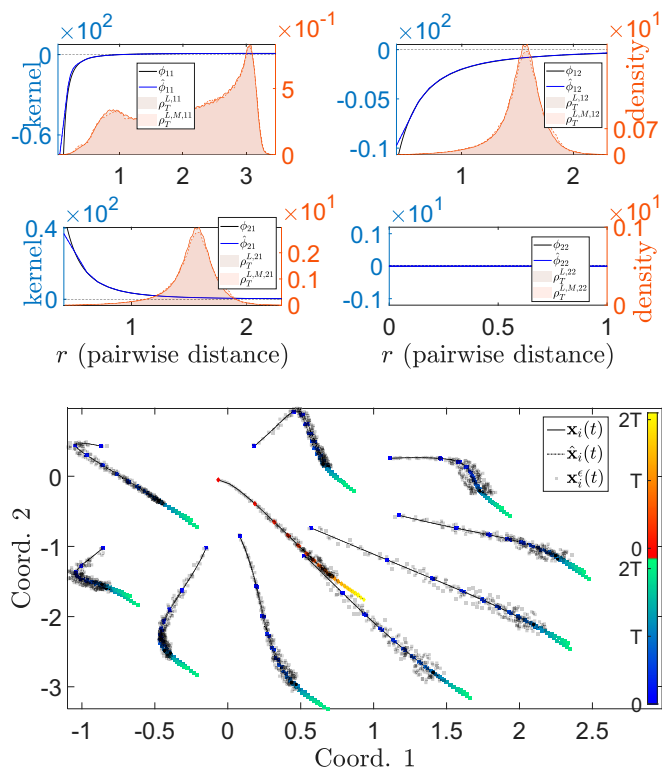
Fig. 8. Kernel estimation for PS1st from noisy observations. (*Upper*) Interaction kernels learned with Unif.$([-\sigma, \sigma])$ multiplicative noise with $\sigma = 0.1$ in the observed positions and velocities, with parameters as in *SI Appendix*, Table S9. The estimated kernels are minimally affected and only in regions with small ρ_T^L. (*Lower*) One of the observed trajectories before and after being perturbed by noise. The solid lines represent the true trajectory, the dashed semitransparent lines represent the noisy trajectory used as training data (together with noisy observations of the derivative), and the dashed-dotted lines are the predicted trajectory learned from the noisy trajectory.

6. Discussion and Conclusion

We proposed a nonparametric estimator for learning interaction kernels from observations of agent systems, implemented by computationally efficient algorithms. We applied the estimator to several classes of systems, including first- and second-order, with single- and multiple-type agents, and with simple environments. We have also considered observation data from different sampling regimes: many short-time trajectories, a single large-time trajectory, and intermediate time scales.

Our inference approach is nonparametric, does not rely on a dictionary of hypotheses (such as in refs. 6–8), exploits the structure of dynamics, and enjoys optimal rates of convergence (which we proved here for first-order systems), independent of the dimension of the state space of the system. Having techniques with solid statistical guarantees is fundamental in establishing trust in data-driven models for these systems and in using them as an aide to the researcher in formulating and testing conjectures about models underlying observed systems. In this vein, we presented two examples of model selection, showing that our estimators can reliably identify the order of a system and identify whether a system is driven by energy- or alignment-type interactions.

We expect further generalizations to the case of stochastic dynamical systems and to the cases of more general interaction kernels that depend on more general types of interaction between agents, beyond pairwise, distance-based interactions. Other future directions include (but are not limited to) a

better understanding of learnability, model selection based on the theory, learning from partial observations, and learning reduced models for large systems.

1. J. A. Carrillo, Y. Choi, S. Perez, "A review on attractive–repulsive hydrodynamics for consensus in collective behavior" in *Active Particles*, N. Bellomo, P. Degond, E. T, Eds. (Birkhäuser, Cham, Switzerland, 2017), Vol 1, pp. 259–298.
2. T. Kolokolnikov, H. Sun, D. Uminsky, A. Bertozzi, Stability of ring patterns arising from two-dimensional particle interactions. *Phys. Rev. E* **84**, 015203(R) (2011).
3. T. Vicsek, A. Zafeiris, Collective motion. *Phys. Rep.* **517**, 71–140 (2012).
4. Y. Shoham, K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundation* (Cambridge University Press, Cambridge, UK, 2009).
5. S. M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900* (Harvard Univ Press, Cambridge, MA, ed. 1, 1986).
6. H. Schaeffer, R. Caflisch, C. D. Hauck, S. Osher, Sparse dynamics for partial differential equations. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6634–6639 (2013).
7. S. Brunton, J. Proctor, J. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 3932–3937 (2016).
8. G. Tran, R. Ward, Exact recovery of chaotic systems from highly corrupted data. *Multi Model Simul.* **15**, 1108–1129 (2017).
9. W. Bialek *et al.*, Statistical mechanics for natural flocks of birds. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 4786–4791 (2012).
10. Y. Li, J. Wu, R. Tedrake, J. B. Tenenbaum, A. Torralba, Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. arXiv:1810.01566 (3 October 2018).
11. M. Ballerini *et al.*, Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1232–1237 (2008).
12. R. Lukeman, Y. X. Li, L. Edelstein-Keshet, Inferring individual rules from collective behavior. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12576–12580 (2010).
13. Y. Katz, K. Tunstrom, C. Ioannou, C. Huepe, I. Couzin, Inferring the structure and dynamics of interactions in schooling fish. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18720–8725 (2011).
14. U. Krause, A discrete nonlinear and non-autonomous model of consensus formation. *Commun. Part. Differ. Equation* **2000**, 227–236 (2000).
15. I. Couzin, J. Krause, N. Franks, S. Levin, Effective leadership and decision-making in animal groups on the move. *Nature* **433**, 513–516 (2005).
16. L. Györfi, M. Kohler, A. Krzyzak, H. Walk, *A Distribution-Free Theory of Nonparametric Regression* (Springer, New York, NY, 2002).
17. M. Bongini, M. Fornasier, M. Hansen, M. Maggioni, Inferring interaction rules from observations of evolutive systems I: The variational approach. *Math. Mod. Methods Appl. Sci.* **27**, 909–951 (2017).
18. H. Schaeffer, G. Tran, R. Ward, Extracting high-dimensional dynamics from limited data. *SIAM J. Appl. Math.* **78**, 3279–3295 (2017).
19. N. J. Brunel, Parameter estimation of ODE's via nonparametric estimators. *Electron. J. Stat.* **2**, 1242–1267 (2008).
20. H. Liang, H. Wu, Parameter estimation for differential equation models using a framework of measurement error in regression models. *J. Am. Stat. Assoc.* **103**, 1570–1583 (2008).
21. J. Cao, L. Wang, J. Xu, Robust estimation for ordinary differential equation models. *Biometrics* **67**, 1305–1313 (2011).
22. J. O. Ramsay, G. Hooker, D. Campbell, J. Cao, Parameter estimation for differential equations: A generalized smoothing approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69**, 741–796 (2007).
23. R. Bellman, R. S. Roth, The use of splines with unknown end points in the identification of systems. *J. Math. Anal. Appl.* **34**, 26–33 (1971).
24. J. M. Varah, A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Comput.* **3**, 28–46 (1982).
25. J. O. Ramsay, Principal differential analysis: Data reduction by differential operators. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 495–508 (1996).
26. M. Pascual, S. P. Ellner, Linking ecological patterns to environmental forcing via nonlinear time series models. *Ecology* **81**, 2767–2780 (2000).
27. J. Timmer, H. Rust, W. Horbelt, H. Voss, Parametric, nonparametric and parametric modelling of a chaotic circuit time series. *Phys. Lett. A* **274**, 123–134 (2000).
28. H. Miao, X. Xia, A. S. Perelson, H. Wu, On identifiability of nonlinear ode models and applications in viral dynamics. *SIAM Rev.* **53**, 3–39 (2011).
29. J. Ramsay, G. Hooker, *Dynamic Data Analysis: Modeling Data with Differential Equations* (Springer Series in Statistics, Springer, New York, NY, 2018).
30. J. von Brecht, D. Uminsky, On soccer balls and linearized inverse statistical mechanics. *J. Nonlinear Sci.* **22**, 935–959 (2012).
31. R. Simione, D. Slepčev, I. Topaloglu, Existence of ground states of nonlocal-interaction energies. *J. Stat. Phys.* **159**, 972–986 (2015).
32. F. Cucker, J. G. Dong, A general collision-avoiding flocking framework. *IEEE Trans. Automat. Contr.* **56**, 1124–1129 (2011).
33. F. Cucker, E. Mordecki, Flocking in noisy environments. *J. Math. Pure Appl.* **89**, 278–296 (2008).
34. G. Grégoire, H. Chaté, Onset of collective and cohesive motion. *Phys. Rev. Lett.* **92**, 025702 (2004).
35. J. Ke, J. W. Minett, C. P. Au, W. S. Y. Wang, Self-organization and selection in the emergence of vocabulary. *Complexity* **7**, 41–54 (2002).
36. J. A. Carrilo, Y. P. Choi, M. Haurray, "The derivation of swarming models: Mean-field limit and Wasserstein distances" in *Collective Dynamics from Bacteria to Crowds: An Excursion Through Modeling, Analysis and Simulation*, A. Muntean, F. Toschi, Eds. (CISM International Centre for Mechanical Sciences Courses and Lectures, Springer, Wien, Austria, Vol. 553, 2014), pp. 1–46.
37. J. A. Carrilo, M. Fornasier, G. Toscani, F. Vecil, "Particle, kinetic, and hydrodynamic models of swarming" in *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences, Modeling and Simulation in Science, Engineering and Technology*, G. Naldi, L. Pareschi, G. Toscani, N. Bellom, Eds. (Springer, Birkhäuser Boston, MA, 2010), pp. 297–336.
38. S. Mostch, E. Tadmor, Heterophilious dynamics enhances consensus. *SIAM Rev.* **56**, 577–621 (2014).
39. Y. Chen, T. Kolokolnikov, A minimal model of predator-swarm interactions. *J. R. Soc. Interf.* **11**, 20131208 (2013).
40. J. Jeschke, R. Tollrian, Prey swarming: Which predators become confused and why? *Anim. Behav.* **74**, 387–393 (2007).
41. M. Zheng, Y. Kashimori, O. Hoshino, K. Fujita, T. Kambara, Behavior pattern (innate action) of individuals in fish schools generating efficient collective evasion from predation. *J. Theor. Biol.* **235**, 13–167 (2005).
42. S. Ha, D. Levy, Particle, kinetic and fluid models for phototaxis. *Discrete Contin. Dyn. Syst. Ser. B* **12**, 77–108 (2009).
43. J. M. Skerker, H. C. Berg, Direct observation of extension and retraction of type IV pili. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 6901–6904 (2001).
44. D. Bhaya, A. Takahashi, A. R. Grossman, Light regulation of type IV pilus-dependent motility by chemosensor-like elements in synechocystis PCC6803. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 7540–7545 (2001).
45. T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, O. Shochet, Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.* **75**, 1226–1229 (1995).

APPLIED MATHEMATICS