Data-driven discovery of interacting particle systems using Gaussian processes

Jinchao Feng
Department of Applied Mathematics and Statistics
Johns Hopkins University
jfeng34@jhu.edu

Yunxiang Ren
Department of Physics
Harvard University
yren@g.harvard.edu

Sui Tang
Department of Mathematics
University of California, Santa Barbara
suitang@ucsb.edu

Abstract
Interacting particle or agent systems that display a rich variety of collective motions are ubiquitous in science and engineering. A fundamental and challenging goal is to understand the link between individual interaction rules and collective behaviors. In this paper, we study the data-driven discovery of distance-based interaction laws in second-order interacting particle systems. We propose a learning approach that models the latent interaction kernel functions as Gaussian processes, which can simultaneously fulfill two inference goals: one is the nonparametric inference of interaction kernel function with the pointwise uncertainty quantification, and the other one is the inference of unknown parameters in the non-collective forces of the system. We formulate learning interaction kernel functions as a statistical inverse problem and provide a detailed analysis of recoverability conditions, establishing that a coercivity condition is sufficient for recoverability. We provide a finite-sample analysis, showing that our posterior mean estimator converges at an optimal rate equal to the one in the classical 1-dimensional Kernel Ridge regression. Numerical results on systems that exhibit different collective behaviors demonstrate efficient learning of our approach from scarce noisy trajectory data.

1 Introduction
Interacting particle/agent systems are ubiquitous in science and engineering. The individual interactions among agents produce a rich variety of collective motions with visually compelling patterns, such as crystallization of particles, clustering of peoples’ opinions on social events, and coordinated movements of ants, fish, birds, and cars. For many interacting agent systems arising in biological, ecological, and social sciences, interaction laws are still the subject of investigation. A common belief in scientific discovery is that the complicated collective behaviors are consequences of simple interaction rules, e.g., only depending on pairwise distances \([1]\).

There have been tremendous research efforts of modeling collective dynamics using interaction laws based on pairwise distances. The goal is to induce desired collective behaviors from relatively simple parametric families of functions, such as flocking \([5]\), crystallization \([6, 7]\), clustering \([8, 9]\), milling \([11, 13]\). These efforts have been put to establish the well-posedness of the system and to prove that asymptotic behaviors of solutions reproducing similar qualitative macroscopic patterns. In this paper, we consider the inverse problem: given the observational data of the system, can we learn their interaction laws and even the governing equations?
We consider a system of $N$ homogeneous agents in $\mathbb{R}^d$, interacting according to a system of second-order ODEs,

$$m_i \ddot{x}_i = F^v(x_i, \dot{x}_i, \alpha) + \sum_{i' = 1}^{N} \frac{1}{N} \phi(||x_{i'} - x_i||)(x_{i'} - x_i), \quad i = 1, \ldots, N,$$

(1)

where the equations are derived from Newton’s second law; $m_i$ is the mass of agent $i$; $F^v$ is a parametric non-collective force with scalar parameters $\alpha$, modeling frictions of agents with the environment; the function $\phi$ is the energy-based interaction kernel function, derived from a potential energy function $U$ depending on pairwise distances:

$$U(X(t)) := \sum_{i,i'} \frac{1}{2N} \Phi(||x_{i'}(t) - x_i(t)||), \quad \Phi'(r) = \phi(r)r.$$  

(2)

The system (1) has applications in various disciplines to model a rich variety of collective behaviors, where the scalar parameters $\alpha$, for example, can be referred to stubbornness and biased opinions in opinion dynamics, or strength of self-propulsion and nonlinear drag in fish-milling systems.

To ensure the well-posedness of the system (1) given any initial condition, i.e., there is a unique solution that can be extended to all time, we assume that $\phi$ lies in the admissible space

$$W^{1,\infty}_c(0, R) := \{ \phi : \text{supp}(\phi) \subset [0, R]; \| \phi \|_\infty + \| \phi' \|_\infty < \infty \}.$$  

(3)

The restriction $\text{supp}(\phi) \subset [0, R]$ models the finite range of interactions between agents, and it may be relaxed to $\phi \in W^{1,\infty}(\mathbb{R}_+)$ with a suitable decay. We refer to [16] for the qualitative analysis of this type of systems.

In general, little information about the analytical form of the interaction kernel functions is available and it is often the case that the parameters $\alpha$ in the non-collective force $F^v$ are also unknown. We investigate whether they can be learned from observations of trajectory data. In particular, we would like to infer $\phi$, in a nonparametric fashion, i.e., without assuming parametric forms.

**Problem Statement.** For simplicity, we assume that the masses of agents are the same and have been normalized to be 1. We then write the second-order system (1) in a compact form:

$$Z(t) = F^v_\alpha(Y(t)) + f_\phi(X(t)),$$

(4)

where $Y(t) := [X(t), V(t)]^T \in \mathbb{R}^{2dN}$ represents the state variable for the system, $Z(t) = \dot{V}(t)$, $f_\phi(X(t)) : \mathbb{R}^{dN} \to \mathbb{R}^{dN}$ represents the distance-based interactions governed by $\phi$ as in (1). Our observations of the trajectory data consist of \{ $Y^{(m)}(t_i), Z_{\alpha^2}^{(m)}(t_i)$\}_{i=1}^{L,M}$, sampled along trajectories at $0 = t_1 < \cdots < t_L = T$, where the $M$ initial conditions $Y^{(m)}(0)$ are drawn i.i.d from a probability measure $\mu_0 = [\mu_0^X, \mu_0^\dot{X}]^T$ on $\mathbb{R}^{2dN}$; the acceleration $Z_{\alpha^2}^{(m)}(t_i)$ is observed with i.i.d additive Gaussian noise $\epsilon^{(m)} \sim N(0, \sigma^2 I_{dN \times dN})$ so that $Z_{\sigma^2}^{(m)}(t_i) = Z_{\alpha^2}^{(m)}(t_i) + \epsilon^{(m)}$. Our goal is to learn $\phi$ as well as the unknown scalar parameters $\alpha$ from the trajectory data, and then use learned estimators to predict the collective behavior of the system.

In applications, we may only have access to scarce and possibly noisy trajectory data, i.e., $M, L$ are small. In such scenarios, it is useful to obtain quantitative predictive uncertainties in estimated interaction kernel functions, that quantify the reliability of estimators. This information would be helpful to design a data acquisition plan, often referred to as the active learning, where the goal is to optimally enhance our knowledge about the system. Gaussian processes regression (GPR) is a non-parametric Bayesian machine learning technique with built-in quantifications of uncertainties encoded in the posterior variances of estimators. In this paper, we employ and modify Gaussian Processes (GPs) to interacting particle systems for our inference task and provide a thorough quantitative analysis of the performance of the estimators for interaction kernel functions.

### 1.1 Overview of the results

We begin by proposing a learning approach based on GPR that consists of three steps. We first assign a Gaussian prior to $\phi$, where the covariance function is parameterized by a set of scalar parameters
We show that it is well-posed and equivalent to a classical 1-dimensional nonparametric regression. We then predict the values of $\phi$ at test locations by the mean of its posterior distribution, with the marginal posterior variance that quantifies the reliability. Our approach relies on two key ingredients: (1) the observational noise is Gaussian; (2) the joint distribution of $f_\phi$ and $\phi$ according to the prior is still Gaussian (since $f_\phi$ is linear in $\phi$). So the posterior distribution of $\phi$ can be computed by conditioning on the observations of $f_\phi$.

We also develop a systematic learning theory for the prediction step of the proposed approach. We consider the regime where $L$ is fixed and $M \to \infty$. We introduce relevant Hilbert spaces and operators to formulate learning the interaction kernel function $\phi$ as a statistical inverse learning problem. We show that it is well-posed and equivalent to a classical 1-dimensional nonparametric regression problem by introducing a coercivity condition. We then generalize the well-known connection between GPR and Kernel Ridge Regression (KRR) to our setting: by deriving a Representer theorem, we prove that the posterior mean estimator can be viewed as a KRR estimator. Levering this connection and recent advancement of KRR theory, we provide non-asymptotic analyses on the reconstruction error of posterior mean estimator, using a Reproducing Kernel Hilbert Space (RKHS) norm that corresponds to the Gaussian prior. In particular, by choosing a Gaussian prior that is adaptive to data, we prove in Theorem 5 that, the posterior mean estimators converge in the RKHS norm at an optimal rate of $M$ to $\phi$, as if we were in the in-principle-easier (both in terms of dimension and observed quantities) 1-dimensional KRR setting with noisy observations.

1.2 Comparison with previous work

Our work is built on the recent progress on the data-driven discovery of interacting agent systems, where the problem of learning interaction kernel functions was introduced and a learning approach based on Vanilla least squares was proposed. Compared to [29,30], both our learning problem and approach are new: Here, the inference is not only done for interaction kernel function but also parameters in non-collective forces. Moreover, our approach has new features: (1) estimates $\alpha$ by a powerful training procedure that allows flexible modeling; (2) provides us with the uncertainty quantifications of our estimators; (3) enjoys superior performances and extrapolation properties when the data is scarce and noisy.

The learning theory developed in this paper is related to but significantly departs from the ones for the previous approach. All of these works investigate the performance of estimators as $M \to \infty$. However, it is in this work that we first establish a rigorous operator-theoretic framework to treat this learning problem as a statistical inverse learning problem. Within this framework, we see the connection to the previous work: the posterior mean estimators obtained in the prediction step of our approach can be viewed as KRR estimators, whose risk functionals are the regularized version of those proposed in previous work and underlying hypothesis space is an RKHS corresponding to the covariance kernel of the prior. However, the previous work deals with the noise-free trajectory data and studied the convergence of least square estimators in a $L^2$ space. We study the convergence of estimators in the RKHS norm for noisy trajectory data. From the perspective of the inverse problem, we analyze the reconstruction error while the previous work analyzed the residual error, where the coercivity condition is the bridge (see Appendix C for more details).

1.3 Notation and preliminaries

Let $\rho$ be a Borel positive measure on $\mathbb{R}^D$. We use $L^2(\mathbb{R}^D; \rho; \mathbb{R}^n)$ to denote the set of $L^2(\rho)$ integrable vector-valued functions that mapping $\mathbb{R}^D$ to $\mathbb{R}^n$. For a function $f \in L^2(\mathbb{R}^D; \rho; \mathbb{R}^n)$, let $X = [x_1, \cdots, x_m]^T \in \mathbb{R}^{mD}$ with $x_i \in \mathbb{R}^D$, then $f(X) = [f(x_1), \cdots, f(x_m)]^T \in \mathbb{R}^{mn}$, i.e. acting componentwisely. Let $S_1$ be a measurable subset of $\mathbb{R}^m$, the restriction of the measure $\rho$ on $S_1$, denote by $\rho|_{S_1}$, is defined as $\rho|_{S_1}(S_2) = \rho(S_1 \cap S_2)$ for any measurable subset $S_2$ of $\mathbb{R}^D$.

Let $\mathcal{H}$ be a Hilbert space. We denote by $B(\mathcal{H})$ the set of bounded linear operators on $\mathcal{H}$. Let $A \in B(\mathcal{H})$, the notation $\text{Im}(A)$ denotes its image space. For two self-adjoint operators $A, B \in B(\mathcal{H})$, we say that $A \geq B$ if $A - B$ is a positive operator, i.e. $(A - B)h, h)_{\mathcal{H}} \geq 0$ for all $h \in \mathcal{H}$. If $A$ is a
compact positive operator, and \( \lambda_n \) represents the \( n \)th eigenvalue in a decreasing order. By the spectral theory of compact operators, the eigenfunctions \( \{ \varphi_n \}_{n=1}^N \) (\( N \) could be \( \infty \)) of \( A \) form an orthonormal basis for \( \mathcal{H} \) so that \( A = \sum_{n=1}^N \lambda_n \varphi_n \varphi_n^* \). For \( \tau < 0 \), we define \( A' = \sum_{n=1}^N \lambda_n' \varphi_n \varphi_n^* \) on the subspace \( S_\tau \) of \( \mathcal{H} \) given by \( S_\tau = \{ \sum_{n=1}^N a_n \varphi_n \mid \sum_{n=1}^N (a_n \lambda_n')^2 \text{ is convergent} \} \). If \( h \not\in S_\tau \), then \( \| A' h \|_\mathcal{H} = \infty \).

**Preliminaries on RKHS** Let \( D \) be a compact domain of \( \mathbb{R}^D \). We say that \( K : D \times D \to \mathbb{R} \) is a Mercer Kernel if it is continuous, symmetric and positive semidefinite, i.e., for any finite set of distinct points \( \{ x_1, \ldots, x_M \} \subset D \), the matrix \( (K(x_i, x_j))_{i,j=1}^M \) is positive semidefinite. For \( x \in \mathbb{R}^D \), \( K_x \) is a function defined on \( D \) such that \( K_x(y) = K(x, y) \). The Moore–Aronszajn theorem proves that there is a RKHS \( \mathcal{H}_K \) associated with the kernel \( K \), which is defined to be the closure of the linear span of the set of functions \( \{ K_x : x \in D \} \) with respect to the inner product \( \langle \cdot, \cdot \rangle_{\mathcal{H}_K} \) satisfying \( (K_x, K_y)_{\mathcal{H}_K} = K(x, y) \).

**2 Methodology**

We start by modeling \( \phi \) as a Gaussian process \([18,20]\), i.e., consider the prior \( \phi \sim \mathcal{GP}(0, K_\theta(r, r')) \), with mean zero and covariance kernel function \( K_\theta \) which depends on hyper-parameters \( \theta \). From the perspective of function space, a favorable situation is that \( \phi \) lies in (or is well-approximated by) the RKHS that is associated with \( K_\theta \).

We introduce the vectorized notation for the training data \( \mathcal{Y} = [Y^{(1,1)}; \ldots; Y^{(M,L)}]' \), and let \( \mathcal{Z} := [Z_\sigma^{(1,1)}; \ldots; Z_\sigma^{(M,L)}]' \). By [4] and the properties of GPs, \( \mathcal{Z} \) follows a multivariate Gaussian distribution with the mean vector \( F_\alpha^\mathcal{Z}(\mathcal{Y}) \in \mathbb{R}^{d\text{NML}} \) and the covariance matrix \( K_{\mathcal{F}_\phi}(\mathcal{X}, \mathcal{X}; \theta) + \sigma^2 I_{\text{NML}} \in \mathbb{R}^{d\text{NML} \times d\text{NML}} \), where \( K_{\mathcal{F}_\phi}(\mathcal{X}, \mathcal{X}; \theta) = (\text{Cov}(f_\phi(X^{(m,l)}), f_\phi(X^{(m',l')})))_{m,M,L,L} \in \mathbb{R}^{d\text{NML} \times d\text{NML}} \) with \((i,j)\)th block of \( \text{Cov}(f_\phi(X), f_\phi(X')) \) computed by

\[
\text{Cov}(f_\phi(X)_i, f_\phi(X')_j) = \frac{1}{N^2} \sum_{k \neq i, k' \neq j} (K_\theta(r_{ik}, r_{j'k'})r_{ik}r_{j'k'})^T \in \mathbb{R}^{d \times d},
\]

where \( r_{ik} := \| r_{ik} \| = \| x_k - x_i \| \) and \( r_{ik} \) is defined similarly. Therefore, we can train the hyper-parameters \( \alpha \) and \( \theta \) by maximizing the probability of the observational data, which is equivalent to minimize the negative log marginal likelihood (see Chapter 4 in [18])

\[
- \log P(\mathcal{Z} | \mathcal{Y}, \alpha, \theta, \sigma^2) = \frac{1}{2} (\mathcal{Z} - F_\alpha^\mathcal{Z}(\mathcal{Y}))^T (K_{\mathcal{F}_\phi}(\mathcal{X}, \mathcal{X}; \theta) + \sigma^2 I) (\mathcal{Z} - F_\alpha^\mathcal{Z}(\mathcal{Y})) + \frac{1}{2} \log |K_{\mathcal{F}_\phi}(\mathcal{X}, \mathcal{X}; \theta) + \sigma^2 I| + \frac{d\text{NML}}{2} \log 2\pi.
\]

To solve for the hyper-parameters \((\alpha, \theta)\), we can apply a gradient based method, Quasi-Newton optimizer L-BFGS [34], to minimize the negative log marginal likelihood. The marginal likelihood induces an automatic trade-off between data-fit and model complexity [35]. This flexible training procedure distinguishes GPs from other kernel-based methods [36,38] and regularization-based approaches [39,41].

After the training procedure, we obtain an updated prior on \( \phi \), we can then predict the value \( \phi(r^*) \) using the mean of its posterior distribution. Note that

\[
\begin{bmatrix} f_\phi(X) \\ \phi(r^*) \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} K_{\mathcal{F}_\phi}(\mathcal{X}, \mathcal{X}) & K_{\mathcal{F}_\phi}(\mathcal{X}, r^*) \\ K_{\mathcal{F}_\phi}(r^*, \mathcal{X}) & K_{\theta}(r^*, r^*) \end{bmatrix} \right),
\]

where \( K_{\mathcal{F}_\phi}(\mathcal{X}, r^*) = K_{\phi,\mathcal{F}_\phi}(r^*, \mathcal{X}) \) denotes the covariance matrix between \( f_\phi(X) \) and \( \phi(r^*) \). By conditioning on \( f_\phi(X) \), we obtain that

\[
p(\phi(r^*) | \mathcal{Y}, \mathcal{Z}, r^*) \sim \mathcal{N}(\bar{\phi}^*, \text{Var}(\phi^*)),
\]

where

\[
\bar{\phi}^* = K_{\phi,\mathcal{F}_\phi}(r^*, \mathcal{X})(K_{\mathcal{F}_\phi}(\mathcal{X}, \mathcal{X}) + \sigma^2 I)^{-1}(\mathcal{Z} - F_\alpha^\mathcal{Z}(\mathcal{Y}))
\]

\[
\text{Var}(\phi^*) = K_{\theta}(r^*, r^*) - K_{\phi,\mathcal{F}_\phi}(r^*, \mathcal{X})(K_{\mathcal{F}_\phi}(\mathcal{X}, \mathcal{X}) + \sigma^2 I)^{-1}K_{\mathcal{F}_\phi,\phi}(\mathcal{X}, r^*),
\]
and the detailed derivation is shown in Appendix A. The posterior variances $\text{Var}(\phi^*)$ can be used as a good indicator for the uncertainty of the estimation $\hat{\phi}(r^*) := \hat{\phi}^*$ based on our Bayesian approach. Moreover, using the estimated parameters $\hat{\alpha}$ and interaction kernels $\hat{\phi}$, we can predict the dynamics based on the equations $\hat{Z}(t) = F_{\hat{\alpha}}(Y(t)) + f_{\hat{\phi}}(X(t))$.

3 Learning theory

Numerical results in Section 4 show that $\alpha$ are accurately recovered from a small amount of noisy data in the training step. In this section, we assume that $\phi$ is the only unknown term in the governing equation (1) and shall focus on the prediction step of the proposed approach. Our goal is to establish a learning theory, which analyzes the performance of the posterior mean (9) that approximates $f$ when $L$ is fixed and $M \to \infty$. For an easy presentation, we restrict our attention to first-order systems, which is a special case of second-order systems by assuming the masses of the agents are zero:

$$\hat{X}(t) = f_{\hat{\phi}}(X(t)).$$

But our analysis can be extended to second-order systems with (known) non-collective force terms with slight modifications.

In first-order systems, we denote the vectorized training dataset as $\mathcal{Y}_{\sigma^2,M} = \{X_M, \mathcal{V}_{\sigma^2,M}\}$, where $X_M = \text{Vec}(\{X^{(m)}(t_l)\}_{m=1}^{M,L}) \in \mathbb{R}^{dNML}$, $\mathcal{V}_{\sigma^2,M} = \text{Vec}(\{X^{(m)}(t_l) + \epsilon^{(m,l)}\}_{m=1}^{M,L}) \in \mathbb{R}^{dNML}$; $X^{(m)}(t_l) \sim \mu_0$, that is a probability measure on $\mathbb{R}^{dN}$; the noise term $\epsilon^{(m,l)} \sim \mathcal{N}(0, \sigma^2 I_{dN})$; we assume that $\mu_0$ is independent of the distribution of noise.

Suppose that $\phi \sim \mathcal{GP}(0, \tilde{K})$, where $\tilde{K}$ is a Mercer kernel defined on $[0, R] \times [0, R]$ and may depend on the size of observational data. In the prediction step, we approximate $\phi(r^*)$ at a new point $r^* \in [0, R]$ by the posterior mean

$$\hat{\phi}_M(r^*) = \tilde{K}_{\phi, X_M}(r^*, X_M) (\tilde{K}_{\phi, X_M} X_M + \sigma^2 I)^{-1} \mathcal{V}_{\sigma^2,M},$$

where the matrices $\tilde{K}_{\phi, X_M}(r^*, X_M)$ and $\tilde{K}_{\phi, X_M} X_M$ denote the covariance matrix between $f_{\phi}(X_M)$ and $\phi(r^*)$, and $f_{\phi}(X_M)$ and $f_{\phi}(X_M)$ respectively.

3.1 Connection with nonparametric regression and inverse problem

Our trajectory data is of the type for nonparametric regression of $f_{\phi}$. The recent works [42-51] that use GPs in learning differential equations typically model $f_{\phi} : \mathbb{R}^{dN} \to \mathbb{R}^{dN}$ as a Gaussian process. However, the straightforward application of existing approaches may suffer from the well-known curse of dimensionality since $f_{\phi}$ is of dimension $dN$. In the most favorable situation where $f_{\phi}$ is sufficiently smooth (at least belongs to the RKHS associated with the covariance kernel), one could get a learning rate that is independent of the dimension. However, this requires $f_{\phi}$ to be $s$-times differentiable with bounded $s$-th order derivative and $s > \frac{dN}{2}$ (due to Sobolev Embedding theorem), which is too restrictive in many applications. Furthermore, our trajectory data are correlated in time due to the underlying dynamics, which departs from the classical nonparametric regression setting and causes hurdles in determining the sample complexity.

We proceed in a different direction. We shift the regression target to $\phi$, and the Gaussian process has incorporated the structure of the system of equations (11), as well as its symmetries (over permutations of agents of each type, and over translations). Since $\phi$ is an 1-dimensional function, one just needs to simply choose covariance kernels to incorporate smoothness of $\phi$, ranging from Lipschitz continuity to infinitely differentiability. However, the problem of learning $\phi$ becomes an inverse problem that may be ill-posed. Below, we give a rigorous formulation by introducing relevant function spaces.

**Formulation of the inverse problem.** We introduce a probability measure on $\mathbb{R}^{dN}$:

$$\rho_X := \mathbb{E} \left[ \frac{1}{L} \sum_{l=1}^{L} \delta_{X(t_l)} \right],$$

(13)
where $\delta$ is the Dirac $\delta$-distribution, $\mathbf{X}(t_i) \in \mathbb{R}^{dN}$ is the position vector of all agents at time $t_i$, and the expectation is taken with respect to random initial conditions drawn from $\mu_0$. The $L^2$ space associated to the probability measure is denote by $L^2(\mathbb{R}^{dN}; \rho_\mathbf{X}; \mathbb{R}^{dN})$ and we have $f_\phi \in L^2(\mathbb{R}^{dN}; \rho_\mathbf{X}; \mathbb{R}^{dN})$.

Let $K$ be a Mercer kernel that is defined on $[0, R] \times [0, R]$ and $\mathcal{H}_K$ be the RKHS associated to $K$. Recall that the admissible space is introduced in (1) to guarantee the well-posedness of the system (11). It is reasonable to make the following assumption:

**Assumption 1.** Suppose that $\kappa^2 = \sup_{r \in [0, R]} K(r, r) < \infty$, and $\phi \in \mathcal{H}_K$.\footnote{One can choose a Matérn kernel whose associated RKHS containing the admissible space as a subspace.}

Now we define a linear operator $A : \mathcal{H}_K \rightarrow L^2(\mathbb{R}^{dN}; \rho_\mathbf{X}; \mathbb{R}^{dN})$ by

$$A\phi = f_\phi,$$

(14)

where $f_\phi$ is the right hand side of system (11) by replacing $\phi$ with $\varphi$. Then $A$ is a bounded linear operator (see Appendix C). In the case of “infinite data”, i.e., $M = \infty$, our learning problem is equivalent to solve the linear equation (14) in $\mathcal{H}_K$ given $A$ and $f_\phi$ in $L^2(\mathbb{R}^{dN}; \rho_\mathbf{X}; \mathbb{R}^{dN})$, and therefore becomes an inverse problem.

However, this inverse problem may be ill-posed. The heuristic argument below reveals the possible non-uniqueness of the solution: due to the structure of interaction force shown in r.h.s of (1), the interaction kernel function $\phi$ is observed through a collection of non-independent linear measurements with values $\hat{x}_i$, the l.h.s. of (11), at locations $r_{i'i'} := |\hat{x}_{i'} - \hat{x}_i|$, with coefficients $r_{i'i'} := |x_{i'} - x_i|$. One could attempt to recover $\{\phi(r_{i'i'}))\}_{i'i'}$ from the equations of $\hat{x}_i$’s by solving the corresponding linear system. Unfortunately, this linear system is usually underdetermined as $dN$ (number of known quantities) $\leq N(N-1)/2$ (number of unknowns) and in general one will not be able to recover the values of $\phi$ at locations $\{r_{i'i'}\}_{i,i'}$.

Before we further discuss the performance of the proposed estimators using GPs, we must first establish that it is well-posed so one can asymptotically identify $\phi$ from sampled trajectory data. Otherwise the obtained estimator will have limited value as a scientific and predictable tool.

**Well-posedness by a coercivity condition** In our numerical experiments, we find that our estimators produced faithful aproximations to the ground truth and the accuracy got improved once we have more data. This motivated us to study under which condition the inverse problem is well-posed and verify this condition is generically satisfied. Note that the observational variables for $\phi$ consist of pairwise distances. In (19)\footnote{e.g., choose $\mu_0 := \text{Unif}[-R/2, R/2]^{dN}$. Then $\text{Supp}(\rho^1) = [0, R]$ and $\text{Supp}(\rho^2) \subset \text{Supp}(\rho^1)$ for $L > 1$.}, a probability measure on $\mathbb{R}^+$ that encodes the information about the dynamics marginalized to pairwise distances was introduced as

$$\rho^2(d\mathbf{r}) := \frac{1}{(\pi N^2)^{d/2}} \sum_{l=1}^{N} \sum_{i,i'=1,i \neq i'}^{N} \mathbb{E}_{\mu_0}[\delta_{r_{i'i'}(t_i)}],$$

(15)

where $\delta$ is the Dirac $\delta$-distribution, and thus $\mathbb{E}_{\mu_0}[\delta_{r_{i'i'}(t_i)}(d\mathbf{r})]$ is the distribution of the random variable $r_{i'i'}(t) = |(x_{i'}(t) - x_i(t)|$, with $x_i(t)$ being the position of particle $i$ at time $t$.

The probability measure $\rho^2$ depends on both the distribution of initial conditions $\mu_0$, while it is independent of the observed data. Note that it is on the support of $\rho^2$ that $\phi$ could be learned. Without loss of generality, we assume that $\rho^2$ is non-degenerate on $[0, R]$. Due to the structure of the interaction force (see r.h.s of (1)), we introduce a positive measure that appear naturally in estimating the error of estimators

$$\hat{\rho}^2(r) = r^2 \rho^2(d\mathbf{r}) \subset [0, R], r \in \mathbb{R}^+.$$  

(16)

Now we introduce a sufficient condition to guarantee the injectivity of the operator $A$. Since functions in $\mathcal{H}_K$ are continuous, $\mathcal{H}_K$ can be naturally embedded as a subspace of $L^2([0, R]; \rho^2; \mathbb{R})$.

**Definition 2.** We say that the system (11) satisfies the coercivity condition on $\mathcal{H}_K$, if $\forall \phi \in \mathcal{H}_K$, there exists $c_{\mathcal{H}_K} > 0$ such that

$$\|A\phi\|^2_{L^2(\rho_\mathbf{X})} = \|f_\phi\|^2_{L^2(\rho_\mathbf{X})} \geq c_{\mathcal{H}_K} \|\phi\|^2_{L^2(\hat{\rho}^2)}.$$  

(17)

We choose the largest $c_{\mathcal{H}_K}$ that satisfies (17) and refer it as the coercivity constant.
We prove Theorem 4 by deriving a Representer theorem (see Appendix C) for the empirical risk we first decompose the reconstruction error as the sum of two types of errors which generalizes the classical facts.

We conjecture that the coercivity condition is generally satisfied for various systems and initial setting of inverse problem, that decouples in an efficient way the contribution of the noise from the other error terms (bias, variance) that appear routinely in statistical learning theory. More specifically, we first decompose the reconstruction error as the sum of two types of errors

\[
\phi^\lambda_M - \phi = \phi^\lambda_M - \phi^\lambda_\infty + \phi^\lambda_\infty - \phi,
\]

Sample error Approximation error

Proposition 3 below shows that, if the coercivity condition holds, our inverse problem will become well-posed as if we were in a classical one dimensional KRR setting: let us consider learning \( \phi \) from i.i.d. noisy samples: \( y_i = \phi(r_i) + \epsilon_i, r_i \sim \tilde{\rho}_F^0, \epsilon_i \sim \mathcal{N}(0, \sigma^2) \). Denote by \( J_{\tilde{\rho}_F^0}^*: \mathcal{H}_K \rightarrow L^2([0, R]; \tilde{\rho}_F^0; \mathbb{R}) \) the canonical embedding map, and its adjoint operator \( J_{\tilde{\rho}_F^0}^*: L^2([0, R]; \tilde{\rho}_F^0; \mathbb{R}) \rightarrow \mathcal{H}_K \) is an integral operator with respect to the kernel \( K \), i.e., for \( \varphi \in L^2([0, R]; \tilde{\rho}_F^0; \mathbb{R}) \) and \( r \in [0, R] \),

\[
(J_{\tilde{\rho}_F^0}^* \varphi)(r) = \int_0^R K(r, r') \varphi(r') d\rho^0_F(r').
\]

**Proposition 3.** Assume that the coercivity condition \([17]\) holds. Then we have that

\[
c_{\mathcal{H}_K} J_{\tilde{\rho}_F^0}^* J_{\tilde{\rho}_F^0} \leq A^* A \leq \kappa^2 R^2 J_{\tilde{\rho}_F^0}^* J_{\tilde{\rho}_F^0}.
\]  

(18)

The decay rate of eigenvalues of \( J_{\tilde{\rho}_F^0}^* J_{\tilde{\rho}_F^0} \) plays an important role in the derivation of learning theory for KRR. The Courant–Fischer–Weyl min-max principle (see Appendix C) indicates that \( A^* A \) and \( J_{\tilde{\rho}_F^0}^* J_{\tilde{\rho}_F^0} \) have the same decay rate. So the coercivity condition is the bridge that allows us to generalize the classical learning theory to our setting.

The coercivity condition \([18]\) introduces constraints on \( \mathcal{H}_K \) and on the distribution of the solutions of the system, and it is therefore natural that it depends on the distribution \( \mu_0 \) of the initial condition \( X(0) \), and the true interaction kernel \( \phi \). The positivity of \( c_{\mathcal{H}_K} \) is related to the positivity of relevant integral operators. Leveraging the recent advancement in \([31, 32]\), our ongoing work shows that, in the case of \( L = 1 \), the coercivity constant \( c_{\mathcal{H}_K} \geq \frac{N-1}{N} \) for general systems and \( \mathcal{H}_K \). Furthermore, it can be a positive constant independent of \( N \) if \( \mu_0 \) is exchangeable Gaussian. The generalization for the case \( L > 1 \) requires significant effort, since solutions to the systems are implicit and their behavior are of rich variety. However, as the numerical results and relevant discussions in \([29, 32]\), we conjecture that the coercivity condition is generally satisfied for various systems and initial distributions.

### 3.2 Finite sample analysis of posterior mean estimators

There is a well-known connection between GPR and KRR in the classical nonparametric regression \([23, 24]\) setting. Our learning problem shifts the regression target to \( \phi \) with dependent observational data, and therefore departs from the classical setting. We show that the posterior mean estimator in our approach still coincides with KRR estimators for suitable regularized least square risk functional, which generalizes the classical facts.

**Theorem 4.** (Connection with the Kernel Ridge Regression) Given the noisy trajectory data \( \mathbb{Y}^{\sigma^2, M} \), if \( \phi \sim \mathcal{G}(0, \tilde{K}) \) with \( \tilde{K} = \frac{\sigma^2 K}{MN} \), for some \( \lambda > 0 \), the posterior mean \( \bar{\phi}_M \) in \([12]\) coincides with the KRR estimator \( \phi^\lambda_M \) to the regularized empirical least square risk functional \( \phi^\lambda_M := \arg \min_{\varphi \in \mathcal{H}_K} \mathcal{E}^\lambda_M(\varphi), \) where

\[
\mathcal{E}^\lambda_M(\varphi) = \frac{1}{LM} \sum_{l=1}^{L} \sum_{m=1}^{M} \| f_{\varphi}(X^{(m,l)}) - Y_{\sigma^2}^{(m,l)} \|_2^2 + \lambda \| \varphi \|_{\mathcal{H}_K}^2.
\]  

(19)

We prove Theorem 4 by deriving a Representer theorem (see Appendix C) for the empirical risk functional \([19]\). The proof is done by developing an operator-theoretic approach that derives the operator representation of the minimizers. From now on, we assume that \( \phi \sim \mathcal{G}(0, \tilde{K}) \) with \( \tilde{K} = \frac{\sigma^2 K}{MN\lambda} \) so that our posterior mean estimator is \( \phi^\lambda_M \). It suffices to analyzing the performance of KRR estimators \( \phi^\lambda_M \).

One of our key technical contributions is splitting the reconstruction error of the estimator in the setting of inverse problem, that decouples in an efficient way the contribution of the noise from the other error terms (bias, variance) that appear routinely in statistical learning theory. More specifically, we first decompose the reconstruction error as the sum of two types of errors
where $\phi_{H_K}^{\lambda, M}$ is the unique minimizer to the expectation functional of $\phi_{H_K}$ in (19) ($M = \infty$).

The sample error comes from the randomness in the initial conditions of observed trajectories and the randomness in noise term. Our analysis relies on decoupling the sample error into noise part and noise-free part. We refer to Appendix C.

To estimate the approximation error $\|\phi_{H_K}^{\lambda,M} - \phi\|_{H_K}$, we follow the standard routine in the literature of Tikhonov regularization for inverse problem (see section 5 in [26]). This is done by assuming $\phi \in \text{Im}(B^\dagger)$, where $B = A^*A$ and $0 < \gamma \leq \frac{1}{2}$. Below we show that, given the data $\hat{Y}_{\sigma^2,M}$, with appropriate choice of $\lambda$, our posterior mean estimator achieves the convergence rate in $M$, that coincides with the optimal rate in classical KRR setting when $\phi$ lies in $\text{Im}(J_y^\alpha, J_y^\beta)^\gamma$ (see Table 1 in [22]).

**Theorem 5.** If we choose $\lambda \asymp M^{-\frac{1}{2 + \gamma}}$. For any $\delta \in (0,1)$, it holds with probability at least $1 - \delta$ that

$$\|\phi_{H_K}^{\lambda,M} - \phi\|_{H_K} \lesssim C(\phi, \kappa, R, c_{H_K}, \sigma) \log(\frac{8}{\delta}) M^{-\frac{1}{2 + \gamma}},$$

where $C = \max\{\frac{\kappa^2 R^2 \|\phi\|_{L^\infty}}{\sqrt{\kappa n}}, \frac{2\kappa R \sigma}{\sqrt{L N d}}, \|B^\dagger - \gamma \|_{H_K}\}.$

One shortcoming of our result is that the bound is independent of $L$ and $N$ (we have $LN^2/2$ pairwise distances for each of the $M$ trajectories): our result only sees trajectories: our result only sees the randomness in noise term. Our analysis relies on decoupling the sample error into noise part and noise-free part. However, obtaining this optimal rate is satisfactory, because we do not observe the values of $\phi$ and the pairwise distances are in general correlated.

### 4 Numerical examples

In this section, we report the empirical performance of the proposed approach with the Matérn covariance function in three systems exhibiting clustering, milling, and flocking behaviors, motivated by applications in social science, biology, and ecology. The first two examples used the synthetic covariance function in three systems exhibiting clustering, milling, and flocking behaviors, motivated by applications in social science, biology, and ecology.

The first example is a first-order system that models the collective dynamics of continuous opinion exchange in the presence of stubborn agents [52]. For agent $i$, $x_i \in \mathbb{R}$ represents its opinion. If it is stubborn with a bias $P_i \in \mathbb{R}$, we have $F^v(x_i, \dot{x}_i, \alpha) = -\dot{x}_i - \kappa_i(x_i - P_i)$, where $\kappa_i$ describes the rate of convergence. The regular agent $i$ has $\kappa_i = 0$. The interaction kernel $\phi$ is a piecewise linear function (Figure 1(a)) modeling opinion influence. We consider a system consisting of 10 agents with 3 stubborn agents, $\alpha = (P_3, P_2, P_3, \kappa) = (1, 0, -1, 10)$. We choose $\mu_0 = \text{Unif}[-1, 1]^9$ and observe the system at the time interval $t \in [0, 15]$ with $M = 6, L = 4$, and $\sigma = 0.05$. As time evolves, the opinions of agents will merge into clusters. We predict the system at the interval $[0, 20]$ for two sets of initial conditions, one is from training data and the other one is randomly generated from $\mu_0$. The result is summarized in the top panel of Figure 1.

The second example is a second-order system that models the motion of $N$ self-propelled particles powered by biological or mechanical motors (Dorsogma model [14]), where $F^v(x_i, \dot{x}_i, \alpha) = (\gamma - \beta |\dot{x}_i|^{p})\dot{x}_i$ and $\phi(r) = \frac{1}{r}(-e^{-2r} + e^{-\gamma r})$. We consider a system with 10 agents in $\mathbb{R}^2$, and $\alpha = (\gamma, \beta) = (1.5, 0.5)$. We choose $\mu_0 = \text{Unif}([0, 5])^9$, and consider the system at the time interval $[0, 5]$ with $M = 3, L = 3$, and $\sigma = 0.1$. As time evolves, the true system will display the milling pattern. We predict the system at the time interval $[0, 10]$ for two set of initial conditions in the same setting as in the first example. The result is summarized in the bottom panel of Figure 1.

In the third example, we consider a real data application where we fit a system of 50 fish that swim in a shallow tank [14] into a Cucker-Smale system. We have $F^v(x_i, \dot{x}_i, \alpha) = \kappa \dot{x}_i (1 - ||\dot{x}_i||^p)$ with $\alpha = (\kappa, p)$, and the interaction force is alignment-based by replacing $(x_i - x_j)$ with $(\dot{x}_i - \dot{x}_i)$ in the Cucker-Smale system. The original position dataset consists of 201 frames and fish trajectories exhibit flocking behavior (the velocities reach the same direction) from frame 0 to frame 95. We first normalized the position data into the region $[0, 1]$ and applied smoothing by using a moving window average.
We would like to leverage the recent advancement in approximate GPs ([53–55]) to improve the computational efficiency of the current approach to deal with abundant trajectory data. Another future work is to enhance our knowledge about the interaction law using the least amount of trajectory data by looking at their marginal pairwise distance distributions. We shall also consider extensions of our learning approach to more general systems, such as heterogeneous systems with multiple types of agents and external potentials and stochastic systems.

5 Conclusion and Future work

In this work, we propose a robust learning approach based on GPs with a rigorous learning theory for data-driven discovery of interacting particle systems. Our approach enjoys flexible model selections via a training step, and has shown the potential to handle real-world systems with scarce noisy observational data well.

The present work could be extended in different venues. A well-known computational limitation of GPs is that inverting dense covariance matrices scales cubically with the number of training data. We would like to leverage the recent advancement in approximate GPs ([53–55]) to improve the computational efficiency of the current approach to deal with abundant trajectory data. Another future work is to apply the quantitative framework developed in this paper to design a data acquisition plan. The goal is to enhance our knowledge about the interaction law using the least amount of trajectory data by looking at their marginal pairwise distance distributions. We shall also consider extensions of our learning approach to more general systems, such as heterogeneous systems with multiple types of agents and external potentials and stochastic systems.
Acknowledgement

Sui Tang would like to thank Fei Lu, Mauro Maggioni and Cheng Zhang for helpful discussions. We also thank computational facilities administered by the Center for Scientific Computing (CSC) at UC Santa Barbara.

Appendix

A Learning approach

In this section, we provide more technical details for our learning approach.

Table 1: Notation for second-order systems

| Variable          | Definition                                                                 |
|-------------------|-----------------------------------------------------------------------------|
| $X \in \mathbb{R}^{dN}$ | vectorization of position vectors $(x_i)_{i=1}^N$                          |
| $V \in \mathbb{R}^{dN}$ | vectorization of velocity vectors $(v_j)_{j=1}^N = (x_i)_{i=1}^N$          |
| $Y \in \mathbb{R}^{2dN}$ | $Y = (X, V)^T$                                                              |
| $Z \in \mathbb{R}^{dN}$ | vectorization of $(m_i\dot{x}_i)_{i=1}^N$                                 |
| $r_{ij}', r_{ij}'' \in \mathbb{R}^d$ | $x_j - x_i, x'_j - x'_i$                                                   |
| $r_{ij}', r_{ij}'' \in \mathbb{R}^+$ | $\|r_{ij}'\|, \|r_{ij}''\|$                                                |
| $F_{N}^{\alpha}$ | the non-collective force with parameter $\alpha$                           |
| $f_\phi$ | energy-based interaction forced field                                       |

Lemma 1. Let $\phi$ be a Gaussian process with mean zero and covariance function $K_\phi : [0, R] \times [0, R] \to \mathbb{R}$, i.e. $\phi \sim \mathcal{GP}(0, K_\phi(r, r'))$, and $Z(t) = F_N^{\alpha}(Y(t)) + f_\phi(X(t))$ as defined in the main text (4). Then for any $t, t' \in [0, T]$, we have that,

$$
\begin{bmatrix}
Z(t) \\
Z(t')
\end{bmatrix}
\sim \mathcal{N}
\left(
\begin{bmatrix}
F_N^{\alpha}(Y(t)) \\
F_N^{\alpha}(Y(t'))
\end{bmatrix}, K_{f_\phi}(X(t), X(t'))
\right),
$$

(1)

where $K_{f_\phi}(X(t), X(t'))$ is the covariance matrix $\text{Cov}(f_\phi(X(t)), f_\phi(X(t'))) = (\text{Cov}([f_\phi(X(t))]_i, [f_\phi(X(t'))]_j))_{i,j=1}^{N,N}$ with the $(i, j)$-th block

$$
\text{Cov}([f_\phi(X)]_i, [f_\phi(X')]_j) = \frac{1}{N^2} \sum_{k\neq i,k\neq j} K_\phi(r_{ik}', r_{jk}') r_{ik} r_{jk}' r_{ik} r_{jk}'^T.
$$

(2)

Proof. Since $\phi \sim \mathcal{GP}(0, K_\phi(r, r'))$, for any $r, r' \in [0, R]$, we have that,

$$
\mathbb{E}[\phi(r)] = 0,
$$

(3)

$$
\text{Cov}[\phi(r), \phi(r')] = K_\phi(r, r').
$$

(4)

Therefore, for any collection of states $\{r_i\}_{i=1}^n \subset [0, R]$, and $\{a_i\}_{i=1}^n, \{b_i\}_{i=1}^n \subset \mathbb{R}$, the linear operator on function values $\mathcal{L}(\{\phi(r_i)\}_{i=1}^n) := (a_i \phi(r_i) + b_i)_{i=1}^n$ satisfies

$$
\mathcal{L}(\{\phi(r_i)\}_{i=1}^n) \sim \mathcal{N}([\text{vec}(b_i)_{i=1}^n], \Sigma_{\mathcal{L}(\phi)}),
$$

(5)

where $\mathcal{N}$ denotes the Gaussian distribution and the covariance matrix $\Sigma_{\mathcal{L}(\phi)} = \{a_i a_j K_\phi(r_i, r_j)\}_{i,j=1}^n \in \mathbb{R}^{n \times n}$.

Note that

$$
f_\phi(X(t)) = \sum_{i'=1}^N \frac{1}{N} \phi(\|x_{i'} - x_i\|)(x_{i'} - x_i),
$$

(6)
With the updated prior $\phi^*$. As mentioned in the main text, we can apply the gradient based method, Quasi-Newton optimizer where the mean vector $F_{\phi^*}(X(t), X(t'))$

Applying Lemma 1, we can now derive the negative log marginal likelihood for training parameters $\phi^*$. Therefore, since

This completes the proof.

Applying Lemma 1, we can now derive the negative log marginal likelihood for training parameters $\alpha, \theta$, and $\sigma$, with given observational data as demonstrated in the main text Section 2.

**Proposition 2.** Denote $Y^{(m,l)} = Y^{(m)}(t_l)$ and $Z_{\sigma^2}(m,l) = Z^{(m)}(t_l) + \epsilon^{(m,l)}$. Suppose we are given the training data set $(Y, Z) := \{ (Y^{(m,l)}, Z_{\sigma^2}^{(m,l)}) \}_{m,l=1}^{M,L}$ for $M, L \in \mathbb{N}$, such that

where $F_{\alpha^*}^u$ defined in Table 1 and i.i.d noise $\epsilon^{(m,l)} \sim \mathcal{N}(0, \sigma^2 I_{dN})$. Then the marginal likelihood of $Z$ given $Y$ and parameters $\alpha, \theta, \sigma$ satisfies

- $\log p(Z|Y, \alpha, \theta, \sigma^2) = \frac{1}{2} (Z - F_{\alpha^*}^u(Y))^T (K_{f_{\phi^*}}(X, X; \theta) + \sigma^2 I)^{-1} (Z - F_{\alpha^*}^u(Y))$ + $\frac{1}{2} \log |K_{f_{\phi^*}}(X, X; \theta) + \sigma^2 I| + \frac{dNML}{2} \log 2\pi.$

**Proof.** Using Lemma 1 since $\epsilon^{(m,l)}$ is i.i.d Gaussian noise and is independent of the initial distributions, we have that

where the mean vector $F_{\alpha^*}^u(Y) = \text{Vec}([F_{\alpha^*}^u(Y^{(m,l)})]_{m,l=1}^{M,L}) \in \mathbb{R}^{dNML}$, and the covariance matrix $K_{f_{\phi^*}}(X, X; \theta) = \text{Cov}(f_{\phi^*}(X^{(m,l)}), f_{\phi^*}(X^{(m', l)}))_{m,m',l,l'=1,1,1,1}$ can be computed component-wise by using (2). According to the properties of the Gaussian distribution, given $Y$ and parameters $\alpha, \theta, \sigma$, we have the negative log marginal likelihood function as shown in (11).

As mentioned in the main text, we can apply the gradient based method, Quasi-Newton optimizer L-BFGS [34], to minimize the negative log marginal likelihood and solve for the hyper-parameters $(\alpha, \theta, \sigma)$.

**Proposition 3.** Let $\gamma = (K_{f_{\phi^*}}(X, X; \theta) + \sigma^2 I)^{-1}(Z - F_{\alpha^*}^u(Y))$. The partial derivatives of the marginal likelihood w.r.t. the parameters $\alpha, \theta$, and $\sigma$ can be computed as follows:

With the updated prior $\phi^*$ from $\theta$, and the parameters $\alpha, \sigma$, we show the detailed derivation of our estimators for the prediction $\phi(r^*)$ at $r^* \in [0, \mathcal{R}]$. 11.
Theorem 4. Suppose we are given the training data set \( \{(Y^{(m),l}, Z^{(m),l})\}_{m,l=1} \) defined in Theorem 3 then for any \( r^* \in [0, R] \), \( \phi(r^*) \) satisfies 
\[
 p(\phi(r^*))|Y, Z, r^*) \sim \mathcal{N}(\tilde{\phi}^*, \text{Var}(\phi^*)),
\]
where 
\[
 \tilde{\phi}^* = K_{\phi, f_0}(r^*, X)(K_{f_0}(X, X) + \sigma^2 I)^{-1}(Z - F^\phi_\alpha(Y)),
\]
\[
 \text{Var}(\phi^*) = K_{\theta}(r^*, r^*) - K_{\phi, f_0}(r^*, X)(K_{f_0}(X, X) + \sigma^2 I)^{-1}K_{f_0, \phi}(X, r^*).
\]

Proof. Since \( f_0(X) \) is defined componentwisely by (6), for any \( r^* \in [0, R] \), we have 
\[
 \begin{bmatrix} f_0(X) \\ \phi(r^*) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} 0 \\ K_{f_0}(X, X) & K_{f_0, \phi}(X, r^*) \\ K_{\phi, f_0}(r^*, X) & K_{\phi, \phi}(r^*, r^*) \end{bmatrix} \right).
\]
(19)
where \( K_{f_0, \phi}(X, r^*) \) is the covariance matrix between \( f_0(X) \) and \( \phi(r^*) \) as we defined in Theorem 2 and \( K_{\phi, f_0}(X, r^*) = K_{\phi, f_0}(r^*, X)^T \) denotes the covariance matrix between \( f_0(X) \) and \( \phi(r^*) \), i.e. 
\[
 K_{f_0, \phi}(X, r^*) = (\text{Cov}(f_0(X^{(m),l}), \phi(r^*)))_{m,l=1} \text{ and the } i\text{th component of Cov}(f_0(X), \phi(r^*)) \text{ is computed by}
\]
\[
 \text{Cov}(f_0(X^{(m),l}), \phi(r^*)) = \frac{1}{N} \sum_{k \neq i} K_{\theta}(X_k, r^*)r^*_{ik}.
\]
(20)
Therefore, based on the properties of joint Gaussian distribution (see Lemma 16), conditioning on \( f_0(X) \), we have that 
\[
 p(\phi(r^*))|Y, Z, r^*) \sim \mathcal{N}(\tilde{\phi}^*, \text{var}(\phi^*)),
\]
(21)
where \( \tilde{\phi}^* \) and \( \text{Var}(\phi^*) \) are defined as (17), (18). \( \square \)

B Numerical implementation

Numerical setup. We simulate the trajectory data on the time interval \([0, T]\) with given i.i.d initial conditions generated from the probability measures specified for each system. For the training data sets, we generate \( M \) trajectories and observe each trajectory at \( L \) equidistant times \( 0 = t_1 < t_2 < \cdots < t_L = T \). We construct empirical approximation to the probability measure \( \tilde{\rho}^\mu_T \) defined in the main text equation (16), with 2000 trajectories, and let \([0, R]\) be its support. All ODE systems are evolved using ode15s in MATLAB®2020a with a relative tolerance at \( 10^{-5} \) and absolute tolerance at \( 10^{-8} \). We apply the minimize function in the GPML package \(^3\) to train the parameters using conjugate gradient optimization with the partial derivatives shown in Proposition 3 and set maximum number of function evaluations equal to 600.

Choice of covariance function. We choose the Matérn covariance function defined on \([0, R] \times [0, R]\) for the Gaussian process priors in our numerical experiments, i.e., 
\[
 K_\theta(r, r') = s_0^{2(1-\nu)/\nu} \left( \frac{\sqrt{2\nu}|r - r'|}{\omega_\phi} \right) \Gamma(\nu) B_\nu \left( \frac{\sqrt{2\nu}|r - r'|}{\omega_\phi} \right),
\]
where the parameter \( \nu > 0 \) determines the smoothness; \( \Gamma(\nu) \) is the Gamma function; \( B_\nu \) is the modified Bessel function of the second kind; the hyper-parameters \( \theta = \{s_0^2, \omega_\phi\} \) quantify the amplitude and scales.

The Reproducing Kernel Hilbert Space (RKHS), \( \mathcal{H}_{\text{Matérn}} \), associated with this Matérn kernel is norm-equivalent to the Sobolev space \( W^{\nu+1/2}_2([0, R]) \) defined by 
\[
 W^{\nu+1/2}_2([0, R]) := \left\{ f \in L^2([0, R]) : \|f\|_{W^{\nu+1/2}_2}^2 := \sum_{\beta \in \mathbb{N}_0^d, \|\beta\| \leq \nu + 1/2} \|D^\beta f\|^2_{L^2} < \infty \right\}.
\]
(23)
That is to say, \( \mathcal{H}_{\text{Matérn}} = W^{\nu+1/2}_2([0, R]) \) as a set of functions, and there exist constants \( c_1, c_2 > 0 \) such that 
\[
 c_1 \|f\|_{W^{\nu+1/2}_2} \leq \|f\|_{\mathcal{H}_{\text{Matérn}}} \leq c_2 \|f\|_{W^{\nu+1/2}_2} \quad \forall f \in \mathcal{H}_{\text{Matérn}}.
\]
(24)
In other words, \( \mathcal{H}_{\text{Matérn}} \) consists of functions that are differentiable up to order \( \nu \) and weak differentiable up to order \( s = \nu + 1/2 \).

\(^3\)Carl Edward Rasmussen & Hannes Nickisch \( \text{http://gaussianprocess.org/gpml/code} \)
We are interested in learning the parameters $\alpha$. We consider the Taylor model [52], which models the collective dynamics of continuous opinion with noisy training data. A set of $n = 10$ independent learning trials and report the estimation errors in the form of (mean value) ± (standard deviation). We use the tables and figures to summarize the results. Specifically,

- **Scalar parameters estimation.** The unknown scalar parameters in $\alpha$ can be both linear (Sec B.2) or nonlinear (Sec B.1) with respect to the non-collective force functions. We report the estimated parameters $\hat{\alpha}$ for $F^\nu$, given by (11). We also report the estimated noise $\hat{\sigma}$ in the cases with noisy observation data.

- **Interaction kernel estimation.** We approximate the interaction kernel $\phi \in \mathcal{H}_{\text{Matérn}}$ (Sec B.2) or $\phi \notin \mathcal{H}_{\text{Matérn}}$ (Sec 5.1) by the posterior mean obtain by (17). We compare the true and estimated interaction kernels in relative $L^\infty((0, \hat{R})$-norm (reconstruction error) and relative $L^2(p_0, \hat{T})$-norm (residual error). We plot the posterior mean estimator with an uncertainty band given by twice the marginal posterior standard deviation, over the region larger than $[0, \hat{R}]$, to illustrate the extrapolation properties of our estimators.

- **Trajectory prediction.** We compare the discrepancy between the true trajectories (evolved using $\alpha, \phi$) and predicted trajectories (evolved using $\hat{\alpha}, \hat{\phi}$) on both the training time interval $[0, \hat{T}]$ and on the future time interval $[\hat{T}, T_f]$, over two different sets of initial conditions – one taken from the training data, and one consisting of new samples from the same initial distribution. We used the max-in-time trajectory error over time interval $[T_0, T_1]$ that is defined as

$$
\|Y - \hat{Y}\|_{T \in [T_0, T_1]} = \sup_{t \in [T_0, T_1]} \|Y(t) - \hat{Y}(t)\|.
$$

**B.1 Opinion dynamics with stubborn agents**

We consider the Taylor model [52], which models the collective dynamics of continuous opinion exchange in the presence of stubborn agents. It is a first-order system of $N$ interacting agents, each agent $i$ is characterized by a continuous opinion variable $x_i \in \mathbb{R}$. The dynamics of opinion exchange are governed by the following first-order equation,

$$
\dot{x}_i = F^\phi(x_i, \alpha) + \sum_{i'=1}^N \frac{1}{N} \phi(||x_{i'} - x_i||)(x_{i'} - x_i),
$$

where

$$
\phi(r) = \begin{cases}
2.5r & \text{if } 0 \leq r < 0.4 \\
1 & \text{if } 0.4 \leq r < 0.6 \\
2.5 - 2.5r & \text{if } 0.6 \leq r < 1 \\
0 & \text{if } r \geq 1
\end{cases}
$$

and

$$
F^\phi(x, \alpha) = \begin{cases}
-\kappa(x_i - P_i) & \text{if agent } i \text{ is stubborn with bias } P_i \\
0 & \text{otherwise}
\end{cases}
$$

The interaction kernel $\phi$ encodes the non-repulsive interactions between agents: all agents aim to align their opinions to their connected neighbors according to distance-based attractive influences. The non-collective force $F^\phi(x, \alpha)$ describes the additional influence induced by the stubbornness: the stubborn agents have strong desires to follow his/her bias $P_i$, and $\kappa$ controls the rate of convergence towards his/her bias. The stubborn agents may cause a major effect on the collective opinion formation process. If $\kappa = 0$, then stubborn agents do not follow their biases and behave as regular agents.

We are interested in learning the parameters $\alpha = (P_1, P_2, P_3, \kappa)$ and interaction kernel $\phi$ from trajectory data. Note that this first-order system is a special case of the second-order system (defined in the main text equation (1)) as we discussed in previous sections with $m_i = 0$ for all $i$, and $F^\nu(x_i, \dot{x}_i, \alpha) = -\dot{x}_i + F^\phi(x_i, \alpha)$. Therefore, we can apply our strategies proposed in the main text Section 2 to learn the parameters $\alpha$ in $F^\phi$, the interaction kernel $\phi(r)$, and predict the trajectories with noisy training data.

The training data $(X, \mathcal{Y})$ is generated with parameters shown in Table 2 and the observations are made at the time interval $[0, 15]$ with $M = 6, L = 4$, and noise level $\sigma = 0.05$. 

Table 2: System parameters in the opinion dynamics

| d   | N   | \([0; T; T_f]\) | \(\alpha = (P_1, P_2, P_3, \kappa)\) | \(\mu_0\) |
|-----|-----|----------------|--------------------------------------|-----------|
| 1   | 10  | \([0, 15, 20]\) | \((1, 0, -1, 10)\)                  | \(\text{Unif}([-1, 1])\) |

We initialize the parameters \((\sigma, P_1, P_2, P_3, \kappa) = (1/2, 1/2, 1/2, 1/2, 1/2)\). Table 3 shows the returned estimations for \(\alpha\) and the errors of resulting estimations for \(\phi(r)\) in 10 independent trails of experiments. The results demonstrate that the algorithm produce accurate estimation of the parameters \((P_1, P_2, P_3, \kappa)\) from the noisy training data. For the estimation of \(\phi(r)\), even though \(\phi\) is not in the RKHS generated by the Matérn kernel, our algorithm still provides us with faithful prediction in the region where the training data covers (see Figure 1(a) in the main text). At the region around \(r = 0\), we see the approximation is not as good as other regions. We impute this phenomenon to the fact that \(\phi(r)\) is weighted by \(r^2\) in the model (26), thus we lose information of \(\phi\) when \(r\) is close to zero. However, we expect that our estimators will produce accurate trajectories since they are generated by \(\phi(r)r^2\), and Figure 1(b) in the main text convinced our intuition.

Table 3: Means and standard deviations of estimations for parameters \(\sigma\) and \(\alpha\), and predicted errors for \(\phi\) on \([0, 1.99]\) in opinion dynamics with noisy data when \(M = 6, L = 4, \sigma = 0.05\)

| \(\sigma\) | \(P_1\) | \(P_2\) | \(P_3\) | \(\kappa\) | Relative \(L^2\) error | Relative \(L^2(\phi)\) error |
|-----------|---------|---------|---------|-----------|------------------------|-----------------------------|
| \(0.0482 \pm 2.8 \cdot 10^{-4}\) | \(1.0009 \pm 1.4 \cdot 10^{-4}\) | \(-0.0004 \pm 6.9 \cdot 10^{-4}\) | \(-1.0002 \pm 3.2 \cdot 10^{-4}\) | \(9.9928 \pm 2.2 \cdot 10^{-4}\) | \(1.9 \cdot 10^{-4} \pm 1.3 \cdot 10^{-4}\) | \(1.9 \cdot 10^{-4} \pm 1.4 \cdot 10^{-4}\) |

Table 4: Means and standard deviations of predicted errors for the trajectories \(X\) in opinion dynamics with \(\hat{\alpha}, \hat{\beta}\) and \(\hat{\phi}\) estimated from noisy data when \(M = 6, L = 4, \sigma = 0.05\)

| Training \([0, 15]\) | Testing \([0, 15]\) | Training \([15, 20]\) | Testing \([15, 20]\) |
|---------------------|---------------------|---------------------|---------------------|
| \(1.1 \cdot 10^{-1} \pm 4.2 \cdot 10^{-2}\) | \(1.2 \cdot 10^{-1} \pm 5.9 \cdot 10^{-2}\) | \(8.0 \cdot 10^{-2} \pm 6.4 \cdot 10^{-2}\) | \(6.9 \cdot 10^{-2} \pm 4.1 \cdot 10^{-2}\) |

### B.2 Fish-Milling dynamics with friction force

We consider the Dorsogna model [13] which describes the motion of \(N\) self-propelled particles powered by biological or mechanical motors under the frictional forces: for \(i = 1, \cdots, N\)

\[
m_i \ddot{x}_i = F^u(x_i, \dot{x}_i, \alpha) + \sum_{i' \neq i} \frac{1}{N} \phi(\|x_{i'} - x_i\|)(x_{i'} - x_i) \tag{29}
\]

\[
F^u(x_i, \dot{x}_i, \alpha) = (\gamma - \beta|\dot{x}_i|^2)\dot{x}_i. \tag{30}
\]

The form of (29) is derived using Newton’s law: the right hand side of (29) describing three forces acting on each agent: self-propulsion with strength \(\gamma\), nonlinear drag with strength \(\beta\), and social interactions by \(\phi\). This system can produce a rich variety of collective patterns: in our numerical example, we consider the interaction kernel that is derived from the Morse-type potential

\[
\phi(r) = \frac{1}{r} \left[ -\frac{C_{rp}}{l_{rp}} e^{-\frac{r}{l_{rp}}} + \frac{C_a}{l_a} e^{-\frac{r}{l_a}} \right] \tag{31}
\]

where \(l_a, l_{rp}\) represent the attractive and repulsive potential ranges; \(C_a, C_{rp}\) represent the respective amplitudes. Since this kernel is singular at \(r = 0\), we truncate it at \(r_t = 0.05\) with a function of the form \(ae^{-br}\) to ensure that the new function has a continuous derivative. When we have velocity independent forces, i.e. \(\gamma = \beta = 0\), the system (29) forms a typical Hamiltonian system with conserved energy. We assume that we do not have the knowledge of the parametric form of \(\phi, \gamma\) and \(\beta\). Our goal is to learn them from the trajectory data.

The training data \((Y, Z)\) is generated with parameters shown in Table 5 (we make correction of the initial conditions announced in the main text), and the observations are made at the time interval \([0, 5]\) with \(M = 3, L = 3\), and noise level \(\sigma = 0.1\).
We use the Cucker-Smale system [5] to model the flocking behavior:

\[ \dot{x}_i = F^v(x_i, \dot{x}_i, \alpha) + \frac{1}{N} \sum_{i' = 1}^{N} \left[ \phi(\|x_{i'} - x_i\|)(\dot{x}_{i'} - \dot{x}_i) \right], \quad \text{for } i = 1, \ldots, N, \]  

(32)

where \( \phi \) is a communication kernel, or an influence function, that makes the agents flock, and \( F^v \) measures the effects of additional attraction/repulsion forces. We shall consider a Rayleigh-type friction force given by

\[ F^v(x_i, \dot{x}_i, \alpha) = \kappa |\dot{x}_i| (1 - |\dot{x}_i|^p), \]  

(33)

We initialize the parameters \((\sigma, \gamma, \beta) = (1, 1, 1)\). Table 6 summarizes the returned estimations for \( \alpha \) and relative estimation errors for \( \phi \). In this model, \( \phi \) is in the RKHS generated by the Matérn kernel we pick. Although the errors are slightly larger near \( r = 0 \) due to the same reason as we mentioned in previous case, the true interaction kernel \( \phi(r) \) is still fully covered in the uncertainty region we constructed using the posterior variances (see Figure 1(d) in the main text).

Table 6: Means and standard deviations of estimations for parameters \( \sigma \) and \( \alpha \), and predicted errors for \( \phi \) on \([0, 1.76]\) in fishing milling dynamics with noisy data when \( M = 3, L = 3, \sigma = 0.1 \)

| Estimated parameters | Errors in \( \phi \) |
|----------------------|---------------------|
| \( \hat{\alpha} \)    | \( \hat{\gamma} \)   | \( \beta \)                        |
| 0.9991 ± 5.6 \cdot 10^{-3} | 1.5057 ± 2.7 \cdot 10^{-2} | 0.5036 ± 1.1 \cdot 10^{-2} |
| Relative \( L^\infty \) error | Relative \( L^2(\hat{\mu}_f) \) error |
| 6.5 \cdot 10^{-2} ± 3.2 \cdot 10^{-2} | 3.0 \cdot 10^{-2} ± 1.7 \cdot 10^{-2} |

The errors of predictive trajectories with estimated \( \phi \) and \( \alpha \) are shown in Table 7. Even if the trajectory prediction errors can go up to \( O(10^{-1}) \) with the presence of a relatively large noise, our estimators provided faithful predictions to most of the agents in the system, and the milling pattern as shown in Figure 1(e) in the main text.

Table 7: Means and standard deviations of predicted errors for the trajectories \( \mathcal{X} \) in fishing milling dynamics with \( \hat{\gamma}, \hat{\beta} \) and \( \hat{\phi} \) estimated from noisy data when \( M = 3, L = 3, \sigma = 0.1 \)

|                  | Training [0, 5] | Testing [0, 5] | Training [5, 10] | Testing [5, 10] |
|------------------|----------------|----------------|------------------|----------------|
| \( \|v_i - \bar{v}_c\| \approx 0 \) for all \( i \) and some common velocity \( \bar{v}_c \). |

\[ 2.6 \cdot 10^{-1} \pm 1.0 \cdot 10^{-1} \quad 2.2 \cdot 10^{-1} \pm 1.1 \cdot 10^{-1} \quad 7.0 \cdot 10^{-1} \pm 3.7 \cdot 10^{-1} \quad 5.8 \cdot 10^{-1} \pm 2.4 \cdot 10^{-1} \]

B.3 Real Data Application

In this example, we use the real dataset of swimming fish by Couzin et al that is available at ScholarsArchive of Oregon State University[4]. The experimental arena consisted of a 2.1 × 1.2 m (7 × 4 ft) white shallow tank surrounded by a floor-to-ceiling white curtain. Water depth was chosen to be 4.5-5cm so the schools would be approximately 2D. We choose the data from frame 2201 to frame 2296 that consists of 50 fish. We relabel them as frame 0 to frame 95 and refer more details of the dataset to the supplementary information of [4].

In this dataset, the fish will eventually follow approximately the same direction as time evolves. We preprocess the position data as described in the main text. In this case, the magnitude of velocity data of fish are relatively small. The velocities can be thought of the same, as long as their normalized direction vectors are very close. So the fish can be thought of exhibiting the flocking behavior (i.e. \( |v_i - v_c| \approx 0 \) for all \( i \) and some common velocity \( v_c \)).

We use the Cucker-Smale system [5] to model the flocking behavior:

\[ \dot{x}_i = F^v(x_i, \dot{x}_i, \alpha) + \frac{1}{N} \sum_{i' = 1}^{N} \left[ \phi(\|x_{i'} - x_i\|)(\dot{x}_{i'} - \dot{x}_i) \right], \quad \text{for } i = 1, \ldots, N, \]  

(32)

where \( \phi \) is a communication kernel, or an influence function, that makes the agents flock, and \( F^v \) measures the effects of additional attraction/repulsion forces. We shall consider a Rayleigh-type friction force given by

\[ F^v(x_i, \dot{x}_i, \alpha) = \kappa |\dot{x}_i| (1 - |\dot{x}_i|^p), \]  

(33)

---

[4] Katz, Yael, Kolbjörn Tunstrom, Christos C Ioannou, Cristian Huepe, and Iain D Couzin. 2021. Fish Schooling Data Subset. Oregon State University.
with parameter $\alpha = (\kappa, p)$, $\kappa, p > 0$ and $\kappa$ is a strength parameter. This non-collective force pushes all magnitudes of the velocities $|v_i|$ towards the same value, and counteracts the directional alignment forces governed by $\phi$. Note that (32) is slightly different from the system (1), we modify our learning approach to incorporate this change.

The training data consists of frame 0 and frame 29. The initialization of hyperparameters for $\theta, \sigma$ and $\alpha$ are $(1, 1), 0.001, (1, 1)$. In the training procedure, we set the length of runs in the minimizer solver to be 100. After we obtained the estimators, we run the learned dynamical system (32) at the time interval $[0, 20]$ with the frame 0 as the initial condition. We find that the simulated position data at $t = 19$ matched the position data at frame 95 very well. We then compare the original position data set with the simulated ones at $t = 0 : 0.2 : 19$.

To measure the performance, we compute the flocking direction vector $v_i$ at time $t$ by solving

$$
\max_{\|w_i\|=1} \frac{1}{N} \sum_{i=1}^{N} |\langle v_i(t), \frac{v_i(t) - v_i(\tau)}{\|v_i(\tau)\|} \rangle|^2
$$

and introduce the flocking score at time $t$ by computing $E_t = \frac{1}{N} \sum_{i=1}^{N} \langle v_i, \frac{v_i(t)}{\|v_i(t)\|} \rangle$. If $E_t = 1$, then all the velocities reach the same direction. We then take the average of the flocking scores at all times and denote it by $E$.

| Table 8: Estimated hyperparameters using the real data set and trajectory prediction error |
|----------------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| estimated parameters                  | estimated parameters for $\phi$ | trajectory prediction | The average flocking scores |
| $\hat{n}$                             | $\hat{\kappa}$   | $\hat{p}$       | $\varepsilon_0$   | $\omega_0$      | $\varepsilon$    |
| $\hat{r}$                             | $\hat{\kappa}$   | $\hat{p}$       | $\varepsilon_0$   | $\omega_0$      | $\varepsilon$    |
| $2.0 \cdot 10^{-3}$                   | $2.7 \cdot 10^{-2}$ | $8.0 \cdot 10^{-1}$ | $1.0$             | $1.4 \cdot 10$  | $7.0 \cdot 10^{-2}$ |
| $2.0 \cdot 10^{-3}$                   | $2.7 \cdot 10^{-2}$ | $8.0 \cdot 10^{-1}$ | $1.0$             | $1.4 \cdot 10$  | $7.0 \cdot 10^{-2}$ |

We compare the above learning results with the ones without the training step. That is, we fit the real data into the system with the initialized hyper-parameters. We summarize the results in Figure 1. Compare with Figure 2 in the main text, we see that the training procedure of our learning approach is very effective in improving the prediction.

![Figure 1: Fitting into a Cucker-Smale system (dim=100). (a): predictive mean $\phi$, and two-standard-deviation band (light blue color) around the mean. The grey bars represent the empirical density of the $\rho^f_z$. (b): the real trajectory (left) versus prediction (right) with frame 0 as the initial condition. (c): the total relative error in position for frame 0-95 and comparison of unit norm flocking direction vector using the inner product.](image)

C Technical details for the learning theory

C.1 Additional notations and terminologies

Let $\mathcal{H}$ be a Hilbert space. We denote by $\mathcal{B}(\mathcal{H})$ the set of bounded linear operators mapping $\mathcal{H}$ to itself. We use $\langle \cdot, \cdot \rangle_\mathcal{H}$ to denote its inner product, and still use $\langle \cdot, \cdot \rangle$ to denote the inner product on the Euclidean space. For $d, N, M, L \in \mathbb{N}^+$, let $w = (w_{m,l,i})_{m,l,i=1}^{M,L,N}$, $z = (z_{m,l,i})_{m,l,i=1}^{M,L,N} \in \mathbb{R}^{dNML}$

We add the constraint $\langle v_i, \frac{v_i(t)}{\|v_i(t)\|} \rangle > 0$ to pick the right sign of the vector.
with \(w_{m,l,i}, z_{m,l,i} \in \mathbb{R}^d\), we define
\[
\langle w, z \rangle = \frac{1}{MLN} \sum_{m,l,i=1}^{M,L,N} \langle w_{m,l,i}, z_{m,l,i} \rangle
\]  
(34)
where \(\langle w_{m,l,i}, z_{m,l,i} \rangle\) is the canonical inner product on \(\mathbb{R}^d\).

We use \(|A|_H\) to denote its operator norm if \(A \in \mathcal{B}(H)\). If \(A\) is a Hilbert Schmidt operator, then \(|A|_{HS}\) denotes its Hilbert–Schmidt norm that satisfies \(|A|_{HS}^2 = \text{Tr}(A^*A)\). \(A\) is said to be in the trace class if \(\text{Tr}(|A|) < \infty\) for \(|A| = \sqrt{A^*A}\).

**C.2 Operator-theoretical framework for the statistical inverse problem**

In the paper, we make the following assumption.

**Assumption 5.** For \(i, i' \in N\), we have \(r_{ii'} \in L^2(\mathbb{R}^{dN}; \rho_X; \mathbb{R}^d)\).

Assumption 5 is mild. It holds as long as the distribution of the initial conditions is compactly supported or decays sufficiently fast.

**Proposition 6.** Let \(A\) be a linear operator defined by \(A \varphi = f \varphi\), that maps \(\mathcal{H}_K\) to \(L^2(\mathbb{R}^{dN}; \rho_X; \mathbb{R}^{dN})\). Then \(A\) is bounded and its adjoint operator \(A^*\) satisfies
\[
A^*g = \int_X \frac{1}{N^2} \sum_{i=1}^{N} K_{r_{i'i'}} \langle r_{ii'}, g_i(X) \rangle d\rho_X,
\]  
(35)
where \(g = [g_1^T, \ldots, g_N^T]^T\) with \(g_i : \mathbb{R}^{dN} \to \mathbb{R}^d\). As a consequence,
\[
B \varphi := A^*A \varphi = \frac{1}{N^3} \int_X \sum_{i,i',i''} K_{r_{i'i'}} \langle \varphi, K_{r_{ii''}} \rangle \langle \varphi, r_{ii''} \rangle d\rho_X
\]  
(36)
is a trace class operator mapping \(\mathcal{H}_K\) to \(\mathcal{H}_K\). In addition, \(B\) can be also viewed as a bounded linear operator from \(L^2(\tilde{\rho}_X^T)\) to \(L^2(\tilde{\rho}_X^T)\).

**Proof.** Since \(\mathcal{H}_K\) can be naturally embedded as a subspace of \(L^2(\tilde{\rho}_X^T)\), using Lemma 17 and Lemma 18, we have that
\[
\|A \varphi\|^2_{L^2(\rho_X)} = \|f \varphi\|^2_{L^2(\tilde{\rho}_X^T)} \leq \frac{N}{N - 1} \|\varphi\|^2_{L^2(\tilde{\rho}_X^T)} < R^2 \|\varphi\|^2_{L^2(\tilde{\rho}_X^T)} \leq \kappa^2 R^2 \|\varphi\|^2_{L^2(\mathcal{H}_K)}.
\]  
(37)
This shows that \(A\) is a bounded linear operator mapping \(\mathcal{H}_K\) to \(L^2(\mathbb{R}^{dN}; \rho_X; \mathbb{R}^{dN})\).

Next, we prove (35). We first show that the map for each \((i, i')\)
\[
X \to K_{r_{ii'}} \in \mathcal{H}_K
\]
is continuous since \(\|K_{r_{ii'}} - K_{r_{i'i'}}\|^2_{\mathcal{H}} = K(r_{ii'}, r_{ii'}) + K(r_{i'i'}, r_{i'i'}) - 2K(r_{ii'}, r_{i'i'})\) for all \(r_{ii'} = \|x_i - x_{i'}\|, \|r_{ii'} = \|x_i' - x_{i'}\|\), and \(X, X' \in \mathbb{R}^{dN}\), and both \(K\) and \(\|\cdot\|\) are continuous. Hence given a function \(g \in L^2(\mathbb{R}^{dN}; \rho_X; \mathbb{R}^{dN})\), the map
\[
X \to \frac{1}{N^2} \sum_{i=1}^{N} K_{r_{ii'}} \langle r_{ii'}, g_i(X) \rangle
\]
is measurable from \(\mathbb{R}^{dN}\) to \(\mathcal{H}_K\). Moreover,
\[
\|\frac{1}{N^2} \sum_{i=1}^{N} K_{r_{ii'}} \langle r_{ii'}, g_i(X) \rangle\|_{\mathcal{H}_K} \leq \frac{K}{N^3} \sum_{i=1}^{N} \|\langle r_{ii'}, g_i(X) \rangle\|.
\]
By Assumption 3 we have both \(r_{ii'}, g_i(X) \in L^2(\mathbb{R}^{dN}; \rho_X; \mathbb{R}^d)\). By Hölder inequality, \(\langle r_{ii'}, g_i(X) \rangle\) is in \(L^1(\mathbb{R}^{dN}; \rho_X; \mathbb{R})\), and hence \(\frac{1}{N^3} \sum_{i=1}^{N} \|K_{r_{ii'}}(r_{ii'}, g_i(X))\|\) is integrable as a vector-valued map.
Finally, for all $\psi \in \mathcal{H}_K$,
\[
\langle A\psi, g \rangle_{L^2(\rho_X)} = \frac{1}{N} \sum_{i=1}^N \int_X \langle [f(X)]_i, g_i(X) \rangle \, d\rho_X
\]
\[
= \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N \int_X \psi(r_{ii'}) \langle r_{ii'}, g_i(X) \rangle \, d\rho_X
\]
\[
= \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N \int_X \langle \psi, K_{r_{ii'}} \rangle_{\mathcal{H}_K} \langle r_{ii'}, g_i(X) \rangle \, d\rho_X
\]
\[
= \langle \psi, \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N K_{r_{ii'}} \langle r_{ii'}, g_i(X) \rangle \, d\rho_X \rangle_{\mathcal{H}_K} = \langle \psi, A^*g \rangle_{\mathcal{H}_K},
\]
so by uniqueness of the integral, (35) holds. Equation (36) is a consequence of (35) and the fact that the integral commutes with the scalar product.

We now prove that $B$ is a trace class operator. Let $(e_n)_{n \in \mathbb{N}}$ be an orthonormal basis of $\mathcal{H}_K$. Since $B = A^* A$ is a positive operator,
\[
\sum_n \langle \sqrt{B}e_n, e_n \rangle_{\mathcal{H}_K} \leq \text{Tr}(A^* A) = \sum_n \|e_n\|_{L^2(\tilde{\rho}_F)}^2 \leq \sum_n \|e_n\|_{L^2(\tilde{\rho}_F)}^2 \leq R^2 \sum_n \|e_n\|_{L^2(\tilde{\rho}_F)}^2 \leq R^2 \int \langle K_r, K_r \rangle_{\mathcal{H}_K} \, d\rho_F(r) \leq \kappa^2 R^2.
\]
Lastly, if $\varphi \in L^2(\tilde{\rho}_F)$, using the fact that $\|f_{K_\varphi}(X)\| \leq \sqrt{N}R \|K\|_\infty \leq \sqrt{N}R\kappa^2$ and Lemma [17]
one can show that
\[
\|B\varphi\|_{L^2(\tilde{\rho}_F)} \leq R\|B\varphi\|_\infty \leq \frac{N-1}{N} \kappa^2 R^2 \|\varphi\|_{L^2(\tilde{\rho}_F)}.
\]
As a result, $B\varphi \in L^2(\tilde{\rho}_F)$, and $B$ can be viewed as a bounded linear operator from $L^2(\tilde{\rho}_F)$ to $L^2(\tilde{\rho}_F)$ with $\|B\|_{L^2(\tilde{\rho}_F)} \leq \kappa^2 R^2$.

**Proof of Proposition 3 in the main text**. Recall that $J_{\tilde{\rho}_F}^*: \mathcal{H}_K \to L^2([0, R]; \tilde{\rho}_F; \mathbb{R})$ is the canonical embedding map, and its adjoint operator $J_{\tilde{\rho}_F}: L^2([0, R]; \tilde{\rho}_F; \mathbb{R}) \to \mathcal{H}_K$ is an integral operator with respect to the kernel $K$, i.e., for $\varphi \in L^2([0, R]; \tilde{\rho}_F; \mathbb{R})$ and $r \in [0, R]$,
\[
(J_{\tilde{\rho}_F}^*\varphi)(r) = \int_0^R K(r, r') \varphi(r') \, d\rho_F(r').
\]
Then $J_{\tilde{\rho}_F}^*: \mathcal{H}_K \to \mathcal{H}_K$ is a compact, and positive operator. Note that
\[
\langle J_{\tilde{\rho}_F}^* J_{\tilde{\rho}_F}^* \varphi, \varphi \rangle_{\mathcal{H}_K} = \|\varphi\|_{L^2(\rho_X)}^2 \quad \text{and} \quad \langle A^* A \varphi, \varphi \rangle_{\mathcal{H}_K} = \|A\varphi\|_{L^2(\rho_X)}^2,
\]
the coercivity condition (defined in the main text equation (17) ) together with Proposition[6] imply that
\[
c_{\mathcal{H}_K} J_{\tilde{\rho}_F}^* J_{\tilde{\rho}_F} \leq A^* A \leq \kappa^2 R^2 J_{\tilde{\rho}_F}^* J_{\tilde{\rho}_F}.
\]

**Theorem 7** (Courant–Fischer–Weyl min-max principle, see [56]). Let $U$ be a compact, self-adjoint, positive operator on a Hilbert space $\mathcal{H}$, whose eigenvalues are listed in decreasing order $\lambda_1 \geq \lambda_2 \cdots$. Letting $S_k \subset \mathcal{H}$ be a $k$ dimensional subspace. Then:
\[
\max_{S_k} \min_{x \in S_k, \|x\| = 1} \langle Ux, x \rangle = \lambda_k^k,
\]
\[
\min_{S_{k-1}} \max_{x \in S_{k-1}, \|x\| = 1} \langle Ux, x \rangle = \lambda_k^k.
\]
Therefore (40) indicates that $A^* A$ and $J_{\tilde{\rho}_F}^* J_{\tilde{\rho}_F}$ have the same eigenvalue decay rate.
Operator representations for minimizers. When the trajectory data is infinite \((M \to \infty)\), the expected risk functional of \(\mathcal{E}^{\lambda,M}(\cdot)\) is
\[
\mathcal{E}^{\lambda,\infty}(\varphi) := \mathbb{E}\left[\frac{1}{LM} \sum_{l=1, m=1}^{L, M} \|f_{\varphi}(X^{(m,l)}) - V_{\sigma^2}^{m,l}\|^2\right] + \lambda \|\varphi\|^2_{\mathcal{H}K} \tag{43}
\]
where the expectation is taken with respect to the joint distribution of \(\mu\), and
\[
\mathcal{E}^{\lambda,\infty}(\varphi) = \|A_\varphi - A\phi\|^2_{L^2(\rho_X)} + \lambda \|\varphi\|^2_{\mathcal{H}K}, \tag{44}
\]
where \(\lambda > 0\), and it is given by
\[
\phi_{\mathcal{H}K}^{\lambda,\infty} := (B + \lambda)^{-1} A^*_\phi \in \mathbb{H}K. \tag{45}
\]
Proposition 8. Consider the expected risk \(\mathcal{E}^{\lambda,\infty}(\cdot)\) in (43) with a possible regularization term determined by \(\lambda \geq 0\). We solve the minimization problem
\[
\arg\min_{\varphi \in \mathbb{H}K} \mathcal{E}^{\lambda,\infty}(\varphi).
\]
- Case \(\lambda = 0\). Then its minimizer \(\phi_{\mathcal{H}K}^{0,\infty}\) always exists and satisfies
  \[
  B\phi_{\mathcal{H}K}^{0,\infty} = A^*f_\phi.
  \]
- Case \(\lambda > 0\). Then a unique minimizer exists and it is given by
  \[
  \phi_{\mathcal{H}K}^{\lambda,\infty} := (B + \lambda)^{-1} A^*f_\phi.
  \]
Corollary 9. For any \(\varphi \in \mathbb{H}K\), we have that
\[
\mathcal{E}^{0,\infty}(\varphi) - \mathcal{E}^{0,\infty}(\phi_{\mathcal{H}K}^{0,\infty}) = \|A_\varphi - A\phi_{\mathcal{H}K}^{0,\infty}\|^2_{L^2(\rho_X)} = \|\sqrt{B}(\varphi - \phi_{\mathcal{H}K}^{0,\infty})\|^2_{\mathcal{H}K}.
\]
Remark 1. In the context of learning theory, \(\mathcal{E}^{0,\infty}(\varphi) - \mathcal{E}^{0,\infty}(\phi_{\mathcal{H}K}^{0,\infty})\) is called the residual error \([27]\). Assuming the coercivity condition (defined in the main text equation (17)), we have that \(\phi_{\mathcal{H}K}^{0,\infty} = \phi\) and that \(\mathcal{E}^{0,\infty}(\varphi) - \mathcal{E}^{0,\infty}(\phi_{\mathcal{H}K}^{0,\infty})\) is equivalent to \(\|\varphi - \phi\|_{L^2(\rho_X)}\) by (40).

Now we consider the empirical setting.
Proposition 10. Given the empirical noisy trajectory data \(\mathbb{V}_{\sigma^2,M} = \{X_M, \mathbb{V}_{\sigma^2,M}\}\) as in the main text Section 3. We define the sampling operator \(A_\mathcal{M} : \mathbb{H}K \to \mathbb{R}^{dNML}\) by
\[
A_\mathcal{M}\varphi = f_\varphi(X_M) := \text{Vec}(\{f_\varphi(X^{m,l})\}_{m=1, l=1}^{M, L}), \tag{46}
\]
where \(\mathbb{R}^{dNML}\) is equipped with the inner product defined in (34).
1. The adjoint operator \(A_\mathcal{M}^*\) is a finite rank operator. For any \(\mathbb{W}\) in \(\mathbb{R}^{dNML}\), let \(\mathbb{W}_{m,l,i} \in \mathbb{R}^d\) denote the \(i\)-th component of the \((m, l)\)th block of \(\mathbb{W}\) as in the main text Section 3, then we have
\[
A_\mathcal{M}^*\mathbb{W} = \frac{1}{LM} \sum_{l=1, m=1}^{L, M} \sum_{i=1, i' \neq i}^{N} \frac{1}{N^2} K_{l, l'}^{(m, l)} \langle r_{i' i}^{(m, l)}, \mathbb{W}_{m,l,i} \rangle, \tag{47}
\]
For any function \(\varphi \in \mathbb{H}K\), we have that
\[
B_\mathcal{M}\varphi := A_\mathcal{M}^*A_\mathcal{M}\varphi = \frac{1}{LM} \sum_{l=1, m=1}^{L, M} \sum_{i=1, i' \neq i}^{N} \frac{1}{N^3} K_{l, l'}^{(m, l)} \langle \varphi, K_{l, l'}^{(m, l)} \rangle \mathbb{H}K \langle r_{i' i}^{(m, l)}, r_{i' i'}^{(m, l)} \rangle. \tag{48}
\]
2. If \(\lambda > 0\), there is a unique minimizer \(\phi_{\mathcal{H}K}^{\lambda,M}\) that solves
\[
\arg\min_{\varphi \in \mathbb{H}K} \mathcal{E}^{\lambda,M}(\varphi)
\]
and it is given by
\[
\phi_{\mathcal{H}K}^{\lambda,M} = (B_\mathcal{M} + \lambda)^{-1} A_\mathcal{M}^*\mathbb{V}_{\sigma^2,M}. \tag{49}
\]
Moreover, we have that
\[ E^{\lambda,M}(\varphi) = \| A_M \varphi - \nabla_{\sigma^2,M} \|_2^2 + \lambda \| \varphi \|_H^2 \]
and solving its normal equation.

**Theorem 11** (Representer theorem). If \( \lambda > 0 \), the minimizer of the regularized empirical risk functional \( E^{\lambda,M}(\cdot) \) (defined in the main text equation (19)) has the form
\[ \phi_{H,K}^{\lambda,M} = \sum_{r \in r_K} \hat{c}_r K_r, \tag{47} \]
where \( r_K \subseteq \mathbb{R}^{MLN^2} \) is the set contains all the pair distances in \( \mathcal{X}_M \), i.e.
\[ r_K = \left[ \mathbf{r}_{11}^{(1,1)}, \mathbf{r}_{1N}^{(1,1)}, \ldots, \mathbf{r}_{NN}^{(1,1)}, \mathbf{r}_{11}^{(M,L)}, \mathbf{r}_{1N}^{(M,L)}, \ldots, \mathbf{r}_{NN}^{(M,L)} \right]^T. \tag{48} \]
Moreover, we have that
\[ \hat{c} = \frac{1}{N} \mathbf{r}_K^T \cdot (K_{\mathbf{f}_\lambda}(\mathcal{X}_M, \mathcal{X}_M) + \lambda NMLI)^{-1} \nabla_{\sigma^2,M}, \tag{49} \]
where the block-diagonal matrix \( \mathbf{r}_K \in \mathbb{R}^{MLdN \times MLN^2} \) and \( \mathbf{r}_{m,l} \in \mathbb{R}^{dN \times N^2} \) defined by
\[ \mathbf{r}_{m,l} = \begin{bmatrix} r_{11}^{(m,l)} & \cdots & r_{N1}^{(m,l)} \\ 0 & r_{21}^{(m,l)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{NN}^{(m,l)} \end{bmatrix}. \tag{50} \]

**Proof.** Let \( H_{K,M} \) be the subspace of \( H_K \) spanned by the set of functions \( \{ K_r : r \in r_K \} \). By Proposition 10, we know that \( B_M(H_{K,M}) \subseteq H_{K,M} \). Since \( B_M \) is self-adjoint and compact, by spectral theory of self-adjoint compact operator (see [57]), \( H_{K,M} \) is also an invariant subspace for the operator \( (B_M + \lambda I)^{-1} \). Then by (46), there exists a vector \( \hat{c} \) such that
\[ \phi_{H,K}^{\lambda,M} = \sum_{r \in r_K} \hat{c}_r K_r. \tag{51} \]

Then, multiplying \( (B_M + \lambda I) \) on both sides of (46) and plugging in (51), we can obtain
\[ (\mathbf{r}_K^T \mathbf{r}_K + \lambda N^3 MLI) \hat{c} = N\mathbf{r}_K^T \nabla_{\sigma^2,M}, \tag{52} \]
where we use the matrix representation of \( (B_M + \lambda I) \) with respect to the spanning set \( \{ K_r : r \in r_K \} \).

Recall that we have \( K(r_{m,l}, r_{m,l}) = \langle K_{\mathbf{f}_\lambda}(\mathcal{X}_M, \mathcal{X}_M), \mathbf{f}_{\mathbf{r}_{m,l}}(\mathcal{X}_M) \rangle = \text{Cov}(\mathbf{f}_{\mathbf{r}_{m,l}}(\mathcal{X}_M), \mathbf{f}_{\mathbf{r}_{m,l}}(\mathcal{X}_M)) \), by the identity
\[ r_{m,l} K(r_{m,l}, r_{m,l}) = N^2 K_{\mathbf{f}_\lambda}(\mathcal{X}_M, \mathcal{X}_M) \tag{53} \]
and the fact that the matrix in the left hand side of (52) is invertible, one can verify that
\[ \hat{c} = \frac{1}{N} \mathbf{r}_K^T \cdot (K_{\mathbf{f}_\lambda}(\mathcal{X}_M, \mathcal{X}_M) + \lambda NMLI)^{-1} \nabla_{\sigma^2,M} \tag{54} \]
is the solution.

Now we are ready to present the proof for Theorem 4 in the main text.
Proof of Theorem 4 in the main text. Let \( \tilde{K} = \frac{\sigma^2 K}{MN^2} \). Since \( \phi \sim \mathcal{GP}(0, \tilde{K}) \), the posterior mean in the main text equation (12) will become

\[
\hat{\phi}_M(r^*) = \tilde{K}_{\phi, \Phi_0}(r^*, \mathcal{X}_M)(\tilde{K}_{\phi_0}^{-1}(X_M, X_M) + \sigma^2 I)^{-1}V_{\sigma^2, M} \]

\[
= \frac{1}{N} K_{r^*_M}(r^*) \mathcal{R}_M^T(\mathcal{K}_{\phi_0}^{-1}(X_M, X_M) + \sigma^2 I)^{-1}V_{\sigma^2, M} \]

\[
= \frac{1}{N} K_{r^*_M}(r^*) \mathcal{R}_M^T(K_{\phi_0}^{-1}(X_M, X_M) + NML)\mathcal{V}_{\sigma^2, M} \]

\[
= K_{\phi, \Phi_0}(r^*, \mathcal{X}_M)(K_{\phi_0}^{-1}(X_M, X_M) + NML)\mathcal{V}_{\sigma^2, M} \]

\[
= \sum_{r \in r_M} \hat{c}_r K_r,
\]

where \( \hat{c} \) is defined in (49) and we use the identity \( K_{\phi, \Phi_0}(r^*, \mathcal{X}_M) = \frac{1}{N} K_{r^*_M}(r^*) \mathcal{R}_M^T \) (also for \( \tilde{K} \)) in the proof.

\[ \square \]

C.3 Finite sample analysis of reconstruction error

In this subsection, we shall assume that \( \phi \sim \mathcal{GP}(0, \tilde{K}) \) with \( \tilde{K} = \frac{\sigma^2 K}{MN^2} (\lambda > 0) \) and that the coercivity condition (defined in the main text equation (17)) holds.

Analysis of sample errors. We employ the operator representation:

\[
\phi^{\lambda, M}_{\mathcal{H}_K} = (B_M + \lambda)^{-1} A_M^* \mathcal{V}_{\sigma^2, M} = \phi^{\lambda, M}_{\mathcal{H}_K} + \text{Noise term}
\]

\[
\phi^{\lambda, \infty}_{\mathcal{H}_K} = (B + \lambda)^{-1} B \phi,
\]

where \( \phi^{\lambda, M}_{\mathcal{H}_K} \) is the empirical minimizer of \( \mathcal{E}^{\lambda, M}(\cdot) \) for noise-free observations and \( \mathcal{V}_{\sigma^2, M} \) denotes the noise vector. We first provide non-asymptotic analysis of sample error \( \| \langle (B_M + \lambda)^{-1} B_M \phi - (B + \lambda)^{-1} B \phi \rangle \mathcal{H}_K \| \) for any \( \phi \in \mathcal{H}_K \) and then apply it to \( \phi \). This allows us to obtain a bound on \( \| \phi^{\lambda, M}_{\mathcal{H}_K} - \phi^{\lambda, M}_{\mathcal{H}_K} \| \mathcal{H}_K \).

Lemma 12. For any bounded function \( \phi \in L^2(\hat{\rho}^2_M) \) and any positive integer \( M \), we have that

\[
\| B_M \phi \|_{\mathcal{H}_K} \leq R^2 \| \phi \|_{\infty}, a.s., \tag{55}
\]

\[
\mathbb{E} \| B_M \phi \|_{\mathcal{H}_K}^2 \leq \| \phi \|_{L^2(\hat{\rho}^2_M)}^2 \kappa^2 R^2. \tag{56}
\]

Proof. Note that \( \| K_r \|_{\mathcal{H}_K} \leq \kappa \) for any \( r \in [0, R] \), we have that

\[
\| B_M \phi \|_{\mathcal{H}_K} \leq \frac{1}{LM} \sum_{i=1}^{L_M} \sum_{i=1}^{M} \sum_{i=1}^{N} \frac{1}{N} \| K_{N M}^{(m,i)} \|_{\mathcal{H}_K} \| \phi \|_{\infty} R^2
\]

\[
\leq \kappa \| \phi \|_{\infty} R^2, a.s.
\]

For the second inequality, we have that

\[
\mathbb{E} \| B_M \phi \|^2_{\mathcal{H}_K} = \mathbb{E} \langle A_M^* A_M \phi, A_M^* A_M \phi \rangle_{\mathcal{H}_K} = \langle A \phi, A \phi \rangle_{L^2(\rho_X)} \leq \| A \phi \|_{L^2(\rho_X)} \| A \|_{L^2(\rho_X)} \| \phi \|_{L^2(\hat{\rho}^2_M)} \leq \kappa^2 R^2 \| \phi \|_{L^2(\hat{\rho}^2_M)}^2,
\]

where we use the Lemma[17] and [38].

\[ \square \]

Theorem 13. For a bounded function \( \phi \in L^2(\hat{\rho}^2_M) \) and 0 < \( \delta < 1 \), with probability at least 1 - \( \delta \), there holds

\[
\| B_M \phi - B \phi \|_{\mathcal{H}_K} \leq \frac{4\kappa R^2 \| \phi \|_{\infty} \log(2/\delta)}{M} + \kappa R\| \phi \|_{L^2(\hat{\rho}^2_M)} \sqrt{\frac{2\log(2/\delta)}{M}}, \tag{57}
\]
Proof. Define the $\mathcal{H}_K$-valued random variable

$$\xi^{(m)} = \frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{N} \frac{1}{N^2} K_{r^{(m,l)}}(\varphi, K_{r^{(m,l)}}) \mathcal{H}_K \langle \xi^{(m,l)}, r^{(m,l)} \rangle.$$

Then the random variables $\{\xi^{(m)}\}_{m=1}^{M}$ are i.i.d. According to Lemma 12, we have that

$$\|\xi^{(m)}\|_{\mathcal{H}_K} \leq \kappa R^2 \|\varphi\|_{\infty},$$

$$\mathbb{E}\|\xi^{(m)}\|^2_{\mathcal{H}_K} \leq \kappa^2 R^2 \|\varphi\|_{L^2(\rho)}.$$

Note that $B_M\varphi - B\varphi = \frac{1}{M} \sum_{m=1}^{M} (\xi^{(m)} - \mathbb{E}(\xi^{(m)}))$. The conclusion follows by applying Lemma 19 to $\{\xi^{(m)}\}_{m=1}^{M}$.

Theorem 14. For any bounded function $\varphi \in L^2(\rho)$. Let $0 < \delta < 1$, with probability at least $1 - \delta$, there holds

$$\|(B_M + \lambda)^{-1} B_M \varphi - (B + \lambda)^{-1} B\varphi\|_{\mathcal{H}_K} \leq \frac{\kappa R^2 \|\varphi\|_{\infty} \sqrt{2 \log(4/\delta)}}{\sqrt{M\lambda} (C_{\kappa,\mathcal{H}_K} + C_{\kappa,R,\lambda} \frac{2 \log(4/\delta)}{\sqrt{M\lambda}})},$$

where $C_{\kappa,\mathcal{H}_K} = (\kappa + 1) \frac{2}{\sqrt{M\lambda}}$ and $C_{\kappa,R,\lambda} = \kappa R + \sqrt{\lambda}$.

Proof. We introduce an intermediate quantity $(B_M + \lambda)^{-1} B\varphi$ and decompose

$$(B_M + \lambda)^{-1} B_M \varphi - (B + \lambda)^{-1} B\varphi = (B_M + \lambda)^{-1} B_M \varphi - (B_M + \lambda)^{-1} B\varphi + (B_M + \lambda)^{-1} B\varphi - (B + \lambda)^{-1} B\varphi.$$

Since $\|(B_M + \lambda)^{-1}\|_{\mathcal{H}_K} \leq \frac{1}{\lambda}$, we have that

$$\|(B_M + \lambda)^{-1} B_M \varphi - (B + \lambda)^{-1} B\varphi\|_{\mathcal{H}_K} \leq \frac{1}{\lambda} \|B_M \varphi - B\varphi\|_{\mathcal{H}_K}.$$

Applying Theorem 13 to $B_M \varphi - B\varphi$, we obtain with probability at least $1 - \delta/2$

$$\frac{1}{\lambda} \|B_M \varphi - B\varphi\|_{\mathcal{H}_K} \leq \frac{4\kappa R^2 \|\varphi\|_{\infty} \log(4/\delta)}{\lambda M} + \kappa R \|\varphi\|_{L^2(\rho)} \sqrt{\frac{2 \log(4/\delta)}{\lambda^2 M}}.$$

On the other hand, we have

$$\|(B_M + \lambda)^{-1} B\varphi - (B + \lambda)^{-1} B\varphi\|_{\mathcal{H}_K} = \|(B_M + \lambda)^{-1}(B - B_M)(B + \lambda)^{-1} B\varphi\|_{\mathcal{H}_K} \leq \frac{1}{\lambda} \|(B - B_M)(B + \lambda)^{-1} B\varphi\|_{\mathcal{H}_K}.$$

Since $\varphi^{\lambda,\infty}_{\mathcal{H}_K} = (B + \lambda)^{-1} B\varphi$ is the unique minimizer of the expected risk functional $\mathcal{E}(\psi) = \|A\psi - A\varphi\|_{L^2(\rho_X)}^2 + \lambda \|\psi\|_{\mathcal{H}_K}^2$, plugging $\psi = 0$, we obtain that

$$\|A\varphi^{\lambda,\infty}_{\mathcal{H}_K} - A\varphi\|_{L^2(\rho_X)}^2 + \lambda \|\varphi^{\lambda,\infty}_{\mathcal{H}_K}\|_{\mathcal{H}_K}^2 < \|A\varphi\|_{L^2(\rho_X)}^2,$$

which implies that

$$\|\varphi^{\lambda,\infty}_{\mathcal{H}_K}\|_{\mathcal{H}_K} \leq \frac{1}{\sqrt{\lambda}} \|A\varphi\|_{L^2(\rho_X)},$$

$$\|A\varphi^{\lambda,\infty}_{\mathcal{H}_K}\|_{L^2(\rho_X)}^2 \leq 2 \|A\varphi\|_{L^2(\rho_X)}^2.$$ (58)

By Lemma 18 and (58), it follows that

$$\|\varphi^{\lambda,\infty}_{\mathcal{H}_K}\|_{\mathcal{H}_K} \leq \kappa \|\varphi^{\lambda,\infty}_{\mathcal{H}_K}\|_{\mathcal{H}_K} \leq \frac{\kappa}{\sqrt{\lambda}} \|A\varphi\|_{L^2(\rho_X)}.$$(60)
Suppose the coercivity condition \([17]\) holds. We have that
\[
\|\varphi_{\mathcal{H}_K}^\infty\|_{L^2(\rho_T^*_M)} \leq \frac{1}{c_{\mathcal{H}_K}} \|A\varphi_{\mathcal{H}_K}^\infty\|_{L^2(\rho_T^*_M)} \leq \frac{2}{c_{\mathcal{H}_K}} \|A\varphi\|_{L^2(\rho_T^*_M)}.
\] (61)

Note that \(\|A\varphi\|_{L^2(\rho_T^*_M)} < R^2 \|\varphi\|_\infty^2\) (see \([37]\)). Applying theorem \([13]\) to \(\varphi_{\mathcal{H}_K}^\lambda = (B + \lambda)^{-1} B \varphi\), and using \((60), (61)\), we obtain that with probability at least \(1 - \delta/2\),
\[
\frac{1}{\lambda} \|(B - B_M)(B + \lambda)^{-1} B \varphi\|_{\mathcal{H}_K} \leq \frac{4\kappa R^2 \|\varphi_{\mathcal{H}_K}^\lambda\|_\infty \log(4/\delta)}{\lambda M} + \frac{\kappa R \|\varphi_{\mathcal{H}_K}^\lambda\|_{L^2(\rho_T^*_M)} \sqrt{2 \log(4/\delta)}}{\lambda^2 M} \leq \frac{4\kappa^2 R^2 \|\varphi\|_\infty \log(4/\delta)}{\lambda^2 M} + \frac{\sqrt{2} \kappa R \|\varphi\|_{L^2(\rho_T^*_M)} \sqrt{2 \log(4/\delta)}}{\lambda^2 M} \leq \frac{4\kappa^2 R^2 \|\varphi\|_\infty \log(4/\delta)}{\lambda^2 M} + \frac{\sqrt{2} \kappa R \|\varphi\|_{L^2(\rho_T^*_M)} \sqrt{2 \log(4/\delta)}}{\lambda^2 M}.
\]

Finally, by combining two bounds, we obtain that with a probability at least \(1 - \delta\)
\[
\|(B + \lambda)^{-1} B_M \varphi - (B + \lambda)^{-1} B \varphi\|_{\mathcal{H}_K} \leq \frac{\kappa R^2 \|\varphi\|_\infty \sqrt{2 \log(4/\delta)}}{\sqrt{M\lambda}} \left[(\kappa + 1) + \frac{\kappa R + \sqrt{\lambda}}{\sqrt{M\lambda}} \frac{2 \log(4/\delta)}{2 \log(8/\delta)} \right] \leq \frac{\kappa R^2 \|\varphi\|_\infty \sqrt{2 \log(4/\delta)}}{\sqrt{M\lambda}} \left(C_{\mathcal{H}_K} + \frac{C_{\mathcal{H}_K, R, \lambda}}{\sqrt{M\lambda}} \sqrt{2 \log(4/\delta)} \right).
\]
where \(C_{\mathcal{H}_K} = (\kappa + 1) \sqrt{2 \frac{\lambda}{c_{\mathcal{H}_K}}}\) and \(C_{\mathcal{H}_K, R, \lambda} = \kappa R + \sqrt{\lambda}\).

**Theorem 15** (Sample error). For any \(\delta \in (0, 1)\), it holds with a probability at least \(1 - \delta\) that
\[
\|\phi_{\mathcal{H}_K}^\lambda - \phi_{\mathcal{H}_K}^\infty\|_{\mathcal{H}_K} \leq \frac{\kappa R^2 \|\varphi\|_\infty \sqrt{2 \log(8/\delta)}}{\sqrt{M\lambda}} \left(C_{\mathcal{H}_K} + \frac{C_{\mathcal{H}_K, R, \lambda}}{\sqrt{M\lambda}} \sqrt{2 \log(8/\delta)} \right) + \frac{2\kappa R \sigma \log(8/\delta)}{\sqrt{c\lambda d M N L}}
\]
where \(c\) is a absolute constant appearing in the Hanson-Wright inequality (Theorem \([20]\)), \(C_{\mathcal{H}_K} = (\kappa + 1) \sqrt{2 \frac{\lambda}{c_{\mathcal{H}_K}}}\) and \(C_{\mathcal{H}_K, R, \lambda} = \kappa R + \sqrt{\lambda}\).

**Proof.** We decompose \(\phi_{\mathcal{H}_K}^\lambda - \phi_{\mathcal{H}_K}^\infty = \phi_{\mathcal{H}_K}^\lambda - \tilde{\phi}_{\mathcal{H}_K}^\lambda + \tilde{\phi}_{\mathcal{H}_K}^\lambda - \phi_{\mathcal{H}_K}^\infty\) where \(\tilde{\phi}_{\mathcal{H}_K}^\lambda\) is the empirical minimizer for noise-free observations. Then applying Theorem \([13]\) to the term \(\phi_{\mathcal{H}_K}^\lambda - \tilde{\phi}_{\mathcal{H}_K}^\lambda\), we obtain that with probability at least \(1 - \delta\),
\[
\|\tilde{\phi}_{\mathcal{H}_K}^\lambda - \phi_{\mathcal{H}_K}^\infty\|_{\mathcal{H}_K} \leq \frac{\kappa R^2 \|\varphi\|_\infty \sqrt{2 \log(4/\delta)}}{\sqrt{M\lambda}} \left(C_{\mathcal{H}_K} + \frac{C_{\mathcal{H}_K, R, \lambda}}{\sqrt{M\lambda}} \sqrt{2 \log(4/\delta)} \right).
\] (62)

We now just need to estimate the “noise part” \(\phi_{\mathcal{H}_K}^\lambda - \tilde{\phi}_{\mathcal{H}_K}^\lambda\). According to \([46]\),
\[
\tilde{\phi}_{\mathcal{H}_K}^\lambda - \phi_{\mathcal{H}_K}^\lambda = (B_M + \lambda)^{-1} A_M^* \mathbb{W}_M
\] (63)
where the noise vector \(\mathbb{W}_M\) follows a multivariate Gaussian distribution with zero mean and variance \(\sigma^2 J_{NML}\). Note that
\[
\|\tilde{\phi}_{\mathcal{H}_K}^\lambda - \phi_{\mathcal{H}_K}^\lambda\|_{\mathcal{H}_K}^2 = \langle \mathbb{W}_M, A_M (B_M + \lambda)^{-2} A_M^* \mathbb{W}_M \rangle = \mathbb{W}_M^T \Sigma_M \mathbb{W}_M,
\]
where the matrix
\[
\Sigma_M = (K_{\mathcal{E}}(\mathbb{X}_M, \mathbb{X}_M) + \lambda N d M L I)^{-1} K_{\mathcal{E}}(\mathbb{X}_M, \mathbb{X}_M)(K_{\mathcal{E}}(\mathbb{X}_M, \mathbb{X}_M) + \lambda d N M L I)^{-1}.
\]
The matrix $\Sigma_M$ is the matrix form of the operator $A_M(B_M + \lambda)^{-2}A_M^*$, as is derived from (46), (49) and (53). We have that

$$\text{Tr}(\Sigma_M) \leq \frac{1}{\lambda^4(MLN)^2} \text{Tr}(K_{f_0}(X_M, X_M))$$

$$= \frac{1}{\lambda^2(MLN)^2} \sum_{m=1,d=1,m' = 1}^{M,L,N} \frac{1}{N^2} \sum_{k \neq i, k' \neq i} K(r_{ik, i'}, r_{ik', i'})^T r_{ik}$$

$$\leq \frac{1}{\lambda^2d^2MLN} \kappa^2R^2, \text{a.s.}$$

and

$$\text{Tr}(\Sigma_M^2) \leq \frac{1}{\lambda^4(MLN)^2} \text{Tr}(K_{f_0}(X_M, X_M)^2)$$

$$= \frac{1}{\lambda^4(MLN)^2} \sum_{m,m',i,i' = 1}^{M,L,N} \left\| \frac{1}{N^2} \sum_{k \neq i, k' \neq i} K(r_{ik, i'}, r_{ik', i'}) r_{ik}(r_{ik', i'})^T \right\|_F^2$$

$$\leq \frac{\kappa^4R^4}{\lambda^4d^4(MLN)^2}, \text{a.s.}$$

Now we apply the Hanson-Wright inequality (Theorem 20) for the Gaussian random vector $\mathbb{W}_M$ with $S_0 = \sigma^2$. Note that for any $\epsilon > 0$,

$$\min \left\{ \frac{1}{\sigma^4\|\Sigma\|_F^2}, \frac{\epsilon}{\sigma^2\|\Sigma\|} \right\} \geq \min \left\{ \frac{1}{\sigma^4\text{Tr}(\Sigma_M)}, \frac{\epsilon}{\sigma^2\text{Tr}(\Sigma_M)} \right\},$$

we obtain that, with a probability at least $1 - e^{-t^2}$,

$$\mathbb{W}_M^T \Sigma_M \mathbb{W}_M \leq \frac{1}{c} \sigma^2 \max \{ \text{Tr}(\Sigma_M), \sqrt{\text{Tr}(\Sigma_M^2)} \} (1 + 2t + t^2)$$

$$\leq \frac{\kappa^2R^2\sigma^2}{c\lambda^2d^2MLN} (1 + 2t + t^2)$$

for any $t > 0$, where $c$ is an absolute positive constant appearing in Hanson-Wright inequality. Therefore, with a probability at least $1 - \delta$, there holds

$$\| \phi_{\mathcal{H}_K}^{\lambda,M} - \phi_{\mathcal{H}_K}^{\lambda,M} \|_{\mathcal{H}_K} \leq \frac{\kappa R\sigma (\log(1/\delta) + 1)}{\sqrt{c\lambda d} \sqrt{MLN}} < \frac{2\kappa R\sigma \log(4/\delta)}{\sqrt{c\lambda d} \sqrt{MLN}}$$

(64)

Now combining (62) and (64), we obtain that with probability at least $1 - \delta$,

$$\| \phi_{\mathcal{H}_K}^{\lambda,M} - \phi_{\mathcal{H}_K}^{\lambda,\infty} \|_{\mathcal{H}_K} \leq \frac{\kappa R^2\| \phi \|_{\infty} \sqrt{2\log(8/\delta)}}{\sqrt{ML}} \left( C_{\mathcal{H}_K} + \frac{C_{\mathcal{H}_K}R^{2}\sqrt{2\log(8/\delta)}}{\sqrt{ML}} \right) + \frac{2\kappa R\sigma \log(4/\delta)}{\sqrt{c\lambda d} \sqrt{MLN}}$$

$$\Box$$

**Analysis of approximation error** $\| \phi_{\mathcal{H}_K}^{\lambda,\infty} - \phi \|_{\mathcal{H}_K}$. To get a convergence rate for the reconstruction error $\| \phi_{\mathcal{H}_K}^{\lambda,M} - \phi \|_{\mathcal{H}_K}$, we need to get an estimation of the approximation error $\| \phi_{\mathcal{H}_K}^{\lambda,\infty} - \phi \|_{\mathcal{H}_K}$. Assume the coercivity condition holds. Then $B \in \mathcal{B}(\mathcal{H}_K)$ is a strictly positive operator. We follow the standard argument in the literature of Tikhonov regularization (see Section 5 in [25]). Let $B = \sum_{n=1}^{N} \lambda_n (\cdot, e_n)e_n$ (possibly $N = \infty$) be the spectral decomposition of $B$ with $0 < \lambda_{n+1} < \lambda_n$ and $\{e_n\}_{n=1}^{N}$ be an orthonormal basis of $\mathcal{H}_K$. Then

$$\| \phi_{\mathcal{H}_K}^{\lambda,\infty} - \phi \|_{\mathcal{H}_K}^2 = \| (B + \lambda)^{-1} B \phi - \phi \|_{\mathcal{H}_K}^2 = \| \lambda (B + \lambda)^{-1} \phi \|_{\mathcal{H}_K}^2$$

$$= \sum_{n=1}^{N} (\frac{\lambda}{\lambda_n + \lambda})^2 \langle \phi, e_n \rangle_{\mathcal{H}_K}^2.$$ 

(65)

Assume now that $\phi \in \text{Im} B^\gamma$ with $0 < \gamma \leq \frac{1}{2}$. Since the function $x^\gamma$ is concave on $[0, \infty]$, $\frac{\lambda}{\lambda_n + \lambda} \leq \frac{\lambda_n^\gamma}{\lambda_n}$, Then we have $\| \phi_{\mathcal{H}_K}^{\lambda,\infty} - \phi \|_{\mathcal{H}_K} \leq \lambda^\gamma \| B^{-\gamma} \phi \|_{\mathcal{H}_K}$ where $B^{-\gamma} \phi$ represents the pre-image of $\phi$. 

24
Proof of Theorem 5 in the main text. Without loss of generality, let \( \lambda = M^{-\frac{1}{2\pi+1}} \). By Theorem 15 and approximation error (65), with a probability at least \( 1 - \delta \), we have that
\[
\| \phi^M_{H_K} - \phi \|_{H_K} \leq \| C^\lambda_{H_K} \|_{H_K} + \| \phi^\lambda_{H_K} - \phi \|_{H_K}
\]
\[
\leq \frac{kR^2\|\phi\|_{\infty} \sqrt{2 \log(8/\delta)}}{\sqrt{M\lambda}} (C_{\varepsilon, H_K} + C_{\varepsilon, R, \lambda} 2 \log(8/\delta) \sqrt{M\lambda}) + \frac{2kR\sigma \log(8/\delta)}{\sqrt{c\lambda d\sqrt{MN}}} + \lambda^\gamma B^{-\gamma} \| \phi \|_{H_K}
\]
\[
\lesssim C \log(8/\delta) M^{-\frac{\gamma}{2\pi+1}},
\]
where \( C = \max\{s^2 R^2 \| \phi \|_{\infty}, \frac{2kR\sigma}{\sqrt{c\lambda dN}} \} \| B^{-\gamma} \|_{H_K} \}, \) and the symbol \( \lesssim \) means that the inequality holds up to a multiplicative constant that are independent from the listed parameters.

\[ \square \]

C.4 Auxiliary lemmas and theorems

Lemma 16. Let \( x \) and \( y \) be jointly Gaussian random vectors
\[
[ \begin{array}{c} x \\ y \end{array} ] \sim \mathcal{N} ( \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} A & C \\ CT & B \end{bmatrix} ),
\]
then the marginal distribution of \( x \) and the conditional distribution of \( x \) given \( y \) are
\[
x \sim \mathcal{N} ( \mu_x, A ), \quad \text{and} \quad x|y \sim \mathcal{N} ( \mu_x + CB^{-1}(y - \mu_y), A - CB^{-1}CT ) .
\]

Proof. See, e.g. [18], Appendix A.

\[ \square \]

Lemma 17. For any function \( \varphi \in L^2(\hat{\rho}_T^L) \), we have that
\[
\| f_{\varphi} \|_{L^2(\hat{\rho}_T^L)} \leq \frac{N - 1}{N} \| \varphi \|_{L^2(\hat{\rho}_T^L)} .
\]

Proof. See the proof of Proposition 16 in [32] by taking \( K = 1 \).

\[ \square \]

Lemma 18. By Assumption 1, we have that, for any \( \varphi \in H_K \), there holds \( \| \varphi \|_{\infty} \leq \kappa \| \varphi \|_{H_K} \).

Proof. By the reproducing property of \( K \), we have that
\[
| \langle \varphi, K \rangle_{H_K} | \leq \| \varphi \|_{H_K} \| K \|_{H_K} \leq \kappa \| \varphi \|_{H_K} .
\]
The conclusion follows.

\[ \square \]

Lemma 19 (Lemma 8 in [21]). Let \( H \) be a Hilbert space and \( \xi \) be a random variable on \( (Z, \rho) \) with values in \( H \). Suppose that, \( \| \xi \|_{H} \leq S < \infty \) almost surely. Let \( z_m \) be i.i.d drawn from \( \rho \). For any \( 0 < \delta < 1 \), with confidence \( 1 - \delta \),
\[
\left\| \frac{1}{M} \sum_{m=1}^{M} (\xi(z_m) - \mathbb{E}(\xi)) \right\| \leq \frac{4S \log(2/\delta)}{M} + \sqrt{\frac{2\mathbb{E}(\| \xi \|_{H}^2) \log(2/\delta)}{M}} .
\]
The original version of Lemma 19 is presented in [58].

Theorem 20 (Hanson-Wright inequality [59]). Let \( X = (X_1, \cdots, X_n) \in \mathbb{R}^n \) be a random vector with independent components \( X_i \) which satisfy \( \mathbb{E}X_i = 0 \) and \( \| X_i \|_{\psi_2} \leq S_0 \), where \( \| \cdot \|_{\psi_2} \) is the subGaussian norm. Let \( A \) be an \( n \times n \) matrix and \( \| A \|_{HS} \) denotes the Hilbert-Schmidt norm. Then, for every \( \epsilon \geq 0 \)
\[
\mathbb{P} \left\{ \left\| X^T A X \right\| - \mathbb{E}X^T AX \right\| \geq \epsilon \right\} \leq 2 \exp \left\{ - c \min \left\{ \frac{\epsilon^2}{S_0^2 \| A \|_{HS}^2}, \frac{\epsilon}{S_0^2 \| A \|} \right\} \right\} ,
\]
where \( c \) is an absolute positive constant.
References

[1] Michele Ballerini, Nicola Cabibbo, Raphael Candelier, Andrea Cavagna, Evaristo Cisbani, Irene Giardina, Vivien Lecomte, Alberto Orlandi, Giorgio Parisi, Andrea Procaccini, et al. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. Proceedings of the national academy of sciences, 105(4):1232–1237, 2008.

[2] Ryan Lukeman, Yue-Xian Li, and Leah Edelstein-Keshet. Inferring individual rules from collective behavior. Proceedings of the National Academy of Sciences, 107(28):12576–12580, 2010.

[3] David JT Sumpter. Collective animal behavior. Princeton University Press, 2010.

[4] Yael Katz, Kolbjørn Tunstrøm, Christos C Ioannou, Cristián Huepe, and Iain D Couzin. Inferring the structure and dynamics of interactions in schooling fish. Proceedings of the National Academy of Sciences, 108(46):18720–18725, 2011.

[5] Felipe Cucker and Steve Smale. Emergent behavior in flocks. IEEE Transactions on automatic control, 52(5):852–862, 2007.

[6] Mathieu Lewin and Xavier Blanc. The crystallization conjecture: a review. EMS Surveys in Mathematical Sciences, 2(2):255–306, 2015.

[7] Elena Vedmedenko. Competing interactions and pattern formation in nanoworld. John Wiley & Sons, 2007.

[8] Ulrich Krause et al. A discrete nonlinear and non-autonomous model of consensus formation. Communications in difference equations, 2000:227–236, 2000.

[9] Vincent D Blondel, Julien M Hendrickx, and John N Tsitsiklis. On krause’s multi-agent consensus model with state-dependent connectivity. IEEE transactions on Automatic Control, 54(11):2586–2597, 2009.

[10] Sebastien Motsch and Eitan Tadmor. Heterophilious dynamics enhances consensus. SIAM review, 56(4):577–621, 2014.

[11] Yao-Li Chuang, Maria R D’orsogna, Daniel Marthaler, Andrea L Bertozzi, and Lincoln S Chayes. State transitions and the continuum limit for a 2d interacting, self-propelled particle system. Physica D: Nonlinear Phenomena, 232(1):33–47, 2007.

[12] Nicole Abaid and Maurizio Porfiri. Fish in a ring: spatio-temporal pattern formation in one-dimensional animal groups. Journal of The Royal Society Interface, 7(51):1441–1453, 2010.

[13] Giacomo Albi, D Balagué, José A Carrillo, and J32150701305 von Brecht. Stability analysis of flock and mill rings for second order models in swarming. SIAM Journal on Applied Mathematics, 74(3):794–818, 2014.

[14] Maria R D’Orsogna, Yao-Li Chuang, Andrea L Bertozzi, and Lincoln S Chayes. Self-propelled particles with soft-core interactions: patterns, stability, and collapse. Physical review letters, 96(10):104302, 2006.

[15] Fabian Baumann, Igor M Sokolov, and Melvyn Tyloo. A laplacian approach to stubborn agents and their role in opinion formation on influence networks. Physica A: Statistical Mechanics and its Applications, 557:124869, 2020.

[16] Theodore Kolokolnikov, José A Carrillo, Andrea Bertozzi, Razvan Fetecau, and Mark Lewis. Emergent behaviour in multi-particle systems with non-local interactions. Physica D: Nonlinear Phenomena, (260):1–4, 2013.

[17] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. Journal of artificial intelligence research, 4:129–145, 1996.

[18] Christopher KL Williams and Carl Edward Rasmussen. Gaussian processes for machine learning, volume 2. MIT press Cambridge, MA, 2006.
19) Kevin P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

20) Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

21) Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, Francesca Odone, and Peter Bartlett. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6(5), 2005.

22) Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

23) Yun Yang, Anirban Bhattacharya, and Debdeep Pati. Frequentist coverage and sup-norm convergence rate in gaussian process regression. *arXiv preprint arXiv:1708.04753*, 2017.

24) Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

25) Steve Smale and Ding-Xuan Zhou. Shannon sampling ii: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, 2005.

26) Andrea Caponnetto and Ernesto De Vito. Fast rates for regularized least-squares algorithm. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE COMPUTER SCIENCE AND ARTIFICIAL, 2005.

27) Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

28) Zejian Liu and Meng Li. Non-asymptotic analysis in kernel ridge regression. *arXiv preprint arXiv:2006.01350*, 2020.

29) Fei Lu, Ming Zhong, Sui Tang, and Mauro Maggioni. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proceedings of the National Academy of Sciences*, 116(29):14424–14433, 2019.

30) Ming Zhong, Jason Miller, and Mauro Maggioni. Data-driven discovery of emergent behaviors in collective dynamics. *Physica D: Nonlinear Phenomena*, 411:132542, 2020.

31) Fei Lu, Mauro Maggioni, and Sui Tang. Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories. *arXiv preprint arXiv:2007.15174*, 2020.

32) Fei Lu, Mauro Maggioni, and Sui Tang. Learning interaction kernels in heterogeneous systems of agents from multiple trajectories. *Journal of Machine Learning Research*, 22(32):1–67, 2021.

33) Jason Miller, Sui Tang, Ming Zhong, and Mauro Maggioni. Learning theory for inferring interaction kernels in second-order interacting agent systems. *arXiv preprint arXiv:2010.03729*, 2020.

34) Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

35) Carl Edward Rasmussen and Zoubin Ghahramani. Occam’s razor. *Advances in neural information processing systems*, pages 294–300, 2001.

36) Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

37) Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

38) Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.
[39] Andrei Nikolajevits Tihonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math.*, 4:1035–1038, 1963.

[40] Andrei Nikolaevich Tikhonov, AV Goncharsky, VV Stepanov, and Anatoly G Yagola. *Numerical methods for the solution of ill-posed problems*, volume 328. Springer Science & Business Media, 2013.

[41] Tomaso Poggio and Federico Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

[42] Markus Heinonen, Cagatay Yildiz, Henrik Mannerström, Jukka Intosalmi, and Harri Lähdesmäki. Learning unknown ode models with gaussian processes. In *International Conference on Machine Learning*, pages 1959–1968. PMLR, 2018.

[43] Cedric Archambeau, Dan Cornford, Manfred Opper, and John Shawe-Taylor. Gaussian process approximations of stochastic differential equations. In *Gaussian Processes in Practice*, pages 1–16. PMLR, 2007.

[44] Cagatay Yildiz, Markus Heinonen, Jukka Intosalmi, Henrik Mannerström, and Harri Lähdesmäki. Learning stochastic differential equations with gaussian processes without gradient matching. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2018.

[45] Zheng Zhao, Filip Tronarp, Roland Hostettler, and Simo Särkkä. State-space gaussian process for drift estimation in stochastic differential equations. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5295–5299. IEEE, 2020.

[46] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Machine learning of linear differential equations using gaussian processes. *Journal of Computational Physics*, 348:683–693, 2017.

[47] Zhiping Mao, Zhen Li, and George Em Karniadakis. Nonlocal flocking dynamics: Learning the fractional order of pdes from particle simulations. *Communications on Applied Mathematics and Computation*, 1(4):597–619, 2019.

[48] Jiuhai Chen, Lulu Kang, and Guang Lin. Gaussian process assisted active learning of physical laws. *Technometrics*, pages 1–14, 2020.

[49] Hongqiao Wang and Xiang Zhou. Explicit estimation of derivatives from data and differential equations by gaussian process regression. *International Journal for Uncertainty Quantification*, 11(4), 2021.

[50] Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart. Solving and learning nonlinear pdes with gaussian processes. *arXiv preprint arXiv:2103.12959*, 2021.

[51] Seungjoon Lee, Mahdi Kooshkbaghi, Konstantinos Spiliotis, Konstantinos I Siettos, and Ioannis G Kevrekidis. Coarse-scale pdes from fine-scale observations via machine learning. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(1):013141, 2020.

[52] Michael Taylor. Towards a mathematical theory of influence and attitude change. *Human Relations*, 21(2):121–139, 1968.

[53] Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18:1257–1264, 2005.

[54] Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[55] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

[56] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.
[57] Jiri Blank, Pavel Exner, and Miloslav Havlicek. *Hilbert space operators in quantum physics*. Springer Science & Business Media, 2008.

[58] V Yurinsky. Sums and gaussian vectors. *lecture notes in mathematics*. 1995.

[59] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.