Physica D 439 (2022) 133406

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

Learning mean-field equations from particle data using WSINDy

Daniel A. Messenger *, David M. Bortz

Department of Applied Mathematics, University of Colorado Boulder, 11 Engineering Dr, Boulder, CO 80309, USA

ARTICLE INFO

Article history:
Received 30 November 2021
Received in revised form 18 May 2022
Accepted 2 June 2022
Available online 18 June 2022
Communicated by R. Kuske

Keywords:
Data-driven modeling
Interacting particle systems
Weak form
Mean-field limit
Sparse regression

ABSTRACT

We develop a weak-form sparse identification method for interacting particle systems (IPS) with the primary goals of reducing computational complexity for large particle number N and offering robustness to either intrinsic or extrinsic noise. In particular, we use concepts from mean-field theory of IPS in combination with the weak-form sparse identification of nonlinear dynamics algorithm (WSINDy) to provide a fast and reliable system identification scheme for recovering the governing stochastic differential equations for an IPS when the number of particles per experiment N is on the order of several thousands and the number of experiments M is less than 100. This is in contrast to existing work showing that system identification for N less than 100 and M on the order of several thousand is feasible using strong-form methods. We prove that under some standard regularity assumptions the scheme converges with rate $\mathcal{O}(N^{-1/2})$ in the ordinary least squares setting and we demonstrate the convergence rate numerically on several systems in one and two spatial dimensions. Our examples include a canonical problem from homogenization theory (as a first step towards learning coarse-grained models), the dynamics of an attractive–repulsive swarm, and the IPS description of the parabolic–elliptic Keller–Segel model for chemotaxis. Code is available at https://github.com/MathBioCU/WSINDy_IPS.

1. Problem statement

Recently there has been considerable interest in the methodology of data-driven discovery for governing equations. Building on the Sparse Identification of Nonlinear Dynamics (SINDy) [1], we developed a weak form version (WSINDy) for ODEs [2] and for PDEs [3]. In this work, we develop a formulation for discovering governing stochastic differential equations (SDEs) for interacting particle systems (IPS). To promote clarity and for reference later in the article, we first state the problem of interest. Subsequently, we will provide a discussion of background concepts and current results in the literature.

Consider a particle system $\mathbf{X}_t = (X_t^{(1)},\dots,X_t^{(N)}) \in \mathbb{R}^{Nd}$ where on some fixed time window $t \in [0,T]$, each particle $X_t^{(i)} \in \mathbb{R}^d$ evolves according to the overdamped dynamics

$$dX_t^{(i)} = \left(-\nabla K * \mu_t^N\left(X_t^{(i)}\right) - \nabla V\left(X_t^{(i)}\right)\right)dt + \sigma(X_t^{(i)})\,dB_t^{(i)} \qquad (1.1)$$

with initial data $X_0^{(i)}$ each drawn independently from some probability measure $\mu_0 \in \mathcal{P}_p(\mathbb{R}^d)$, where $\mathcal{P}_p(\mathbb{R}^d)$ is the space probability measures on $\mathbb{R}^d$ with finite pth moment.[1] Here, K is the *interaction potential* defining the pairwise forces between particles, V is the *local potential* containing all exogenous forces, σ is the *diffusivity*, and $\left(B_t^{(i)}\right)_{i=1,\dots,N}$ are independent Brownian motions each adapted to the same filtered probability space $(\Omega, \mathcal{B}, \mathbb{P}, (\mathcal{F}_t)_{t\geq 0})$. The *empirical measure* is defined

$$\mu_t^N := \frac{1}{N}\sum_{i=1}^N \delta_{X_t^{(i)}},$$

and the convolution $\nabla K * \mu_t^N$ is defined

$$\nabla K * \mu_t^N(x) = \nabla \int_{\mathbb{R}^d} K(x-y)\,d\mu_t^N(y) = \frac{1}{N}\sum_{i=1}^N \nabla K\left(x - X_t^{(i)}\right)$$

where we set $\nabla K(0) = 0$ whenever $\nabla K(0)$ is undefined. The recovery problem we wish to solve is the following.

(P) Let $\mathbb{X} = (\mathbf{X}_\mathbf{t}^{(1)},\dots,\mathbf{X}_\mathbf{t}^{(M)})$ be discrete-time data at L timepoints $\mathbf{t} := (t_1,\dots,t_L)$ for M i.i.d. trials of the process (1.1) with $K = K^\star$, $V = V^\star$, and $\sigma = \sigma^\star$ and let $\mathbb{Y} = \mathbb{X} + \varepsilon$ be a corrupted dataset. For some fixed compact domain $\mathcal{D} \subset \mathbb{R}^d$ containing supp $(\mathbb{Y})$, and finite-dimensional hypothesis spaces[2]

* Corresponding author.
E-mail addresses: daniel.messenger@colorado.edu (D.A. Messenger), dmbortz@colorado.edu (D.M. Bortz).

[1] We define the pth moment of a probability measure μ for $p \geq 0$ by $M_p(\mu) := \int_{\mathbb{R}^d} |x|^p d\mu(x)$.

[2] The set $\mathcal{D} - \mathcal{D}$ is defined $\mathcal{D} - \mathcal{D} = \{x - y \;:\; (x,y) \in \mathcal{D}\times\mathcal{D}\}$.

https://doi.org/10.1016/j.physd.2022.133406

$\mathcal{H}_K \subset L^2(\mathcal{D}-\mathcal{D})$, $\mathcal{H}_V \subset L^2(\mathcal{D})$, and $\mathcal{H}_\sigma \subset L^2(\mathcal{D})$, solve

$$\left(\widehat{K}, \widehat{V}, \widehat{\sigma}\right) = \underset{K\in\mathcal{H}_K, V\in\mathcal{H}_V, \sigma\in\mathcal{H}_\sigma}{\text{argmin}} \left\|\nabla K - \nabla K^\star\right\|_{L^2(\mathcal{D}-\mathcal{D})} + \left\|\nabla V - \nabla V^\star\right\|_{L^2(\mathcal{D})} + \left\|\sigma - \sigma^\star\right\|_{L^2(\mathcal{D})}.$$

The problem **(P)** is clearly intractable because we do not have access to $K^\star$, $V^\star$, or $\sigma^\star$, and moreover the interactions between these terms render simultaneous identification of them ill-posed. We consider two cases: (i) $\varepsilon \neq 0$ and $\sigma^\star = 0$, corresponding to purely *extrinsic noise*, and (ii) $\varepsilon = 0$ and $\sigma^\star \neq 0$, corresponding to purely *intrinsic noise*. The extrinsic noise case is important for many applications, such as cell tracking, where uncertainty is present in the position measurements. In this case we examine ε representing i.i.d. Gaussian noise with mean zero and variance[3] $\epsilon^2\mathbf{I}_d$ added to each particle position in $\mathbb{X}$. In the case of purely intrinsic noise, identification of the diffusivity $\sigma^\star$ is required as well as the deterministic forces on each particle as defined by $K^\star$ and $V^\star$. A natural next step is to consider the case with both extrinsic and intrinsic noise. However, the combined noise case is sufficiently nuanced as to render it beyond the scope of the article, and we leave it for future work.

2. Background

Interacting particle systems (IPS) such as (1.1) are used to describe physical and artificial phenomena in a range of fields including astrophysics [4,5], molecular dynamics [6], cellular biology [7–9], and opinion dynamics [10]. In many cases the number of particles N is large, with cell migration experiments often tracking $10^3 - 10^6$ cells and simulations in physics (molecular dynamics, particle-in-cell, etc.) requiring N in the range $10^6 - 10^{12}$. Inference of such systems from particle data thus requires efficient means of computing pairwise forces from $\mathcal{O}(N^2)$ interactions at each timestep for multiple candidate interaction potentials K. Frequently, so-called *mean-field* equations at the continuum level are sufficient to describe the evolution of the system, however in many cases (e.g. chemotaxis in biology [11]) only phenomenological mean-field equations are available. Moreover, it is often unclear how many particles N are needed for a mean-field description to suffice. Many disciplines are now developing machine learning techniques to extract coarse-grained dynamics from high-fidelity simulations (see [12] for a recent review in molecular dynamics). In this work we provide a means for inferring governing mean-field equations from particle data assumed to follow the dynamics (1.1) that is highly efficient for large N, and is effective in learning mean-field equations when N is in range $10^3 - 10^5$.

Inference of the drift and diffusion terms for stochastic differential equations (SDEs) is by now a mature field, with the primary method being maximum-likelihood estimation, which uses Girsanov's theorem together with the Radon–Nikodym derivative to arrive at a log-likelihood function for regression. See [13,14] for some early works and [15] for a textbook on this approach. More recently, sparse regression approaches using the Kramers–Moyal expansion have been developed [16–18] and the authors of [19] use sparse regression to learn population level ODEs from agent-based modeling simulations. The authors of [20] also derived a bias-correcting regression framework for inferring the drift and diffusion in underdamped Langevin dynamics, and in [21] a neural network-based algorithm for inferring SDEs was developed.

Only in the last few years have significant strides been made towards parameter inference of *interacting* particle systems such as (1.1) from data. Apart from some exceptions, such as a Gaussian process regression algorithm recently developed in [22], applications of maximum likelihood theory are by far the most frequently studied. An early but often overlooked work by Kasonga [23] extends the maximum-likelihood approach to inference of the interaction potential K, assuming full availability of the continuous particle trajectories and the diffusivity σ. Two decades later, Bishwal [24] further extended this approach to discrete particle observations in the specific context of linear particle interactions. In both cases, a sequence of finite-dimensional subspaces is used to approximate the interaction function, and convergence is shown as the dimension of the subspace J and number of particles N both approach infinity. More recently, the maximum likelihood approach has been carried out in [25,26] in the case of radial interactions and in [27] in the case of linear particle interactions and single-trajectory data (i.e. one instance of the particle system). The authors of [28] recently developed an online maximum likelihood method for inference of IPS, and in [29] maximum likelihood is applied to parameter estimation in an IPS for pedestrian flow. It should also be noted that parameter estimation for IPS is common in biological sciences, with the most frequently used technique being nonlinear least squares with a cost function comprised of summary statistics [7,30].

Problem **(P)** is made challenging by the coupled effects of K, V, and σ. In each of the previously mentioned algorithms, the assumption is made that σ is known and/or that K takes a specific form (radial or linear). In addition, the maximum likelihood-based approach approximates the differential $dX_t^{(i)}$ of particle i using a 1st-order finite difference: $dX_t^{(i)} \approx X_{t+\Delta t}^{(i)} - X_t^{(i)}$, which is especially ill-suited to problems involving *extrinsic* noise in the particle positions. Our primary goal is to show that the weak-form sparse regression framework allows for identification of the full model (K, V, σ), with significantly reduced computational complexity, when N is on the order of several thousands or more. We use a two-step process: the density of particles is approximated using a density kernel G and then the WSINDy algorithm (weak-form sparse identification of nonlinear dynamics) is applied in the PDE setting [2,3]. WSINDy is a modified version of the original SINDy algorithm [1,31] where the weak formulation of the dynamics is enforced using a family of test functions that offers reduced computational complexity, high-accuracy recovery in low-noise regimes, and increased robustness to high-noise scenarios. The feasibility of this approach for IPS is grounded in the convergence of IPS to associated mean-field equations. The reduction in computational complexity follows from the reduction in evaluation of candidate potentials (as discussed in Section 4.2), as well as the convolutional nature of the weak-form algorithm.

To the best of our knowledge, we present here the first *weak-form sparse regression* approach for inference of interacting particle systems, however we now review several related approaches that have recently been developed. In [32], the authors learn local hydrodynamic equations from active matter particle systems using the SINDy algorithm in the strong-form PDE setting. In contrast to [32], our approach learns nonlocal equations using the weak-form, however similarly to [32] we perform model selection and inference of parameters using sparse regression at the continuum level. The weak form provides an advantage because no smoothness is required on the particle density (for requisite smoothness the authors of [32] use a Gaussian kernel, which is more expensive to compute than simple particle binning as done here). The authors of [33] developed an integral formulation for inference of plasma physics models from PIC data using SINDy, however their method involves first computing strong-form derivatives and then averaging, rather than integration by parts against test functions as done here, and as in [32], the

[3] By $\mathbf{I}_d$ we mean the identity in $\mathbb{R}^d$.

Table 1
Notations used throughout.

Variable	Definition	Domain
K	Pairwise interaction potential	$L^1_{loc}(\mathbb{R}^d, \mathbb{R})$
V	Local potential	$C(\mathbb{R}^d, \mathbb{R})$
σ	Diffusivity	$C(\mathbb{R}^d, \mathbb{R}^{d\times d})$
N	Number of particles per experiment	$\{2, 3, \dots\}$
d	Dimension of latent space	$\mathbb{N}$
T	Final time	$(0, \infty)$
$(\Omega, \mathcal{B}, \mathbb{P}, (\mathcal{F}_t)_{t\geq 0})$	Filtered probability space	
$(B_t^{(i)})_{i=1}^N$	Independent $\mathbb{R}^d$ Brownian motions on $(\Omega, \mathcal{B}, \mathbb{P}, (\mathcal{F}_t)_{t\geq 0})$	
$X_t^{(i)}$	ith particle in the particle system (1.1) at time t	$\mathbb{R}^d$
$\mathbf{X}_t$	N-particle system (1.1) at time t	$\mathbb{R}^{Nd}$
μ_t^N	Empirical measure of $\mathbf{X}_t$	$\mathcal{P}(\mathbb{R}^d)$
F_t^N	Distribution of $\mathbf{X}_t$	$\mathcal{P}(\mathbb{R}^{Nd})$
X_t	Mean-field process (3.2) at time t	$\mathbb{R}^d$
μ_t	Distribution of X_t	$\mathcal{P}(\mathbb{R}^d)$
$\mathbf{t}$	L discrete timepoints	$[0, T]$
$\mathbb{X}_\mathbf{t}$	Collection of M independent samples of $\mathbf{X}_t$ at $\mathbf{t}$	$\mathbb{R}^{MLNd}$
$\mathbb{Y}_\mathbf{t}$	Sample of $\mathbf{X}_\mathbf{t}$ corrupted with i.i.d. additive noise	$\mathbb{R}^{MLNd}$
U_t	Approximate density from particle positions	$\mathcal{P}(\mathbb{R}^d)$
G	Density kernel mapping μ_t^N to U_t	$L^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$
$\mathcal{D}$	Spatial support of U_t, $t \in [0, T]$	Compact subset of $\mathbb{R}^d$
$\mathbf{C}$	Discretization of $\mathcal{D}$	
$\mathbf{U}_t$	Discrete approximate density $U_t(\mathbf{C})$	
$\langle \cdot, \cdot \rangle_h$	semi-discrete inner product, trapezoidal rule over $\mathbf{C}$	
$\langle \cdot, \cdot \rangle_{h,\Delta t}$	Fully-discrete inner product, trapezoidal rule over $\mathbf{C} \times \mathbf{t}$	
$\mathbb{L}_K$	Library of candidate interaction forces	
$\mathbb{L}_V$	Library of candidate local forces	
$\mathbb{L}_\sigma$	Library of candidate diffusivities	
$\mathbb{L}$	$(\mathbb{L}_K, \mathbb{L}_V, \mathbb{L}_\sigma)$	
Ψ	Set of n test functions $(\psi_k)_{k=1}^n$	$C^2(\mathbb{R}^d \times (0, T))$
$\phi_{m,p}(v;\ \Delta)$	Test functions used in this work (Eq. (4.4))	
$\boldsymbol{\lambda}$	Set of sparsity thresholds	
$\mathcal{L}$	Loss function for sparsity thresholds (Eq. (4.6))	

learned models are local. In [34], the authors apply the maximum likelihood approach in the continuum setting on the underlying nonlocal Fokker–Planck equation and learn directly the nonlocal PDE using strong-form discretizations of the dynamics. While we similarly use the continuum setting for inference (albeit in weak form), our approach differs from [34] in that it is designed for the more realistic setting of discrete-time *particle* data, rather than pointwise data on the particle *density* (assumed to be smooth in [34]).

2.1. Contributions

The purpose of the present article is to show that the weak form provides an advantage in speed and accuracy compared with existing inference methods for particle systems when the number of particles is sufficiently large (on the order of several thousand or more). The key points of this article include:

(I) Formulation of a weak-form sparse recovery algorithm for simultaneous identification of the particle interaction force K, local potential V, and diffusivity σ from discrete-time particle data.
(II) Convergence with rate $\mathcal{O}(N^{-1/2})$ of the resulting full-rank least-squares solution as the number of particles $N \to \infty$ and timestep $\Delta t \to 0$.
(III) Numerical illustration of (II) along with robustness to either intrinsic randomness (e.g. Brownian motion) or extrinsic randomness (e.g. additive measurement noise).

2.2. Paper organization

In Section 3 we review results from mean-field theory used to show convergence of the weak-form method. In Section 4 we introduce the WSINDy algorithm applied to interacting particles, including hyperparameter selection, computational complexity, and convergence of the method under suitable assumptions in the limit of large N. Section 5 contains numerical examples exhibiting the convergence rates of the previous section and examining the robustness of the algorithm to various sources of corruption, and Section 6 contains a discussion of extensions and future directions. In the Appendix we provide information on the hyperparameters used A.1, derivation of the homogenized equation (5.3) A.2, results and discussion for the case of small N and large M (in comparison with [26]) A.3, and proofs to technical lemmas A.4. Table 1 includes a list of notations used throughout.

3. Review of mean-field theory

Our weak-form approach utilizes that under fairly general assumptions the empirical measure μ_t^N of the process $\mathbf{X}_t$ defined in (1.1) converges weakly to μ_t, the distribution of the associated mean-field process X_t defined in (3.2). Specifically, under suitable assumptions on V, K, σ, and μ_0, there exists $T > 0$ such that for all $t \in [0, T]$, the mean-field limit[4]

$$\lim_{N\to\infty} \mu_t^N = \mu_t$$

holds in the weak topology of measures,[5] where μ_t is a weak-measure solution to the mean-field dynamics

$$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla K * \mu_t) + \nabla \cdot (\mu_t \nabla V) + \frac{1}{2}\sum_{i,j=1}^d \frac{\partial^2}{\partial x_i \partial x_j}\left(\sigma\sigma^T \mu_t\right), \quad \mu_0 \in \mathcal{P}_2(\mathbb{R}^d). \tag{3.1}$$

[4] We use the notation $t \to \mu_t$ to denote the evolution of probability measures. Subscripts will not be used to denote differentiation.

[5] Meaning that for all continuous bounded functions $\phi : \mathbb{R}^d \to \mathbb{R}$, $\int_{\mathbb{R}^d} \phi(x) d\mu_t^N(x) \to \int_{\mathbb{R}^d} \phi(x) d\mu_t(x)$.

Eq. (3.1) describes the evolution of the distribution of the McKean–Vlasov process

$$dX_t = -\nabla K * \mu_t\,(X_t)\,dt - \nabla V\,(X_t)\,dt + \sigma(X_t)\,dB_t. \tag{3.2}$$

This implies that as $N \to \infty$, an initially correlated particle system driven by pairwise interaction becomes uncorrelated and only interacts with its mean-field distribution μ_t. In particular, the following theorem summarizes several mean-field results taken from the review article [35] with proofs in [36,37].[6]

Theorem *([35–37]). Assume that ∇K is globally Lipschitz, $V = 0$, and $\sigma(x) = \sigma = const$. In addition assume that $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then for any $T > 0$, for all $t \leq T$ it holds that*

(i) There exists a unique solution (X_t, μ_t) where X_t is a strong solution to (3.2) and μ_t is a weak-measure solution to (3.1).

(ii) For any $\phi \in C_b^1(\mathbb{R}^d)$,

$$\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \phi(X_t^{(i)}) - \int_{\mathbb{R}^d}\phi(x)d\mu_t(x)\right)^2\right] \leq C\frac{\|\phi\|_{C^1}^2}{N} \tag{3.3}$$

with C depending on $Lip(\nabla K)$ and T.

(iii) For any $k \in \mathbb{N}$, a.e. $-\,t < T$, the k-particle marginal

$$\rho_t^{(k),N}(x_1, \dots, x_k) := \int_{\mathbb{R}^{d(N-k)}} F_t^N(x_1, \dots, x_k, x_{k+1}, \dots, x_N) \times dx_{k+1}\cdots dx_N$$

converges weakly to $\mu_t^{\otimes k}$ as $N \to \infty$, where $F_t^N \in \mathcal{P}(\mathbb{R}^{Nd})$ is the distribution of $\mathbf{X}_t$.

The previous result immediately extends to the case of ∇V and σ both globally Lipschitz and has been extended to ∇K only locally-Lipschitz in [38], ∇K with Coulomb-type singularity at the origin in [39], and domains with boundaries in [40,41]. Analysis of the model (3.1) continues to evolve in various contexts, including analysis of equilibria [42–44] and connections to deep learning [45]. For our convergence result below we simply assume that $K^\star, V^\star, \sigma^\star$ and μ_0 are such that (i) and (ii) from the above theorem hold.

3.1. Weak form

Despite the $\mathcal{O}(N^{-1/2})$ convergence of the empirical measure in previous theorem, it is unclear at what particle number N the mean-field equations become a suitable framework for inference using particle data, due to the complex variance structure at any finite N. A key piece of the present work is to show that the weak form of the mean-field equations does indeed provide a suitable setting when N is at least several thousands. Moreover, since in many cases (3.1) can only be understood in a weak sense, the weak form is the natural framework for identification. We say that μ_t is a weak solution to (3.1) if for any $\psi \in C^2(\mathbb{R}^d \times (0,T))$ compactly supported it holds that

$$\begin{aligned}\int_0^T\int_{\mathbb{R}^d}\partial_t\psi(x,t)\,d\mu_t(x)dt &= \int_0^T\int_{\mathbb{R}^d}\Big(\nabla\psi(x,t)\cdot\nabla K * \mu_t(x) + \nabla\psi(x,t)\cdot\nabla V(x) \\ &\quad - \frac{1}{2}\mathrm{Tr}\left(\nabla^2\psi(x,t)\sigma(x)\sigma^T(x)\right)\Big)\,d\mu_t(x)dt,\end{aligned} \tag{3.4}$$

where $\nabla^2\psi$ denotes the Hessian of ψ and $\mathrm{Tr}(\mathbf{A})$ is the trace of the matrix $\mathbf{A}$. Our method requires discretizing (3.4) for all $\psi \in \Psi$ where $\Psi = (\psi_1, \dots, \psi_n)$ is a suitable test function basis, and approximating the mean-field distribution μ_t with a density U_t constructed from discrete particle data at time t. We then find K, V, and σ within specified finite-dimensional function spaces.

4. Algorithm

We propose the general Algorithm 4.1 for discovery of mean-field equations from particle data. The inputs are a discrete-time sample $\mathbb{Y}$ containing M experiments each with N particle positions over L timepoints $\mathbf{t} = (t_1, \dots, t_L)$. The following hyperparameters are defined by the user: (i) a kernel G used to map the empirical measure μ_t^N to an approximate density U_t, (ii) a spatial grid $\mathbf{C}$ over which to evaluate the approximate density $\mathbf{U}_t = U_t(\mathbf{C})$, (iii) a library of trial functions $\mathbb{L} = \{\mathbb{L}_K, \mathbb{L}_V, \mathbb{L}_\sigma\} = \{(K_j)_{j=1}^{J_K}, (V_j)_{j=1}^{J_V}, (\sigma_j)_{j=1}^{J_\sigma}\}$, (iv) a basis of test functions $\Psi = (\psi_k)_{k=1}^n$, (v) a quadrature rule over the spatiotemporal grid $(\mathbf{C}, \mathbf{t})$ denoted by an inner product $\langle\cdot,\cdot\rangle$, and (vi) sparsity factors $\boldsymbol{\lambda}$ for the modified sequential thresholding least-squares Algorithm 4.2 (MSTLS) reviewed below. We discuss choices of these hyperparameters in Section 4.1, computational complexity of the algorithm in Section 4.2, convergence of the algorithm in Section 4.3. In Section 4.4 we briefly discuss gaps between theory and practice. Table 1 includes a list of notations used throughout.

Algorithm 4.1 WSINDy for identifying mean-field Eq. (3.1) from particle data $\mathbb{Y}$

$(\widehat{\mathbf{w}}, \hat{\lambda}) =$ **WSINDy** $(\mathbb{Y}, \mathbf{t};\ G, \mathbf{C}, \mathbb{L}, \Psi, \langle\cdot,\cdot\rangle, \boldsymbol{\lambda})$

```
for $\ell = 1 : L$ do
    for $m = 1 : M$ do
        $\mathbf{U}_\ell^{(m)} = \int_{\mathbb{R}^d} G(\mathbf{C}, y)d\mu_{t_\ell}^N(y)$ where $\mu_{t_\ell}^N$ is the empirical measure for $\mathbf{Y}_{t_\ell}^{(m)}$
    end for
    $\mathbf{U}_\ell = \frac{1}{M}\sum_{m=1}^M \mathbf{U}_\ell^{(m)}$
end for

for $j = 1 : J_K$ do
    for $k = 1 : n$ do
        $\mathbf{G}_{kj}^K = \langle\nabla\psi_k, \mathbf{U}\nabla K_j * \mathbf{U}\rangle$
    end for
end for

for $j = 1 : J_V$ do
    for $k = 1 : n$ do
        $\mathbf{G}_{kj}^V = \langle\nabla\psi_k, \mathbf{U}\nabla V_j\rangle$
    end for
end for

for $j = 1 : J_\sigma$ do
    for $k = 1 : n$ do
        $\mathbf{G}_{kj}^\sigma = \frac{1}{2}\sum_{p,q=1}^d \langle\partial_{x_p x_q}\psi_k, (\sigma_j\sigma_j^T)_{pq}\mathbf{U}\rangle$
    end for
end for
$\mathbf{G} = [\mathbf{G}^K\ \mathbf{G}^V\ \mathbf{G}^\sigma]$

for $k = 1 : n$ do
    $\mathbf{b}_k = \langle\partial_t\psi_k, \mathbf{U}\rangle$
end for

$(\widehat{\mathbf{w}}, \widehat{\lambda}) = \mathrm{MSTLS}(\mathbf{G}, \mathbf{b};\ \boldsymbol{\lambda})$    (see Algorithm 4.2)
```

[6] For a function $f : \mathbb{R}^d \to Y$, where Y is a metric space with metric ρ, we define Lip(f) by

$$\mathrm{Lip}(f) := \sup_{x,y\in\mathbb{R}^d}\frac{\rho(f(x), f(y))}{|x-y|}$$

where $|\cdot|$ denotes the Euclidean norm. We say f is Lipschitz when Lip$(f) < \infty$. Also, $\|f\|_{C^1} := \|f\|_\infty + \sum_{i=1}^d \left\|\frac{\partial f}{\partial x_i}\right\|_\infty$.

4.1. Hyperparameter selection

4.1.1. Quadrature

We assume that the set of gridpoints $\mathbf{C}$ in Algorithm 4.1 is chosen from some compact domain $\mathcal{D} \subset \mathbb{R}^d$ containing $\operatorname{supp}(\mathbb{Y})$. The choice of $\mathbf{C}$ (and $\mathcal{D}$) must be chosen in conjunction with the quadrature scheme, which includes integration in time using the given timepoints $\mathbf{t}$ as well as space. For completeness, the inner products in lines 10, 16, 22, and 27 of Algorithm 4.1 are defined in the continuous setting by

$$\langle f, g \rangle = \int_0^T \int_{\mathcal{D}} f(x,t)g(x,t)dxdt,$$

and the convolution in line 10 is defined by

$$\nabla K_j * U_t(x) = \int_{\mathcal{D}} \nabla K_j(x-y)U_t(y)dy.$$

In the present work we adopt the scheme used in the application of WSINDy for local PDEs [3], which includes the trapezoidal rule in space and time with test functions ψ compactly supported in $\mathcal{D} \times (0,T)$. We take $\mathcal{D}$ to be a rectangular domain enclosing $\operatorname{supp}(\mathbb{Y})$ and $\mathbf{C} \subset \mathcal{D}$ to be equally-spaced in order to efficiently evaluate convolution terms. In what follows we denote by $\langle \cdot, \cdot \rangle$ the continuous inner product, $\langle \cdot, \cdot \rangle_h$ the inner product over $\mathcal{D} \times [0,T]$ evaluated using the composite trapezoidal rule in space with meshwidth h and Lebesgue integration in time, and by $\langle \cdot, \cdot \rangle_{h,\Delta t}$ the trapezoidal rule in both space and time, with meshwidth h in space and Δt in time. With some abuse of notation, $f*g$ will denote the convolution of f and g, understood to be discrete or continuous by the context. Note also that we denote by μ^N, μ, and U the measures over $\mathbb{R}^d \times [0,T]$ defined by $\mu_t^N \Lambda_{[0,T]}$, $\mu_t \Lambda_{[0,T]}$ and $U_t \Lambda_{[0,T]}$, respectively, where $\Lambda_{[0,T]}$ is the Lebesgue measure on $[0,T]$.

4.1.2. Density kernel

Having chosen the domain $\mathcal{D} \subset \mathbb{R}^d$ containing the particle data $\mathbb{Y}$, let $P^h = \{B_k\}_k$ be a partition of $\mathcal{D}$ ($\cup_k B_k = \mathcal{D}$) with h indicating the size of the atoms B_k. For the remainder of the paper we take B_k to be hypercubes of equal side length h in order to minimize computation time for integration, although this is by no means necessary. For particle positions $\mathbf{X}_t$, we define the histogram[7]

$$U_t = \sum_k \frac{1}{|B_k|} \mathbb{1}_{B_k}(x) \left(\frac{1}{N} \sum_i \mathbb{1}_{B_k}(X_t^{(i)}) \right) = \int_{\mathcal{D}} G(x,y) d\mu_t^N(y). \quad (4.1)$$

Here the *density kernel* is defined

$$G(x,y) = \sum_k \frac{1}{|B_k|} \mathbb{1}_{B_k}(x) \mathbb{1}_{B_k}(y),$$

and in this setting the corresponding spatial grid $\mathbf{C} = (\mathbf{c}_k)_k$ is the set of center-points of the bins B_k, from which we define the discrete histogram data $\mathbf{U}_t = U_t(\mathbf{C})$. The discrete histogram $\mathbf{U}_t$ then serves as an approximation to the mean-field distribution μ_t.

Pointwise estimation of densities from samples of particles usually requires large numbers of particles to achieve reasonably low variance, and in general the variance grows inversely proportional to the bin width h. One benefit of the weak form is that integrating against a histogram U does not suffer from the same increase in variance with small h. In particular,

Lemma 1. *Let $(Y^{(1)}, Y^{(2)}, \dots)$ be a sequence of $\mathbb{R}^d$-valued random variables such that the empirical measure μ^N of $\mathbf{Y} := (Y^{(1)}, \dots, Y^{(N)})$ converges weakly to $\mu \in \mathcal{P}(\mathbb{R}^d)$ according to*

$$\mathbb{E}\left[(\langle \psi, \mu^N \rangle - \langle \psi, \mu \rangle)^2 \right] \leq C \left\| \psi \right\|_{C^1}^2 N^{-1} \quad (4.2)$$

for all $\psi \in C^1(\mathbb{R}^d)$ and C a universal constant. Let U be the histogram computed with kernel G using (4.1) with n bins and equal sidelength h. Then for any ψ in $C^1(\mathbb{R}^d)$ compactly supported in $\mathcal{D}$, we have the mean-squared error (for $\widetilde{C}$ depending on C and d)

$$\mathbb{E}\left[\left(\langle \psi, U \rangle_h - \langle \psi, \mu \rangle \right)^2 \right] \leq \widetilde{C} \left\| \psi \right\|_{C^1}^2 \left(h^2 + N^{-1} \right).$$

Remark 1. We note that (4.2) follows immediately for $Y^{(i)} \sim \mu$ i.i.d.,[8] and also for $\mathbf{Y} = \mathbf{X}_t$ a solution to (1.1) at time t with mean-field distribution $\mu = \mu_t$ according to (3.3) (for suitable K, V, and σ), which is the setting of the current article.

Proof of Lemma 1. First we note that by compact support of ψ, the trapezoidal rule can be written

$$\langle \psi, U \rangle_h = \left\langle \psi, \int_{\mathbb{R}^d} G(\cdot, y) d\mu^N(y) \right\rangle_h = \left\langle \psi^{\mathbf{C}}, \mu^N \right\rangle = \frac{1}{N} \sum_{i=1}^N \psi^{\mathbf{C}}(Y^{(i)})$$

where the midpoint approximation $\psi^{\mathbf{C}}$ of ψ is given by

$$\psi^{\mathbf{C}}(x) = \sum_{k=1}^K \psi(c_k) \mathbb{1}_{B_k}(x). \quad (4.3)$$

Hence we simply split the error and use (4.2):

$$\begin{aligned} \mathbb{E}\left[\left(\langle \psi, U \rangle_h - \langle \psi, \mu \rangle \right)^2 \right] &\leq 2\mathbb{E}\left[\langle \psi^{\mathbf{C}} - \psi, \mu^N \rangle^2 \right] \\ &\quad + 2\mathbb{E}\left[(\langle \psi, \mu^N \rangle - \langle \psi, \mu \rangle)^2 \right] \\ &\leq \left\| \psi \right\|_{C^1}^2 \left(\frac{d}{2} h^2 + 2CN^{-1} \right). \end{aligned}$$

The previous lemma in particular shows that small bin width h does not negatively impact $\langle \psi, U \rangle_h$ as an estimator of $\langle \psi, \mu \rangle$, which is in contrast to $U(x)$ as a pointwise estimator of $\mu(x)$. For example, if we assume that $\mathbf{Y}$ is sampled from a C^1 density μ, it is well known that the mean-square optimal bin width is $h = \mathcal{O}(N^{-1/3})$ [46]. Summarizing this result, elementary computation reveals the pointwise bias for $x \in B_k$,

$$\operatorname{bias}(U(x)) = \mathbb{E}\left[U(x) \right] - \mu(x) = \frac{\mu(B_k)}{|B_k|} - \mu(x) := \mu(\xi) - \mu(x)$$

for some $\xi \in B_k$. Letting $L_k = \max_{x \in B_k} |\nabla \mu(x)|$, we have

$$\operatorname{bias}(U(x))^2 \leq L_k^2 2^{d-1} h^2.$$

For the variance we get

$$\operatorname{Var}\left(U(x) \right) = \frac{1}{N} \frac{\mu(B_k)(1-\mu(B_k))}{|B_k|^2} = \frac{\mu(\xi)}{N} \left(1 - \mu(B_k) \right) \frac{1}{\sqrt{2^{d-1}} h},$$

and hence a bound for the mean-squared error

$$\mathbb{E}\left[(U(x) - \mu(x))^2 \right] \leq L_k^2 2^{d-1} h^2 + \frac{\mu(\xi)}{N\sqrt{2^{d-1}}} h^{-1}.$$

Minimizing the bound over h we find an approximately optimal bin width

$$h^* = \left(\frac{\rho(\xi)}{2^{\frac{3d-1}{2}} L_k^2} \right)^{1/3} N^{-1/3} = \mathcal{O}(N^{-1/3}),$$

[7] The indicator function is defined $\mathbb{1}_A(x) := \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$.

[8] In this case (4.2) is the variance of a Monte-Carlo estimator for $\int \psi(x) d\mu(x)$.

which provides an overall pointwise root-mean-squared error of $\mathcal{O}(N^{-1/3})$. Hence, not only does the weak form remove the inverse h dependence in the variance, but fewer particles are needed to accurately approximate integrals of the density μ.

4.1.3. Test function basis

For the test functions $(\psi_k)_{1\leq k\leq n}$ we use the same approach as the PDE setting [3], namely we fix a *reference test function* ψ and set

$$\psi_k(x,t)=\psi(\mathbf{x}_k-x,t_k-t)$$

where $\mathcal{Q}:=\{(\mathbf{x}_k,t_k)\}_{1\leq k\leq n}$ is a fixed set of *query points*. This, together with a separable representation

$$\psi(x,t)=\phi_1(x_1)\cdots\phi_d(x_d)\phi_{d+1}(t),$$

enables construction of the linear system $(\mathbf{G},\mathbf{b})$ using the FFT. We choose ϕ_j, $1\leq j\leq d+1$, of the form

$$\phi_{m,p}(v;\ \Delta):=\max\left(1-\left(\frac{v}{m\Delta}\right)^2,0\right)^p \tag{4.4}$$

where m is the integer *support parameter* such that $\phi_{m,p}$ is supported on $2m+1$ points of spacing $\Delta\in\{h,\Delta t\}$ and $p\geq 1$ is the *degree* of $\phi_{m,p}$. For simplicity we set $\phi_j=\phi_{m_x,p_x}$ for $1\leq j\leq d$ and $\phi_{d+1}=\phi_{m_t,p_t}$, so that only the numbers m_x,p_x,m_t,p_t need to be specified.

Since $\phi_{m,p}$ has exactly p weak derivatives, p_x and p_t must be at least as large as the maximum spatial and temporal derivatives appearing in the library $\mathbb{L}$, or $p_x\geq 2$, $p_t\geq 1$. Larger p results in higher-accuracy enforcement of the weak form (3.4) in low-noise situations (see Lemma 2 of [2] for details), however the convergence analysis below indicates that smaller $\mathrm{Lip}(\partial^\alpha\psi)$, $|\alpha|\leq 2$, may reduce variance. The support parameter m determines the length and time scales of interest and must be chosen small enough to extract relevant scales yet large enough to sufficiently reduce variance.

In [3, Appendix A] the authors developed a changepoint algorithm to choose m_x,m_t,p_x,p_t automatically from the Fourier spectrum of the data $\mathbf{U}$. Here, for each of the three examples in Section 5, we fix ψ across all particle numbers N, extrinsic noises ε, and intrinsic noises σ, in order to instead focus on convergence in N. To strike a balance between accuracy and small $\mathrm{Lip}(\psi)$ we choose $p_t=3$ and $p_x=5$ throughout. We used a combination of the changepoint algorithm and manual tuning to arrive at m_x and m_t which work well across all noise levels and numbers of particles examined. Query points $\mathcal{Q}$ are taken to be an equally-spaced subgrid of $\mathbf{C}$ with spacing s_x and s_t for spatial and temporal coordinates. The resulting values p_x, p_t, m_x, m_t, s_x, and s_t determine the weak discretization scheme and can be found in Appendix A.1 for each example below.

The results in Section 5 appear robust to $3\leq p_x,p_t\leq 9$. In addition, choosing m_x and m_t specific to each dataset $\mathbb{Y}$ using the changepoint method often improves results. Although automated in the changepoint algorithm, we recommend visualizing the overlap between the Fourier spectra of ψ and $\mathbf{U}$ when choosing p_x,p_t,m_x,m_t in order to directly observe which the modes in the data will experience filtering under convolution with ψ. In general, there is much flexibility in the choice of ψ. Optimizing ψ continues to be an active area of research.

4.1.4. Trial function library

The general Algorithm 4.1 does not impose a radial structure for the interaction potential K, nor does it assume any prior knowledge that the particle system is in fact interacting. In the examples below,[9] the libraries $\mathbb{L}_K,\mathbb{L}_V,\mathbb{L}_\sigma$ are composed of monomial and/or trigonometric terms to demonstrate that sparse regression is effective in selecting the correct combination of nonlocal drift, local drift, and diffusion terms. Rank deficiency can result, however, from naive choices of nonlocal and local bases. Consider the kernel $K(x)=\frac{1}{2}|x|^2$, which satisfies

$$\nabla K*\mu_t=x-M_1(\mu_t)=\nabla V(x)$$

where $V(x)=\frac{1}{2}|x-M_1(\mu_t)|^2$ and $M_1(\mu_t)$ is the first moment of μ_t. Since $M_1(\mu_t)$ is conserved in the model (3.2) posed in free-space,[10] including the same power-law terms in both libraries $\mathbb{L}_K$ and $\mathbb{L}_V$ will lead to rank deficiency. This is easily avoided by incorporating known symmetries of the model (3.2), however in general we recommend that the user builds the library $\mathbb{L}$ incrementally and monitors the condition number of $\mathbf{G}$ while selecting terms.

4.1.5. Sparse regression

As in [3], we enforce sparsity using a *modified* sequential thresholding least-squares algorithm (MSTLS), included as Algorithm 4.2, where the "modifications" are two-fold. First, we incorporate into the thresholding step the magnitude of the overall term $\left\|\mathbf{w}_j\mathbf{G}_j\right\|_2$ as well as the coefficient magnitude $|\mathbf{w}_j|$, by defining non-uniform lower and upper thresholds

$$\begin{cases} L_j^\lambda=\lambda\max\left\{1,\ \dfrac{\|\mathbf{b}\|}{\|\mathbf{G}_j\|}\right\} \\ U_j^\lambda=\dfrac{1}{\lambda}\min\left\{1,\ \dfrac{\|\mathbf{b}\|}{\|\mathbf{G}_j\|}\right\}\end{cases},\qquad 1\leq j\leq\mathfrak{J}, \tag{4.5}$$

where $\mathfrak{J}=J_K+J_V+J_\sigma$ is the number of columns in $\mathbf{G}$. Second, we perform a grid search[11] over candidate sparsity parameters $\boldsymbol{\lambda}$ and choose the parameter $\widehat{\lambda}$ that is the smallest minimizer over $\boldsymbol{\lambda}$ of the cost function

$$\mathcal{L}(\lambda)=\frac{\left\|\mathbf{G}(\mathbf{w}^\lambda-\mathbf{w}^0)\right\|_2}{\left\|\mathbf{G}\mathbf{w}^0\right\|_2}+\frac{\left\|\mathbf{w}^\lambda\right\|_0}{\mathfrak{J}} \tag{4.6}$$

where $\mathbf{w}^\lambda$ is the output of the sequential thresholding algorithm with non-uniform thresholds (4.5) and $\mathbf{w}^0=\mathbf{G}^\dagger\mathbf{b}$ is the least-squares solution.[12] The final coefficient vector is then set to $\widehat{\mathbf{w}}=\mathbf{w}^{\widehat{\lambda}}$.

We now review some aspects of Algorithm 4.2. Results from [47] on the convergence of STLS carry over for the inner loop of Algorithm 4.2, namely if $\mathbf{G}$ is full-rank, the inner loop terminates in at most $\mathfrak{J}$ iterations with a resulting coefficient vector $\mathbf{w}^\lambda$ that is a local minimizer of the cost function $F(\mathbf{w})=\|\mathbf{G}\mathbf{w}-\mathbf{b}\|_2^2+\lambda^2\|\mathbf{w}\|_0$. This implies that the full algorithm terminates in at-most $m\mathfrak{J}$ least-squares solves (each on a subset of columns of $\mathbf{G}$).

When considering recovery of the true weight vector $\mathbf{w}^\star$, Theorem 1 implies convergence in particle number N of $\widehat{\mathbf{w}}$ to $\mathbf{w}^\star$ when $\mathbf{G}$ is full-rank. The rate of convergence depends implicitly on the condition number of $\mathbf{G}$, hence it is recommended that one builds the library $\mathbb{L}$ incrementally, stopping before the conditional number $\kappa(\mathbf{G})$ grows too large. If $\mathbf{G}$ is rank deficient, classical recovery guarantees from compressive sensing do not necessarily apply, due to high correlations between the columns of $\mathbf{G}$

[9] Details of the libraries used in examples can be found in Tables 2–4 in Appendix A.1.

[10] This is not true in domains with boundaries, where nonlocalities can be seen to impart mean translation [42].

[11] Note that this is feasible because the STLS algorithm terminates in finitely many iterations.

[12] The Moore–Penrose inverse $\mathbf{A}^\dagger$ is defined for a rank-r matrix $\mathbf{A}$ using the reduced SVD $\mathbf{A}=U_r\Sigma_rV_r^*$ as $\mathbf{A}^\dagger:=V_r\Sigma_r^{-1}U_r^*$. The subscript r denotes restriction to the first r columns.

Table 2
Trial function library for local 2D example (Section 5.1).

Mean-field term	Trial function library
$\nabla\cdot(U\nabla K*U)$	$\nabla\cdot(U\nabla\|x\|^m*U)$, $m\in\{1,2,3,4,5,6,7\}$
$\nabla\cdot(U\nabla V)$	$\partial_{x_i}(U\cos(mx_1)\cos(nx_2))$, $(m,n)\in\{0,1,2,3,4,5\}$, $i\in\{1,2\}$
$\frac{1}{2}\sum_{i,j=1}^d\frac{\partial^2(U\sigma\sigma^T)_{ij}}{\partial x_i\partial x_j}$	$\varDelta(U\cos(mx_1)\cos(nx_2))$, $(m,n)\in\{0,1,2,3,4,5\}$

Table 3
Trial function library for nonlocal 1D example (Section 5.2).

Mean-field term	Trial function library
$\nabla\cdot(U\nabla K*U)$	$\partial_x\cdot(U\partial_x\|x\|^m*U)$, $m\in\{1,2,3,4,5,6,7\}$
$\nabla\cdot(U\nabla V)$	$\partial_x(Ux^m)$, $m\in\{0,2,3,4,5,6,7,8\}$
$\frac{1}{2}\sum_{i,j=1}^d\frac{\partial^2(U\sigma\sigma^T)_{ij}}{\partial x_i\partial x_j}$	$\partial_{xx}(Ux^m)$, $m\in\{0,1,2,3,4,5,6,7,8\}$

Table 4
Trial function library for nonlocal 2D example (Section 5.3). Interaction potentials $[K]_\delta$ indicate cutoff potentials of the form (5.6) with $\delta=0.01$ such that the resulting potential is Lipschitz.

Mean-field term	Trial function library	
$\nabla\cdot(U\nabla K*U)$	$\nabla\cdot(U\nabla\|x\|^m*U)$ $\nabla\cdot(U\nabla\left[\|x\|^{1/2}\right]_\delta*U)$ $\nabla\cdot(U\nabla[\|x\|(\log\|x\|-1)]_\delta*U)$ $\nabla\cdot(U\nabla[\log\|x\|]_\delta*U)$	$m\in\{2,3,4,5,6\}$
$\nabla\cdot(U\nabla V)$	$\partial_{x_i}\left(Ux_1^mx_2^n\right)$	$0\le m+n\le 5$, $i\in\{1,2\}$
$\frac{1}{2}\sum_{i,j=1}^d\frac{\partial^2(U\sigma\sigma^T)_{ij}}{\partial x_i\partial x_j}$	$\varDelta(U\cos(mx_1)\cos(nx_2))$, $(m,n)\in\{0,1,2\}$	

Algorithm 4.2 Modified sequential thresholding with automatic threshold selection
$(\widehat{\mathbf{w}},\hat{\lambda})=\mathbf{MSTLS}(\mathbf{G}\in\mathbb{R}^{n\times\mathfrak{J}},\mathbf{b}\in\mathbb{R}^n,\boldsymbol{\lambda}\in\mathbb{R}^m,\text{maxits})$

```
$\mathbf{W}=\mathbf{0}\in\mathbb{R}^{\mathfrak{J}\times m}$
$\mathbf{w}^0=\mathbf{G}^\dagger\mathbf{b}$
for $i=1:m$ do
    $\lambda=\boldsymbol{\lambda}_i$
    $\ell=0$
    while $\ell<$ maxits do
        $\mathcal{I}^\ell=\{1\le j\le\mathfrak{J}\ :\ L_j^\lambda\le|\mathbf{w}_j^\ell|\le U_j^\lambda\}$   (Thresholding step: see Eq. (4.5))
        $\mathbf{w}^{\ell+1}=\text{argmin}_{\text{supp}(\mathbf{w})\subset\mathcal{I}^\ell}\ \|\mathbf{G}\mathbf{w}-\mathbf{b}\|_2^2$
        $\ell=\ell+1$
    end while
    $\mathbf{w}^\lambda=\mathbf{w}^\ell$
    $\mathbf{W}_{:,i}=\mathbf{w}^\lambda$
end for
$\widehat{\lambda}=\min\left(\text{argmin}_{\lambda\in\boldsymbol{\lambda}}\mathcal{L}(\lambda)\right)$   (Identificaiton of best $\lambda$: see Eq. (4.6))
$\widehat{\mathbf{w}}=\mathbf{w}^{\widehat{\lambda}}$
```

(recall each column is constructed from the same dataset $\mathbf{U}$).[13] One may employ additional regularization (e.g. Tikhonov regularization as in [31]); however, in general, improvements to existing sparse regression algorithms for rank-deficient, noisy, and highly-correlated matrices is an active area of research.

The bounds (4.5) enforce a quasi-dominant balance rule, such that $\left\|\mathbf{w}_j\mathbf{G}_j\right\|_2$ is within $-\log_{10}(\lambda)$ orders of magnitude from $\|\mathbf{b}\|_2$ and $|\mathbf{w}_j|$ is within $-\log_{10}(\lambda)$ orders of magnitude from 1 (the coefficient of time derivative $\partial_t\mu_t$). This is specifically designed to handle poorly-scaled data (see the Burgers and Korteweg–de Vries examples in [3]), however we leave a more thorough examination of the thresholding requirements necessary for models with multiple scales to future work.

As the sum of two relative errors, minimizers of the cost function $\mathcal{L}$ equally weight the accuracy and sparsity of $\mathbf{w}^{\widehat{\lambda}}$. By choosing $\widehat{\lambda}$ to be the smallest minimizer of $\mathcal{L}$ over $\boldsymbol{\lambda}$, we identify the thresholds $\lambda\in\boldsymbol{\lambda}$ such that $\lambda<\widehat{\lambda}$ as those resulting in an overfit model. We commonly choose $\boldsymbol{\lambda}$ to be log-equally spaced (e.g. 50 points from 10^{-4} to 1), and starting from a coarse grid, refine $\boldsymbol{\lambda}$ until the minimum of $\mathcal{L}$ is stationary.

4.2. Computational complexity

To compute convolutions against ∇K for each $K\in\mathbb{L}_K$, we first evaluate $(\partial_{x_i}K)_{1\le i\le d}$ at the grid $\mathbf{C}-\mathbf{C}$ defined by

$$\mathbf{C}-\mathbf{C}:=\{x\in\mathbb{R}^d\ :\ x=(i_1h,\dots,i_dh),\quad -n_\ell\le i_\ell\le n_\ell\},$$

where h is the spacing of $\mathbf{C}$ and n_ℓ, $1\le\ell\le d$, is the number of points in $\mathbf{C}$ along the ℓth coordinate. Computing[14] $\partial_{x_i}\mathbf{K}:=\partial_{x_i}K(\mathbf{C}-\mathbf{C})$ requires $2^d|\mathbf{C}|$ evaluations of K, where $|\mathbf{C}|=\prod_{\ell=1}^d n_\ell$ is the number of points in $\mathbf{C}$. We then use the d-dimensional FFT to compute the convolutions

$$\partial_{x_i}\mathbf{K}*\mathbf{U}_t\approx\partial_{x_i}K*U_t(\mathbf{C}),\quad t\in\mathbf{t}$$

where only entries corresponding to particle interactions within $\mathbf{C}$ are retained. For $d=1$ this amounts to $\mathcal{O}(|\mathbf{C}|\log|\mathbf{C}|)$ flops per timestep. For $d=2$ and higher dimensions, the d-dimensional FFT is considerably slower unless one of the arrays is separable. To enforce separability, trial interaction potentials in $\mathbb{L}_K$ can be chosen to be a sum of separable functions,

$$K(x)=\sum_{q=1}^Q k_{1,q}(x_1)\cdots k_{d,q}(x_d),\tag{4.7}$$

in which case only a series of one-dimensional FFTs are needed to compute $\partial_{x_i}\mathbf{K}*\mathbf{U}_t$, and again the cost is $\mathcal{O}(|\mathbf{C}|\log|\mathbf{C}|)$ per timestep. When K is not separable, a low-rank approximation can be computed from $\partial_{x_i}\mathbf{K}$,

$$\partial_{x_i}\mathbf{K}\approx\sum_{q=1}^Q\sigma_q\mathbf{k}_{1,q}\otimes\cdots\otimes\mathbf{k}_{d,q}\tag{4.8}$$

which again reduces convolutions to a series of one-dimensional FFTs. For $d=2$, this is accomplished using the truncated SVD, while for higher dimensions there does not exist a unique *best* rank-Q tensor approximation, although several efficient algorithms are available to compute a sufficiently accurate decomposition [49–51] (and the field of fast tensor decompositions is advancing rapidly).

We propose to compute convolutions by first computing a low-rank decomposition of $\partial_{x_i}\mathbf{K}$ using the randomized truncated SVD [52] or a suitable randomized tensor decomposition and then

[13] In particular, correlations result in large mutual incoherence, which renders algorithms such as Basis Pursuit, Orthogonal Matching Pursuit, and Hard Thresholding Pursuit useless (see [48, Chapter 5] for details).

[14] Note that $\mathbf{C}-\mathbf{C}$ is simply $\mathbf{C}$ shifted to lie in the positive orthant $\{x\in\mathbb{R}^d\ :\ x_\ell\ge 0,\quad 1\le\ell\le d\}$ and reflected through each coordinate plane $x_\ell=0$. In this way $\mathbf{C}-\mathbf{C}$ discretizes the set $\mathcal{D}-\mathcal{D}:=\{x-y\in\mathbb{R}^d\ :\ (x,y)\in\mathcal{D}\times\mathcal{D}\}$ containing all observed interparticle distances.

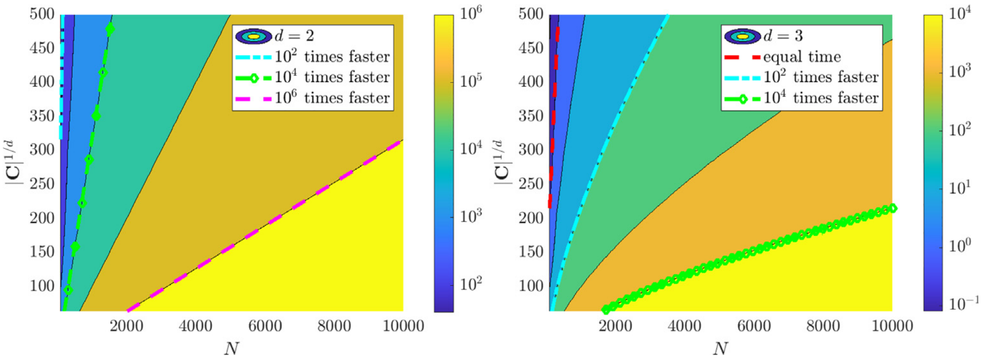

Fig. 1. Factor by which the mean-field evaluation of interaction forces using histograms reduces total function evaluations as a function of particle number N and average gridpoints per dimension $|\mathbf{C}|^{1/d}$ for data with $M = 10$ experiments each with $L = 100$ timepoints. For example, with $d = 2$ spatial dimensions (left) and $N > 2000$ particles, the number of function evaluations is reduced by at least a factor of 10^4.

applying the d-dimensional FFT as a series of one-dimensional FFTs. In the examples below we consider only $d = 1$ and $d = 2$, and leave extension to higher dimensions to future work.

Using low-rank approximations, the mean-field approach provides a significant reduction in computational complexity compared to direct evaluations of particle trajectories when N is sufficiently large. A particle-level computation of the nonlocal force in weak-form requires evaluating terms of the form

$$\sum_{\ell=1}^{L}\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\partial_x\psi(X^{(i)}_{t_\ell},t_\ell)\partial_xK(X^{(i)}_{t_\ell}-X^{(j)}_{t_\ell})\right)\Delta t$$
$$=\left\langle\partial_x\psi,\mu^N(\partial_xK*\mu^N)\right\rangle_{h,\Delta t}.$$

For a single candidate interaction potential K, a collection of J test functions ψ, and M experiments, this amounts to $MLN^2 + MLNJ$ function evaluations in $\mathbb{R}^d$ and $\mathcal{O}(MLN^2J)$ flops. If we use the proposed method, employing the convolutional weak form with a separable reference test function ψ (as in WSINDy for PDEs [3]) and exploiting a rank Q approximation of $\partial_x\mathbf{K}$ when computing convolutions against interaction potential, we instead evaluate

$$\partial_x\psi*(U(\partial_xK*U))$$

using $\mathcal{O}(LQ|\mathbf{C}|\log(|\mathbf{C}|))$ flops and only $2^d|\mathbf{C}|$ evaluations of ∂_xK, reused at each of the L timepoints.[15] Fig. 1 provides a visualization of the reduction in function evaluations for $L = 100$ timepoints and $M = 10$ experiments over a range of N and $|\mathbf{C}|^{1/d}$ (points along each spatial dimension when $|\mathbf{C}|$ is a hypercube) in $d = 2$ and $d = 3$ spatial dimensions. Table 5 in Appendix A.1 lists walltimes for the examples below, showing that with $N =$ 64,000 particles the full algorithm implemented in MATLAB runs in under 10 s with all computations in serial on a laptop with an AMD Ryzen 7 pro 4750u processor, and requiring less than 8 Gb of RAM. The dependence on N is only through the $\mathcal{O}(N)$ computation of the histogram, hence this approach may find applications in physical coarse-graining (e.g. of molecular dynamics or plasma simulations).

4.3. Convergence

We now show that the estimators $\widehat{K}$, $\widehat{V}$, and $\widehat{\sigma}$ of the weak-form method converge with a rate $\mathcal{O}(h + N^{-1/2} + \Delta t^{\eta})$ when ordinary least squares are used (i.e. $\boldsymbol{\lambda} = 0$) and only $M = 1$ experiment is available. Here $\eta > 0$ is the Hölder exponent of the sample paths of the process $\mathbf{X}_t$. We assume that $\mathcal{D}$, $\mathbf{C}$, G, P^h and the resulting histogram $\mathbf{U} = (\mathbf{U}_t)_{t\leq T}$ are as in Section 4.1.2. We make the following assumptions on the true model and resulting linear system throughout this section.

Assumption H. Let $p \geq 1$ be fixed.

(H.1) For each $N \geq 2$, $\mathbf{X}_t = (X^{(1)}_t, \dots, X^{(N)}_t)$ is a strong solution to (1.1) for $t \in [0, T]$, and for some $\eta > 0$ the sample paths $t \to X^{(i)}_t(\omega)$ are almost-surely η-Hölder continuous, i.e. for some $C_\eta > 0$,

$$|X^{(i)}_t(\omega) - X^{(i)}_s(\omega)| \leq C_\eta|t-s|^\eta, \quad \forall\, 0 \leq s \leq t \leq T,$$
$$\forall\, 1 \leq i \leq N, \quad \text{for a.e. } \omega \in \Omega.$$

(H.2) The initial particle distribution μ_0 satisfies the moment bound

$$\int_{\mathbb{R}^d}|x|^p d\mu_0(x) := M_p < \infty.$$

(H.3) $\nabla K^\star$ and $\nabla V^\star$ satisfy for some $C_p > 0$ the growth bound:

$$|\nabla V^\star(x) - \nabla V^\star(y)| + |\nabla K^\star(x) - \nabla K^\star(y)|$$
$$\leq C_p|x-y|(1+\max\{|x|,|y|\}^{p-1}), \quad x, y \in \mathbb{R}^d.$$

(H.4) For the same constant $C_p > 0$, it holds that[16]

$$\left\|\sigma^\star(x) - \sigma^\star(y)\right\|_F \leq C_p|x-y|^{1/2}$$
$$\times\;(1+\max\{|x|,|y|\}^{p/2-1/2}), \quad x, y \in \mathbb{R}^d$$

(H.5) The test functions $(\psi_k)_{1\leq k\leq n} \subset C^2(\mathbb{R}^d \times (0, T))$ are compactly supported and together with the library $\mathbb{L}$ are such that $\mathbf{G}$ has full column rank with[17] $\left\|\mathbf{G}^\dagger\right\|_1 \leq C_{\mathbf{G}}$ almost surely for some constant $C_{\mathbf{G}} > 0$.

(H.6) The true functions $K^\star$, $V^\star$, and $\sigma^\star$ are in the span of $\mathbb{L}$.

We will now define some notation and state some technical lemmas with proofs found in Appendix A.4. Define the weak-form operator

$$\mathscr{L}(\rho,\psi,\langle\cdot,\cdot\rangle)$$
$$:=\left\langle\partial_t\psi - \nabla\psi\cdot\nabla K^\star*\rho - \nabla\psi\cdot\nabla V^\star + \frac{1}{2}\mathrm{Tr}\left(\nabla^2\psi\sigma^\star(\sigma^\star)^T\right),\rho\right\rangle,\tag{4.9}$$

where $\rho = (\rho_t)_{t\leq T}$ is a curve in $\mathcal{P}_p(\mathbb{R}^d)$, ψ is a C^2 function compactly supported over $\mathbb{R}^d\times(0,T)$, and $\langle\cdot,\cdot\rangle$ is an inner product

[15] We neglect the cost of computing the histogram $\mathbf{U}$ and evaluating $\psi(\mathbf{C})$, together amounting to an additional $\mathcal{O}(NML + |\mathbf{C}|)$ flops, as these terms are lower order and reused in each column of $\mathbf{G}$ and $\mathbf{b}$.

[16] For $\mathbf{A} \in \mathbb{R}^{d\times d}$ the Frobenius norm is defined $\|\mathbf{A}\|_F = \sqrt{\mathrm{Tr}(\mathbf{A}^T\mathbf{A})}$

[17] $\left\|\mathbf{G}^\dagger\right\|_q$ is the induced matrix q-norm of $\mathbf{G}^\dagger$.

over $\mathbb{R}^d \times (0, T)$. If $\rho = (\mu_t)_{t\leq T}$ is a weak solution to (3.1) and $\langle\cdot,\cdot\rangle$ is the $L^2(\mathbb{R}^d)$ inner product then $\mathscr{L}(\rho, \psi, \langle\cdot,\cdot\rangle) = 0$. If instead $\rho = (\mu_t^N)_{t\leq T}$, then by Itô's formula $\mathscr{L}(\rho, \psi, \langle\cdot,\cdot\rangle)$ takes the form of an Itô integral, and we have the following:

Lemma 2. *Under Assumptions (H.1)–(H.5), there exists a constant $C > 0$ independent of N such that*

$$\mathbb{E}\left[\left|\mathscr{L}(\mu^N, \psi, \langle\cdot,\cdot\rangle)\right|\right] \leq \frac{C}{\sqrt{N}}.$$

Proof. See Appendix A.4.

With the following lemma, we can relate the histogram U to the empirical measure μ^N through $\mathscr{L}$ using the inner product $\langle\cdot,\cdot\rangle_h$ defined by trapezoidal-rule integration in space and continuous integration in time.

Lemma 3. *Under Assumptions (H.1)–(H.5), for C independent of N and h, it holds that*

$$\mathbb{E}\left[|\mathscr{L}(U, \psi, \langle\cdot,\cdot\rangle_h) - \mathscr{L}(\mu^N, \psi, \langle\cdot,\cdot\rangle)|\right] \leq Ch.$$

Proof. See Appendix A.4.

To incorporate discrete-time effects, we consider the difference between $\mathscr{L}(U, \psi, \langle\cdot,\cdot\rangle_h)$ and $\mathscr{L}(U, \psi, \langle\cdot,\cdot\rangle_{h,\Delta t})$, where recall that $\langle\cdot,\cdot\rangle_{h,\Delta t}$ denotes trapezoidal rule integration in space with meshwidth h and in time with sampling rate Δt.

Lemma 4. *Under Assumptions (H.1)–(H.5), for C independent of N, h, and Δt, it holds that*

$$\mathbb{E}\left[|\mathscr{L}(U, \psi, \langle\cdot,\cdot\rangle_h) - \mathscr{L}(U, \psi, \langle\cdot,\cdot\rangle_{h,\Delta t})|\right] \leq C(h + \Delta t^\eta).$$

Proof. See Appendix A.4.

The previous estimates directly lead to the following bound on the model coefficients $\widehat{\mathbf{w}}$:

Theorem 1. *Assume that Assumption H holds. Let $\widehat{\mathbf{w}}$ be the learned model coefficients and $\mathbf{w}^\star$ the true model coefficients. For C independent of N, h, and Δt it holds that*

$$\mathbb{E}\left[\left\|\widehat{\mathbf{w}} - \mathbf{w}^\star\right\|_1\right] \leq C\left(h + N^{-1/2} + \Delta t^\eta\right).$$

Proof. Using that $K^\star$, $V^\star$, and $\sigma^\star$ are in the span of $\mathbb{L}$ (H.6), we have that

$$\mathbf{b}_k = \langle \partial_t \psi_k, \mathbf{U}\rangle_{h,\Delta t} = \mathscr{L}(U, \psi_k, \langle\cdot,\cdot\rangle_{h,\Delta t}) + \mathbf{G}_k^T \mathbf{w}^\star := \mathbf{L}_k + \mathbf{G}_k^T\mathbf{w}^\star,$$

where $\mathbf{G}_k^T$ is the kth row of $\mathbf{G}$. From Lemmas 2–4 we have

$$\begin{aligned}\mathbb{E}\left[|\mathbf{L}_k|\right] &\leq \mathbb{E}\left[|\mathscr{L}(U, \psi_k, \langle\cdot,\cdot\rangle_{h,\Delta t}) - \mathscr{L}(U, \psi_k, \langle\cdot,\cdot\rangle_h)|\right] \\ &+ \mathbb{E}\left[|\mathscr{L}(U, \psi_k, \langle\cdot,\cdot\rangle_h) - \mathscr{L}(\mu^N, \psi_k, \langle\cdot,\cdot\rangle)|\right] \\ &+ \mathbb{E}\left[|\mathscr{L}(\mu^N, \psi_k, \langle\cdot,\cdot\rangle)|\right] \\ &\leq C'\left(h + N^{-1/2} + \Delta t^\eta\right).\end{aligned}$$

Using that $\mathbf{G}$ is full rank, it holds that $\widehat{\mathbf{w}} = \mathbf{G}^\dagger\mathbf{b} = \mathbf{G}^\dagger\mathbf{L} + \mathbf{w}^\star$, hence the result follows from the uniform bound on $\left\|\mathbf{G}^\dagger\right\|_1$ (H.5):

$$\mathbb{E}\left[\left\|\widehat{\mathbf{w}} - \mathbf{w}^\star\right\|_1\right] \leq \mathbb{E}\left[\|\mathbf{G}^\dagger\|_1 \|\mathbf{L}\|_1\right] \leq C' C_{\mathbf{G}}\left(h + N^{-1/2} + \Delta t^\eta\right). \quad \square$$

Under the assumption (H.6), an immediate corollary is

$$\begin{aligned}\mathbb{E}\Bigg[\ &\left\|K^\star - \widehat{K}\right\|_{L^2(\mathcal{D}-\mathcal{D})} + \left\|V^\star - \widehat{V}\right\|_{L^2(\mathcal{D})} \\ &+ \left\|\left\|\sigma^\star(\sigma^\star)^T - \widehat{\sigma}(\widehat{\sigma})^T\right\|_F\right\|_{L^2(\mathcal{D})}\Bigg] \\ &\leq C\left(h + N^{-1/2} + \Delta t^\eta\right),\end{aligned} \tag{4.10}$$

This follows from

$$\begin{aligned}\left\|K^\star - \widehat{K}\right\|_{L^2(\mathcal{D}-\mathcal{D})} &\leq \sum_{j=1}^{\mathfrak{J}} |\mathbf{w}_j^\star - \widehat{\mathbf{w}}_j|\, \left\|K_j\right\|_{L^2(\mathcal{D}-\mathcal{D})} \\ &\leq \left(\sup_j \left\|K_j\right\|_{L^2(\mathcal{D}-\mathcal{D})}\right)\left\|\mathbf{w}^\star - \widehat{\mathbf{w}}\right\|_1,\end{aligned}$$

and similarly for $\widehat{V}$ and $\widehat{\sigma}$. Finally, setting $h = N^{-\alpha}$ for $\alpha > 0$ will ensure convergence as $N \to \infty$ and $\Delta t \to 0$.

4.4. Theory vs. Practice

We now make several remarks about the practical performance of Algorithm 4.1 with respect to the theoretical convergence of Theorem 1.

Remark 2. An important case of Theorem 1 is $\sigma^\star = 0$, in which case μ_t^N itself is a weak-measure solution to the mean-field Eq. (3.1) and the algorithm returns, for $\eta \geq 2$, $\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_1 \leq C(h + \Delta t^\eta)$. This partially explains the accuracy observed for purely-extrinsic noise examples in Figs. 5 and 9. We note further that in the absence of noise ($\varepsilon = 0$ and $\sigma^\star = 0$, not included in this work) Algorithm 4.1 recovers systems to high accuracy similarly to WSINDy applied to local dynamical systems [2,3].

Remark 3. Algorithm 4.1 in general implements sparse regression, yet Theorem 1 deals with ordinary least squares. Since least squares is a common subroutine of many sparse regression algorithms (including the MSTLS algorithm used here), the result is still relevant to sparse regression. Lastly, the full-rank assumption on $\mathbf{G}$ implies that as $N \to \infty$ sequential thresholding reduces to least squares.

Remark 4. Theorem 1 assumes data from a single experiment ($M = 1$), while the examples below show that $M > 1$ experiments improve results. For any fixed $M > 1$, the $N \to \infty$ limit results in convergence, however, the N-fixed and $M \to \infty$ limit does not result in convergence, as this does not lead to the mean-field equations.[18] The examples below show that using $M > 1$ has a practical advantage, and in Appendix A.3 we demonstrate that even for small particle systems ($N = 10$) the large M regime yields satisfactory results.

Remark 5. Many interesting examples have non-Lipschitz ∇K, in particular a lack of smoothness at $x = 0$. If μ_t^N does not converge to a singular measure as $N \to \infty$, then the bound (A.4) holds for ∇K with a jump discontinuity at $x = 0$, where an additional $\mathcal{O}(h)$ term arises from pairwise interactions within an $\mathcal{O}(h)$ distance. The examples below are chosen in part to show that $\mathcal{O}(N^{-1/2})$ convergence holds for ∇K with jumps at the origin.

5. Examples

We now demonstrate the successful identification of several particle systems in one and two spatial dimensions as well as the $\mathcal{O}(N^{-1/2})$ convergence predicted in Theorem 1. In each case we use Algorithm 4.1 to discover a mean-field equation of the form (3.1) from discrete-time particle data. For each dataset we simulate the associated interacting particle system $\mathbf{X}_t$ given by (1.1) using the Euler–Maruyama scheme (initial conditions and timestep are given in each example). We assess the ability of

[18] Note that the opposite convergence holds for the algorithm introduced in [26]: N-fixed, $M \to \infty$ results in recovery of K.

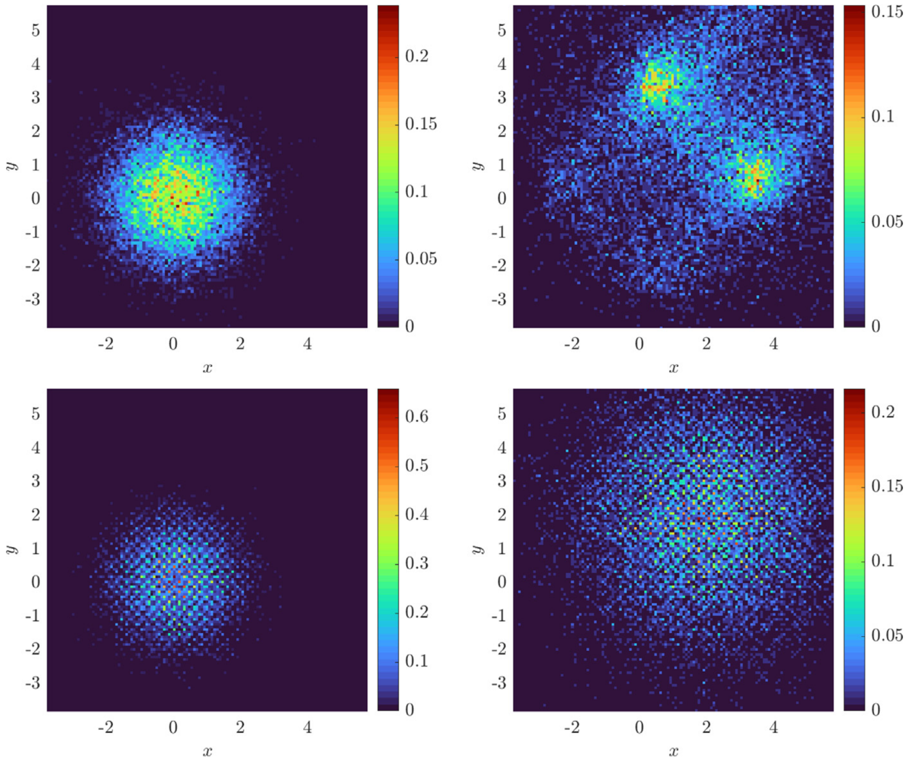

Fig. 2. Snapshots at time $t = 2\Delta t = 0.06$ (left) and $t = 100\Delta t = 2$ (right) of histograms computed with 128 bins in x and y from 16,384 particles evolving under (5.2) with $\omega = 1$ (top) and $\omega = 20$ (bottom).

WSINDy to select the correct model using the *true positivity ratio*[19]

$$\mathrm{TPR}(\widehat{\mathbf{w}}) = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} \tag{5.1}$$

where TP is the number of correctly identified nonzero coefficients, FN is the number of coefficients falsely identified as zero, and FP is the number of coefficients falsely identified as nonzero [53]. To demonstrate the $\mathcal{O}(N^{-1/2})$ convergence given by (4.10), for correctly identified models (i.e. $\mathrm{TPR}(\widehat{\mathbf{w}}) = 1$) we compute the relative ℓ_2-error of the recovered interaction force $\nabla\widehat{K}$, local force $\nabla\widehat{V}$, and diffusivity $\widehat{\sigma}$ over $\mathbf{C} - \mathbf{C}$ and $\mathbf{C}$, respectively, denoting this by $\|\cdot\|$ in the plots below. Results are averaged over 100 trials.

For the computational grid $\mathbf{C}$ we first compute the sample standard deviation s of $\mathbb{Y}$ and we choose $\mathcal{D}$ to be the rectangular grid extending $3s$ from the mean of $\mathbb{Y}$ in each spatial dimension. We then set $\mathbf{C}$ to have 128 points in x and y for $d = 2$ dimensions, and 256 points in x for $d = 1$, noting that these numbers are fairly arbitrary, and used to show that the grid need not be too large. We set the sparsity factors so that $\log_{10}(\boldsymbol{\lambda})$ contains 100 equally spaced points from -4 to 0. More information on the specifications of each example can be found in Appendix A.1. (MATLAB code used to generate examples is available at https://github.com/MathBioCU/WSINDy_IPS.)

5.1. Two-dimensional local model and homogenization

The first system we examine is a local model ($K^\star(x,y) = 0$) defined by the local potential $V^\star(x,y) = -x - y$ and diffusivity $\sigma^\star(x,y) = \sqrt{2\,(1+0.95\cos(\omega x)\cos(\omega y))}\mathbf{I}_2$, where $\mathbf{I}_2$ is the identity in $\mathbf{R}^2$. This results in a constant advection, variable diffusivity mean-field model[20]

$$\partial_t\mu_t = -\partial_x\mu_t - \partial_y\mu_t + \Delta\left[(1+0.95\cos(\omega x)\cos(\omega y))\,\mu_t\right]. \tag{5.2}$$

The purpose of this example is three-fold. First, we are interested in the ability of Algorithm 4.1 to correctly identify a local model from a library containing both local and nonlocal terms. Next, we evaluate whether the $\mathcal{O}(N^{-1/2})$ convergence is realized. Lastly, we investigate whether for large ω the weak-form identifies the associated homogenized equation (derived in Appendix A.2)

$$\partial_t\mu_t = -\partial_x\mu_t - \partial_y\mu_t + \overline{\omega}\Delta\mu_t, \tag{5.3}$$

where $\overline{\omega}$ is given by the harmonic mean of diffusivity:

$$\overline{\omega} = \left(\int_{\mathcal{D}} \frac{dxdy}{1+0.95\cos(x)\cos(y)}\right)^{-1}.$$

For $\omega \in \{1, 20\}$ we evolve the particles from an initial Gaussian distribution with mean zero and covariance $\mathbf{I}_2$ and record particle positions for 100 timesteps with $\Delta t = 0.02$ (subsampled from a simulation with timestep 10^{-4}). We use a rectangular domain $\mathcal{D}$ of approximate sidelength 10 and compute histograms with 128 bins in x and y for a spatial resolution of $\Delta x \approx 0.078$ (see Fig. 2 for solution snapshots), over which $\overline{\omega} \approx 0.62$. For $\omega = 1$ we compare recovered equations with the full model (5.2), while for $\omega = 20$ we compare with (5.3), for comparison computing $\overline{\omega}$ over each domain $\mathcal{D}$ using MATLAB's `integral2`. Fig. 3 shows that as the particle number increases, we do in fact recover the desired equations, with $\mathrm{TPR}(\widehat{\mathbf{w}})$ approaching one as N increases. For $\omega = 1$ we observe $\mathcal{O}(N^{-1/2})$ convergence of the local potential $\widehat{V}$ and the diffusivity $\widehat{\sigma}$. For $\omega = 20$, we observe

[19] For example, identification of the true model ($\mathrm{supp}\,(\widehat{\mathbf{w}}) = \mathrm{supp}\,(\mathbf{w}^\star)$) results in a $\mathrm{TPR}(\widehat{\mathbf{w}}) = 1$, while identification of only half of the correct nonzero terms and no additional falsely identified terms results in $\mathrm{TPR}(\widehat{\mathbf{w}}) = 0.5$.

[20] Since the model is local, (5.2) is the Fokker–Planck equation for the distribution of each particle, rather than only in the limit of infinite particles.

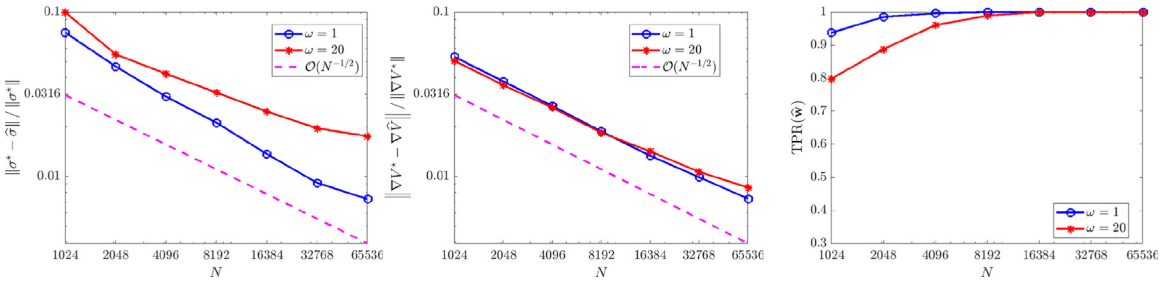

Fig. 3. Convergence of $\widehat{\sigma}$ (left) and $\nabla\widehat{V}$ (middle), recall $\|\cdot\|$ denotes the ℓ_2 norm, for (5.2) with $\omega \in \{1, 20\}$, as well as TPR($\widehat{\mathbf{w}}$) (right). For $\omega = 1$, results are compared to the exact model (5.2), while for $\omega = 20$ results are compared to the homogenized equation (5.3).

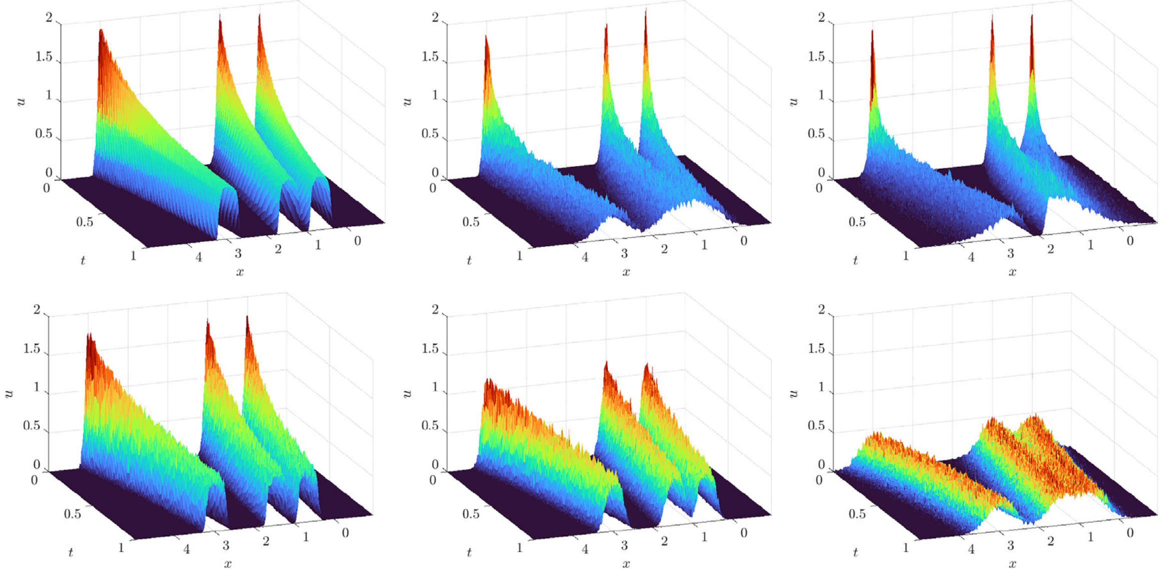

Fig. 4. Histograms computed with 256 bins width $h = 0.0234$ from 8000 particles in 1D evolving under $K^\star = K_{\text{QANR}}(x)$ (5.4). Top left to top right: $\sigma^\star(x) = 0$, $\sigma^\star(x) = \sqrt{2(0.1)}$, $\sigma^\star(x) = \sqrt{2(0.1)}|x-2|$. Bottom: deterministic particles with i.i.d. Gaussian noise added to particle positions with resulting noise ratios (left to right) $\epsilon = 0.0316, 0.1, 0.316$.

approximate $\mathcal{O}(N^{-1/2})$ convergence of $\widehat{V}$, and $\widehat{\sigma}$ converging to within 2% of $\sqrt{2\overline{\omega}}$, the homogenized diffusivity (higher accuracy can hardly be expected for $\omega = 20$ since (5.3) is itself an approximation in the limit of infinite ω).

5.2. One-dimensional nonlocal model

We simulate the evolution of particle systems under the quadratic attraction/Newtonian repulsion potential

$$K_{\text{QANR}}(x) = \frac{1}{2}x^2 - |x| \tag{5.4}$$

with no external potential ($V = 0$). The $-|x|$ portion of K_{QANR}, leading to a discontinuity in ∇K, is the one-dimensional free-space Green's function for $-\Delta$. For $d \geq 1$, when replaced by the corresponding Green's function in d dimensions, the distribution of particles evolves under K_{QANR} into the characteristic of the unit ball in $\mathbb{R}^d$, which has implications for design and control of autonomous systems [54]. We compare three diffusivity profiles, $\sigma(x) = 0$ corresponding to zero intrinsic noise, $\sigma(x) = \sqrt{2(0.1)}$ leading to constant-diffusivity intrinsic noise, and $\sigma(x) = \sqrt{2(0.1)}|x-2|$ leading to variable-diffusivity intrinsic noise. With zero intrinsic noise ($\sigma(x) = 0$), we examine the effect of extrinsic noise on recovery, and assume uncertainty in the particle positions due to measurement noise at each timestep, $\mathbb{Y} = \mathbb{X} + \varepsilon$, for $\varepsilon \sim \mathcal{N}(0, \epsilon^2 \|\mathbf{X_t}\|^2_{\text{RMS}})$ i.i.d. and $\epsilon \in \{0.01, 0.0316, 0.1, 0.316\}$. In this way ϵ is the *noise ratio*, such that $\|\varepsilon\|_F \,/\, \|\mathbb{X}\|_F \approx \epsilon$ (computed with ε and $\mathbb{X}$ stretched into column vectors).

Measurement data consists of 100 timesteps at resolution $\Delta t = 0.01$, coarsened from simulations with timestep 0.001. Initial particle positions are drawn from a mixture of three Gaussians each with standard deviation 0.005. Histograms are constructed with 256 bins of width $h = 0.0234$. Typical histograms for each noise level are shown in Fig. 4 computed one experiment with $N = 8000$ particles.

For the case of extrinsic noise (Fig. 5), we use only one experiment ($M = 1$) and examine the number of particles N and the noise ratio ϵ. We find that recovery is accurate and reliable for $\epsilon \leq 0.1$, yielding correct identification of K_{QANR} with less than 1% relative error in at least 98/100 trials. Increasing N from 500 to 8000 leads to minor improvements in accuracy for $\epsilon \leq 0.1$, but otherwise has little effect, implying that for low to moderate noise levels the mean-field equations are readily identifiable even from smaller particle systems. For $\epsilon = 10^{-1/2} \approx 0.3162$ (see Fig. 4 (bottom right) for an example histogram), we observe a decrease in TPR($\widehat{\mathbf{w}}$) (Fig. 5 middle panel) resulting from the generic identification of a linear diffusion term $\nu\partial_{xx}u$ with $\nu \approx 0.05$. Using that $\sqrt{2\nu} \approx \sqrt{2(0.05)} = \epsilon$, we can identify this as the best-fit *intrinsic* noise model. Furthermore, increases in N lead to reliable identification of the drift term, as measured by TPR($\widehat{\mathbf{w}}_{drift}$) (rightmost panel Fig. 5) which is the restriction of TPR to drift terms $\mathbb{L}_K$ and $\mathbb{L}_V$.

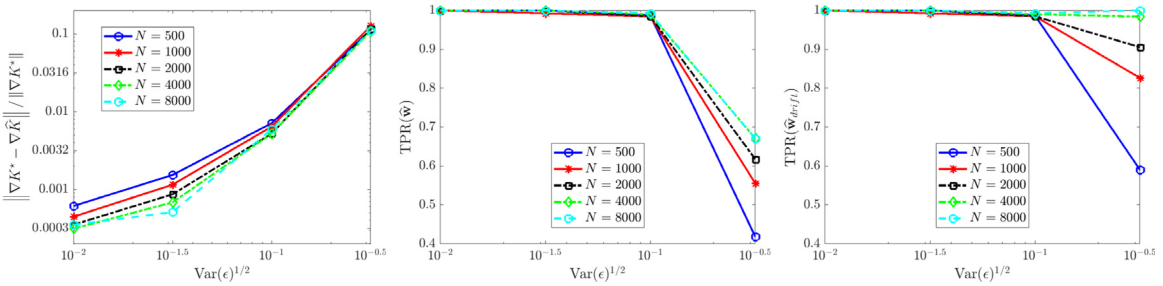

Fig. 5. Recovery of (3.1) in one spatial dimension for $K^\star = K_{\mathrm{QANR}}$ and $\sigma^\star = 0$ under different levels of observational noise ϵ. Left: relative error in learned interaction kernel $\widehat{K}$. Middle: true positivity ratio for full model (3.1). Right: true positivity ratio for drift term.

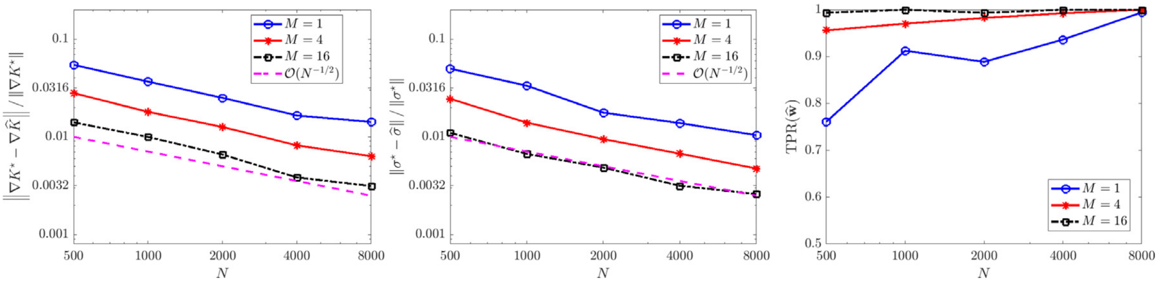

Fig. 6. Recovery of (3.1) in one spatial dimension for $K^\star = K_{\mathrm{QANR}}$ and $\sigma^\star = \sqrt{2(0.1)}$.

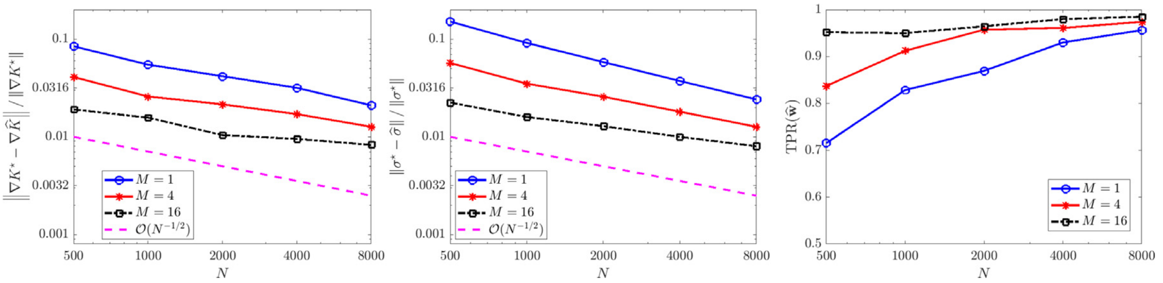

Fig. 7. Recovery of (3.1) in one spatial dimension for $K^\star = K_{\mathrm{QANR}}$ and $\sigma^\star = \sqrt{2(0.1)}|x-2|$.

For constant diffusivity $\sigma(x) = \sqrt{2(0.1)}$ (Fig. 6), the full model is recovered with less than 3% errors in $\hat{K}$ and $\hat{\sigma}$ in at least 98/100 trials when the total particle count NM is at least 8000, and yields errors less than 1% for $NM \geq 16{,}000$. The error trends for $\hat{K}$ and $\hat{\sigma}$ in this case both strongly agree with the predicted $\mathcal{O}(N^{-1/2})$ rate. For non-constant diffusivity $\sigma(x) = \sqrt{2(0.1)}|x-2|$ (Fig. 7), we also observe robust recovery ($\mathrm{TPR}(\widehat{\mathbf{w}}) \geq 0.95$) for $NM \geq 8000$ with error trends close to $\mathcal{O}(N^{-1/2})$, although the accuracy in $\hat{K}$ and $\hat{\sigma}$ is diminished due to the strong order $\Delta t^{1/2}$ convergence of Euler–Maruyama applied to diffusivities σ that are unbounded in x [55].

5.3. Two-dimensional nonlocal model

We now discuss an example of singular interaction in two spatial dimensions using the logarithmic potential

$$K(x) = \frac{1}{2\pi}\log|x| \tag{5.5}$$

with constant diffusivity $\sigma(x) = \sigma \in \{0, \frac{1}{\sqrt{4\pi}}\}$. This example corresponds to the parabolic–elliptic Keller–Segel model of chemotaxis, where $\sigma_c := \frac{1}{\sqrt{4\pi}}$ is the critical diffusivity such that $\sigma > \sigma_c$ leads diffusion-dominated spreading of particles throughout the domain (vanishing particle density at every point in $\mathbb{R}^2$) and $\sigma < \sigma_c$ leads to aggregation-dominated concentration of the particle density to the dirac-delta located at the center of mass of the initial particle density [44,56]. For $\sigma = 0$ we examine the affect of additive i.i.d. measurement noise $\varepsilon \sim \mathcal{N}(0, \epsilon^2\|\mathbf{X_t}\|^2_{\mathrm{RMS}})$ for $\epsilon \in \{0.01, 0.0316, 0.1, 0.316, 1\}$.

We simulate the particle system with a cutoff potential

$$K_\delta(x) = \begin{cases} \dfrac{1}{2\pi}\left(\log(\delta) - 1 + \dfrac{|x|}{\delta}\right), & |x| < \delta \\[2ex] \dfrac{1}{2\pi}\log|x|, & |x| \geq \delta \end{cases} \tag{5.6}$$

with $\delta = 0.01$, so that K_δ is Lipschitz and ∇K_δ has a jump discontinuity at the origin. Initial particle positions are uniformly distributed on a disk of radius 2 and the particle position data consists of 81 timepoints recorded at a resolution $\Delta t = 0.1$, coarsened from 0.0025. Histograms are created with 128 × 128 bins in x and y of sidelength $h = 0.0469$ (see Fig. 8 for histogram snapshots over time). We examine $M = 2^0, \ldots, 2^6$ experiments with $N = 2000$ or $N = 4000$ particles.

In Fig. 9 we observe a similar trend in the $\sigma = 0$ case as in the 1D nonlocal example, namely that recovery for $\epsilon \leq 0.1$ is robust with low errors in $\widehat{K}$ (on the order of 0.0032), only in this case the full model is robustly recovered up to $\epsilon = 0.316$. At $\epsilon = 1$, with $N = 4000$ the method frequently identifies a diffusion term $\nu\Delta u$ with $\nu \approx 0.5 = \epsilon^2/2$, and for $N = 2000$ the method occasionally identifies the backwards diffusion equation $\partial_t\mu_t = -\alpha\Delta\mu_t$, $\alpha > 0$. This is easily prevented by enforcing

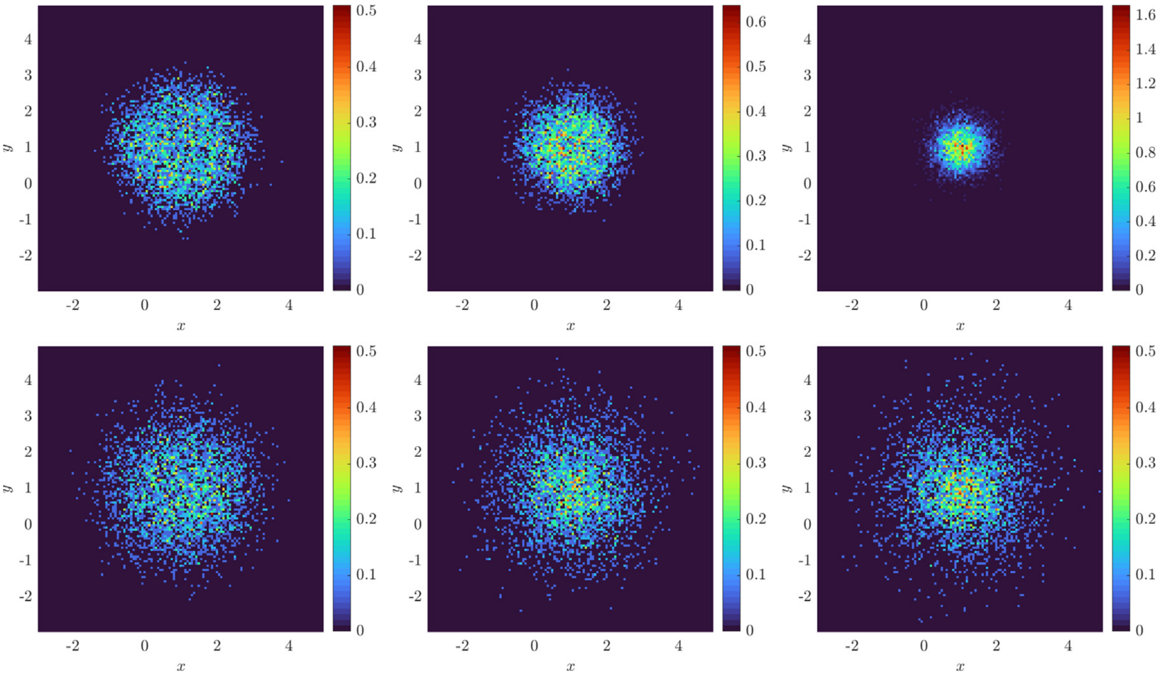

Fig. 8. Histograms created from 4000 particles evolving under logarithmic attraction (Eq. (5.5)) with varying noise levels at times (left to right) $t = 4$, $t = 8$, and $t = 12$. Top: $\epsilon = 0.316$, $\sigma = 0$ (extrinsic only). Bottom: $\epsilon = 0$, $\sigma = (4\pi)^{-1/2} \approx 0.28$ (intrinsic only).

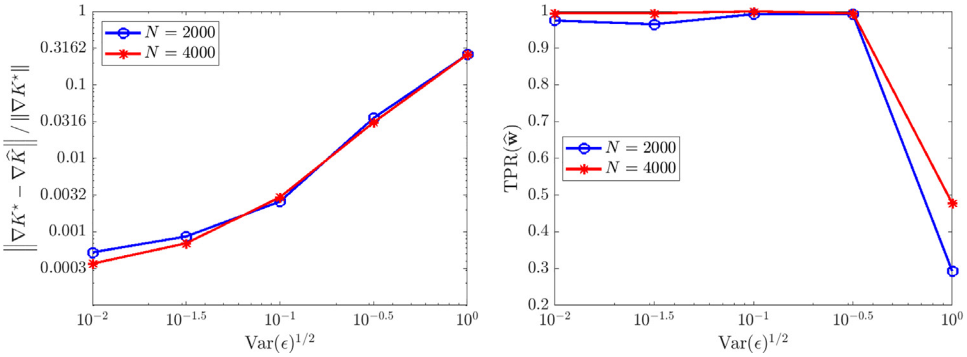

Fig. 9. Recovery of (3.1) in two spatial dimensions with $K^\star$ given by (5.5) from deterministic particles ($\sigma^\star = 0$) with extrinsic noise ϵ.

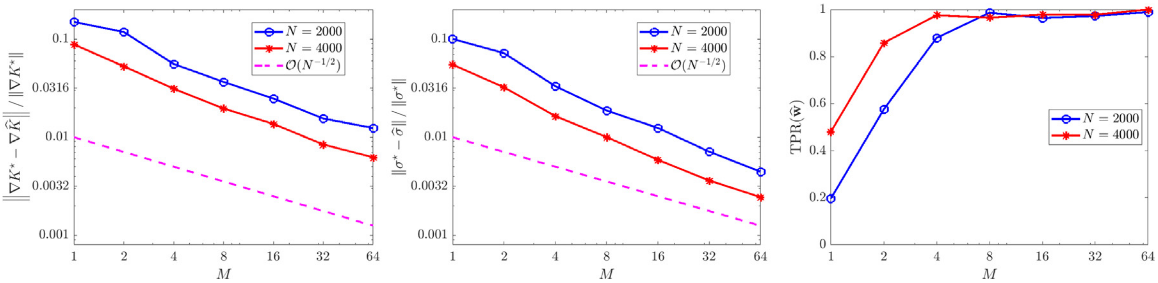

Fig. 10. Recovery of (3.1) in two spatial dimensions with $K^\star$ given by (5.5) and $\sigma^\star = \frac{1}{\sqrt{4\pi}}$.

positivity of σ, however we leave this and other constraints as an extension for future work.

With diffusivity $\sigma = \frac{1}{\sqrt{4\pi}}$, we obtain $\mathrm{TPR}(\widehat{\mathbf{w}})$ approximately greater than 0.95 for $NM \geq 16,000$ (Fig. 10, right), with an error trend in $\widehat{K}$ following an $\mathcal{O}(N^{-1/2})$ rate, and a trend in $\widehat{\sigma}$ of roughly $\mathcal{O}(N^{-2/3})$. Since convergence in M for any fixed N is not covered by the theorem above, this shows that combining multiple experiments may yield similar accuracy trends for moderately-sized particle systems.

6. Discussion

We have developed a weak-form method for sparse identification of governing equations for interacting particle systems using the formalism of mean-field equations. In particular, we have investigated two lines of inquiry, (1) is the mean-field setting applicable for inference from medium-size batches of particles? And (2) can a low-cost, low-regularity density approximation such as a histogram be used to enforce weak-form agreement with the mean-field PDE? We have demonstrated on several examples that the answer is yes to both questions, despite the fact that the mean-field equations are only valid in the limit of infinitely many particles ($N \to \infty$). This framework is suitable for systems of several thousand particles in one and two spatial dimensions, and we have proved convergence in N for the associated least-squares problem using simple histograms as approximate particle densities. In addition, the sparse regression approach allows one to identify the full system, including interaction potential K, local potential V, and diffusivity σ.

It was initially unclear whether the mean-field setting could be utilized in weak form for finite particle batches, hence this can be seen as a proof of concept for particle systems with N in the range $10^3 - 10^5$. With convergence in N and low computational complexity, our weak-form approach is well-suited *as is* for much larger particle systems. In the opposite regime, for small fixed N, the authors of [26] show that their maximum likelihood-based method converges as $M \to \infty$ (i.e. in the limit of infinite experiments). While the same convergence does not hold for our weak-form method, the results in Section 5 suggest that in practice, combining M independent experiments each with N particles improves results. Furthermore, we include evidence in Appendix A.3 that even for small N, our method correctly identifies the mean-field model when M is large enough, with performance similar to that in [26]. We leave a full investigation of the interplay between M and N to future work.

In the operable regime of $N > 10^3$, there is potential for improvements and extensions in many directions. On the subject of density estimation, histograms are highly efficient, yet they lead to piecewise-constant approximations of μ_t and hence $\mathcal{O}(h)$ errors. Choosing a density kernel G to achieve high-accuracy quadrature without sacrificing the $\mathcal{O}(N)$ runtime of histogram computation seems prudent, although one must be cautious about making assumptions on the smoothness of mean-field distribution μ_t. For instance, in the 1D nonlocal example 5.2, discontinuities develop in μ_t for the case $\sigma = 0$, hence a histogram approximation is more appropriate than using e.g. a Gaussian kernel.

The computational grid $\mathbf{C}$, quadrature method $\langle \cdot, \cdot \rangle_{h,\Delta t}$, and reference test function ψ may also be optimized further or adapted to specific problems. The approach chosen here of $\mathbf{C}$ equally-spaced and separable piecewise-polynomial ψ, along with integration using the trapezoidal quadrature, has several advantages, including high accuracy and fast computation using convolutions. However, this may need adjustment for higher dimensions. It might be advantageous to adapt $\mathbf{C}$ to the data $\mathbb{Y}$, however this may prevent one from evaluating $(\mathbf{G}, \mathbf{b})$ using the FFT if a non-uniform grid results, hence increases the overall computational complexity. One could also use multiple reference test functions ψ. The possibilities of varying the test functions (within the smoothness requirements of the library $\mathbb{L}$) have been largely unexplored in weak-form identification methods.

Several theoretical questions remain unanswered, namely model recovery statistics for finite N. As a consequence of Theorem 1, as well as convergence results on sequential thresholding [47], we have that $\mathbf{G}$ being full-rank and $\mathbb{L}$ containing the true model is sufficient to guarantee convergence $\widehat{\mathbf{w}} \to \mathbf{w}^\star$ as $N \to \infty$ at the rate $\mathcal{O}(N^{-1/2})$. Noise, whether extrinsic or intrinsic, for finite N may result in identification of an incorrect model when $\mathbf{G}$ is poorly-conditioned. The effect is more severe if the true model has a small coefficient, which requires a small threshold λ, which correspondingly may lead to a non-sparse solution. These are sensitivities of any sparse regression algorithm (see e.g. [57]) and accounting for the effect of noise and poor conditioning is an active area of research in equation discovery.

We also note that several researchers have focused on the uniqueness in kernel identifiability [34,58]. This issue does not directly apply to our scenario[21] of identifying the triple (K, V, σ). Moreover, in the cases we considered, we do not see any identifiability issues (e.g. rank deficiency) even in the high noise case with low particle number. Quantifying the transition to identifiability as $N \to \infty$ as a function of the condition number $\kappa(\mathbf{G})$ is an important subject for future work.

For extensions, the example system (5.2) and resulting homogenization motivates further study of effective equations for systems with complex microstructure. In other fields this is described as *coarse-graining*. A related line of study is inference of 2nd-order particle systems, as explored in [32], which often lead to an infinite hierarchy of mean-field equations. Our weak-form approach may provide a principled method for truncated and closing such hierarchies using particle data. Another extension is to enforce convex constraints in the regression problem, such as lower bounds on diffusivity, or K with long-range attraction depending on the distribution $\rho_{rr} \in \mathcal{P}([0, \infty))$ of pairwise distances (see [26] for further use of ρ_{rr}). Finally, the framework we have introduced can easily be used to find nonlocal models from continuous solution data (e.g. given $\mathbf{U}$ instead of $\mathbb{Y}$), whereby questions of nonlocal representations of models can be investigated.

Lastly, we note that MATLAB code is available at https://github.com/MathBioCU/WSINDy_IPS.

CRediT authorship contribution statement

Daniel A. Messenger: Concept for the article, Wrote the first draft, Editing, Performed the mathematical analysis, Wrote software selected examples, Ran simulations, Analyzed the data. **David M. Bortz:** Concept for the article, Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported in part by the NSF Mathematical Biology MODULUS grant 2054085, in part by the NSF/NIH Joint DMS/NIGMS Mathematical Biology Initiative grant R01GM126559, and in part by the NSF Computing and Communications Foundations grant 1815983. This work also utilized resources from the University of Colorado Boulder Research Computing Group, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The authors would also like to thank Prof. Vanja Dukić (University of Colorado at Boulder, Department of Applied Mathematics) for insightful discussions and helpful suggestions of references.

Appendix

A.1. Specifications for examples

In Tables 2–5 we include hyperparameter specifications and resulting attributes of Algorithm 4.1 applied to the three examples in Section 5. In particular, we report the typical walltime in Table 5, showing that on each example Algorithm 4.1 learns the mean-field equation from a dataset with ∼64,000 particles in under 10 s.

[21] E.g. due to multiple representations of the drift combining both nonlocal and local terms — see Section 4.1.4

Table 5
Discretization parameters and general information for examples. The number of nonzeros in the true weight vector $\|\mathbf{w}^\star\|_0$ is given for each parameter set examined. Namely, for the local 2D example, $\omega = 1$ results in a 4-term model, while the homogenized case $\omega = 20$ results in a three-term model. For the nonlocal 1D example, $\sigma \in \{0, \sqrt{2(0.1)}, \sqrt{2(0.1)}|x-2|\}$ result in 2-term, 3-term, and 5-term models, respectively, and for the nonlocal 2D example $\sigma \in \{0, (4\pi)^{-1}\}$ results in 1-term and 2-term models. The norm $\left\|\mathbf{G}^\dagger\right\|_1$, condition number $\kappa_2(\mathbf{G})$ and walltime are listed for representative samples with 64,000 total particles.

Example	(m_x, m_t)	(p_x, p_t)	(s_x, s_t)	size($\mathbf{U}$)	$(h, \Delta t)$
Local 2D	(31,16)	(5,3)	(10,5)	128 × 128 × 101	(0.078, 0.02)
Nonlocal 1D	(29,8)	(5,3)	(5,1)	256 × 101	(0.023, 0.01)
Nonlocal 2D	(25,8)	(5,3)	(8,1)	128 × 128 × 81	(0.047, 0.1)

Example	$\|\mathbf{w}^\star\|_0$	size($\mathbf{G}$)	$\|\mathbf{G}^\dagger\|_1$	$\kappa_2(\mathbf{G})$	Walltime
Local 2D	{4, 3}	686 × 85	2.0×10^3	3.0×10^7	9.2 s
Nonlocal 1D	{2, 3, 5}	3400 × 24	1.3×10^5	8.7×10^8	0.7 s
Nonlocal 2D	{1, 2}	6500 × 59	1.1×10^4	6.4×10^6	8.5 s

A.2. Derivation of homogenized equation (5.3)

We briefly provide a derivation of the homogenized equation (5.3) in the static case. Let $\Omega \in \mathbb{R}^d$ be an open bounded domain with smooth boundary and $\mathbb{T}^d$ be the d-dimensional torus. Let $a(x, y) : \Omega \times \mathbb{T}^d \to \mathbb{R}$ be continuous and uniformly bounded below,

$$a(x, y) \geq \alpha > 0, \quad (x, y) \in \Omega \times \mathbb{T}^d.$$

Then for any $f \in L^2(\Omega)$, the equation

$$-\Delta\left(a(x, x/\epsilon)u^\epsilon(x)\right) = f(x), \qquad u^\epsilon\big|_{\partial\Omega} = 0$$

has a unique weak solution $u^\epsilon \in L^2(\Omega)$ given by

$$u^\epsilon(x) = \frac{(Gf)(x)}{a(x, x/\epsilon)},$$

where G is the Green's function for $(-\Delta)^{-1}$ with homogeneous Dirichlet boundary conditions on $\partial\Omega$. By the coercivity of a we have that $\|u^\epsilon\|_{L^2(\Omega)}$ is uniformly bounded in ϵ. By the lemma in [59, Section 2.4], up to a subsequence $\{\epsilon_j\}_{j\in\mathbb{N}}$, there exists a function $u(x, y)$ periodic in its second variable such that for any continuous function $\phi(x, x/\epsilon)$, we have

$$\lim_{\epsilon\to 0} \int u^\epsilon(x)\phi(x, x/\epsilon)dx = \iint u(x, y)\phi(x, y)dydx.$$

Setting $\phi(x, y) = \phi(x)$, we see that on the same subsequence, $u^\epsilon \rightharpoonup \int u(x, y)dy$. Applying the same lemma to the constant series $u^\epsilon = 1$ and letting $\phi(x, x/\epsilon) = \phi(x)a^{-1}(x, x/\epsilon)$, we see that (up to possibly a second subsequence),

$$a^{-1}(x, x/\epsilon) \rightharpoonup \int \frac{dy}{a(x, y)}.$$

Letting $a^*(x) := \left(\int \frac{dy}{a(x,y)}\right)^{-1}$ and putting together the previous limits, we see that

$$u^\epsilon(x) \rightharpoonup u^*(x) := \int u(x, y)dy = (Gf)(x)\int \frac{dy}{a(x, y)} =: \frac{(Gf)(x)}{a^*(x)},$$

and hence u^* solves the homogenized equation

$$\Delta\left(a^* u^*\right) = f.$$

A.3. Recovery for small N and large M

The related maximum-likelihood approach [26] is shown to be suitable for small N and large M, hence a natural line of inquiry is the performance of Algorithm 4.1 in this regime. Theorem 1 does not apply to this regime, and in fact convergence of the algorithm is not expected: letting $U_t^{M,N} = \frac{1}{M}\sum_{m=1}^M U_t^{(m),N}$ where $U_t^{(m),N}$ is the approximate density constructed from experiment m with N particles, we have the weak-measure convergence $U_t^{M,N} \to \rho_t^{(1),N}$ as $M \to \infty$, where $\rho_t^{(1),N}$ is the 1-particle marginal of the distribution of $\mathbf{X}_t$ in $\mathbb{R}^{Nd}$. Unlike the mean-field distribution μ_t, $\rho_t^{(1),N}$ is not a weak solution to the mean-field Fokker–Planck equation (3.1), instead we have

$$\partial_t\rho_t^{(1),N} = \frac{N-1}{N}\nabla\cdot\int_{\mathbb{R}^d}\nabla K(x-y)\rho_t^{(2),N}(x, y)dy + \nabla\cdot\left(\nabla V\rho_t^{(1),N}\right)$$
$$+\frac{1}{2}\sum_{i,j=1}^d \partial_{x_ix_j}(\sigma\sigma^T\rho_t^{(1),N}),$$

holding weakly, which depends on the 2-particle marginal $\rho_t^{(2),N}$ [35]. Nevertheless, using the 1D nonlocal example in Section 5.2 with $\sigma = \sqrt{2(0.1)} \approx 0.45$, we observe in Fig. 11 (right panel) that our weak-form algorithm correctly identifies the model in $>$ 96% of trials with just $N = 10$ particles per experiment when $M \in [2^{10}, 2^{12}]$, and that error in K (left panel) follows a $\mathcal{O}(M^{-1/2})$ trend. At $M = 4096 \approx 10^{3.61}$ experiments the error[22] in K is less than 1% and the runtime is approximately 0.9 s. The lack of convergence in M is reflected in the diffusivity (middle panel of Fig. 11), where the error appears to plateau at around 1.7% for $h \approx 0.0468$ and at 3.5% for $h \approx 0.0234$. The lower resolution (larger binwidth h) appears to yield slightly better results, possibly indicating that larger h produces a coarse-graining effect such that $\rho^{(2),N} \approx \rho^{(1),N} \otimes \rho^{(1),N}$ over larger distances, although this effect deserves more thorough study in future work.

A.4. Technical lemmas

We now prove Lemmas 2–4 under Assumption H. First, some consequences of Assumption H. (I) The η-Hölder continuity of sample paths (H.1) implies that for each $t \in [0, T]$,

$$\int_{\mathbb{R}^d}|x|^p d\mu_t^N = \frac{1}{N}\sum_{i=1}^N |X_t^{(i)}|^p \leq \frac{2^p}{N}\sum_{i=1}^N |X_0^{(i)}|^p + C_\eta 2^p t^{p\eta}.$$

Together with the pth moment bound on μ_0 (H.2), this implies

$$\mathbb{E}\left[\sup_{t\leq T}\int_{\mathbb{R}^d}|x|^p d\mu_t^N\right] \leq 2^p(M_p + C_\eta T^{p\eta}), \tag{A.1}$$

independent of N. (II) The growth bounds on $\nabla K^\star$, $\nabla V^\star$, and $\sigma^\star$ (H.3)–(H.4) imply that for some $C > 0$,

$$|\nabla K^\star(x)| + |\nabla V^\star(x)| + \left\|\sigma^\star(x)(\sigma^\star(x))^T\right\|_F \leq C(1 + |x|^p), \tag{A.2}$$

where $\|\cdot\|_F$ is the Frobenius norm.

Proof of Lemma 2. Applying Itô's formula to the process $\frac{1}{N}\sum_{i=1}^N \psi(X^{(i)}, t)$, we get that

$$\mathscr{L}(\mu^N, \psi, \langle\cdot,\cdot\rangle) = \frac{1}{N}\sum_{i=1}^N\int_0^T \nabla\psi(X_t^{(i)}, t)^T\sigma^\star(X_t^{(i)})dB_t^{(i)}.$$

Note that each integral on the right-hand side is a local martingale, since (A.2) and (H.5) ensure boundedness of $\nabla\psi(x, t)^T\sigma^\star(x)$ over any compact set in $\mathbb{R}^d$, hence has mean zero. By independence of the Brownian motions $B_t^{(i)}$, exchangeability of $X_t^{(i)}$, the moment bound (A.1), and the growth bounds on σ (H.4), the Itô

[22] For comparison, in [26] Fig. 4 the error in recovering K using the maximum-likelihood approach on an opinion dynamics example for $M = 10^{3.6}$, $N = 10$, and $\sigma = 0.5$ is approximately $100 \times 10^{-1.2}\% = 6.3\%$.

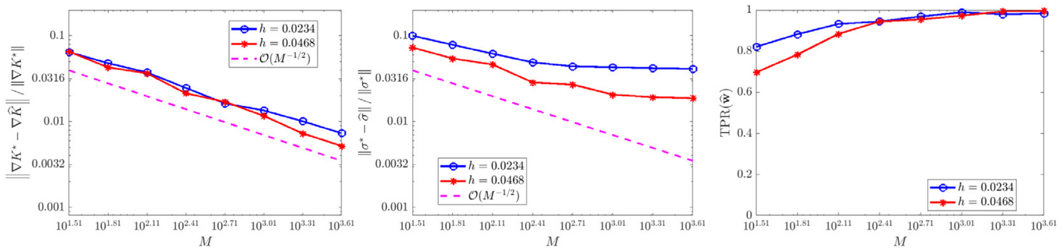

Fig. 11. Recovery of (3.1) in one spatial dimension for $K^\star = K_{\text{QANR}}$ and $\sigma^\star = \sqrt{2(0.1)}$ with only $N = 10$ particles per experiment.

isometry gives us

$$\begin{aligned}
&\mathbb{E}\left[\mathscr{L}(\mu^N, \psi, \langle\cdot,\cdot\rangle)^2\right]\\
&\quad= \frac{1}{N}\int_0^T \mathbb{E}_{X\sim\rho_t^{(1)}}\left[\left|\nabla\psi(X,t)^T\sigma^\star(X)\right|^2\right]dt\\
&\quad= \frac{1}{N}\int_0^T \mathbb{E}\left[\int_{\mathbb{R}^d}\left|\nabla\psi(x,t)^T\sigma^\star(x)\right|^2 d\mu_t^N(x)\right]dt\\
&\quad\leq \frac{C'}{N}\|\nabla\psi\|_{2,\infty}^2\int_0^T\mathbb{E}\left[1+\int_{\mathbb{R}^d}|x|^p d\mu_t^N(x)\right]dt\\
&\quad\leq CN^{-1}
\end{aligned}$$

where C depends on M_p, C_p, T, and ψ. The result follows from Jensen's inequality.[23]

Proof of Lemma 3. Using the notation $f^{\mathbf{C}}$ from Lemma 1 to denote piecewise constant approximation of a function f over the domain $\mathcal{D}$ using the grid $\mathbf{C}$, we have

$$\begin{aligned}
&\mathscr{L}(U,\psi,\langle\cdot,\cdot\rangle_h) - \mathscr{L}(\mu^N,\psi,\langle\cdot,\cdot\rangle)\\
&\quad= -\underbrace{\left(\left\langle(\nabla\psi\cdot((\nabla K^\star)^{\mathbf{C}}*\mu^N))^{\mathbf{C}},\mu^N\right\rangle - \left\langle\nabla\psi\cdot\nabla K^\star*\mu^N,\mu^N\right\rangle\right)}_{E_{\text{interact}}}\\
&\qquad+\left\langle\partial_t\psi^{\mathbf{C}} - \partial_t\psi,\mu^N\right\rangle - \left\langle(\nabla\psi\cdot\nabla V^\star)^{\mathbf{C}} - \nabla\psi\cdot\nabla V^\star,\mu^N\right\rangle\\
&\qquad+\frac{1}{2}\left\langle\operatorname{Tr}\left(\nabla^2\psi\sigma^\star(\sigma^\star)^T\right)^{\mathbf{C}} - \operatorname{Tr}\left(\nabla^2\psi\sigma^\star(\sigma^\star)^T\right),\mu^N\right\rangle\\
&\quad= E_{\text{interact}} + E_{\text{linear}}.
\end{aligned}$$

The right-hand side includes an interaction error E_{interact} followed by a sum E_{linear} of terms that are linear in the difference between a locally Lipschitz function and its piecewise constant approximation. Hence, we can bound E_{linear} using smoothness of ψ (H.5), the moment assumptions on μ_t^N (H.2), and the growth assumptions on V and σ (H.3)–(H.4). Specifically, for $x \in B_k$ with center $\mathbf{c}_k$, the growth assumptions imply

$$\begin{aligned}
&|\nabla\psi(x)\cdot\nabla V^\star(x) - \nabla\psi(\mathbf{c}_k)\cdot\nabla V^\star(\mathbf{c}_k)|\\
&\quad\leq Ch\Big((\|\nabla\psi\|_{2,\infty} + \operatorname{Lip}(\nabla\psi))(1+|x|^p)\Big)\\
&\left|\operatorname{Tr}\left(\nabla^2\psi(x)\sigma^\star(x)(\sigma^\star(x))^T\right) - \operatorname{Tr}\left(\nabla^2\psi(\mathbf{c}_k)\sigma^\star(\mathbf{c}_k)(\sigma^\star(\mathbf{c}_k))^T\right)\right|\\
&\quad\leq C'h\Big((\left\|\nabla^2\psi\right\|_{F,\infty} + \operatorname{Lip}(\nabla^2\psi))(1+|x|^p)\Big)
\end{aligned}$$

for C and C' depending on p, d, and C_p, hence

$$|E_{\text{linear}}| \leq C''\sup_{|\alpha|\leq 2}\operatorname{Lip}(\partial^\alpha\psi)\left(T + \int_0^T\int_{\mathbb{R}^d}|x|^p d\mu_t^N dt\right)h. \tag{A.3}$$

Similarly, for the interaction error we use that for $x \in B_k$ and $y \in B_j$ with centers $\mathbf{c}_k$ and $\mathbf{c}_j$, we have

$$\begin{aligned}
&\left|\nabla\psi(\mathbf{c}_k)\cdot\nabla K^\star(\mathbf{c}_k - \mathbf{c}_j) - \nabla\psi(x)\cdot\nabla K^\star(x-y)\right|\\
&\quad\leq |\nabla\psi(\mathbf{c}_k)|\left|\nabla K^\star(\mathbf{c}_k-\mathbf{c}_j) - \nabla K^\star(x-y)\right|\\
&\qquad+|\nabla\psi(\mathbf{c}_k) - \nabla\psi(x)|\left|\nabla K^\star(x-y)\right|\\
&\quad\leq C'''h\left(\|\nabla\psi\|_{2,\infty} + \operatorname{Lip}(\nabla\psi)\right)(1+|x-y|^p)
\end{aligned}$$

with C''' also depending on p, d, and C_p. From this we have

$$|E_{\text{interact}}| \leq C''''\left(T + \int_0^T\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}|x-y|^p d\mu_t^N(y)d\mu_t^N(x)dt\right)h. \tag{A.4}$$

The result follows from taking expectation and using the moment bound (A.1), where the final constant C depends on p, d, C_p, M_p, T, η, and ψ.

Proof of Lemma 4. Again rewriting the spatial trapezoidal-rule integration in the form $\int_{\mathbb{R}^d}\varphi^{\mathbf{C}}(x)d\mu_t^N$, we see that

$$\mathscr{L}(U,\psi,\langle\cdot,\cdot\rangle_h) - \mathscr{L}(U,\psi,\langle\cdot,\cdot\rangle_{h,\Delta t}) \tag{A.5}$$

reduces to four terms of the form

$$\begin{aligned}
A(\varphi) := \frac{1}{N}\sum_{i=1}^N&\left(\int_0^T\varphi^{\mathbf{C}}(X_t^{(i)})dt\right.\\
&\left.-\frac{\Delta t}{2}\sum_{\ell=1}^L\left(\varphi^{\mathbf{C}}(X_{t_{\ell+1}}^{(i)}) + \varphi^{\mathbf{C}}(X_{t_\ell}^{(i)})\right)\right),
\end{aligned}$$

for $\varphi \in \left\{\partial_t\psi, \nabla\psi\cdot\nabla V^\star, \operatorname{Tr}(\nabla^2\psi\sigma^\star(\sigma^\star)^T), \nabla\psi\cdot\nabla K^\star*\mu_t^N\right\}$. Similarly to the bounds derived for $|\varphi(x) - \varphi^{\mathbf{C}}(x)|$ in Lemma 3, the growth bounds on $V^\star$, $K^\star$ and $\sigma^\star$ imply in general that

$$|\varphi(x) - \varphi(y)| \leq C|x-y|\left(1+\max\{|x|,|y|\}^p\right).$$

Rewriting the summands in $A(\varphi)$,

$$\begin{aligned}
&\int_0^T\varphi^{\mathbf{C}}(X_t^{(i)})dt - \frac{\Delta t}{2}\sum_{\ell=1}^L\left(\varphi^{\mathbf{C}}(X_{t_{\ell+1}}^{(i)}) + \varphi^{\mathbf{C}}(X_{t_\ell}^{(i)})\right)\\
&\quad=\sum_{\ell=1}^L\underbrace{\int_{t_\ell}^{t_{\ell+1}}\left(\frac{t-t_\ell}{\Delta t}\right)(\varphi^{\mathbf{C}}(X_t^{(i)}) - \varphi^{\mathbf{C}}(X_{t_{\ell+1}}^{(i)}))dt}_{I_1}\\
&\qquad+\underbrace{\int_{t_\ell}^{t_{\ell+1}}\left(\frac{t_{\ell+1}-t}{\Delta t}\right)(\varphi^{\mathbf{C}}(X_t^{(i)}) - \varphi^{\mathbf{C}}(X_{t_\ell}^{(i)}))dt}_{I_2},
\end{aligned}$$

and using

$$\begin{aligned}
|\varphi^{\mathbf{C}}(x) - \varphi^{\mathbf{C}}(y)| &\leq |\varphi(x) - \varphi(\mathbf{c}_k)| + |\varphi(x) - \varphi(y)| + |\varphi(y) - \varphi(\mathbf{c}_\ell)|\\
&\leq C(2h + |x-y|)(1+\max\{|x|,|y|\}^p)
\end{aligned}$$

where $x \in B_k$ and $y \in B_\ell$, we see that for I_1,

$$\left|\int_{t_\ell}^{t_{\ell+1}}\left(\frac{t-t_\ell}{\Delta t}\right)(\varphi^{\mathbf{C}}(X_t^{(i)}) - \varphi^{\mathbf{C}}(X_{t_{\ell+1}}^{(i)}))dt\right|$$

[23] $\|f\|_{p,q}$ for vector-valued functions $f : \mathbb{R}^d \to \mathbb{R}^d$ denotes the L^q norm over x of the ℓ^p norm of $f(x)$. Also recall that $\rho_t^{(1)}$ is the $X_t^{(1)}$-marginal of the process $\mathbf{X}_t \in \mathbb{R}^{dN}$.

$$\leq \int_{t_\ell}^{t_{\ell+1}} \left(\frac{t-t_\ell}{\Delta t}\right) C(2h+|X_t^{(i)} - X_{t_{\ell+1}}^{(i)}|)(1+\max\{|X_t^{(i)}|, |X_{t_{\ell+1}}^{(i)}|\}^p)dt$$

$$\leq \int_{t_\ell}^{t_{\ell+1}} \left(\frac{t-t_\ell}{\Delta t}\right) C'(2h+|t_{\ell+1}-t|^\eta|)(1+\max\{|X_t^{(i)}|, |X_{t_{\ell+1}}^{(i)}|\}^p)dt.$$

Taking expectation on both sides and using the moment bound (A.1), we get

$$\mathbb{E}\left[\left|\int_{t_\ell}^{t_{\ell+1}} \left(\frac{t-t_\ell}{\Delta t}\right)(\varphi^{\mathbf{c}}(X_t^{(i)}) - \varphi^{\mathbf{c}}(X_{t_{\ell+1}}^{(i)}))dt\right|\right]$$
$$\leq C\left(\Delta t h + \Delta t^{1+\eta}\right).$$

We get the same bound for I_2. Summing over ℓ, and taking the average in i, we then get

$$\mathbb{E}\left[|A(\varphi)|\right] \leq C(h + \Delta t^\eta),$$

which implies the desired bound on the difference (A.5).

References

[1] Steven L. Brunton, Joshua L. Proctor, J. Nathan Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. 113 (15) (2016) 3932–3937.

[2] Daniel A. Messenger, David M. Bortz, Weak SINDy: Galerkin-based data-driven model selection, Multiscale Model. Simul. 19 (3) (2021) 1474–1497.

[3] Daniel A. Messenger, David M. Bortz, Weak SINDy for partial differential equations, J. Comput. Phys. (2021) 110525.

[4] Michael S. Warren, John K. Salmon, Astrophysical N-body simulations using hierarchical tree data structures, Proc. Supercomput. (1992).

[5] Jiawei Guo, The progress of three astrophysics simulation methods: Monte-Carlo, PIC and MHD, J. Phys. Conf. Ser. 2012 (1) (2021) 012136, IOP Publishing.

[6] Tony Lelievre, Gabriel Stoltz, Partial differential equations and stochastic methods in molecular dynamics, Acta Numer. 25 (2016) 681–880.

[7] Néstor Sepúlveda, Laurence Petitjean, Olivier Cochet, Erwan Grasland-Mongrain, Pascal Silberzan, Vincent Hakim, Collective cell motion in an epithelial sheet can be quantitatively described by a stochastic interacting particle model, PLoS Comput. Biol. 9 (3) (2013) e1002944.

[8] Paul Van Liedekerke, M.M. Palm, N. Jagiella, Dirk Drasdo, Simulating tissue mechanics with agent-based models: concepts, perspectives and some novel results, Comput. Part. Mech. 2 (4) (2015) 401–444.

[9] Dapeng Bi, Xingbo Yang, M Cristina Marchetti, M Lisa Manning, Motility-driven glass and jamming transitions in biological tissues, Phys. Rev. X 6 (2) (2016) 021011.

[10] Vincent D. Blondel, Julien M. Hendrickx, John N. Tsitsiklis, Continuous-time average-preserving opinion dynamics with opinion-dependent communications, SIAM J. Control Optim. 48 (8) (2010) 5214–5240.

[11] Evelyn F. Keller, Lee A. Segel, Model for chemotaxis, J. Theoret. Biol. 30 (2) (1971) 225–234.

[12] Paraskevi Gkeka, Gabriel Stoltz, Amir Barati Farimani, Zineb Belkacemi, Michele Ceriotti, John D Chodera, Aaron R Dinner, Andrew L Ferguson, Jean-Bernard Maillet, Hervé Minoux, et al., Machine learning force fields and coarse-grained variables in molecular dynamics: application to materials and biological systems, J. Chem. Theory Comput. 16 (8) (2020) 4757–4775.

[13] Bo Martin Bibby, Michael Sørensen, Martingale estimation functions for discretely observed diffusion processes, Bernoulli (1995) 17–39.

[14] Andrew W. Lo, Maximum likelihood estimation of generalized Itô processes with discretely sampled data, Econom. Theory 4 (2) (1988) 231–247.

[15] Jaya P.N. Bishwal, Parameter Estimation in Stochastic Differential Equations, Springer, 2007.

[16] Lorenzo Boninsegna, Feliks Nüske, Cecilia Clementi, Sparse learning of stochastic dynamical equations, J. Chem. Phys. 148 (24) (2018) 241723.

[17] Jared L Callaham, J-C Loiseau, Georgios Rigas, Steven L Brunton, Nonlinear stochastic modelling with langevin regression, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 477 (2250) (2021) 20210092.

[18] Yang Li, Jinqiao Duan, Extracting governing laws from sample path data of non-Gaussian stochastic dynamical systems, 2021, arXiv preprint arXiv:2107.10127.

[19] John T Nardini, Ruth E Baker, Matthew J Simpson, Kevin B Flores, Learning differential equation models from stochastic agent-based model simulations, J. R. Soc. Interface 18 (176) (2021) 20200987.

[20] David B. Brückner, Pierre Ronceray, Chase P. Broedersz, Inferring the dynamics of underdamped stochastic systems, Phys. Rev. Lett. 125 (5) (2020) 058103.

[21] Xiaoli Chen, Liu Yang, Jinqiao Duan, George Em Karniadakis, Solving inverse stochastic problems from discrete particle observations using the Fokker–Planck equation and physics-informed neural networks, SIAM J. Sci. Comput. 43 (3) (2021) B811–B830.

[22] Jinchao Feng, Yunxiang Ren, Sui Tang, Data-driven discovery of interacting particle systems using Gaussian processes, 2021, arXiv preprint arXiv:2106.02735.

[23] Raphael A. Kasonga, Maximum likelihood theory for large interacting systems, SIAM J. Appl. Math. 50 (3) (1990) 865–875.

[24] Jaya Prakash Narayan Bishwal, et al., Estimation in interacting diffusions: Continuous and discrete sampling, Appl. Math. 2 (9) (2011) 1154–1158.

[25] Mattia Bongini, Massimo Fornasier, Markus Hansen, Mauro Maggioni, Inferring interaction rules from observations of evolutive systems I: The variational approach, Math. Models Methods Appl. Sci. 27 (05) (2017) 909–951.

[26] Fei Lu, Mauro Maggioni, Sui Tang, Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories, Foundations of Computational Mathematics (2021) 1–55.

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

[28] Louis Sharrock, Nikolas Kantas, Panos Parpas, Grigorios A Pavliotis, Parameter estimation for the mckean-vlasov stochastic differential equation, 2021, arXiv preprint arXiv:2106.13751.

[29] Susana N. Gomes, Andrew M. Stuart, Marie-Therese Wolfram, Parameter estimation for macroscopic pedestrian dynamics models from microscopic data, SIAM J. Appl. Math. 79 (4) (2019) 1475–1500.

[30] Ryan Lukeman, Yue-Xian Li, Leah Edelstein-Keshet, Inferring individual rules from collective behavior, Proc. Natl. Acad. Sci. 107 (28) (2010) 12576–12580.

[31] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, J Nathan Kutz, Data-driven discovery of partial differential equations, Sci. Adv. 3 (4) (2017) e1602614.

[32] Rohit Supekar, Boya Song, Alasdair Hastewell, Alexander Mietke, Jörn Dunkel, Learning hydrodynamic equations for active matter from particle simulations and experiments, 2021, arXiv preprint arXiv:2101.06568.

[33] E Paulo Alves, Frederico Fiuza, Data-driven discovery of reduced plasma physics models from fully-kinetic simulations, 2020, arXiv preprint arXiv:2011.01927.

[34] Quanjun Lang, Fei Lu, Learning interaction kernels in mean-field equations of 1st-order systems of interacting particles, 2020, arXiv preprint arXiv:2010.15694.

[35] Pierre-Emmanuel Jabin, Zhenfu Wang, Mean field limit for stochastic particle systems, in: Active Particles, Volume 1, Springer, 2017, pp. 379–402.

[36] Alain-Sol Sznitman, Topics in propagation of chaos, in: Ecole D'été de Probabilités de Saint-Flour XIX—1989, Springer, 1991, pp. 165–251.

[37] Sylvie Méléard, Asymptotic behaviour of some interacting particle systems; McKean-Vlasov and Boltzmann models, in: Probabilistic Models for Nonlinear Partial Differential Equations, Springer, 1996, pp. 42–95.

[38] François Bolley, José A Canizo, José A Carrillo, Stochastic mean-field limit: non-Lipschitz forces and swarming, Math. Models Methods Appl. Sci. 21 (11) (2011) 2179–2210.

[39] Niklas Boers, Peter Pickl, On mean field limits for dynamical systems, J. Stat. Phys. 164 (1) (2016) 1–16.

[40] Razvan C. Fetecau, Hui Huang, Weiran Sun, Propagation of chaos for the keller–segel equation over bounded domains, J. Differential Equations 266 (4) (2019) 2142–2174.

[41] Razvan C Fetecau, Hui Huang, Daniel Messenger, Weiran Sun, Zero-diffusion limit for aggregation equations over bounded domains, 2018, arXiv preprint arXiv:1809.01763.

[42] Daniel A. Messenger, Razvan C. Fetecau, Equilibria of an aggregation model with linear diffusion in domains with boundaries, Math. Models Methods Appl. Sci. 30 (04) (2020) 805–845.

[43] Razvan C. Fetecau, Mitchell Kovacic, Swarm equilibria in domains with boundaries, SIAM J. Appl. Dyn. Syst. 16 (3) (2017) 1260–1308.

[44] J.A. Carrillo, M.G. Delgadino, F.S. Patacchini, Existence of ground states for aggregation-diffusion equations, Anal. Appl. 17 (03) (2019) 393–423.

[45] Dyego Araújo, Roberto I. Oliveira, Daniel Yukimura, A mean-field limit for certain deep neural networks, 2019, arXiv preprint arXiv:1906.00193.

[46] David Freedman, Persi Diaconis, On the histogram as a density estimator: L2 theory, Z. Wahrscheinlichkeitstheor. Verwandte Gebiete 57 (4) (1981) 453–476.

[47] Linan Zhang, Hayden Schaeffer, On the convergence of the SINDy algorithm, Multiscale Model. Simul. 17 (3) (2019) 948–972.

[48] Simon Foucart, Holger Rauhut, A Mathematical Introduction to Compressive Sensing, Birkhäuser Basel, 2013.

[49] Osman Asif Malik, Stephen Becker, Low-rank tucker decomposition of large tensors using tensorsketch, Adv. Neural Inf. Process. Syst. 31 (2018) 10096–10106.

[50] Yiming Sun, Yang Guo, Charlene Luo, Joel Tropp, Madeleine Udell, Low-rank tucker approximation of a tensor from streaming data, SIAM J. Math. Data Sci. 2 (4) (2020) 1123–1150.

[51] Jun-Gi Jang, U. Kang, D-tucker: Fast and memory-efficient tucker decomposition for dense tensors, in: 2020 IEEE 36th International Conference on Data Engineering (ICDE), IEEE, 2020, pp. 1850–1853.

[52] Wenjian Yu, Yu Gu, Yaohang Li, Efficient randomized algorithms for the fixed-precision low-rank matrix approximation, SIAM J. Matrix Anal. Appl. 39 (3) (2018) 1339–1359.

[53] John H. Lagergren, John T. Nardini, G. Michael Lavigne, Erica M. Rutter, Kevin B. Flores, Learning partial differential equations for biological transport models from noisy spatio-temporal data, Proc. R. Soc. A. 476 (2234) (2020) 20190800.

[54] Razvan C. Fetecau, Yanghong Huang, Theodore Kolokolnikov, Swarm dynamics and equilibria for a nonlocal aggregation model, Nonlinearity 24 (10) (2011) 2681.

[55] Grigorii Noikhovich Milstein, Numerical Integration of Stochastic Differential Equations, Vol. 313, Springer Science & Business Media, 1994.

[56] Jean Dolbeault, Benoît Perthame, Optimal critical mass in the two dimensional Keller–Segel model in R2, C. R. Math. 339 (9) (2004) 611–616.

[57] T Tony Cai, Lie Wang, Orthogonal matching pursuit for sparse signal recovery with noise, IEEE Trans. Inform. Theory 57 (7) (2011) 4680–4688.

[58] Zhongyang Li, Fei Lu, Mauro Maggioni, Sui Tang, Cheng Zhang, On the identifiability of interaction functions in systems of interacting particles, Stoch. Processes Appl. 132 (2021) 135–163.

[59] E. Weinan, Principles of Multiscale Modeling, Cambridge University Press, 2011.