Learning interaction kernels in mean-field equations of 1st-order systems of interacting particles

Quanjun Lang*, Fei Lu*

Abstract

We introduce a nonparametric algorithm to learn interaction kernels of mean-field equations for 1st-order systems of interacting particles. The data consist of discrete space-time observations of the solution. By least squares with regularization, the algorithm learns the kernel efficiently on data-adaptive hypothesis spaces. A key ingredient is a probabilistic error functional derived from the likelihood of the mean-field equation's diffusion process. The estimator converges in a weighted L2 space at a rate for the trade-off between the numerical error and approximation error. We demonstrate our algorithm on three typical examples: the opinion dynamics with a piecewise linear kernel, the granular media model with a quadratic kernel, and the aggregation-diffusion with a repulsive-attractive kernel.

Keywords: nonparametric inference, mean-field equation, interacting particles/agents, inverse problem, reproducing kernel Hilbert space
AMS subject classifications(MSC2010): 68Q32, 35R30, 70F17, 62M99

Contents

*Department of Mathematics, Johns Hopkins University. Email: qlang1@math.jhu.edu; feilu@math.jhu.edu

1 Introduction

We study the inverse problem of estimating the radial interaction kernel ϕ of the mean-field equation

$$\begin{aligned} \partial_t u &= \nu\Delta u + \nabla\cdot[u(K_\phi * u)], x \in \mathbb{R}^d, t > 0, \\ u(x,t) &\geqslant 0, \quad \int_{\mathbb{R}^d} u(x,t)dx = 1, \quad \forall x, t, \end{aligned} \tag{1.1}$$

from observations of a solution discrete in space-time. For simplicity, we assume that $\Omega := \bigcup_{t\in[0,T]} \text{supp}(u(\cdot,t)) \subseteq \mathbb{R}^d$ is bounded with smooth boundary. Then $\partial_t u|_{\partial\Omega} = 0, \nabla_x u|_{\partial\Omega} = 0$. Here ν is the viscosity constant and $K_\phi : \mathbb{R}^d \to \mathbb{R}^d$ is the gradient of a *radial interaction potential* Φ, whose derivative ϕ is called the *interaction kernel*,

$$K_\phi(x) = \nabla(\Phi(|x|)) = \phi(|x|)\frac{x}{|x|}, \text{ with } \phi(r) = \Phi'(r).$$

We denote $K_\phi * u(x,t) = \int_\Omega K_\phi(x-y)u(y,t)dy$. Since only the derivative of Φ affects the equation, we assume, without lost of generality, that the potential satisfies $\Phi(0) = 0$.

Equation (1.1) is the mean-field of the 1st-order stochastic interacting particle system

$$\frac{d}{dt}X_t^i = \frac{1}{N}\sum_{i'=1}^{N} \phi(|X_t^j - X_t^i|)\frac{X_t^j - X_t^i}{|X_t^j - X_t^i|} + \sqrt{2\nu}dB_t^i, \quad \text{for } i = 1,\dots,N \tag{1.2}$$

when $N \to \infty$, where X_t^i represents the i-th particle's position (or agent's opinion), and B_t^i is a standard Brownian motion. Such systems arise in many disciplines: particles or molecules in microscopic models in statistical physics and quantum mechanics [12] and in granular media [26], cells [17, 11, 4] and neural networks [2] in biology, opinions of agents in social science [27], and in Monte Carlo sampling [9], to name just a few, and we refer to [27, 15] for the considerable literature.

Motivated by these applications, there has been increasing interests in the inverse problem of estimating the interaction kernel (or the interaction potential) of the mean-field equation. However, except for ideal situations in physics, little information on the interaction kernel is available, which may vary largely from smooth functions in granular media [6] to piece-wise constant function in opinion dynamics [27] or singular kernel in the Keller-Segel model [4]. Thus, it is crucial to develop new methods beyond parametric estimation (see e.g., [10]). Towards this direction, recent efforts [3, 24, 22, 23, 32] estimate the kernel by nonparametric regression for systems with finitely many particles from multiple trajectories. For large systems, data of trajectories of all particles are often unavailable, instead, it is practical to consider data consisting of a macroscopic concentration density of the particles, i.e., the solution of the mean-field equation.

We introduce a nonparametric learning algorithm (see Algorithm 1) to estimate the interaction kernel ϕ from data with a performance guarantee. The algorithm learns ϕ on a data-adaptive hypothesis space by least squares with regularization. A key ingredient is a probabilistic error functional derived from the likelihood of the diffusion process whose Fokker-Planck equation is the mean-field equation (see Theorem 2.1). The error functional is quadratic, thus we can compute its minimizer by least squares. Furthermore, it does not require spatial derivatives, thus it is suitable for discrete data (see $\mathcal{E}_{M,L}$ in (2.17)).

Our estimator converges as the space-time mesh size decreases, in a weighted L2 space, at a rate for the trade-off between the numerical error and approximation error. More precisely, with space dimension $d = 1$, we consider data consisting of a solution observed on space-time mesh: $\{u(x_m,t_l)\}_{m,l=1}^{M,L}$, where $x_m - x_{m-1} = \Delta x$ and $t_l - t_{l-1} = \Delta t$. We introduce a probability measure $\bar{\rho}_T$ (see (2.2)) for for the kernel's variable, representing the region and intensity of exploration by data. Denote $\mathcal{H}$ a hypothesis space with dimension n and denote $\widehat{\phi}_n$ the projection of ϕ on it in $L^2(\bar{\rho}_T)$. Our estimator $\widehat{\phi}_{n,M,L}$ in (2.21), based on Riemann sum approximation to integrals in the error functional, converges as $(\Delta x, \Delta t)$ decreases (see Theorem 3.5),

$$\|\widehat{\phi}_{n,M,L} - \widehat{\phi}_n\|_{L^2(\bar{\rho}_T)} \leqslant C(\Delta x^\alpha + \Delta t),$$

Table 1 Notations

Notation	Description
ϕ and Φ	true interaction kernel and potential, $\phi = \Phi'$
ψ and Ψ'	a generic interaction kernel and potential, $\psi = \Psi'$
$K_\psi(x) = \psi(\lvert x\rvert)\frac{x}{\lvert x\rvert}$	interaction function with kernel ψ
$x \cdot y$	the inner product between vector x, y
$\lVert\cdot\rVert_\infty$ and $\lVert\cdot\rVert_{k,\infty}$	the L^∞ norm and $W^{k,\infty}$ norm, $k \geqslant 1$
$L^2(\bar{\rho}_T)$	L2 space with $\bar{\rho}_T$ in (2.2)
$H_{\overline{R}_T}$	the RKHS with reproducing kernel $\overline{R}_T$ in (2.8)
$\mathcal{H} = \text{span}\,\{\phi_i\}_{i=1}^n$	hypothesis space with basis functions ϕ_i
$\mathcal{E}_{M,L}(\psi)$	error functional in (2.17), from data $\{u(x_m, t_l)\}_{m,l=0}^{M,L}$
$\widehat{\phi}_{n,M,L}$	estimator: minimizer of $\mathcal{E}_{M,L}$ on $\mathcal{H}$ with dimension n

assuming suitable identifiability conditions and regularity on the solution. Here, α is the order of the numerical integrator and we have $\alpha = 1$ for the Riemann sum integrator. We further consider the rate of convergence as $\Delta x \to 0$ when $\Delta t = 0$, assuming that we can enlarge the hypothesis space $\mathcal{H}$ to control the approximation error by $\lVert\widehat{\phi}_n - \phi\rVert_{L^2(\bar{\rho}_T)} \lessapprox n^{-s}$ with $s \geqslant 1$. With a data-adaptive n for the trade-off between the numerical error and the approximation error, we have

$$\lVert\widehat{\phi}_{n,M,\infty} - \phi\rVert_{L^2(\bar{\rho}_T)} \leqslant \underbrace{\lVert\widehat{\phi}_{n,M,\infty} - \widehat{\phi}_n\rVert_{L^2(\bar{\rho}_T)}}_{\text{numerical error}} + \underbrace{\lVert\widehat{\phi}_n - \phi\rVert_{L^2(\bar{\rho}_T)}}_{\text{approximation error}} \lessapprox (\Delta x)^{\alpha s/(s+1)}, \tag{1.3}$$

(see Theorem 3.7). That is, we achieve a rate of convergence $\frac{\alpha s}{s+1}$, optimal in the sense that it approaches the numerical integrator's order α when the kernel becomes smooth ($s \to \infty$).

We demonstrate the efficiency of the algorithm on three typical examples: the granular media model with a quadratic kernel (Section 4.2), the opinion dynamics with a piecewise linear kernel (Section 4.3), and the repulsion-attraction system with a singular kernel (Section 4.4). In each example, our algorithm leads to accurate estimators that can reproduce highly accurate solutions and free energy. For the smooth quadratic kernel, our estimator achieves the theoretical rate of convergence. For non-smooth piecewise linear kernel and the singular repulsive-attractive kernel, our estimator converges at lower rates.

The remainder of the paper is organized as follows. We present the learning algorithm in Section 2, where we introduce the error functional and the estimator, discuss the choice of basis functions for the hypothesis space, provide practical guidance on regularization and dimension selection. Section 3 studies the rate of convergence of the estimator when the space mesh refines, with the technical proofs postponed in Appendix A. Numerical examples in Sections 4 demonstrate the efficiency of our algorithm on the three examples. We discuss the limitations of this study and directions for future research in Section 5.

Notation We will use the notations in Table 1. We denote by $\lVert\cdot\rVert_\infty$ and $\lVert\cdot\rVert_{k,\infty}$ the L^∞ norm and the $W^{k,\infty}$ norm, respectively, on the corresponding domains. For example, $\lVert u\rVert_\infty$ and $\lVert u\rVert_{1,\infty}$ denote the $L^\infty(\Omega \times [0,T])$ and $W^{1,\infty}(\Omega \times [0,T])$ norms,

$$\lVert u\rVert_\infty = \sup_{x\in\Omega, t\in[0,T]} |u(x,t)|, \quad \lVert u\rVert_{1,\infty} = \lVert\nabla_{x,t} u\rVert_\infty + \lVert u\rVert_\infty .$$

Similarly, $\lVert\phi\rVert_\infty$ and $\lVert\phi\rVert_{k,\infty}$ denote the L^∞ and $W^{k,\infty}$) norms on $\text{supp}(\bar{\rho}_T)$, respectively.

2 Inference of the interaction kernel

We introduce an efficient algorithm estimating the interaction kernel by least squares in a nonparametric fashion. The key is a probabilistic error functional, which is the expectation of the negative likelihood ratio of the diffusion

process whose Fokker-Planck equation is the mean-field equation. Our estimator, the minimizer of the error functional, is then an extension of the maximal likelihood estimator. Remarkably, we can compute the estimator and the error functional without using any spatial derivative of the solution, allowing us to recover the interaction kernel from discrete data. We also discuss the function space of learning, the choice of basis functions and selection of dimension for the hypothesis space, and regularization.

2.1 The error functional and estimator

Suppose first that the data is a continuous space-time solution u on $[0,T]$, we derive an error functional from the likelihood of the diffusion process $(\bar{X}_t, t \in [0,T])$ described by the mean-field equation. More precisely, Eq.(1.1) is the Fokker-Planck equation (or the Kolmogorov forward equation) of the nonlinear stochastic differential equation (see e.g.[26])

$$\begin{cases} d\bar{X}_t = -K_\phi * u(\bar{X}_t, t)dt + \sqrt{2\nu}dB_t, \\ \mathcal{L}(\bar{X}_t) = u(\cdot, t), \end{cases} \tag{2.1}$$

for $t \geqslant 0$. Here $\mathcal{L}(\bar{X}_t)$ denotes the probability density of $\bar{X}_t$ if u is a regular solution, or the probability measure of $\bar{X}_t$ if u is a distribution solution, depending on the initial condition and the interaction kernel. In either case, we can write the convolution as

$$K_\phi * u(\bar{X}_t, t) = \mathbb{E}[K_\phi(\bar{X}_t - \bar{X}_t') \mid \bar{X}_t] = \mathbb{E}[\phi(|\bar{X}_t - \bar{X}_t'|)\frac{\bar{X}_t - \bar{X}_t'}{|\bar{X}_t - \bar{X}_t'|} \mid \bar{X}_t],$$

where $\bar{X}_t'$ is an independent copy of $\bar{X}_t$.

We start from the ambient function space for the interaction kernel: $L^2(\bar{\rho}_T)$, where $\bar{\rho}_T$ is the average-in-time distribution of $|\bar{X}_t' - \bar{X}_t|$ (denoted by ρ_t) on $[0,T]$:

$$\bar{\rho}_T(dr) := \frac{1}{T}\int_0^T \rho_t(dr)dt, \quad \rho_t(dr) := \mathbb{E}[\delta(|\bar{X}_t' - \bar{X}_t| \in dr)]. \tag{2.2}$$

Note that $\bar{\rho}_T$ depends on the initial distribution $u(\cdot, 0)$ and the true interaction kernel ϕ. We point out that the measure $\bar{\rho}_T$ is different from the empirical measure of pairwise distances in particle systems [24, 23, 21], because $\bar{X}_t$ and $\bar{X}_t'$ are independent copies and are no longer interacting particles. However, in view of inference, the high probability region of $\bar{\rho}_T$ is where $|\bar{X}_t - \bar{X}_t'|$ explores the interaction kernel the most, as such, the natural function space of learning is $L^2(\bar{\rho}_T)$. Also, the space $L^2(\bar{\rho}_T)$ ensures that our error functional below is well-defined.

Theorem 2.1 (Error functional)**.** *Let u be a solution to (1.1) on [0,T] with interaction kernel ϕ. Let $\psi \in L^2(\bar{\rho}_T)$ with $\bar{\rho}_T$ in (2.2), $\Psi(r) = \int_0^r \psi(s)ds$ and $K_\psi(x) = \nabla\Psi(|x|)$. The error functional*

$$\mathcal{E}(\psi) := \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \left[\left| K_\psi * u \right|^2 u + 2\partial_t u(\Psi * u) + 2\nu\nabla u \cdot (K_\psi * u) \right] dx\, dt \tag{2.3}$$

is the expectation of the average-in time negative log-likelihood of the process $\bar{X}_t$ in (2.1). Furthermore, if $\psi \in W^{1,\infty}$, we can replace the integrand $\nabla u \cdot (K_\psi * u)$ by $-u(\Delta\Psi * u)$.

Proof. We denote by $\mathcal{P}_\phi$ the law of the process $(\bar{X}_t)$ on the path space with initial condition $\bar{X}_0 \sim u(\cdot, 0)$, with the convention that $\mathcal{P}_0$ denotes the Winner measure. Then, the negative log-likelihood ratio of a trajectory $\bar{X}_{[0,T]}$ from $\mathcal{P}_\psi$ relative to $\mathcal{P}_\phi$ is (see e.g., [19, Section 1.1.4] or [16, Section 3.5])

$$\mathcal{E}_{\bar{X}_{[0,T]}}(\psi) = -2\log\frac{d\mathcal{P}_\psi}{d\mathcal{P}_\phi} = \frac{1}{T}\int_0^T \left(|K_\psi * u(\bar{X}_t)|^2 dt + 2\langle K_\psi * u(\bar{X}_t), d\bar{X}_t \rangle \right),$$

where $\frac{d\mathcal{P}_\psi}{d\mathcal{P}_\phi}$ is the Radon-Nikodym derivative.

Taking expectation and noting that $d\bar{X}_t = -K_\phi * u(\bar{X}_t)dt + \sqrt{2\nu}dB_t$, we obtain

$$\begin{aligned}\mathbb{E}\mathcal{E}_{\bar{X}_{[0,T]}}(\psi) =& \frac{1}{T}\int_0^T \mathbb{E}\left[|K_\psi * u(\bar{X}_t)|^2 - 2K_\psi * u(\bar{X}_t)\cdot K_\phi * u(\bar{X}_t)\right] dt.\\ =& \frac{1}{T}\int_0^T \int_{\mathbb{R}^d}\left[\left|K_\psi * u\right|^2 u - 2u(K_\phi * u)\cdot(K_\psi * u)\right]\, dx\, dt,\end{aligned} \tag{2.4}$$

where we used that the fact that for any $\psi, \phi \in L^2(\bar{\rho}_T)$ (recalling that $u(\cdot,t)$ is the law of $\bar{X}_t$),

$$\mathbb{E}[K_\psi * u(\bar{X}_t)\cdot K_\phi * u(\bar{X}_t)] = \int_{\mathbb{R}^d} u(K_\phi * u)\cdot(K_\psi * u)\, dx.$$

Noticing that $K_\psi * u = \nabla\Psi * u = \nabla(\Psi * u)$ for any $\psi = \Psi'$ and using (1.1), we have

$$\begin{aligned}&\int_0^T\int_{\mathbb{R}^d} u(K_\phi * u)\cdot(K_\psi * u)dx\, dt = \int_0^T\int_{\mathbb{R}^d} u(K_\phi * u)\cdot(\nabla\Psi * u)dx\, dt\\ =& -\int_0^T\int_{\mathbb{R}^d}(\nabla\cdot[u(K_\phi * u)])\Psi * u dx\, dt \quad (\text{ by integration by parts (IbP)})\\ =& -\int_0^T\int_{\mathbb{R}^d}(\partial_t u - \nu\Delta u)(\Psi * u)dx\, dt \quad (\text{ by Equation (1.1)})\\ =& -\int_0^T\int_{\mathbb{R}^d}\partial_t u(\Psi * u)dx\, dt - \nu\int_0^T\int_{\mathbb{R}^d}\nabla u\cdot(K_\psi * u)dx\, dt, \text{ (by IbP).}\end{aligned} \tag{2.5}$$

Combining (2.4), (2.5) and (2.3), we obtain

$$\mathbb{E}\mathcal{E}_{\bar{X}_{[0,T]}}(\psi) = \frac{1}{T}\int_0^T\int_{\mathbb{R}^d}\left[\left|K_\psi * u\right|^2 u - 2u(K_\phi * u)\cdot(K_\psi * u)\right]\, dx\, dt = \mathcal{E}(\psi). \tag{2.6}$$

At last, from integration by parts, we can replace the integrand $\nabla u\cdot(K_\psi * u)$ by $-u(\Delta\Psi * u)$ if $\psi \in W^{1,\infty}$. □

To simplify the notation, we introduce the following bilinear form: for any $\phi, \psi \in \mathcal{H}$,

$$\begin{aligned}\langle\!\langle \phi, \psi\rangle\!\rangle :=& \frac{1}{T}\int_0^T\int_{\mathbb{R}^d}(K_\phi * u)\cdot(K_\psi * u)u(x,t)dx\, dt\\ =& \frac{1}{T}\int_0^T\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}K_\phi(y)\cdot K_\psi(z)\int_{\mathbb{R}^d}u(x-y,t)u(x-z,t)u(x,t)dxdydz\, dt\\ =& \int_{\mathbb{R}^+}\int_{\mathbb{R}^+}\phi(r)\psi(s)\,\overline{R}_T(r,s)\bar{\rho}_T(dr)\bar{\rho}_T(ds),\end{aligned} \tag{2.7}$$

where the kernel $\overline{R}_T(r,s)$, obtained by a change of variables to polar coordinates, is

$$\overline{R}_T(r,s) = \frac{(rs)^{d-1}}{\bar{\rho}_T'(r)\bar{\rho}_T'(s)T}\int_0^T\int_{\mathbb{S}^d}\int_{\mathbb{S}^d}\int_{\mathbb{R}^d}\xi\cdot\eta\, u(x-r\xi,t)u(x-s\eta,t)u(x,t)dxd\xi d\eta\, dt. \tag{2.8}$$

Here $\bar{\rho}_T'$ denotes the probability density of the measure $\bar{\rho}_T$ and $\mathbb{S}^d$ denotes the unit sphere in $\mathbb{R}^d$. The integral kernel $\overline{R}_T$ is a Mercer kernel on $L^2(\bar{\rho}_T)$ and it defines a reproducing kernel Hilbert space (RKHS), which is a function space of learning and we denote it by $H_{\overline{R}_T}$ (see Remark 2.4 for details).

Then, in view of (2.6), we can write the error functional as

$$\mathcal{E}(\psi) = \langle\!\langle\psi,\psi\rangle\!\rangle - 2\langle\!\langle\psi,\phi\rangle\!\rangle = \langle\!\langle\psi-\phi,\psi-\phi\rangle\!\rangle - \langle\!\langle\phi,\phi\rangle\!\rangle. \tag{2.9}$$

Note that it is quadratic, so we can compute its minimizer on a finite-dimensional hypothesis space by least squares.

Theorem 2.2 (Estimator from continuous data)**.** *For any space $\mathcal{H} = \operatorname{span}\{\phi_i\}_{i=1}^n \subseteq L^2(\bar{\rho}_T)$ such that the normal matrix A in (2.11) is invertible, the unique minimizer of the error functional $\mathcal{E}$ on $\mathcal{H}$ is given by*

$$\widehat{\phi}_n = \sum_{i=1}^n \widehat{c}_i \phi_i, \text{ with } \widehat{c} = A^{-1} b. \tag{2.10}$$

where the normal matrix A and vector b are given by (with $\Phi_i(r) = \int_0^r \phi_i(s)ds$)

$$A_{ij} = \langle\!\langle \phi_i, \phi_j \rangle\!\rangle = \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} (K_{\phi_i} * u) \cdot (K_{\phi_j} * u) u(x,t) dx\, dt, \tag{2.11}$$

$$b_i = \langle\!\langle \phi, \phi_i \rangle\!\rangle = -\frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \left[\partial_t u\, \Phi_i * u - \nu \nabla u \cdot (K_{\phi_i} * u)\right] dx\, dt. \tag{2.12}$$

Again, we can replace $-\nabla u \cdot (K_{\phi_i} * u)$ by $u((\nabla \cdot K_{\phi_i}) * u)$ if $\phi_i \in W^{1,\infty}$.

Remark 2.3. *We set $\Phi_i(0) = 0$ for the anti-derivative of ϕ_i for simplicity. In general, the constant $\Phi_i(0)$ does not affect the integral $\int_0^T \int_{\mathbb{R}^d} \partial_t u\, \Phi_i * u dx dt$ in (2.12), because for any constant c, we have $c * u = c$ and $\int_0^T \int_{\mathbb{R}^d} \partial_t c dx dt = c \int_{\mathbb{R}^d} u(x,0) - u(x,T) dt = c - c = 0$.*

Proof. Recall that with the bilinear form (2.7), the error functional can be written as (2.9). For each $\psi = \sum_{i=1}^n c_i \phi_i \in \mathcal{H}$, we can write the error functional as,

$$\mathcal{E}(\psi) = \mathcal{E}(c) = c^\top A c - 2 b^\top c, \tag{2.13}$$

where A is given by (2.11) and b are given by

$$\begin{aligned} b_i = \langle\!\langle \phi, \phi_i \rangle\!\rangle &= \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} (K_\phi * u) \cdot (K_{\phi_i} * u) u(x,t) dx\, dt \\ &= -\frac{1}{T}\int_0^T \int_{\mathbb{R}^d} (\partial_t u - \nu \Delta u)(\Phi_i * u) dx\, dt \\ &= -\frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \partial_t u (\Phi_i * u) dx\, dt - \nu \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \nabla u \cdot (K_{\phi_i} * u) dx\, dt, \end{aligned}$$

where in the last equality, we used integration by parts to get rid of $\nu \Delta u$. Applying integration by parts to $\int_{\mathbb{R}^d} \nabla u \cdot (K_{\phi_i} * u)$ and note that $\nabla (K_{\phi_i} * u) = (\nabla \cdot K_{\phi_i}) * u$, we obtain (2.12).

Since the error functional is quadratic, the minimizer can be given explicitly by (2.10). □

Remark 2.4 (RKHS, identifiability and ill-posedness)**.** *The integral kernel $\overline{R}_T$ is a Mercer kernel and it defines an RKHS $H_{\overline{R}_T} \subseteq L^2(\bar{\rho}_T)$, a function space on which the loss functional $\mathcal{E}$ has a unique minimizer. Roughly speaking, the RKHS is the image of an operator, i.e., $H_{\overline{R}_T} = \mathcal{L}_{\overline{R}_T}^{1/2} L^2(\bar{\rho}_T)$, where $\mathcal{L}_{\overline{R}_T}$ is the integral operator*

$$\mathcal{L}_{\overline{R}_T} \varphi = \int_{\mathbb{R}^+} \varphi(r) \overline{R}_T(r,s) \bar{\rho}_T(dr), \text{ for } \varphi \in L^2(\bar{\rho}_T). \tag{2.14}$$

Note that we have from the definition of the bilinear form (2.7):

$$\langle\!\langle \varphi, \psi \rangle\!\rangle = \langle \varphi, \mathcal{L}_{\overline{R}_T} \psi \rangle_{L^2(\bar{\rho}_T)}. \tag{2.15}$$

Thus, the Frechét derivative of the loss functional on $L^2(\bar{\rho}_T)$ is $\mathcal{E}'(\varphi) = \mathcal{L}_{\overline{R}_T} \varphi - \mathcal{L}_{\overline{R}_T} \phi$, which has a unique zero on $H_{\overline{R}_T}$. Thus, if the true ϕ is in $H_{\overline{R}_T}$, it is identifiable by the loss functional $\mathcal{E}$ as the unique minimizer of $\mathcal{E}$ in $H_{\overline{R}_T}$. Meanwhile, note that we are actually solving $\widehat{\phi} = \mathcal{L}_{\overline{R}_T}^{-1}[\mathcal{L}_{\overline{R}_T} \phi]$ with $\mathcal{L}_{\overline{R}_T} \phi$ estimated from data. This inverse problem is ill-posed because $\mathcal{L}_{\overline{R}_T}$ is a strictly positive compact operator and its inverse is unbounded. We refer to [20] for more discussions on these issues.

Remark 2.5. *The normal matrix's invertibility depends on the basis functions $\{\phi_i\}_{i=1}^n$ of $\mathcal{H}$. Its smallest eigenvalue is the smallest eigenvalue of the integral operator with kernel $\overline{R}_T$ on $\mathcal{H}$ (see Proposition 3.4). Thus, to make the normal matrix invertible, we need the coercivity condition in Definition 3.3.*

Remark 2.6 (The PDE discrepancy error functional)**.** *Our error functional in (2.3) has two advantages over the PDE discrepancy error functional*

$$\mathcal{E}_0(\psi) = \int_0^T \int_{\mathbb{R}^d} \left|\nabla.(u(K_\psi * u)) - g\right|^2 dx\, dt.$$

where $g = \partial_t u - \nu\Delta u$. First, it requires the derivatives ∇u and Δu, because the integration by parts does not apply. Numerical errors from approximating these derivatives can prevent this approach from working, particularly when the data is not a perfect solution to the PDE. This is because the inverse problem is ill-posed (see Remark 2.4, equivalently, the regression matrix A tends to be ill-conditioned). In our numerical tests with solution generated by SPCC (see Section 4.1), this approach fails to produce any acceptable estimator in various settings. Second, since $u(\cdot, t)$ is a probability density, all component of our error functional can be written as expectations, allowing for Monte Carlo approximations, which is essential in high-dimensional problems (see Remark 2.7).

Estimator from discrete data When data are discrete in space-time, we approximate the integrals in the estimator and the error functional by numerical integrators. For simplicity, we consider only data on a regular mesh for $d = 1$ and use the Riemann sum. In practice, we could use higher-order numerical methods for the integration and convolution, for instance, the trapezoid method for the integrals and Fourier transform for the convolution. Note also that these integrations are expectations, so in general, particularly for high dimensional cases, the data can also be independent samples of the distribution, and we approximate the integrations by the empirical mean.

Suppose that the data are $\{u(x_m, t_l)\}_{m,l=1}^{M,L}$, with $t_l = l\Delta t$ for $L = T/\Delta t$ and with $\{x_m\}_{m=1}^M$ being a uniform mesh of Ω with length/area Δx.

From these data, we approximate all the integrals by Riemann sum. We approximate $\bar{\rho}_T$ in (2.2) by its the empirical measure:

$$\rho_L^M(dr) = \frac{1}{L}\sum_{l=1}^{L}\sum_{m,m'=1}^{M,M} u(x_m, t_l)u(x_{m'}, t_l)\delta_{|x_m - x_{m'}|}(r)dr. \tag{2.16}$$

With $\psi = \sum_{i=1}^n c_i\phi_i$, we compute the error functional in (2.13) by

$$\mathcal{E}_{M,L}(\psi) = \mathcal{E}_{M,L}(c) = c^\top A_{M,L}c - 2b_{M,L}^\top c, \tag{2.17}$$

where the normal matrix $A_{M,L}$ and vector $b_{M,L}$, approximating A and b in (2.11)-(2.12), are

$$A_{n,M,L}^{i,j} := \frac{1}{L}\sum_{l=1}^{L}\sum_{m=1}^{M}\left[\left(P_{n,M,L}^i \cdot P_{n,M,L}^j\right)u\right](x_m, t_l)\Delta x, \tag{2.18}$$

$$b_{n,M,L}^i := -\frac{1}{L}\sum_{l=1}^{L}\sum_{m=1}^{M}\left[\widehat{\partial_t u}\, Q_{n,M,L}^i + \nu u R_{n,M,L}^i\right](x_m, t_l)\Delta x. \tag{2.19}$$

Here $P_{n,M,L}^i$, $Q_{n,M,L}^i$ and $R_{n,M,L}^i$ are Riemann sum approximations of $K_{\phi_i} * u$, $\Phi_i * u$ and $\nabla K_{\phi_i} * u$, respectively,

and $\widehat{\partial_t u}$ is the finite difference approximation of $\partial_t u$

$$\begin{aligned} P^i_{n,M,L}(x,t) &:= \sum_{m=1}^{M} K_{\phi_i}(x_m) u(x-x_m,t)\Delta x, \\ Q^i_{n,M,L}(x,t) &:= \sum_{m=1}^{M} \Phi_i(x_m) u(x-x_m,t)\Delta x, \\ R^i_{n,M,L}(x,t) &:= \sum_{m=1}^{M} \nabla\cdot K_{\phi_i}(x_m) u(x-x_m,t)\Delta x, \\ \widehat{\partial_t u}(x,t) &:= \frac{1}{\Delta t}\sum_{l=1}^{L} [u(x,t_l) - u(x,t_{l-1})]\mathbf{1}_{(t_{l-1},t_l]}(t). \end{aligned} \tag{2.20}$$

A few remarks on the numerical aspects: (1) one can use high-order numerical integrators to increase the accuracy of the spatial integrals; (2) we computed $R^i_{n,M,L}(x,t)$ assuming that the basis functions $\{\phi_i\}$ are in $W^{1,\infty}$. If $\{\phi_i\}$ are not differentiable, we can use ∇u as in (2.12); (3) in practice, we use zero padding for u by setting $u(x_i,t)=0$ if $x_i \in \partial\Omega$; (4) also, we normalize the vector $\{u(x_m,t)\}_{m=1}^M$ so that $\sum_{m=1}^M \Delta x u(x_m,t) = 1$ for each t. This ensures that $\sum_{m=1}^M \left[\widehat{\partial_t u}\, Q^i_{n,M,L}\right](x_m,t_l)$ does not depend on the constant $\Phi(0)$, as discussed in Remark 2.3.

Correspondingly, the estimator is

$$\widehat{\phi}_{n,M,L} = \sum_{i=1}^{n} \widehat{c}^i_{n,M,L}\phi_i, \qquad \text{with } \widehat{c}_{n,M,L} = A^{-1}_{n,M,L} b_{n,M,L}. \tag{2.21}$$

2.2 Basis functions for the hypothesis space

We consider two classes of basis function for the hypothesis space $\mathcal{H}$: the B-spline piecewise polynomials whose knots are uniform partition of $\bar{\rho}_T$'s support; the RKHS basis consisting of eigenfunctions of the integral operator with kernel $\overline{R}_T$ in (2.8). The B-splines are universal local basis, while the RKHS basis functions are global basis adaptive to data.

B-spline basis functions B-spline is a class of piecewise polynomials, and is capable of representing the local information of the interaction kernel. Here we review briefly the recurrence definition and properties of the balanced B-splines, for more details we refer to [29, Chapter 2] and [25].

Given a nondecreasing sequence of real numbers $\{r_0, r_1, \ldots, r_m\}$ (called knots), the B-spline basis functions of degree p, denoted by $\{N_{i,p}\}_{i=0}^{m-1}$, is defined recursively as

$$\begin{aligned} N_{i,0}(r) &= \begin{cases} 1, & r_i \leqslant r < r_{i+1}, \\ 0, & otherwise, \end{cases} \\ N_{i,p}(r) &= \frac{r-r_i}{r_{i+p}-r_i} N_{i,p-1}(r) + \frac{r_{i+p+1}-r}{r_{i+p+1}-r_{i+1}} N_{i+1,p-1}(r). \end{aligned} \tag{2.22}$$

Each B-spline basis function $N_{i,p}$ is a nonnegative piecewise polynomial of degree p, locally supported on $[r_i, r_{i+p+1}]$, and it is $p-k$ times continuously differentiable at a knot, where k is the multiplicity of the knot. Hence, continuity increases when the degree increases, and continuity decreases when knot multiplicity increases. Also, it satisfies partition unity: for each $r \in [r_i, r_{i+1}]$, $\sum_j N_{j,p}(r) = \sum_{j=i-p}^{i} N_{j,p}(r) = 1$.

For $f \in W^{k,\infty}$, denote $f_{\mathcal{H}}$ its projection to the linear space $\mathcal{H}$ of B-splines with degree $p \geqslant k-1$ and denote $D^{(k)}f$ the k-th order derivative. We have [25, p.45, Theorem 17]

$$\|f - f_{\mathcal{H}}\|_\infty \leqslant C_p h^k \|D^{(k)} f\|_\infty, \tag{2.23}$$

where C_p is a constant depending on p, and $h = \max_i |r_i - r_{i-1}|$.

We set the knots to be uniform partition of the support of $\bar{\rho}_T$, with augmented knots at the ends of the support interval, say $[R_{min}, R_{max}]$,

$$R_{min} = r_{-p+1} = \cdots = r_0 \leqslant r_1 \leqslant \cdots \leqslant r_m = \cdots = r_{p+m-1} = R_{min}.$$

We set the basis functions of the hypothesis $\mathcal{H}$, whose dimension is $m + p$, to be

$$\phi_i(r) = N_{i-p,p}(r),\ r > 0,\ i = 1, \ldots, m + p.$$

Thus, the basis functions $\{\phi_i\}$ are degree-p piecewise polynomials with knots adaptive to $\bar{\rho}_T$.

The RKHS basis The RKHS basis functions $\{\phi_i\}$ are the eigenfunctions of the integral operator with kernel $\overline{R}_T(r, s)$ (defined in (2.8)) on $L^2(\bar{\rho}_T)$, that is,

$$\mathcal{L}_{\overline{R}_T}\phi_i(s) = \int_{\mathbb{R}^+} \phi_i(r)\overline{R}_T(r, s)\bar{\rho}_T(dr) = \lambda_i\phi_i(s),\ \text{ in } L^2(\bar{\rho}_T). \tag{2.24}$$

Thus, we have $\langle\!\langle \phi_i, \phi_j \rangle\!\rangle = \lambda_i\langle\phi_i, \phi_j\rangle_{L^2(\bar{\rho}_T)} = \lambda_i\delta_{i-j}$ and it leads to a diagonal normal matrix A in (2.11). We estimate these eigenfunctions from data. Hence, the RKHS basis is adaptive to data.

We compute these eigenfunctions by eigen-decomposition of the matrix $(\overline{R}_T(r_i, r_j))$ on a mesh when its size is manageable. When the mesh size is large, we can compute them using a linear transformation from a more convenient basis. That is, we start from linearly independent functions $\{\psi_i\}_{i=1}^n$ (e.g., B-splines) and evaluate $\tilde{A} := \langle\!\langle \psi_i, \psi_j \rangle\!\rangle$ and $P = \langle\phi_i, \phi_j\rangle_{L^2(\bar{\rho}_T)}$, then we solve the generalized eigenvalue problem $\tilde{A}\alpha = \lambda P\alpha$ for eigenvectors $\{\alpha_k \in \mathbb{R}^n\}_{k=1}^n$ such that $\alpha_k^\top P\alpha_l = \delta_{k-l}$. Then, the eigenfunctions are given by $\phi_i = \sum_{l=1}^n \alpha_{i,l}\psi_l$ (we refer to [20] for more details).

2.3 Regularization

In practice, the approximate normal matrix $A_{n,M,L}$ in (2.18) may be ill-conditioned or invertible, which is likely to happen when the dimension of $\mathcal{H}$ increases because of the vanishing eigenvalues of A and the numerical errors. The ill-conditioned normal matrix may amplify the numerical error in $b_{n,M,L}$ in (2.19). To avoid such an issue, we use the Tikhonov regularization (see, e.g.,[14]), which adds a norm-induced well-conditioned matrix to the normal matrix.

More precisely, we impose a regularization norm $|||\cdot|||$ (to be specified below) such that for any $\psi = \sum_{i=1}^n c_i\phi_i$, the matrix B in $|||\psi|||^2 = c^\top Bc$ is well-conditioned. We then minimize the regularized error functional (recall (2.13))

$$\mathcal{E}_\lambda(\psi) = \mathcal{E}(\psi) + \lambda|||\psi|||^2 = c^\top(A + \lambda B)c - 2b^\top c,$$

and the regularized estimator is

$$\widehat{\phi_\lambda} = \sum_{i=1}^n c_\lambda^i\phi_i,\ c_\lambda = (A + \lambda B)^{-1}b. \tag{2.25}$$

Regularization norm We consider two regularization norms $|||\cdot|||$ and many other options are possible. For the RKHS basis, we use $B_{ij} = \delta_{ij}$, which is the common choice [7]. This is equivalent to having a prior knowledge that the coefficient c is small. For the spline basis, we choose $|||f||| = \|f\|_{H^1(\Omega)}$, and in this case $B_{ij} = \langle\!\langle \phi_i, \phi_j \rangle\!\rangle_{H^1(\Omega)}$. This is equivalent to the prior assumption that K_ϕ of the true interaction kernel has $H^1(\Omega)$ regularity.

Regularization parameter by L-curve We select the parameter λ by the L-curve method [14]. Let l be a parametrized curve in $\mathbb{R}^2$:

$$l(\lambda) = (x(\lambda), y(\lambda)) := (\log(\mathcal{E}(\widehat{\phi_\lambda}), \log(\left\|\widehat{\phi_\lambda}\right\|)).$$

Note that $\mathcal{E}(\widehat{\phi_\lambda}) = c_\lambda^\top A c_\lambda - 2b^\top c_\lambda$, and $\left\|\widehat{\phi_\lambda}\right\| = c_\lambda^\top B c_\lambda$. The optimal parameter is the maximizer of the curvature of l. In practice, we restrict λ in the spectral range of A:

$$\lambda_0 = \underset{\lambda_{\min}(A) \leqslant \lambda \leqslant \lambda_{\max}(A)}{\arg\max} \kappa(l) = \underset{\lambda_{\min}(A) \leqslant \lambda \leqslant \lambda_{\max}(A)}{\arg\max} \frac{x' y'' - x' y''}{(x'^2 + y'^2)^{3/2}}. \tag{2.26}$$

This λ_0 balances the error functional $\mathcal{E}$ and the regularization (see [14] for more details).

2.4 dimension of the hypothesis space

The dimension n must neither be too small nor too large to avoid under-fitting or over-fitting. Theorem 3.7 suggests that the dimension is $n \approx (\Delta x)^{-\alpha/(s+1)}$, where $(\Delta x)^\alpha$ is the order of convergence of the numerical integrator in the evaluation of the normal matrix and normal vector, and s is the order of decay for the distance between the true kernel and the hypothesis space. For B-spline bases, the approximation error bound in (2.23) suggests that $s = k$ for $\phi \in W^{k,\infty}$ when we select the degree $p \geqslant k$. This theoretical dimension provides only an estimate on the magnitude. In practice, to find the dimension, we first select a range $[N_1, N_2]$ for the dimension; then we choose the n that minimizes the regularized error functional.

2.5 The algorithm

We summarize the method in this section in Algorithm 1.

Algorithm 1 Estimation of the interaction kernel

Input: Data $\{u(x_i, t_l)\}_{m=1,l=1}^{M,L}$ on the space mesh $\{x_m\}_{m=1}^{M}$ with width/area Δx and time mesh $\{t_l = l\Delta t\}_{l=0}^{T}$.

Output: Estimated $\widehat{\phi}$

1: Estimate the empirical density ρ_L^M in (2.16) and find its support $[R_{min}, R_{max}]$.
2: Select a basis type, RKHS or B-spline, and estimate a dimension range $[N_1, N_2]$, compute the basis functions as described in Section 2.2.
3: **for** $n = N_1 : N_2$ **do**
4: Compute the normal matrix and vector as in (2.18)–(2.20).
5: Determine the optimal regularization constant λ_0 by (2.26).
6: Solve c_n by (2.25) and record the regularized cost $C(n) = \mathcal{E}_{\lambda_0}(\sum_{i=1}^n c_n^i \phi_i)$.
7: Select the dimension by $n^* = \underset{n \in \{N_1, N_1+1, \ldots, N_2\}}{\arg\max} C(n)$.
8: Return the estimator $\widehat{\phi} = \sum_{i=1}^{n^*} c_{n^*}^i \phi_i$.

Computational complexity Given data $\{u(x_i, t_l)\}_{m=0,l=0}^{M,L}$, where M is the space grid size (it is exponential in the space dimension d) and L is the number of time steps. Let n be the number of basis functions in the hypothesis space. The computational complexity of our algorithm is $O((n^2 + n + 1)M^2 L)$: the computation of the measure $\bar{\rho}_T$, $A_{n,M,L}(i,k)$ and $b_{n,M,L}(i)$ each is $O(M^2 L)$. The computational cost of the solution to the least squares and regularization is negligible since n is orders of magnitudes smaller than $M^2 L$.

Scalability for data consisting of many solutions Our algorithm can efficiently deal with data consists of many solutions $\{u^{(k)}(x_m, t_l)_{m=0,l=0}^{M,L}\}_{k=1}^{K}$ through paralleled computation. It can first compute the normal matrix and vector for each solution in parallel, then assemble the matrices and vectors for least squares regression with regularization.

Remark 2.7 (High dimensional case by Monte Carlo)**.** *When the space variable $x \in \mathbb{R}^d$ is high-dimensional (with $d \geqslant 4$), it becomes impractical to have data on mesh-grids since the data size increases exponentially in d. One would consider the setting when data consists of samples from the probability densities $u(\cdot, t)$. In this setting, our algorithm applies directly because all the spatial integration related elements ($\bar{\rho}_T$, the normal matrix A and vector b) can be written as expectations, which can be approximated by Monte Carlo. The computational complexity will remain the same as above with M being the sample size. We leave this as a future direction of exploration.*

3 Convergence of the estimator in mesh size

We analyze the convergence of the discrete-data estimator (2.21) to the continuous-data estimator (2.10) as the mesh refines.

We denote by $|\Omega|$ the Lebesgue measure of Ω and denote by R_Ω its radius. Note that support of $\bar{\rho}_T$ is $[0, R_\Omega]$ though Ω is in $\mathbb{R}^d$. For simplicity, we consider only the case when $\Omega = [a, b] \in \mathbb{R}^1$ and the generalization to a higher dimension is immediate. We assume that the data are

$$\textbf{Data:} \qquad \{u(x_m, t_l)\}_{m,l=0}^{M,L}, \quad x_m = a + m\Delta x, t_l = l\Delta t \tag{3.1}$$

with $M = (b-a)/\Delta x$ and $L = T/\Delta t$.

In this section, we make the following assumptions on u and the basis functions of $\mathcal{H}$.

Assumption 3.1 (Constraints on hypothesis space)**.** *Let the basis functions of $\mathcal{H} = \operatorname{span}\{\phi_i\}_{i=1}^n$ are uniformly bounded in $W^{2,\infty}([0, R_\Omega])$ with notations*

$$c_{\mathcal{H}}^{\infty} := \max_{1 \leqslant i \leqslant n} \|\phi_i\|_\infty, \; c_{\mathcal{H}}^{1,\infty} := \max_{1 \leqslant i \leqslant n} \|\phi_i\|_{1,\infty}, \; c_{\mathcal{H}}^{2,\infty} := \max_{1 \leqslant i \leqslant n} \|\phi_i\|_{2,\infty} < \infty. \tag{3.2}$$

Assumption 3.2. *Assume that solution $u \in W^{2,\infty}(\Omega \times [0, T])$ satisfies $\|u\|_{2,\infty} < \infty$.*

We remark that the second order derivatives of the solution are necessary to control of the Riemann sum approximation of the integrals. With stronger regularity on the solution and higher-order approximations of the integrals than the Euler scheme, one can obtain higher order convergence in space and time. In the other direction, since these integrals are expectations, they can be approximated Monte Carlo, we expect to remove these regularity assumptions in forthcoming research.

To make the estimators in (2.21) and (2.10) well-defined, the normal matrices must be invertible. We introduce the following coercivity condition to ensure it. It extends of the coercivity condition for N-particle systems defined in [21, 23, 24].

Definition 3.3 (Coercivity condition)**.** *The system (1.1) on $[0, T]$ satisfies a coercivity condition on a finite-dimensional linear subspace $\mathcal{H} \subseteq L^2(\bar{\rho}_T)$ with $\bar{\rho}_T$ defined in (2.2) if*

$$c_{\mathcal{H},T} := \inf_{h \in \mathcal{H}, \, \|h\|_{L^2(\bar{\rho}_T)} = 1} \langle\!\langle h, h \rangle\!\rangle > 0, \tag{3.3}$$

where $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ is defined in (2.7). When $\mathcal{H} \subseteq L^2(\bar{\rho}_T)$ is infinite-dimensional, we say coercivity condition holds on $\mathcal{H}$ if it holds on each of $\mathcal{H}$'s finite dimensional linear subspace.

We show that the coercivity constant is the smallest generalized eigenvalue of the normal matrix A in (2.11). The generalized eigenvalue problem appears because A is a representation of the integral operator in (2.24) on $\mathcal{H}$ [20].

Proposition 3.4. *Assume the coercivity condition holds on $\mathcal{H} = \operatorname{span}\{\phi_i\}_{i=1}^n$. Let A be the normal matrix in (2.11), and let $P = \langle \phi_i, \phi_j \rangle_{L^2(\bar{\rho}_T)}$. Then, the smallest singular value of $Ac = \lambda P c$ is $\lambda_{min} = c_{\mathcal{H},T}$.*

Proof. Let $\psi = \sum_{i=1}^n c_i\phi_i \in \mathcal{H}$ with c satisfying $Ac = \lambda_{min}Pc$, that is, c is an eigenvector for λ_{min}. Thus, by the coercivity condition,

$$\lambda_{min}\|\psi\|^2_{L^2(\bar{\rho}_T)} = \lambda_{min}c^\top Pc = c^T Ac = \langle\!\langle \psi, \psi \rangle\!\rangle \geqslant c_{\mathcal{H},T}\|\psi\|^2_{L^2(\bar{\rho}_T)}.$$

Thus, $\lambda_{min} \geqslant c_{\mathcal{H},T}$. Note that the space $\mathcal{H}$ here is finite dimensional, the infimum in the definition of coercivity constant is attained by some $\psi^* = \sum_{i=1}^n v_i\phi_i$. Since λ_{min} is the minimal eigenvalue, we have $c_{\mathcal{H},T}\|\psi^*\|^2_{L^2(\bar{\rho}_T)} = \langle\!\langle \psi^*, \psi^* \rangle\!\rangle = v^\top Av \geqslant \lambda_{min}v^\top Pv = \lambda_{min}\|\psi^*\|^2_{L^2(\bar{\rho}_T)}$. Hence, we have $\lambda_{min} = c_{\mathcal{H},T}$. □

3.1 Error bounds for the estimator

We show that the estimator $\widehat{\phi}_{n,M,L}$ in (2.21) converges as $M \to \infty$ and $L \to \infty$.

Theorem 3.5 (Error bounds for the estimator)**.** *Suppose that the hypothesis space $\mathcal{H} = \text{span}\{\phi_i\}_{i=1}^n$ satisfies Assumption 3.1 and denote $\widehat{\phi}_n$ the projection of ϕ on $\mathcal{H} \subseteq L^2(\bar{\rho}_T)$. Suppose that the coercivity condition in Definition 3.3 holds on $\mathcal{H}$ with a constant $c_{\mathcal{H},T} > 0$. Then, the estimator $\widehat{\phi}_{n,M,L}$ in (2.21) satisfies*

$$\|\widehat{\phi}_{n,M,L} - \widehat{\phi}_n\|_{L^2(\bar{\rho}_T)} \leqslant 2{c_{\mathcal{H},T}}^{-1}\sigma_{max}\left(c^b\sqrt{n} + c^A n\|\phi\|_{L^2(\bar{\rho}_T)}\right)(\Delta x + \Delta t), \tag{3.4}$$

where $c^A := 2|\Omega|(1+|\Omega|)(c_{\mathcal{H}}^{1,\infty})^2\|u\|^2_{1,\infty}$, $c^b := 3|\Omega|(1+R_\Omega+\nu)c_{\mathcal{H}}^{2,\infty}(\|u\|^2_{1,\infty} + \|u\|_{2,\infty})$, and σ_{max} is the square root of the largest eigenvalue of $P = (\langle\phi_i, \phi_j\rangle_{L^2(\bar{\rho}_T)})$.

Proof of Theorem 3.5. Notice that $\widehat{\phi}_{n,M,L}$ and $\widehat{\phi}_n$ are given by

$$\widehat{\phi}_{n,M,L} = \sum_{i=1}^n \widehat{c}^i_{n,M,L}\phi_i, \quad \widehat{\phi}_n = \sum_{i=1}^n \widehat{c}^i\phi_i,$$

where $\widehat{c}_{n,M,L} = A^{-1}_{n,M,L}b_{n,M,L}$ and $\widehat{c} = A^{-1}b$, we have

$$\begin{aligned}\|\widehat{\phi}_{n,M,L} - \widehat{\phi}_n\|_{L^2(\bar{\rho}_T)} &= \sqrt{(\widehat{c}_{n,M,L} - \widehat{c})^\top P(\widehat{c}_{n,M,L} - \widehat{c})} \\ &\leqslant \sigma_{max}\|\widehat{c}_{n,M,L} - \widehat{c}\| = \left\|A^{-1}_{n,M,L}b_{n,M,L} - A^{-1}b\right\|.\end{aligned}$$

Also, by the formula $A^{-1}_{n,M,L} - A^{-1} = A^{-1}_{n,M,L}(A - A_{n,M,L})A^{-1}$, we have

$$\begin{aligned}&\left\|A^{-1}_{n,M,L}b_{n,M,L} - A^{-1}b\right\| = \left\|A^{-1}_{n,M,L}(b_{n,M,L} - b) + (A^{-1}_{n,M,L} - A^{-1})b\right\| \\ \leqslant &\left\|A^{-1}_{n,M,L}\right\|\left(\|b_{n,M,L} - b\| + \|A_{n,M,L} - A\| \cdot \left\|A^{-1}b\right\|\right).\end{aligned}$$

By Proposition 3.8, for small enough Δx and Δt, we have $\|A - A_{n,M,L}\| \leqslant \frac{c_{\mathcal{H},T}}{2}$. Hence $\left\|A^{-1}_{n,M,L}\right\| \leqslant 2{c_{\mathcal{H},T}}^{-1}$. Note that $\left\|A^{-1}b\right\| = \left\|\widehat{\phi}_n\right\|_{L^2(\bar{\rho}_T)} \leqslant \|\phi\|_{L^2(\bar{\rho}_T)}$ is independent of the mesh size. Thus, we obtain (3.4) by (3.6) in Proposition 3.8. □

Remark 3.6 (High order numerical integrators)**.** *For a fixed n, the rate of convergence is the same as the order of the numerical integrators in the computation of A and b. Hence, when the solution and the kernel are smooth, we can achieve faster convergence by using a higher order numerical integrator.*

The next theorem shows that as the spatial mesh refines, the estimator converges at a rate $\frac{\alpha s}{s+1}$, where α is the order of the numerical integrator and s is the rate of decay of the approximation error by the hypothesis space (thus, s is determined by the smoothness of the true kernel since we can tune the hypothesis space).

Theorem 3.7 (Rate of convergence). *Assume $\Delta t = 0$ and consider the estimator $\widehat{\phi}_{n,M,\infty}$ in (2.21) on $\mathcal{H}$ with dimension n. Denote $\widehat{\phi}_n$ the projection of ϕ on $\mathcal{H} \subseteq L^2(\bar{\rho}_T)$. Assume that as $M = |\Omega|/(\Delta x) \to \infty$, we have $\|\widehat{\phi}_{n,M,\infty} - \widehat{\phi}_n\|_{L^2(\bar{\rho}_T)} \lessapprox n(\Delta x)^\alpha$, (for example, $\alpha = 1$ in Theorem 3.5), and assume*

$$\|\widehat{\phi}_n - \phi\|_{L^2(\bar{\rho}_T)} \lessapprox n^{-s}$$

with $s \geqslant 1$ when n increases. Then, with a data-adaptive dimension $n \approx (\Delta x)^{-\alpha/(s+1)}$, we can achieve the rate

$$\|\widehat{\phi}_{n,M,\infty} - \phi\|_{L^2(\bar{\rho}_T)} \lessapprox (\Delta x)^{\alpha s/(s+1)}.$$

Proof. Note that the total error in the estimator consists of numerical error and approximation error:

$$\|\widehat{\phi}_{n,M,\infty} - \phi\|_{L^2(\bar{\rho}_T)} \leqslant \|\widehat{\phi}_{n,M,\infty} - \widehat{\phi}_n\|_{L^2(\bar{\rho}_T)} + \|\widehat{\phi}_n - \phi\|_{L^2(\bar{\rho}_T)}.$$

Then, the total error is of the order $g(n) = n(\Delta x)^{-\alpha} + n^{-s}$. Minimizing it by solving $g'(z) = (\Delta x)^{-\alpha} - sz^{-s-1} = 0$, we get the dimension $n \approx s^{-1/(s+1)}(\Delta x)^{\alpha/(s+1)}$, and the corresponding rate of convergence is $(\Delta x)^{-\alpha s/(s+1)}$. □

The exponent s depends on both the smoothness of ϕ and the hypothesis space $\mathcal{H} = \operatorname{span}\{\phi_i\}_{i=1}^n$. When $\phi \in W^{s,\infty}$ and $\{\phi_i\}$ are B-splines with degree $p \geqslant s-1$ and uniform knots on $\operatorname{supp}(\bar{\rho}_T)$, we have $\|\phi - \widehat{\phi}_n\|_\infty \lessapprox n^{-s}$ by (2.23) and hence $\|\widehat{\phi}_n - \phi\|_{L^2(\bar{\rho}_T)} \lessapprox n^{-s}$. Note that when the smoothness of ϕ increases ($s \to \infty$), the rate $\frac{\alpha s}{s+1}$ approaches α, the order of the numerical integrator. In high-dimensional case using Monte Carlo (see Remark 2.7), we conjecture the convergence rate to be $s/(2s+1)$ and leave it as future work.

3.2 Numerical error in the normal equation

In the proof of Theorem 3.5, the error in the estimator comes from numerical integrations in space and in time, and the numerical errors are passed through $A_{n,M,L}$ and $b_{n,M,L}$. We outline the main proof for the numerical error bounds and leave the technical results in Appendix A.

Note that for each t, the integrals in space are expectations with respect to $u(\cdot, t)$ and that

$$|\Omega| \|u\|_{1,\infty} \geqslant |\Omega| \|u\|_\infty \geqslant \int_\Omega u(x,t)dx = 1. \tag{3.5}$$

Proposition 3.8. *The numeric error of $A_{n,M,L}$ and $b_{n,M,L}$ in (2.18) and (2.19) are bounded by*

$$\begin{aligned} \|A - A_{n,M,L}\| &\leqslant nc^A(\Delta x + \Delta t), \\ \|b - b_{n,M,L}\| &\leqslant \sqrt{n}c^b(\Delta x + \Delta t), \end{aligned} \tag{3.6}$$

where the norm for matrix is in the Frobenius sense, and the constants $c^A := 2|\Omega|(1+|\Omega|)(c_{\mathcal{H}}^{1,\infty})^2 \|u\|_{1,\infty}^2$ and $c^b := 3|\Omega|(1 + R_\Omega + \nu)c_{\mathcal{H}}^{2,\infty}(\|u\|_{1,\infty}^2 + \|u\|_{2,\infty})$ are given in Theorem 3.5.

Proof. Using the notation $\overline{\mathcal{D}}(f)$ in (A.1) with $f = (K_{\phi_i} * u)(K_{\phi_j} * u)u$, we have

$$\begin{aligned} &\left|A_{i,j} - A_{n,M,L}^{i,j}\right| \leqslant \overline{\mathcal{D}}(f) + I_1^A, \text{ with} \\ &\qquad I_1^A := \frac{1}{L}\sum_{l=1}^{L}\sum_{m=1}^{M} \left| \left[\left((K_{\phi_i} * u) \cdot (K_{\phi_j} * u) - P_{n,M,L}^i \cdot P_{n,M,L}^j \right) u \right] (x_m, t_l) \right| \Delta x. \end{aligned}$$

To apply (A.1) for $\overline{\mathcal{D}}(f)$, we estimate first $\nabla_{x,t} f$ using (A.6) and (A.4):

$$\begin{aligned} &\left\| \nabla_{x,t} \left[(K_{\phi_i} * u)(K_{\phi_j} * u)u(x,t) \right] \right\|_\infty \\ \leqslant &\left\| \nabla_{x,t} \left[(K_{\phi_i} * u)(K_{\phi_j} * u) \right] \right\|_\infty \|u\|_\infty + \left\| \left[(K_{\phi_i} * u)(K_{\phi_j} * u) \right] \right\|_\infty \|\nabla_{x,t} u\|_\infty \\ \leqslant &(2|\Omega| \|u\|_\infty + 1) \|u\|_{1,\infty} (c_{\mathcal{H}}^\infty)^2. \end{aligned}$$

Note that $(2|\Omega|\,\|u\|_\infty+1)\leqslant 3\Omega|\,\|u\|_\infty$ by (3.5). Thus, we have

$$\overline{\mathcal{D}}(f)\leqslant 3|\Omega|^2\,\|u\|_{1,\infty}^2\,(c_{\mathcal{H}}^{1,\infty})^2(\Delta x+\Delta t).$$

To estimate I_1^A, note that by (A.3) and (A.4), we have

$$\left\|P_{n,M,L}^i\right\|_\infty\leqslant\left\|K_{\phi_i}*u-P_{n,M,L}^i\right\|_\infty+\|K_{\phi_i}*u\|_\infty\leqslant c_{\mathcal{H}}^{1,\infty}|\Omega|\,\|u\|_{1,\infty}\,\Delta x+c_{\mathcal{H}}^{\infty}.$$

Thus, $\left\|P_{n,M,L}^i\right\|_\infty+\|K_{\phi_i}*u\|_\infty\leqslant 2c_{\mathcal{H}}^{1,\infty}$ when $\Delta x\leqslant 2\min_i\|\phi_i'\|_\infty/(c_{\mathcal{H}}^{1,\infty}|\Omega|\,\|u\|_{1,\infty})$. Then,

$$\begin{aligned}I_1^A&\leqslant|\Omega|\,\|u\|_\infty\left\|(K_{\phi_i}*u)\cdot(K_{\phi_j}*u)-P_{n,M,L}^i\cdot P_{n,M,L}^j\right\|_\infty\\&\leqslant|\Omega|\,\|u\|_\infty\max_i\left\|K_{\phi_j}*u-P_{n,M,L}^j\right\|_\infty\max_i\left(\left\|P_{n,M,L}^i\right\|_\infty+\|K_{\phi_i}*u\|_\infty\right)\\&\leqslant|\Omega|\,\|u\|_\infty\,c_{\mathcal{H}}^{1,\infty}\,\|u\|_{1,\infty}\,|\Omega|\Delta x2c_{\mathcal{H}}^{1,\infty}\leqslant 2|\Omega|^2\,\|u\|_{1,\infty}^2\,(c_{\mathcal{H}}^{1,\infty})^2\Delta x.\end{aligned}$$

Combine the above estimates of $\overline{\mathcal{D}}(f)$ and I_1^A, we have

$$\left|A_{i,j}-A_{n,M,L}^{i,j}\right|\leqslant 5|\Omega|^2\,\|u\|_{1,\infty}^2\,(c_{\mathcal{H}}^{1,\infty})^2(\Delta x+\Delta t).$$

and the bound for $\|A_{n,M,L}-A\|$ in (3.6) follows.

Next we analyze $\|b_{n,M,L}-b\|$. Using $\overline{\mathcal{D}}(f)$ in (A.1) with $f=u(\nabla\cdot K_{\phi_i})*u$, we have

$$\begin{aligned}&\left|b_i-b_{n,M,L}^i\right|\leqslant\overline{\mathcal{D}}(u(\nabla\cdot K_{\phi_i}*u))+I_1^b+I_2^b,\ \text{with}\\&I_1^b:=\left|\frac{1}{T}\int_\Omega\int_0^T\partial_t u\Phi_i*u\,dxdt-\sum_{l=1}^{L}\sum_{m=1}^{M}\left[\widehat{\partial_t u}\Phi_i*u\right](x_m,t_l)\Delta x\Delta t\right|,\\&I_2^b:=\frac{1}{L}\sum_{l,m=1}^{L,M}\left|\left[\widehat{\partial_t u}\Phi_i*u+\nu u\,(\nabla\cdot K_{\phi_i}*u)-\widehat{\partial_t u}\,Q_{n,M,L}^i+\nu uR_{n,M,L}^i\right](x_m,t_l)\right|\Delta x.\end{aligned}\tag{3.7}$$

By (A.1) and the gradient estimate (A.7), we have

$$\overline{\mathcal{D}}(u(\nabla\cdot K_{\phi_i}*u))\leqslant|\Omega|\,\|u\|_{1,\infty}\,c_{\mathcal{H}}^{1,\infty}(1+\|u\|_\infty)(\Delta x+\Delta t).$$

To estimate I_2^b, note that $\widehat{\partial_t u}$ in (2.20) satisfies $\left\|\widehat{\partial_t u}\right\|_\infty\leqslant\|u\|_{1,\infty}$. Then, by (A.3), we have

$$\begin{aligned}&\|\widehat{\partial_t u}\,(\Phi_i*u)-\widehat{\partial_t u}\,Q_{n,M,L}^i\|_\infty\leqslant\|u\|_{1,\infty}\left|\Phi_i*u-Q_{n,M,L}^i\right|\leqslant|\Omega|(1+R_\Omega)c_{\mathcal{H}}^{\infty}\,\|u\|_{1,\infty}^2\,\Delta x,\\&\left\|u\,(\nabla\cdot K_{\phi_i}*u)-uR_{n,M,L}^i\right\|_\infty\leqslant\|u\|_\infty\left|\nabla\cdot K_{\phi_i}*u-R_{n,M,L}^i\right|\leqslant|\Omega|c_{\mathcal{H}}^{2,\infty}\,\|u\|_{1,\infty}^2\,\Delta x.\end{aligned}$$

Hence

$$I_2^b\leqslant|\Omega|(1+R_\Omega+\nu)c_{\mathcal{H}}^{2,\infty}\,\|u\|_{1,\infty}^2\,\Delta x.$$

Together with the estimates of I_1^b in Lemma A.4, we have

$$\left|b_i-b_{n,M,L}^i\right|\leqslant 3c_{\mathcal{H}}^{2,\infty}|\Omega|(1+R_\Omega+\nu)(\|u\|_{1,\infty}^2+\|u\|_{2,\infty})(\Delta x+\Delta t),$$

The estimate for $\|b-b_{n,M,L}\|$ in (3.6) follows. □

4 Numerical examples

We demonstrate the effectiveness of our algorithm using synthetic data for examples with three typical types of interaction kernels: the granular media model with a smooth kernel (Section 4.2), the opinion dynamics with a piecewise linear kernel (Section 4.3), and the aggregation-diffusion with a singular repulsive-attractive kernel (Section 4.4). In each of these examples, our algorithm leads to accurate estimators that can reproduce solutions and free energy almost perfectly. Our estimator achieves the theoretical rate of convergence for the granular media model, and obtains rates lower than the theoretical rates for the opinion dynamics and the singular repulsive-attractive potential. [1]

4.1 Numerical settings

We first specify the numerical settings: the numerical scheme for data generation, the choice of parameters in the learning algorithm, and the assessment of the estimators.

Settings for data generation. We solve the mean-field equation by the semi-implicit Structure Preserving scheme of Chang-Cooper (SPCC) scheme for nonlinear Fokker-Planck equations introduced in Pareschi and Zenella [28]. The SPCC is second order accurate, preserves the steady state and the density properties of the solution u such as non-negativity. In particular, for the explicit SPCC scheme, we need $dt \leqslant \frac{dx^2}{2(Cdx+\nu)}$, where $C = \sup_{\Omega\times[0,T]} |\nabla\Phi * u|$; for semi-implicit scheme, we need $dt \leqslant \frac{dx}{2C}$.

In all the three examples, the data are discrete observations from the "true" solution on $\Omega \times [0, T]$ with a fine space-time mesh with $dt = 0.001$ and $dx = \frac{b-a}{3000}$, where $T = 1$ and $\Omega = [a, b]$ with $a = -10$ and $b = 10$. This setting preserves the steady state and the non-negativity of the solution u. The data are observed at every 10 space mesh, i.e. $\Delta x = 10dx$ (or equivalently, $M = 300$). To compute the rate of convergence, we down-sample the data further to have a sequence of $\Delta x = kdx$ with $k \in \{10, 12, 15, 20, 24, 30, 50, 60, 75, 100\}$, correspondingly, we have $M \in \{300, 250, 200, 150, 120, 100, 60, 50, 40, 30\}$).

We summarize these settings in Table 2.

Table 2 Numerical settings in data generation and inference for all examples.

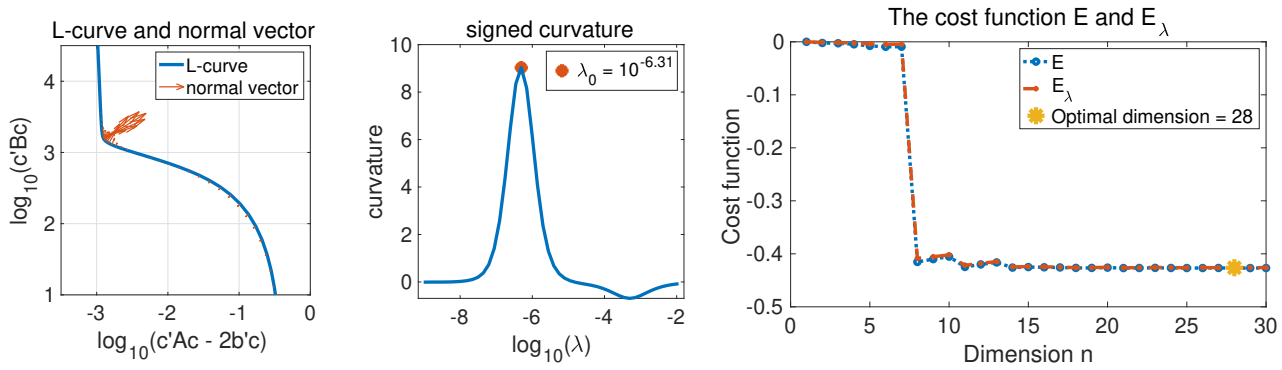
Notation	Description
$[0, T] = [0, 1]$ and $\Omega = [-10, 10]$	time interval and space domain
$dt = 0.001$ and $dx = 20/3000$	time step and space mesh size of true solution
$\Delta t = dt$ and $\Delta x = kdx$	time step and space mesh size of data with $k \in \{10, 12, 15, 20, 24, 30, 50, 60, 75, 100\}$
$r_0 = 0, r_m = 10$	knots for B-spline
$W(u, \widehat{u})$	the Wasserstein distance, defined in (4.1)
$E[u, \phi](t)$	free energy flow, defined in (4.2)

Settings for inference algorithm. We estimate the interaction kennel ϕ by Algorithm 1. We test both spline basis and RKHS basis. For the spline basis in (2.22), we use $r_0 = 0$ and $r_m = 10$. The range of knot number is $m \in [3, 40]$ for all three examples. We set the degree of the spline to be in $\{0, 1, 2, 3\}$ according to the smoothness of each example. We obtain the RKHS basis functions by solving the eigenvalue problem (2.24) as described in Section 2.2, and we set the dimension range to be $[2, 50]$ for all examples.

We select the dimension of the hypothesis space under an adaptive regularization. That is, we find first the optimal regularization constant for each given dimension of hypothesis space as in Section 2.3, then we select the dimension as in Section 2.4. Figure 1 demonstrates this process for the opinion dynamics (the plots are similar for other examples).

Results assessment in a typical estimation. In a typical estimation from data with $\Delta x = 10dx$ (or equivalently $M = 300$), we assess the estimators in three plots: comparison of the estimators with the truth, the Wasserstein

[1] The MATLAB code for this project is available at https://github.com/QuanjunLang/Learning-Interaction-Kernel-From-Mean-Field-Equation.



(a) Use the L-curve to find optimal regularization constant

(b) Find the dimension

Figure 1: Regularization and selection of dimension for the opinion dynamics (see Section 4.3) using degree 1 spline basis with knot number 30. (a) the use of L-curve to find the optimal regularization constant and (b) shows the selection of dimension for the hypothesis space.

distance between the estimated and true solutions, and the estimated and true free energy flows.

- **Estimated and true kernels** We compare the true and estimated kernels by plotting them side-by-side, together with the density $\bar{\rho}_T$. The estimated kernels are from either B-spline basis functions or RKHS basis functions, with dimension provided in the context. We also give the relative RKHS error and relative $L^2(\bar{\rho}_T)$ error.

- **Wasserstein Distance** The solutions from the true and estimated kernels can not be distinguished by eyes. To compare them, we compute the 2-Wasserstein distance between them. We consider two sets of solutions, starting from either the original or a new initial condition $\tilde{u}_0$. We set $\tilde{u}_0$ to be the average of the density functions of $\mathcal{N}(2,1)$ and $\mathcal{N}(-2,1)$, whose major mass is in the support of $\bar{\rho}_T$. Recall that the Wasserstein distance $W_2(f,g)$ of two probability densities f and g over Ω with second order moments is given by

$$W_2(f,g) := \left(\inf_{\gamma \in \Gamma(f,g)} \int_{\Omega \times \Omega} |x-y|^2 d\gamma(x,y) \right)^{1/2}, \tag{4.1}$$

  where $\Gamma(f,g)$ denotes the set of all measures on $\Omega \times \Omega$ with f and g as marginals (see e.g.,[1]). We use the numerical method for Wasserstein distance as in [18]. This method is based on an observation in [5]. More precisely, suppose F is the probability distribution induced by the density f and define its pseudo inverse by setting, for $\alpha \in (0,1)$, $F^{-1}(\alpha) = \inf\{x : F(x) > \alpha\}$. Similarly we have G and G^{-1}. Then the L^2 distance of the pseudo inverse functions $d_2(f,g) = \left(\int_0^1 [F^{-1}(\alpha) - G^{-1}(\alpha)]^2 d\alpha \right)^{1/2}$ is equal to the 2-Wasserstein distance $W_2(f,g)$.

- **Free Energy** We also compare the true and estimated free energy flows. The free energy, whose Wasserstein gradient gives the mean-field equation [4], is defined by

$$E[u,\phi](t) = \nu \int_{\mathbb{R}^d} u \log(u) dx + \int_{\mathbb{R}^d} u(u * \Phi) dx, \text{ with } \Phi(r) = \int_0^r \phi(s) ds. \tag{4.2}$$

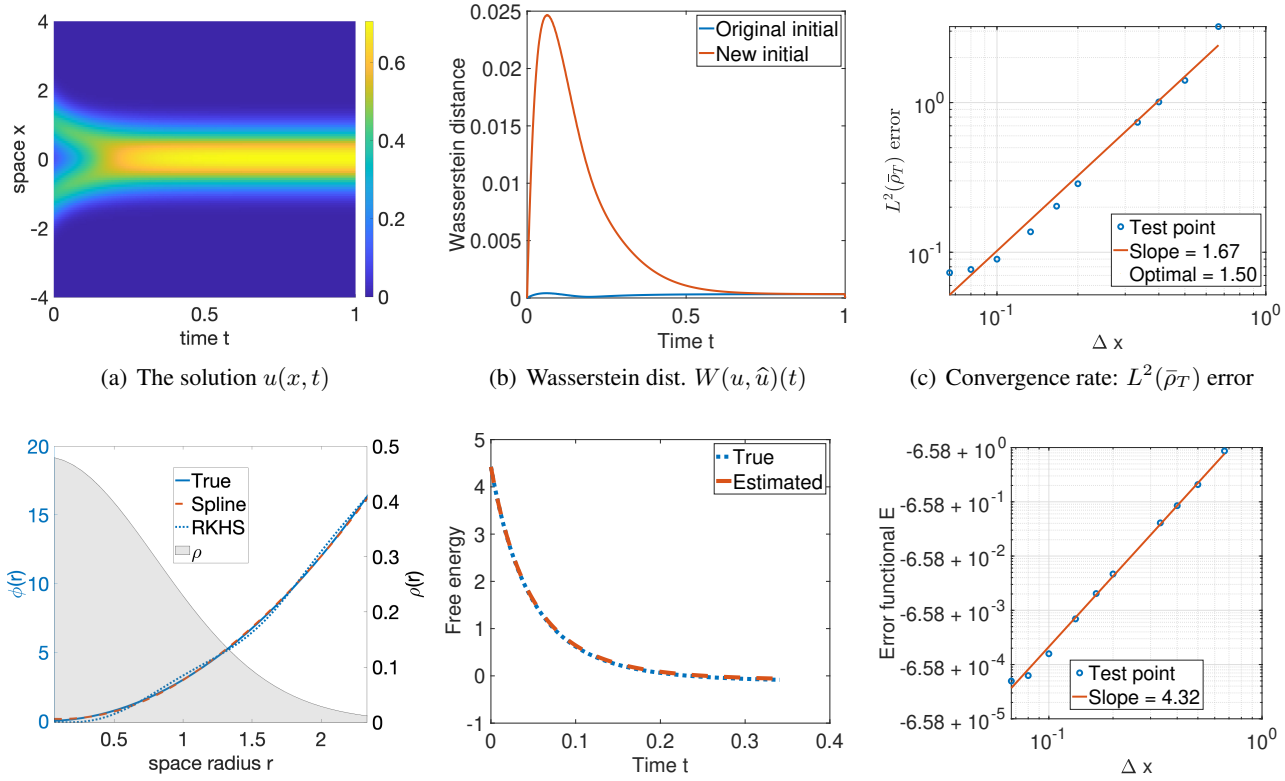
  The true and estimated free energy flows are $E[u,\phi](t)$ and $E[\widehat{u},\widehat{\phi}](t)$, respectively.

Rate of convergence. We test the rate of convergence of the estimator in $L^2(\bar{\rho}_T)$ error and empirical error functional $\mathcal{E}_{M,L}$ in (2.17), as Δx changes. We consider the downsampled data with $\Delta x = kdx$, where $k \in \{10, 12, 15, 20, 24, 30, 50, 60, 75, 100\}$. We use the spline basis, because the data-adaptive RKHS basis is not suitable for such a test.

For each estimator, we compute the $\|\phi - \hat{\phi}\|_{L^2(\bar{\rho}_T)}$ by Riemann sum approximation. The measure $\bar{\rho}_T$, defined in (2.2), is approximated from data by ρ_L^M in (2.16), using the data with the finest mesh $\Delta x = 10dx$ (equivalently, M = 300).

We also compute the decay rate of the loss functional so as to show it decays faster than the $L^2(\bar{\rho}_T)$ error. From data, we obtain a sequence of values for the error functional $\mathcal{E}_{M,L}(\hat{\phi}_{n,M,L})$ as in (2.17). We compute its convergence rate β by optimization: suppose the error functional has the form $E_k = a\Delta x_k^\beta - \gamma$ with the multiplicative constant a and $\gamma = \langle\!\langle \phi, \phi \rangle\!\rangle$ unknown, we compute β by

$$(\beta, \gamma, a) = \arg\min_{\beta,c,a} \sum_k |\log(E_k + \gamma) - \beta \log \Delta x_k - \log a|^2.$$



(a) The solution $u(x,t)$ (b) Wasserstein dist. $W(u,\hat{u})(t)$ (c) Convergence rate: $L^2(\bar{\rho}_T)$ error

(d) Estimators of ϕ (e) Free energy (f) Convergence rate: $\mathcal{E}_{M,L}$ in Δx

Figure 2: Learning results of cubic potential. Subfigure (a) show the solution formulating a steady state. (d) shows the estimated kernels, superimposed with the empirical density $\bar{\rho}_T$. The relative errors are shown in Table 3. (b) shows that the Wasserstein distance between the solutions are small. In particular, they all tend to the same steady state at large time. In (e) the free energy is well-learned. Subfigure (c) and (f) show the rates of convergence of the $L^2(\bar{\rho}_T)$ error and the error functional. The two rates are close to the rates in Theorem 3.7.

4.2 Cubic potential

The cubic potential $\Phi(x) = |x|^3$ (equivalently, $\phi(r) = 3r^2$) is of special interest for modeling of granular media [26, 6]. Since Φ is only non-uniformly convex on a single point, the equation (1.1) possess a unique steady state [6] and thus the SDE (2.1) is ergodic. We set $\nu = 1$ and take $u_0(x)$ to be the average of the densities of $\mathcal{N}(1, 0.25)$ and $\mathcal{N}(-1, 0.25)$.

We use B-spline basis with degree 2, matching the degree of the true kernel.

Figure 2 presents the estimation results. Sub-figure (a) shows the solution $u(x,t)$, which is dominated by the diffusion. Subfigure (d) shows that the estimated kernels, either by B-spline or RKHS basis, are close to the true kernel, with relative errors shown in Table 3.

Table 3 Relative errors of estimators in Figure 2(d).

Basis type	dimension	Relative error in $L^2(\bar{\rho}_T)$ ($\|\phi\| = 3.85$)	Relative error in RKHS ($\sqrt{\langle\!\langle \phi, \phi \rangle\!\rangle} = 2.57$)
B-spline	12	2.18%	0.79%
RKHS	14	5.83%	0.79%

Subfigure (b) plots the Wasserstein distances between true and reproduced solutions, showing that the solution with the original initial condition is accurately reproduced. For the new initial condition $\widetilde{u}_0$, the Wasserstein distance is relatively large at first and then decays to the same level as the original initial condition case. This is because: (1) $\widetilde{u}_0$, the mixture of $\mathcal{N}(2,1)$ and $\mathcal{N}(-2,1)$, has a large probability mass outside of the "well-learned region, the large-probability region of $\bar{\rho}_T$; (2) the system converge to a unique state state for different initial conditions. Subfigure (e) shows that the free energy flow is almost perfectly reproduced. Subfigure (c) shows that we nearly achieve the theoretical rate $\frac{\alpha s}{s+1} = 1.5$ in Theorem 3.7, where $s = 3$ because the B-spline has a degree 2 and $\alpha = 2$ is the order of trapezoidal integration. Subfigure (f) show that the error functional converges at the rate $(\Delta x)^{4.32}$, much faster than the $L^2(\bar{\rho}_T)$ error's rate.

4.3 Opinion dynamics

Opinion dynamics (see [27] and the reference therein) describes the evolution of opinions of agents in social networks. We consider the case when the system formulates clusters of opinions: the interaction function $f(|x|) = \phi(|x|)/|x|$ is piecewise constant,

$$f(r) = \begin{cases} -1, & 0 \leqslant r \leqslant 3, \\ 2, & 3 < r \leqslant 4, \\ 0, & 4 < r. \end{cases}$$

and hence $\phi(|x|) = f(|x|)|x|$ is piecewise linear; the initial value $u_0(x)$ is density of the Gaussian mixture $\frac{1}{3}[\mathcal{N}(-2,1) + \mathcal{N}(-4, 0.5^2) + \mathcal{N}(2,1)]$; the viscosity constant is $\nu = 0.1$.

We set the degree of spline basis to be 1, since ϕ is piecewise linear.

Table 4 Relative errors of estimators in Figure 3(d).

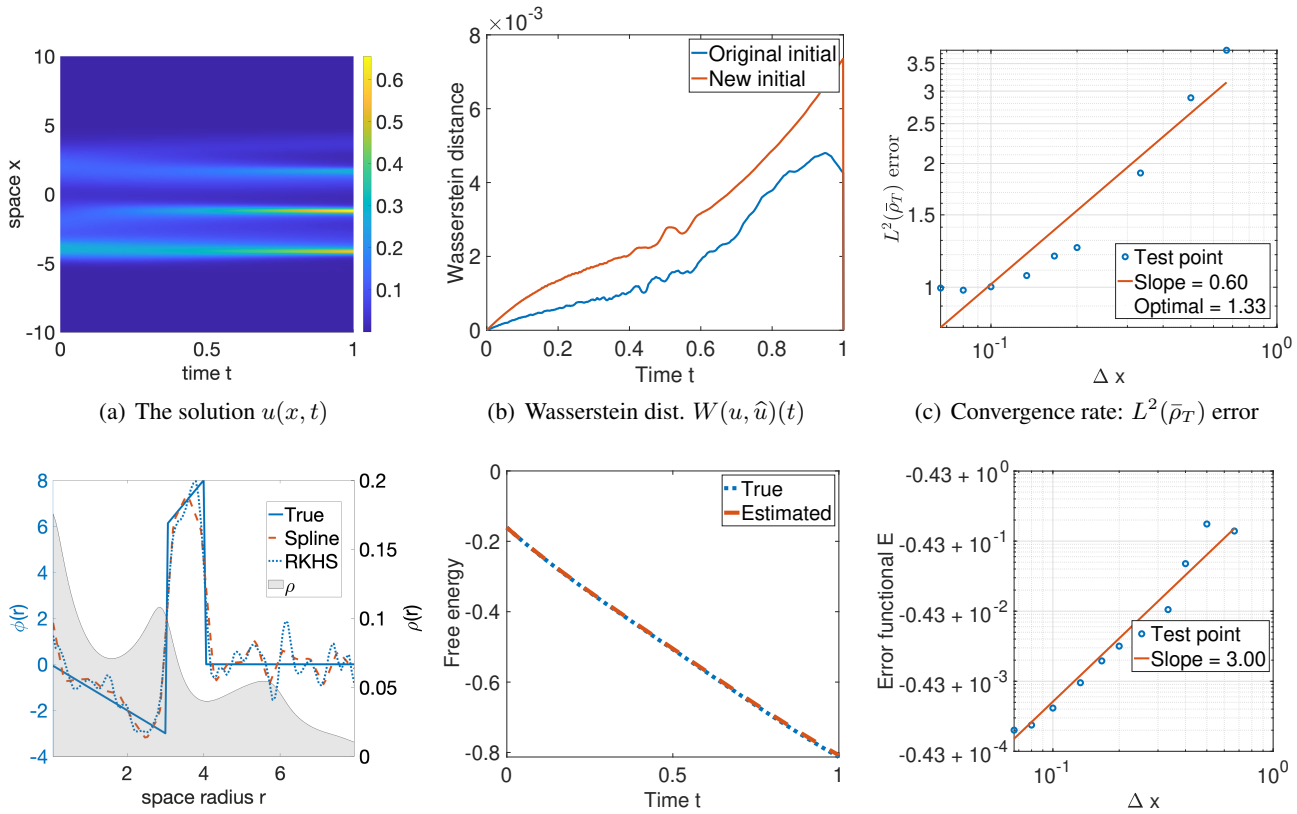
Basis type	dimension	Relative error in $L^2(\bar{\rho}_T)$ ($\|\phi\| = 2.71$)	Relative error in RKHS ($\sqrt{\langle\!\langle \phi, \phi \rangle\!\rangle} = 0.65$)
B-spline	29	36.67%	7.87%
RKHS	40	44.52%	8.17%

Figure 3 presents the estimation results. Subfigure (a) is the solution $u(x,t)$, which shows three clusters forming at time $T = 1$. Subfigure (d) shows the estimated and true kernels, with relative errors shown in Table 4. Subfigure (b) is the Wasserstein distance between true and reproduced solutions, showing that the solutions are accurately reproduced. The Wasserstein distance increases because of the formulating clusters, which lead to singular measures. Subfigure (e) shows that the free energy flow is almost perfectly reproduced. Subfigure (c) and (f) shows the rates of convergence of the estimator in $L^2(\bar{\rho}_T)$ and the error functional. Due to the lack of regularity of ϕ, the rate 0.60 is smaller than the theoretical rate $\frac{\alpha s}{s+1} = \frac{4}{3}$ in Theorem 3.7, where $s = 2$ because the degree of B-spline is 1 and $\alpha = 2$ is the order of trapezoidal integration. The error functional converges at the rate 3.00, which is much larger than the $L^2(\bar{\rho}_T)$ error's rate.

4.4 The repulsion-attraction potential

To model collision free particles, the repulsion attraction (RA) potential with singularity at 0 is widely used. We consider the power law RA potential [4]:

$$\Phi(x) = \frac{|x|^p}{p} - \frac{|x|^q}{q}, \quad 2 \geqslant p > q > -d.$$



(a) The solution $u(x,t)$ (b) Wasserstein dist. $W(u,\widehat{u})(t)$ (c) Convergence rate: $L^2(\bar{\rho}_T)$ error

(d) Estimators of ϕ (e) Free energy (f) Convergence rate: $\mathcal{E}_{M,L}$ in Δx

Figure 3: Learning results of opinion dynamics. The solution in (a) formulates clusters. Subfigure (d) shows the estimated kernels, with relative errors shown in Table 4. The Wasserstein distance in (b) shows that the solution is accurately reproduced by the estimated kernel with spline, for both the original and new initial conditions. Subfigure (e) show that the free energy is almost perfectly reproduced. Subfigure (c) and (f) show the rates of convergence of the estimator $L^2(\bar{\rho}_T)$ and the error functional. Due to the lack of regularity of the true kernel, the rate in Theorem 3.7 does not apply.

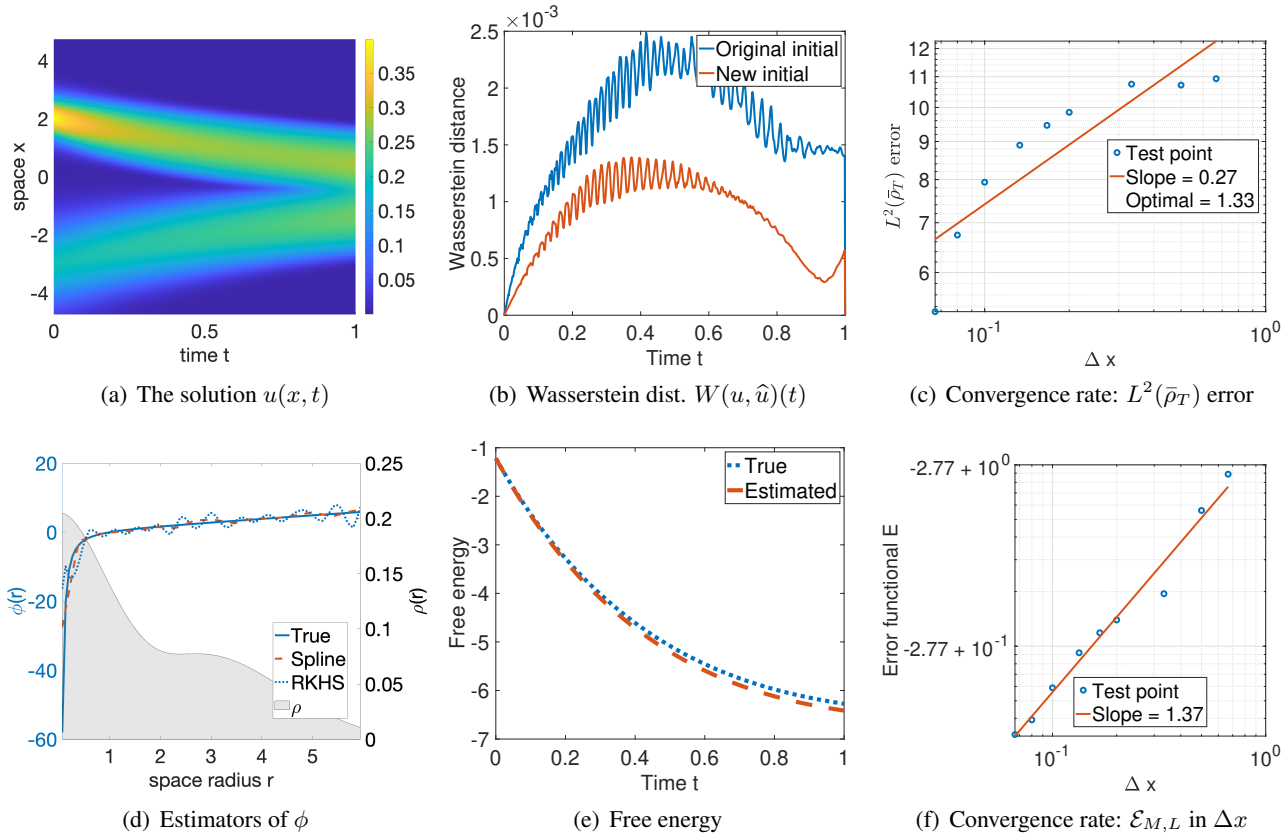
When $0 < q < 1$, 0 is a singular point of ϕ. When $q < 0$, 0 is also a singular point of Φ. We take $p = 2$ and $q = -0.5$. To show the repulsion clearly, we use a small viscosity constant $v = 0.01$. We take the initial value u_0 being the mean of the densities of $\mathcal{N}(2, 0.25)$ and $\mathcal{N}(-3, 1)$. We use B-spline basis with degree 1.

Figure 4 exhibits the estimation results. Sub-figure (a) shows the solution $u(x,t)$, which demonstrate the attraction and repulsion under the influence of diffusion. Subfigure (d) shows that the estimated kernels, either by B-spline or RKHS basis, are close to the true kernel, with relative errors shown in Table 5. The large relative error

Table 5 Relative errors of estimators in Figure 4(d).

Basis type	dimension	Relative error in $L^2(\bar{\rho}_T)$ ($\|\phi\| = 10.72$)	Relative error in RKHS ($\sqrt{\langle\langle \phi, \phi \rangle\rangle} = 1.60$)
B-spline	27	48.83%	3.86%
RKHS	35	75.54%	4.07%

in $L^2(\bar{\rho}_T)$ is due to the singularity at the origin and that the measure $\bar{\rho}_T$ does not reflect the repulsion. Nevertheless, Subfigure (b) shows that the estimated kernel can reproduce accurate solutions, suggesting that the $L^2(\bar{\rho}_T)$ norm may not be suitable for the assessment of the estimator of singular kernels. The slightly oscillating Wasserstein distances indicate that the error in the estimator does not propagate. Subfigure (e) shows that the free energy flow is almost perfectly reproduced. Subfigure (c) and (f) shows relatively low rates of convergence of the estimator in $L^2(\bar{\rho}_T)$ and the error functional. The rates 0.27 and 1.37 are low: due to the singularity of the kernel at the origin, the theoretical rate in Theorem 3.7 ($\frac{\alpha s}{s+1} = \frac{4}{3}$ with $\alpha = 2, s = 2$) does not apply.



(a) The solution $u(x,t)$ (b) Wasserstein dist. $W(u,\widehat{u})(t)$ (c) Convergence rate: $L^2(\bar{\rho}_T)$ error

(d) Estimators of ϕ (e) Free energy (f) Convergence rate: $\mathcal{E}_{M,L}$ in Δx

Figure 4: Learning results of a repulsion-attraction potential. Subfigure (a) demonstrates a repulsion-attraction effect. The two clusters tend to get closer, but not merging because of the repulsion. Estimators in (d) show that our method tries to learn the singularity at the origin. The relative errors are shown in Table 5. (b) shows that the Wasserstein distance between the solutions are small. The free energy estimate in (e) is close. Subfigure (c) and (f) show the rate of convergence for the estimator in $L^2(\bar{\rho}_T)$ and for the error functional. These rates are relatively low due to the singularity of the true kernel (and the rate in Theorem 3.7 does not apply).

5 Conclusion and future work

We have introduced a nonparametric learning algorithm to estimate the interaction kernel from discrete data with a performance guarantee. From the likelihood of the diffusion process whose Fokker-Planck equation is the mean-field equation, we derive a probabilistic error functional. Then, the algorithm learns the kernel on a data-adaptive hypothesis space by least squares with regularization. The algorithm does not require spatial derivatives of the solution, so it is suitable for discrete data. We prove that, as the space-time mesh refines, the estimator converges in a weighted L2 space under an identifiability condition, at a rate $\frac{\alpha s}{s+1}$, where α is the order of the numerical integrator and s is the rate of decay of the approximation error by the hypothesis space.

We demonstrate our algorithm's performance on three typical examples:

- the granular media model with a quadratic kernel;
- the opinion dynamics with a piecewise linear kernel;
- the repulsion-attraction with a singular kernel.

In all the examples, the estimator is accurate, and it can reproduce solutions with a small Wasserstein distance to the truth and with almost perfect free energy. For the granular media with a quadratic kernel, our estimator achieves the theoretical rate of convergence. For the opinion dynamics and the singular repulsion-attraction model, our estimator converges at rates lower than the theoretical rates.

There are many directions to extend the present work. We mention a few here:

- Non-radial interaction kernels. Many applications involve non-radial kernels, such as the Biot-Savart kernel [15] and the local time kernel for viscous Burgers equation [30]. The major issue is the curse of dimensionality

in representing high-dimensional functions. We expect to represent the function using the data adaptive reproducing kernel for the regression.

- High-dimensional space. It becomes impractical to have data on mesh-grids when the dimension d of the space is large, because the size of mesh-grids increases exponentially in d. It is natural to consider data consisting of samples of particles and approximate the error functional by Monte Carlo. Our algorithm applies and the rate convergence would be $s/(2s+1)$.
- Partial observations from large systems. When only partial agents of a large system are observed, it is an ill-posed problem to estimate the position of other agents [31]. By the propagations of chaos, we may view these particles as independent trajectories and estimate the interaction kernel from the SDE of the mean-field equation.

A Appendix: errors bounds for the numerical integrators

We provide technical bounds on the error of the numerical integrator based on Riemann sum (the Euler scheme). Let's start with a reminder about the error of the Euler scheme.

Lemma A.1. *Let $x_m = m\Delta x$ and $t_l = l\Delta t$ be the mesh given in (3.1). Suppose that Assumption 3.2 holds true. Suppose $f \in W^{1,\infty}(\Omega \times [0,T])$. Then the Euler scheme is of order $\sqrt{\Delta x^2 + \Delta t^2}$, i.e.*

$$\overline{\mathcal{D}}(f) := \Big|\frac{1}{T}\int_\Omega \int_0^T f(x,t)\,dxdt - \sum_{l,m=1}^{L,M} f(x_m,t_l)\Delta x\Delta t\Big| \leqslant |\Omega|\,\|\nabla_{x,t} f\|_\infty\,(\Delta x + \Delta t), \tag{A.1}$$

$$\mathcal{D}_t(f) := \Big|\int_\Omega f u dx - \sum_{m=1}^{M} f(x_m,t)u(x_m,t)\Delta x\Big| \leqslant (\|\nabla f\|_\infty + |\Omega|\,\|f\|_\infty\,\|\nabla u\|_\infty)\,\Delta x. \tag{A.2}$$

Proof. Note that for $x \in [x_m, x_{m+1}], t \in [t_l, t_{l+1}]$, there exists (ξ_m, ζ_l) such that

$$|f(x_m,t_l) - f(x,t)| = |\nabla_{x,t} f(\xi_m,\zeta_l)\cdot(x - x_m, t - t_l)| \leqslant \|\nabla_{x,t} f\|_\infty \sqrt{\Delta x^2 + \Delta t^2}.$$

Note that $\sqrt{\Delta x^2 + \Delta t^2} \leqslant \Delta x + \Delta t$. Then, (A.1) follows from

$$\overline{\mathcal{D}}(f) \leqslant \frac{1}{T}\sum_{l=1}^{L}\sum_{m=1}^{M}\int_{x_m}^{x_{m+1}}\int_{t_{l-1}}^{t_l} \big|f(x_m,t_l) - f(x,t)\big| dxdt \leqslant |\Omega|\,\|\nabla_{x,t} f\|_\infty \sqrt{\Delta x^2 + \Delta t^2}.$$

Similarly, (A.2) follows from

$$\begin{aligned}\mathcal{D}_t(f) &\leqslant \sum_{m=1}^{M}\int_{x_m}^{x_{m+1}} \Big[\big|f(x,t) - f(x_m)\big|u(x,t) + \big|f(x_m,t)\big|\big|u(x,t) - u(x_m,t)\big|\Big]\,dx \\ &\leqslant \|\nabla f\|_\infty\,\Delta x + \|f\|_\infty\,\|\nabla u\|_\infty\,|\Omega|\Delta x,\end{aligned}$$

where the last inequality follows from that $\sum_{m=1}^{M}\int_{x_m}^{x_{m+1}} u(x,t)dx = \int_\Omega u(x,t)dx = 1$. □

Lemma A.2. *Suppose that Assumption 3.1 holds true. The errors of $P^i_{n,M,L}$, $Q^i_{n,M,L}$ and $R^i_{n,M,L}$ in (2.20), which approximate $K_{\phi_i} * u$, $\Phi_i * u$ and $\nabla\cdot K_{\phi_i} * u$, are bounded by*

$$\begin{aligned}\big\|P^i_{n,M,L} - K_{\phi_i} * u\big\|_\infty &\leqslant |\Omega|\Delta x\,\|\phi_i\|_{1,\infty}\,\|u\|_{1,\infty} \leqslant c^{1,\infty}_{\mathcal{H}}|\Omega|\,\|u\|_{1,\infty}\,\Delta x, \\ \big\|R^i_{n,M,L} - (\nabla\cdot K_{\phi_i}) * u\big\|_\infty &\leqslant |\Omega|\Delta x\,\|\phi_i'\|_{1,\infty}\,\|u\|_{1,\infty} \leqslant c^{2,\infty}_{\mathcal{H}}|\Omega|\,\|u\|_{1,\infty}\,\Delta x, \\ \big\|Q^i_{n,M,L} - \Phi_i * u\big\|_\infty &\leqslant |\Omega|(1 + R_\Omega)c^{\infty}_{\mathcal{H}}\,\|u\|_{1,\infty}\,\Delta x.\end{aligned} \tag{A.3}$$

Proof. Using the notation $\mathcal{D}_t(f)$ in (A.2) with $f(\cdot) = K_{\phi_i}(x-\cdot)$, we have,

$$\begin{aligned}\left\|P^i_{n,M,L} - K_{\phi_i} * u\right\|_\infty &= \sup_{(x,t)\in\Omega\times[0,T]} |\mathcal{D}_t(K_{\phi_i}(x-\cdot))| \\ &\leqslant \sup_{(x,t)\in\Omega\times[0,T]} \left(\|\nabla K_{\phi_i}(x-\cdot)\|_\infty + |\Omega|\,\|K_{\phi_i}(x-\cdot)\|_\infty \|\nabla u\|_\infty\right)\Delta x.\end{aligned}$$

Recall that we denote $K_{\phi_i}(x) = \phi_i(|x|)\frac{x}{|x|}$. For $x \neq 0$, we have $\frac{d}{dx}|x| = \frac{x}{|x|}$ and

$$\frac{d}{dx}K_{\phi_i}(x) = \frac{d}{dx}\left(\phi_i(|x|)\frac{x}{|x|}\right) = \phi_i'(|x|) + \phi_i(|x|)\frac{|x| - x\frac{x}{|x|}}{|x|^2} = \phi_i'(|x|).$$

Thus, $\|\nabla K_{\phi_i}(x-\cdot)\|_\infty \leqslant \|\phi_i'\|_\infty$ and $\|K_{\phi_i}(x-\cdot)\|_\infty \leqslant \|\phi_i\|_\infty$. Together with (3.5), we have

$$\left\|P^i_{n,M,L} - K_{\phi_i} * u\right\|_\infty \leqslant (\|\phi_i'\|_\infty + |\Omega|\,\|\phi_i\|_\infty \|\nabla u\|_\infty)\Delta x \leqslant |\Omega|\,\|\phi_i\|_{1,\infty}\,\|u\|_{1,\infty}\,\Delta x.$$

Note that $\nabla\cdot K_{\phi_i} = \phi_i'(|x|)$. Then, the same argument leads to the estimate for $R^i_{n,M,L}$.

Similarly, from the definition of $Q^i_{n,M,L}$ and the notation in (A.2), we have

$$\begin{aligned}\left\|Q^i_{n,M,L} - \Phi_i * u\right\|_\infty &\leqslant \sup_{(x,t)\in\Omega\times[0,T]} |\mathcal{D}_t(\Phi_i(|x-\cdot|))| \\ &\leqslant \sup_{x\in\Omega}\left[\|\nabla\Phi_i(|x-\cdot|)\|_\infty + |\Omega|\,\|\Phi_i(|x-\cdot|)\|_\infty\|\nabla u\|_\infty\right]\Delta x \\ &\leqslant (1 + |\Omega|R_\Omega\,\|\nabla u\|_\infty)\,\|\phi_i\|_\infty\Delta x \leqslant |\Omega|(1+R_\Omega)\,\|\phi_i\|_\infty\,\|u\|_{1,\infty}\,\Delta x,\end{aligned}$$

where the second last inequality follows from $\Phi_i(r) = \int_0^r \phi_i(s)ds$ and $\|\Phi_i\|_\infty \leqslant \|\phi_i\|_\infty R_\Omega$. □

Lemma A.3. *Suppose that Assumption 3.1 holds true. Then, for each i, j,*

$$\|K_{\phi_i} * u\|_\infty \leqslant \|\phi_i\|_\infty \leqslant c^\infty_{\mathcal{H}} \tag{A.4}$$

$$\|\nabla_{x,t}\,(K_{\phi_i} * u)\|_\infty \leqslant |\Omega|\,\|\nabla_{x,t}u\|_\infty\,\|\phi_i\|_\infty \leqslant |\Omega|\,\|u\|_{1,\infty}\,c^\infty_{\mathcal{H}}, \tag{A.5}$$

$$\left\|\nabla_{x,t}\left[(K_{\phi_i} * u)(K_{\phi_j} * u)\right]\right\|_\infty \leqslant 2|\Omega|\,\|u\|_{1,\infty}\,(c^\infty_{\mathcal{H}})^2. \tag{A.6}$$

$$\|\nabla_{x,t}\,(u\nabla\cdot K_{\phi_i} * u)\|_\infty \leqslant \|u\|_{1,\infty}\,c^{1,\infty}_{\mathcal{H}}(1 + \|u\|_\infty). \tag{A.7}$$

Proof. Note that $\|u(\cdot,t)\|_{L^1(\Omega)} = 1$ for each t. Then equation (A.4) follows from that

$$\|K_{\phi_i} * u\|_\infty = \sup_{t\in[0,T]} \|K_{\phi_i} * u(\cdot,t)\|_\infty \leqslant \|\phi_i\|_\infty\,\|u(\cdot,t)\|_{L^1(\Omega)} = \|\phi_i\|_\infty,$$

Equation (A.5) follows from that $\nabla_{x,t}\,(K_{\phi_i} * u) = K_{\phi_i} * \nabla_{x,t}u$ and $\left|K_{\phi_i}\right| \leqslant \|\phi_i\|_\infty$. Since

$$\left\|\nabla_{x,t}\left[(K_{\phi_i} * u)(K_{\phi_j} * u)\right]\right\|_\infty \leqslant 2\left(\max_{i=1,\dots,n}\|\nabla_{x,t}\,(K_{\phi_i} * u)\|_\infty\right)\left(\max_{i=1,\dots,n}\|K_{\phi_i} * u\|_\infty\right),$$

we obtain (A.6) from (A.4) –(A.5).

From $\|(\nabla\cdot K_{\phi_i}) * u\|_\infty \leqslant \|\phi_i'\|_\infty$ and $\|\nabla_{x,t}\,(\nabla\cdot K_{\phi_i} * u)\|_\infty \leqslant \|\phi_i'\|_\infty\,\|\nabla_{x,t}u\|_\infty$, we have

$$\begin{aligned}\|\nabla_{x,t}\,(u\nabla\cdot K_{\phi_i} * u)\|_\infty &\leqslant \|\nabla_{x,t}u\|_\infty\,\|(\nabla\cdot K_{\phi_i}) * u\|_\infty + \|u\|_\infty\,\|\nabla_{x,t}\,(\nabla\cdot K_{\phi_i} * u)\|_\infty \\ &\leqslant \|\phi_i'\|_\infty\,\|\nabla_{x,t}u\|_\infty\,(1 + \|u\|_\infty) \leqslant c^{1,\infty}_{\mathcal{H}}\,\|u\|_{1,\infty}\,(1 + \|u\|_\infty).\end{aligned}$$

This gives (A.7). □

Lemma A.4. *For I^b_1 defined in (3.7), we have*

$$I^b_1 \leqslant 2c^\infty_{\mathcal{H}}R_\Omega|\Omega|(\|u\|^2_{1,\infty} + \|u\|_{2,\infty})(\Delta x + \Delta t).$$

Proof. Denote $g(x,t) = \Phi_i * u(x,t)$. Note that $\|\Phi_i\|_\infty \leqslant \|\phi_i\|_\infty R_\Omega \leqslant c_{\mathcal{H}}^\infty R_\Omega$. Then,

$$\|g\|_\infty \leqslant \|\Phi_i\|_\infty \leqslant c_{\mathcal{H}}^\infty R_\Omega; \quad \|\nabla_{x,t} g\|_\infty \leqslant \|\Phi_i\|_\infty \|u\|_{1,\infty} \leqslant c_{\mathcal{H}}^\infty R_\Omega \|u\|_{1,\infty}.$$

Note also that $\widehat{\partial_t u}(x_m, t_l) = \frac{u(x_m,t_l)-u(x_m,t_{l-1})}{\Delta t} = \partial_t u(x_m, t^*)$ for some $t^* \in [t_l, t_{l+1}]$, we have $\left|\partial_t u(x,t) - \widehat{\partial_t u}(x_m,t_l)\right| \leqslant (\|\partial_{xt} u\|_\infty + \|\partial_{tt} u\|_\infty)(\Delta x + \Delta t)$. Thus,

$$\begin{aligned}
&\sup_{x\in(x_m,x_{m+1}),t\in(t_{l-1},t_l)} \left|g(x,t)\partial_t u(x,t) - \widehat{\partial_t u}(x_m,t_l) g(x_m,t_l)\right| \\
\leqslant &\sup_{x\in(x_m,x_{m+1}),t\in(t_{l-1},t_l)} \left[\left|g(x,t) - g(x_m,t_l)\right| \|\partial_t u\|_\infty + \|g\|_\infty \left|\partial_t u(x,t) - \widehat{\partial_t u}(x_m,t_l)\right|\right] \\
\leqslant &(\|\nabla_{x,t} g\|_\infty \|\partial_t u\|_\infty + \|g\|_\infty \|u\|_{2,\infty})(\Delta x + \Delta t).
\end{aligned}$$

The, note that $\|\nabla_{x,t} g\|_\infty \|\partial_t u\|_\infty + \|g\|_\infty \|u\|_{2,\infty} \leqslant 2c_{\mathcal{H}}^\infty R_\Omega(\|u\|_{1,\infty}^2 + \|u\|_{2,\infty})$, we have

$$\begin{aligned}
I_1^b &\leqslant \frac{1}{T}\sum_{l=1}^{L}\sum_{m=1}^{M} \Big| \int_{x_m}^{x_{m+1}} \int_{t_{l-1}}^{t_l} g(x,t)\partial_t u(x,t) - \widehat{\partial_t u}(x_m,t_l) g(x_m,t_l) dx dt\Big| \\
&\leqslant \frac{1}{T}\sum_{l=1}^{L}\sum_{m=1}^{M} \Delta x \Delta t \sup_{x\in(x_m,x_{m+1}),t\in(t_{l-1},t_l)} \left|g(x,t)\partial_t u(x,t) - \widehat{\partial_t u}(x_m,t_l) g(x_m,t_l)\right| \\
&\leqslant 2c_{\mathcal{H}}^\infty R_\Omega |\Omega| (\|u\|_{1,\infty}^2 + \|u\|_{2,\infty})(\Delta x + \Delta t).
\end{aligned}$$

□

Acknowledgments. FL is grateful for supports from NSF-1821211, NSF-1913243 and DE-SC0021361. FL would like to thank Mauro Maggioni, P-E Jabin and Zhenfu Wang for helpful discussions.

References

[1] L. AMBROSIO, N. GIGLI, AND G. SAVARÉ, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008.

[2] J. BALADRON, D. FASOLI, O. FAUGERAS, AND J. TOUBOUL, *Mean-field description and propagation of chaos in networks of Hodgkin-Huxley and FitzHugh-Nagumo neurons*, J. Math. Neurosci., 2 (2012), p. 10.

[3] M. BONGINI, M. FORNASIER, M. HANSEN, AND M. MAGGIONI, *Inferring interaction rules from observations of evolutive systems I: The variational approach*, Math. Models Methods Appl. Sci., 27 (2017), pp. 909–951.

[4] J. A. CARRILLO, K. CRAIG, AND Y. YAO, *Aggregation-Diffusion Equations: Dynamics, Asymptotics, and Singular Limits*, in Active Particles, Volume 2, N. Bellomo, P. Degond, and E. Tadmor, eds., Springer International Publishing, Cham, 2019, pp. 65–108.

[5] J. A. CARRILLO AND G. TOSCANI, *Wasserstein metric and large–time asymptotics of nonlinear diffusion equations*, in New Trends in Mathematical Physics: In Honour of the Salvatore Rionero 70th Birthday, World Scientific, 2004, pp. 234–244.

[6] P. CATTIAUX, A. GUILLIN, AND F. MALRIEU, *Probabilistic approach for granular media equations in the non-uniformly convex case*, Probab. Theory Related Fields, 140 (2007), pp. 19–40.

[7] F. CUCKER AND S. SMALE, *Best Choices for Regularization Parameters in Learning Theory: On the Bias—Variance Problem*, Found. Comput. Math., 2 (2002), pp. 413–428.

[8] F. CUCKER AND D. X. ZHOU, *Learning theory: an approximation theory viewpoint*, vol. 24, Cambridge University Press, 2007.

[9] P. DEL MORAL, *Mean field simulation for Monte Carlo integration*, CRC press, 2013.

[10] M. FEDELE, C. VERNIA, AND P. CONTUCCI, *Inverse problem robustness for multi-species mean-field spin models*, J. Phys. A, 46 (2013), p. 065001.

[11] N. FOURNIER AND B. JOURDAIN, *Stochastic particle approximation of the Keller–Segel equation and two-dimensional generalization of Bessel processes*, Ann. Appl. Probab., 27 (2017), pp. 2807–2861.

[12] F. GOLSE, *On the Dynamics of Large Particle Systems in the Mean Field Limit*, in Macroscopic and Large Scale Phenomena: Coarse Graining, Mean Field Limits and Ergodicity, A. Muntean, J. Rademacher, and A. Zagaris, eds., vol. 3, Springer International Publishing, Cham, 2016, pp. 1–144.

[13] L. GYÖRFI, M. KOHLER, A. KRZYZAK, AND H. WALK, *A distribution-free theory of nonparametric regression*, Springer Science & Business Media, 2006.

[14] P. C. HANSEN, *The L-curve and its use in the numerical treatment of inverse problems*, in in Computational Inverse Problems in Electrocardiology, ed. P. Johnston, Advances in Computational Bioengineering, WIT Press, 2000, pp. 119–142.

[15] P.-E. JABIN AND Z. WANG, *Mean Field Limit for Stochastic Particle Systems*, in Active Particles, Volume 1, N. Bellomo, P. Degond, and E. Tadmor, eds., Springer International Publishing, Cham, 2017, pp. 379–402.

[16] I. KARATZAS AND S. E. SHREVE, *Brownian motion*, in Brownian Motion and Stochastic Calculus, Springer, second ed., 1998.

[17] E. F. KELLER AND L. A. SEGEL, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415.

[18] N. KOLBE, *Wasserstein distance*. URL: https://github.com/nklb/wasserstein-distance, 2020.

[19] Y. A. KUTOYANTS, *Statistical inference for ergodic diffusion processes*, Springer, 2004.

[20] Q. LANG AND F. LU, *Identifiability of interaction kernels in mean-field equations of interacting particles*, arXiv preprint arXiv:2106.05565, (2021).

[21] Z. LI, F. LU, M. MAGGIONI, S. TANG, AND C. ZHANG, *On the identifiability of interaction functions in systems of interacting particles*, arXiv:1912.11965, (2019).

[22] F. LU, M. MAGGIONI, AND S. TANG, *Learning interaction kernels in heterogeneous systems of agents from multiple trajectories*, arXiv:1910.04832, (2019).

[23] F. LU, M. MAGGIONI, AND S. TANG, *Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories*, arXiv:2007.15174, (2020).

[24] F. LU, M. ZHONG, S. TANG, AND M. MAGGIONI, *Nonparametric inference of interaction laws in systems of agents from trajectory data*, Proc. Natl. Acad. Sci. USA, 116 (2019), pp. 14424–14433.

[25] T. LYCHE, C. MANNI, AND H. SPELEERS, *Foundations of Spline Theory: B-Splines, Spline Approximation, and Hierarchical Refinement*, vol. 2219, Springer International Publishing, Cham, 2018, pp. 1–76.

[26] F. MALRIEU, *Convergence to equilibrium for granular media equations and their Euler schemes*, Ann. Appl. Probab., 13 (2003), pp. 540–560.

[27] S. MOTSCH AND E. TADMOR, *Heterophilious Dynamics Enhances Consensus*, SIAM Rev., 56 (2014), pp. 577–621.

[28] L. PARESCHI AND M. ZANELLA, *Structure Preserving Schemes for Nonlinear Fokker–Planck Equations and Applications*, J. Sci. Comput., 74 (2018), pp. 1575–1600.

[29] L. PIEGL AND W. TILLER, *The NURBS Book*, Monographs in Visual Communication, Springer Berlin Heidelberg, Berlin, Heidelberg, 1997.

[30] A.-S. SZNITMAN, *Topics in Propagation of Chaos*, vol. 1464, Springer Berlin Heidelberg, Berlin, Heidelberg, 1991, pp. 165–251.

[31] Z. ZHANG AND F. LU, *Cluster prediction for opinion dynamics from partial observations*, arXiv preprint arXiv:2007.02006, (2020).

[32] M. ZHONG, J. MILLER, AND M. MAGGIONI, *Data-driven discovery of emergent behaviors in collective dynamics*, Phys. D, (2020), p. 132542.