Foundations of Computational Mathematics
https://doi.org/10.1007/s10208-021-09521-z

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Learning Interaction Kernels in Stochastic Systems of Interacting Particles from Multiple Trajectories

Fei Lu[1] · **Mauro Maggioni**[2] · **Sui Tang**[3]

Received: 2 August 2020 / Revised: 16 April 2021 / Accepted: 19 April 2021

Abstract
We consider stochastic systems of interacting particles or agents, with dynamics determined by an interaction kernel, which only depends on pairwise distances. We study the problem of inferring this interaction kernel from observations of the positions of the particles, in either continuous or discrete time, along multiple independent trajectories. We introduce a nonparametric inference approach to this inverse problem, based on a regularized maximum likelihood estimator constrained to suitable hypothesis spaces adaptive to data. We show that a coercivity condition enables us to control the condition number of this problem and prove the consistency of our estimator, and that in fact it converges at a near-optimal learning rate, equal to the min–max rate of one-dimensional nonparametric regression. In particular, this rate is independent of the dimension of the state space, which is typically very high. We also analyze the discretization errors in the case of discrete-time observations, showing that it is of order $1/2$ in terms of the time spacings between observations. This term, when large, dominates the sampling error and the approximation error, preventing convergence of the estimator. Finally, we exhibit an efficient parallel algorithm to construct the estimator from data, and we demonstrate the effectiveness of our algorithm with numerical tests on prototype systems including stochastic opinion dynamics and a Lennard-Jones model.

Keywords Inverse problems · Interacting particle systems · Statistical and machine learning

Mathematics Subject Classification 70F17 · 62G05 · 62M05

Communicated by Hans Munthe Kaas.

Extended author information available on the last page of the article

Published online: 13 July 2021

1 Introduction

We consider a system of particles or agents interacting in a random environment, with their motion described by a first-order stochastic differential equation in the form

$$\mathrm{d}\boldsymbol{x}_{i,t} = \frac{1}{N}\sum_{i'=1}^{N}\phi(\|\boldsymbol{x}_{j,t}-\boldsymbol{x}_{i,t}\|)(\boldsymbol{x}_{j,t}-\boldsymbol{x}_{i,t})\mathrm{d}t + \sigma\mathrm{d}\boldsymbol{B}_{i,t}\,, \quad \text{for } i=1,\ldots,N, \tag{1.1}$$

where $\boldsymbol{x}_{i,t} \in \mathbb{R}^d$ represents the position of particle i at time t, $\phi : \mathbb{R}^+ \to \mathbb{R}$ is an *interaction kernel* dependent on the pairwise distance between particles, and $\boldsymbol{B}_t$ is a standard Brownian motion in $\mathbb{R}^{Nd}$, with $\sigma > 0$ representing the scale of the random noise. This is a gradient system, with the energy potential $V_\phi : \mathbb{R}^{Nd} \to \mathbb{R}$

$$V_\phi(\boldsymbol{X}_t) = \frac{1}{2N}\sum_{i,i'}\Phi(\|\boldsymbol{x}_{i,t}-\boldsymbol{x}_{i',t}\|) \quad \text{with} \quad \Phi'(r) = \phi(r)r\,, \tag{1.2}$$

where $\boldsymbol{X}_t = (\boldsymbol{x}_{i,t})_{i=1,\ldots,N} \in \mathbb{R}^{dN}$ is the state of the system. Letting

$$\boldsymbol{f}_\phi := -\nabla V_\phi\,, \tag{1.3}$$

we can write Eq.(1.1) in vector format as

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{f}_\phi(\boldsymbol{X}_t)\mathrm{d}t + \sigma\mathrm{d}\boldsymbol{B}_t\,. \tag{1.4}$$

The particles interact with each other based on their pairwise distance, with dissipation of the total energy, with the system tending to a stable point of the energy potential, while the random noise injects energy to the system.

Such systems of interacting particles arise in a wide variety of disciplines, including interacting physical particles [22,49] or granular media [1–3,8,12,13] in Physics, opinion aggregation on interacting networks in Social Science [24,43,46], and Monte Carlo sampling [36,39], to name just a few.

Motivated by these applications, the inference of such systems from data gains increasing attention. For deterministic multi-particle systems, various types of learning techniques have been developed (see, e.g., [9,14,40,41,50,55] and the reference therein). When it comes to stochastic multi-particle systems, only a few efforts have been made, e.g., learning reduced Langevin equations on manifolds in [19] (without, however, assuming nor exploiting the structure of pairwise interactions), learning parametric potential functions in [10,15] from single trajectory data, estimating the diffusion parameter in [26], and estimating effective Langevin equations on manifolds in [19].

Our goal is to estimate the interaction kernel ϕ given discrete-time observation data from trajectories $\{\boldsymbol{X}^{(m)}_{t_0:t_L}\}_{m=1}^M$, where the initial conditions $\{\boldsymbol{X}^{(m)}_{t_0}\}_{m=1}^M$ are independent

samples drawn from a distribution μ_0 on $\mathbb{R}^{dN}$, and $t_0 : t_L$ indicates times $0 = t_0 < t_1 < \cdots < t_l < \cdots < t_L = T$, with with $t_l = l\Delta t$.

Since, in general, little information about the analytical form of the kernel is available, we infer it in a nonparametric fashion (e.g., [6,20,23]). We note that the problem we consider is to learn a latent function in the drift term given observations from multiple trajectories, which is different from the ample literature on the inference of stochastic differential equations (see, e.g., [29,34]), focusing either on parameter estimation or on inference for ergodic system. In particular, our learning approach is close in spirit to the nonparametric regression of the drift studied in [44] for ergodic system and in [17] from i.i.d paths. However, for systems of interacting particles one faces the curse of dimensionality when learning the high-dimensional drift directly as a general function on the high-dimensional state space $\mathbb{R}^{dN}$. Instead, we will exploit the structure of the system and learn the latent interaction kernel in the drift, which only depends on pairwise distances, and show that the curse of dimensionality may be avoided, when such inverse problem is well-conditioned.

We introduce a maximum likelihood estimator (MLE), along with an efficient algorithm that can be implemented in parallel over trajectories, with an hypothesis space adaptive to data to reach optimal accuracy. Under a coercivity condition, we prove that the MLE is consistent and converges at the min–max rate for one-dimensional nonparametric regression. We also analyze the discretization errors due to discrete-time observations: we show it leads to an error in the estimator that is of order $\Delta t^{1/2}$ (with $\Delta t = T/L = t_{l+1} - t_l$), and as a result, it prevents us from obtaining the min–max learning rate in sample size. We demonstrate the effectiveness of our algorithm by numerical tests on prototype systems including opinion dynamics and a stochastic Lennard-Jones model (see Sect. 5). Numerical results verify our learning theory in the sense that the min–max rate of convergence is achieved, and the bias due to the numerical error is close to the order $\Delta t^{1/2}$.

1.1 Overview of the Main Results

We consider an approximate maximum likelihood estimator (MLE), which is the maximizer of the approximate likelihood of the observed trajectories, over a suitable hypothesis space $\mathcal{H}$:

$$\widehat{\phi}_{L,T,M,\mathcal{H}} = \arg\min_{\varphi \in \mathcal{H}} \mathcal{E}_{L,T,M}(\varphi),$$

where $\mathcal{E}_{L,T,M}(\varphi)$ is an approximation of the negative log-likelihood of the discrete data $\{\boldsymbol{X}^{(m)}_{t_0:t_L}\}_{m=1}^{M}$. Using the fact that the drift term $\boldsymbol{f}_\phi$ is linear in ϕ and hence $\mathcal{E}_{L,T,M}(\varphi)$ is a quadratic functional, we propose an algorithm (see Algorithm 1) that efficiently computes this MLE by least squares. With a data-adaptive choice of the basis functions $\{\psi_p\}_{p=1}^{n}$ for the hypothesis space $\mathcal{H}$, we obtain the MLE

$$\widehat{\phi}_{L,T,M,\mathcal{H}} = \sum_{p=1}^{n} \widehat{a}_{L,T,M,\mathcal{H}}(p)\psi_p \tag{1.5}$$

by computing the coefficients $\widehat{a}_{L,T,M,\mathcal{H}} \in \mathbb{R}^n$ from normal equations. The algorithm may be implemented by building in parallel the equations for each trajectory.

We develop a systematic learning theory on the performance of this MLE. We propose first a coercivity condition that ensures the robust identifiability of the kernel ϕ, in the sense that the derivative of the pairwise potential defined in (1.2), $\Phi'(r) = \phi(r)r$, can be uniquely identified in the function space $L^2(\mathbb{R}^+, \rho_T)$, where ρ_T is the measure of all pairwise distances between particles. Then, we consider the convergence of the estimator, from both continuous-time and discrete-time observations, under the norm

$$|||\varphi||| := \|\varphi(\cdot)\cdot\|_{L^2(\rho_T)} = \left(\int_{\mathbb{R}^+} |\varphi(r)r|^2 \rho_T(\mathrm{d}r)\right)^{1/2}. \tag{1.6}$$

The case of continuous-time observations (Sect. 3). We consider the MLE

$$\widehat{\phi}_{T,M,\mathcal{H}} = \arg\min_{\varphi\in\mathcal{H}} \mathcal{E}_{T,M}(\varphi),$$

where $\mathcal{E}_{T,M}(\varphi)$ is the exact negative log-likelihood of the continuous-time trajectories $\{\boldsymbol{X}^{(m)}_{[0,T]}\}_{m=1}^M$. We show that the MLE is consistent, that is, converges in probability to the true kernel under the norm $|||\cdot|||$. Furthermore, we show that the MLE converges at a rate which is independent of the dimension of the state space of the system and corresponds to the mini–max rate for one-dimensional nonparametric regression ([6,16,20,23]), when choosing the hypothesis space adaptively according to data, the Hölder continuity s of the true kernel, and with dimension increasing with the amount of observed data. With $\dim(\mathcal{H}) \asymp (\frac{M}{\log M})^{\frac{1}{2s+1}}$, and assuming that the coercivity condition holds on $\mathcal{H}$ with a constant $c_{\mathcal{H}} > 0$, we have, with high probability and in expectation,

$$|||\widehat{\phi}_{T,M,\mathcal{H}} - \phi|||^2 \lesssim \frac{1}{c_{\mathcal{H}}^2}\left(\frac{\log M}{M}\right)^{\frac{2s}{2s+1}}$$

The case of discrete-time observations (Sect. 4). In this case, derivatives and statistics of the trajectories in-between observations need to be approximated, while keeping the estimator efficiently computable: this leads to further approximations of the likelihood and consequently of the MLE. This discretization error of the approximations we use is of order $1/2$ in the observation time gap $\Delta t = T/L$, using an approximation of the likelihood based on the Euler–Maruyama integration scheme. We show that for some $C > 0$, for any $\epsilon > 0$, with high probability

$$|||\widehat{\phi}_{L,T,M,\mathcal{H}} - \phi||| \le |||\widehat{\phi}_{T,\infty,\mathcal{H}} - \phi||| + C\left(\sqrt{\frac{n}{M}}\epsilon + \Delta t^{\frac{1}{2}}\right),$$

where $\widehat{\phi}_{T,\infty,\mathcal{H}}$ is the projection of the true kernel to $\mathcal{H}$ and n is the dimension of hypothesis space $\mathcal{H}$. The discretization error will flatten the learning curve when the sample size is large, overshadowing the sampling error and the approximation error cause by working within the hypothesis space. For some positive constants c_2 and

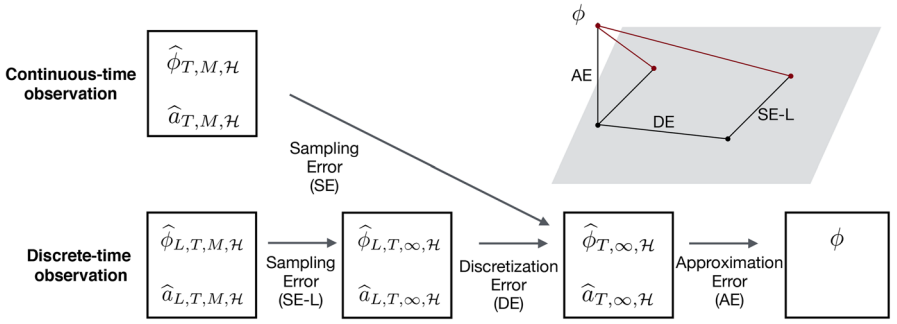


Fig. 1 Diagram of the (regularized) MLE error analysis and convergence. For continuous-time observations, we refer to Proposition 3.4 for analysis of SE: $\hat{\phi}_{T,M,\mathcal{H}} - \hat{\phi}_{T,\infty,\mathcal{H}}$ and Theorem 3.2 for bounding the total estimation error $\hat{\phi}_{T,M,\mathcal{H}} - \phi$. For discrete-time observations, we refer to Proposition 4.2 for SE-L: $\hat{\phi}_{L,T,M,\mathcal{H}} - \hat{\phi}_{L,T,\infty,\mathcal{H}}$, Proposition 4.1 for DE: $\hat{\phi}_{L,T,\infty,\mathcal{H}} - \hat{\phi}_{T,\infty,\mathcal{H}}$, and Theorem 4.2 for $\hat{\phi}_{L,T,M,\mathcal{H}} - \phi$

c_3, where $\widehat{\phi}_{T,\infty,\mathcal{H}}$ is the projection of the true kernel to $\mathcal{H}$. The numerical error may overshadow the sampling error and the approximation error of the hypothesis space.

In both cases, we decompose the error in the MLE into sampling error from the trajectory data, and approximation error from the hypothesis space, as illustrated in the diagram in Fig. 1. In the case of continuous-time observations, the sampling error is the error between $\widehat{\phi}_{T,M,\mathcal{H}}$ and the MLE from infinitely many trajectories (denoted by $\widehat{\phi}_{T,\infty,\mathcal{H}}$): this will be controlled with concentration equalities. The approximation error $\widehat{\phi}_{T,\infty,\mathcal{H}} - \phi$ is adaptively controlled by a proper choice of hypothesis space. The analysis is carried out in the infinite-dimensional space $L^2(\rho_T)$. In the case of discrete-time observations, we provide a finite-dimensional analysis to study directly the MLE in our proposed algorithm, that is, analyzing the error of $\widehat{a}_{L,T,M,\mathcal{H}}$ in (1.5) with proper conditions on the basis functions. The sampling error $\widehat{\phi}_{L,T,M,\mathcal{H}} - \widehat{\phi}_{L,T,\infty,\mathcal{H}}$ is analyzed through $\widehat{a}_{L,T,M,\mathcal{H}} - \widehat{a}_{L,T,\infty,\mathcal{H}}$, and the discretization error between $\widehat{\phi}_{L,T,\infty,\mathcal{H}}$ and $\widehat{\phi}_{T,\infty,\mathcal{H}}$ is analyzed through $\widehat{a}_{L,T,M,\mathcal{H}} - \widehat{a}_{T,\infty,\mathcal{H}}$. The discretization error comes from the discrete-time approximation of the likelihood, and it vanishes when the observation time gap Δt reduces to zero, recovering the convergence of the MLE as in the case of the continuous-time observations.

1.2 Notation and Outline

Throughout this paper, we use bold letters to denote vectors or vector-valued functions. We use the notation in Table 1 for variables in the system of interacting particles.

We call the system of relative positions to a reference particle, say, $(\boldsymbol{r}_{1i} = \boldsymbol{x}_{i,t} - \boldsymbol{x}_{1,t})_{i=2}^N$, by "relative position system". The relative position system can be ergodic under suitable conditions on the potential [37], and these relative positions are the variables we need to learn the interaction kernel. We point out that the interacting particle system (1.1) itself is not ergodic, because the center $\overline{\boldsymbol{x}}_t = \frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_{i,t}$ satisfies $\mathrm{d}\overline{\boldsymbol{x}}_t = \sigma \frac{1}{N}\sum_{i=1}^N \mathrm{d}\boldsymbol{B}_{i,t}$.

Table 1 Notation for the system of interacting particles driven by Eq. (1.1)

Variable	Definition
$\boldsymbol{x}_{i,t} \in \mathbb{R}^d$	Position or opinion of particle i at time t, see (1.1)
$\boldsymbol{X}_t = (\boldsymbol{x}_{1,t}, \ldots, \boldsymbol{x}_{N,t}) \in \mathbb{R}^{dN}$	State vector: position of the N particles, see (1.4)
$\|\cdot\|$	Euclidean norm in $\mathbb{R}^d$ or operator norm of a matrix
$\boldsymbol{r}_{ii'}(t), \boldsymbol{r}_{ii''}(t) \in \mathbb{R}^d$	$\boldsymbol{x}_{i',t} - \boldsymbol{x}_{i,t}$ and $\boldsymbol{x}_{i'',t} - \boldsymbol{x}_{i,t}$,
$r_{ii'}(t), r_{ii''}(t) \in \mathbb{R}^+$	$r_{ii'}(t) = \|\boldsymbol{r}_{ii'}(t)\|$ and $r_{ii''}(t) = \|\boldsymbol{r}_{ii''}(t)\|$,
ϕ	Interaction kernel, see (1.1)
$\mathbf{f}_\phi$	Drift function of the system, see (1.4)
$V_\phi = \frac{1}{2N}\sum_{i,i'} \Phi(\|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i',t}\|)$	Energy potential with $\Phi'(r) = \phi(r)r$, see (1.2)
Relative position system	The system of $(\boldsymbol{r}_{1i}(t) = \boldsymbol{x}_{i,t} - \boldsymbol{x}_{1,t})_{i=2}^N$

We restrict our attention to interaction kernels ϕ in the admissible set

$$\mathcal{K}_{R,S} := \{\varphi \in C^1(\mathbb{R}_+) : \mathrm{Supp}(\varphi) \subset [0, R], \|\varphi\|_\infty + \|\varphi'\|_\infty \le S\}. \tag{1.7}$$

Let Ω be an arbitrary compact (or precompact) set of a Euclidean space (which may be $\mathbb{R}^+$, $\mathbb{R}^d$ or $\mathbb{R}^{dN}$), with the Lebesgue measure unless otherwise specified. We consider the following function spaces

- $L^\infty(\Omega)$: the space of bounded functions on Ω with norm $\|g\|_\infty = \mathrm{ess\,sup}_{x\in\Omega}|g(x)|$;
- $C(\Omega)$: the closed subspace of $L^\infty(\Omega)$ consisting of continuous functions;
- $C_c(\Omega)$: the set of functions in $C(\Omega)$ with compact support;
- $C^{k,\alpha}(\Omega)$ with $k \in \mathbb{N}, 0 < \alpha \le 1$: the space of functions whose k-th derivative is Hölder continuous of order α. In the special case of $k = 0$ and $\alpha = 1$, $g \in C^{0,\alpha}(\Omega)$ is called Lipchitz continuous on Ω; the Lipschitz constant of $g \in \mathrm{Lip}(\Omega)$ is defined as $\mathrm{Lip}[g] := \sup_{x\neq y} \frac{|g(x)-g(y)|}{\|x-y\|}$.

We summarize the notation for the inference of the interaction kernel in Table 2. The function space in which we perform the estimation is the space of functions φ such that $\varphi(\cdot)\cdot \in L^2(\mathbb{R}^+, \rho_T)$, where ρ_T is the measure of pairwise distances between all particles on the time interval $[0, T]$ (see (2.9)). We will focus on learning on the compact (finite- or infinite-dimensional) subset of $L^\infty([0, R])$ (where $[0, R]$ is the support of the functions in the admissible set $\mathcal{K}_{R,S}$) in the theoretical analysis; however, in the numerical implementation we will use finite-dimensional linear subspaces $L^2([0, R], \rho_T)$ spanned by piecewise polynomial functions. While these linear subspaces are not compact, it is shown that the minimizers over the whole linear space are bounded and thus the compactness requirements are not essential (e.g., see Theorem 11.3 in [23]). We shall therefore assume the compactness of the hypothesis space in the theoretical analysis.

The remainder of the paper is organized as follows: We first provide an overview of our learning theory. In Sect. 2, we present a practical learning algorithm with theory-guided optimal settings on the choice of hypothesis spaces and with a practical assessment of the learning results. We then demonstrate the effectiveness of the algo-

Table 2 Notation used for the estimator of the interaction kernel ϕ

Notation	Definition
M	Number of observed trajectories
$t_0 : t_L = \{t_l\}_{l=1}^L$	Observation times in $[0, T]$, $0 = t_0 < \cdots < t_L = T$, $t_l = l\Delta t = lT/L$
μ_0	Probability distribution in $\mathbb{R}^{dN}$ for initial configurations $\boldsymbol{X}_0$
$\mathcal{H}$ and $\{\psi_p\}_{p=1}^n$	The hypothesis space of learning, and a basis for it
$\mathcal{E}_{T,M}(\cdot)$ and $\mathcal{E}_{L,T,M}(\cdot)$	Empirical error functionals from continuous/discrete data, see (3.1) and (2.4)
$\widehat{\phi}_{T,M,\mathcal{H}}$ and $\widehat{\phi}_{L,T,M,\mathcal{H}}$	Minimizers, over $\mathcal{H}$, of $\mathcal{E}_{T,M}(\varphi)$ and $\mathcal{E}_{L,T,M}(\varphi)$, see (3.2) and (2.5)
$\widehat{a}_{L,T,M,\mathcal{H}}$	Coefficient vectors of $\widehat{\phi}_{L,T,M,\mathcal{H}}$ w.r.t. basis $\{\psi_p\}_{p=1}^n$, see (2.8)
$\|\|\|\cdot\|\|\|$	$L^2(\rho_T)$-based norm: $\|\|\|\phi\|\|\| = \|\phi(\cdot)\cdot\|_{L^2(\rho_T)}$, see (1.6)

rithm on prototype systems including a stochastic model for opinion dynamics and a stochastic Lennard-Jones model in Sect. 5. We establish a systematic learning theory analyzing the performance of the MLE, considering continuous-time observations in Sect. 3 and discrete-time observations in Sect. 4. We present in "Appendix" detailed proofs.

2 Nonparametric Inference of the Interaction Kernel

We present in this section the nonparametric technique we study for the inference of the interaction kernel and corresponding algorithms. We discuss the assessment of the performance of the estimator and its performance in trajectory prediction. The proposed estimator is based on maximum likelihood estimation on data-adaptive hypothesis spaces so as to achieve the optimal rate of convergence, guided by our learning theory in Sects. 3–4.

2.1 The Maximum Likelihood Estimator

As a variational approach, we set the error functional to be the negative log-likelihood of the data $\{\boldsymbol{X}_{t_0:t_L}^{(m)}\}_{m=1}^M$ and compute the maximum likelihood estimator (MLE).
The error functional. Recall that by the Girsanov theorem, for a continuous trajectory $\boldsymbol{X}_{[0,T]}$, its negative log-likelihood ratio between the measure induced by system (1.1), with an admissible kernel ϕ, and the Wiener measure is

$$\mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\phi) = \frac{1}{2\sigma^2 TN}\int_0^T \left(\|\boldsymbol{f}_\phi(\boldsymbol{X}_t)\|^2 - 2\langle \boldsymbol{f}_\phi(\boldsymbol{X}_t), \mathrm{d}\boldsymbol{X}_t\rangle\,\mathrm{d}t\right). \tag{2.1}$$

As we do not know the interaction kernel ϕ that generated the trajectory $\boldsymbol{X}_{[0,T]}$, we can let φ be any possible admissible interaction kernel, and upon replacing ϕ by φ in (2.1), observe that $\mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi)$ is the log-likelihood of seeing the trajectory $\boldsymbol{X}_{[0,T]}$ if system (1.1) were driven by the interaction kernel φ. In this case, $\mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi)$ may be

interpreted as a error functional, which we wish to minimize over φ, in order to obtain an estimator for ϕ.

Given only discrete-time observations $\boldsymbol{X}_{t_0:t_L}$, where $(t_l = l\Delta t, l = 0, \ldots, L)$ with $\Delta t = T/L$ (the case of non-equispaced-in-time observation is a straightforward generalization), the error functional $\mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi)$ may be approximated as

$$\mathcal{E}_{\boldsymbol{X}_{t_1:t_L}}(\varphi) := \frac{1}{2\sigma^2 TN} \sum_{l=0}^{L-1} \left(\|\boldsymbol{f}_\varphi(\boldsymbol{X}_{t_l})\|^2 \Delta t - 2\langle \boldsymbol{f}_\varphi(\boldsymbol{X}_{t_l}), \boldsymbol{X}_{t_l} - \boldsymbol{X}_{t_{l-1}} \rangle \right). \tag{2.2}$$

The corresponding approximate likelihood is equivalent to the likelihood based on the Euler–Maruyama (EM) scheme (whose transition probability density is Gaussian):

$$\boldsymbol{X}_{t_{l+1}} = \boldsymbol{X}_{t_l} + \boldsymbol{f}_\varphi(\boldsymbol{X}_{t_l})\Delta t + \sigma\sqrt{\Delta t}\mathbf{W}_l\,, \quad \mathbf{W}_l \sim \mathcal{N}(0, I_{Nd\times Nd}) \tag{2.3}$$

Note that while higher-order approximations of the stochastic integral (or, equivalently, approximations based on higher-order numerical schemes) may be more accurate than the EM scheme, they lead to nonlinear optimization problems in the computation of the MLE defined below, and we shall therefore avoid them. The EM-based approximation preserves the quadratic form of the error functional and leads to an optimization problem that can be solved by least squares. As we show in Theorem 4.2, this discrete-time approximation leads to an error term of order $\Delta t^{1/2}$ in the MLE, which will be small in the regime on which we focus in this work.

Since the observed discrete-time trajectories $\{\boldsymbol{X}^{(m)}_{t_0:t_L}\}_{m=1}^M$ are independent, as the $\boldsymbol{X}^{(m)}_{t_0}$'s are drawn i.i.d. from μ_0, the joint likelihood of the trajectories is the product of the likelihoods of each trajectory. Therefore, the corresponding *empirical error functional* is defined to be

$$\mathcal{E}_{L,T,M}(\varphi) := \frac{1}{M} \sum_{m=1}^{M} \mathcal{E}_{\boldsymbol{X}^{(m)}_{t_1:t_L}}(\varphi). \tag{2.4}$$

A regularized Maximum Likelihood Estimator. The regularized MLE we consider is a minimizer of the above empirical error functional over a suitable hypothesis space $\mathcal{H}$:

$$\widehat{\phi}_{L,T,M,\mathcal{H}} = \operatorname*{arg\,min}_{\varphi\in\mathcal{H}} \mathcal{E}_{L,T,M}(\varphi), \tag{2.5}$$

This regularized MLE is well defined when the minimizer exists and is unique over $\mathcal{H}$. We call this MLE "regularized" to emphasize the constraint $\phi \in \mathcal{H}$, and the fact that $\mathcal{H}$ will change with M, as in nonparametric statistics; this naming is somewhat not standard though. We shall discuss the uniqueness of the minimizer in Sect. 3.1, where we show it is guaranteed by a coercivity condition. When the hypothesis space $\mathcal{H}$ is a finite-dimensional linear space, say, $\mathcal{H} = \operatorname{span}\{\psi_i\}_{i=1}^n$ with basis functions $\psi_i : \mathbb{R}^+ \to \mathbb{R}$, the regularized MLE is the solution of a least squares problem.

To see this, letting $\varphi = \sum_{i=1}^{n} a(i)\psi_i$ and $a := (a(1), \ldots, a(n)) \in \mathbb{R}^n$, we have $\boldsymbol{f}_\varphi(\boldsymbol{X}) = \sum_{i=1}^{n} a(i)\boldsymbol{f}_{\psi_i}(\boldsymbol{X})$, due to the linear dependence of $\boldsymbol{f}_\varphi$ on φ. Then, we can write the error functional in Eq.(2.2) for each trajectory as

$$\mathcal{E}_{\boldsymbol{X}_{t_1:t_L}^{(m)}}(\boldsymbol{a}) := \mathcal{E}_{\boldsymbol{X}_{t_1:t_L}^{(m)}}(\varphi) = a^T A^{(m)} a + a^T b^{(m)},$$

where the matrix $A^{(m)} \in \mathbb{R}^{n\times n}$ and the vector $b^{(m)} \in \mathbb{R}^n$ are given by

$$\begin{aligned} A^{(m)}(i,i') &= \frac{1}{2\sigma^2 LN}\sum_{l=0}^{L-1}\langle \boldsymbol{f}_{\psi_i}(\boldsymbol{X}_{t_l}^{(m)}), \boldsymbol{f}_{\psi_{i'}}(\boldsymbol{X}_{t_l}^{(m)})\rangle, \\ b^{(m)}(i) &= -\frac{1}{\sigma^2 L\Delta t N}\sum_{l=0}^{L-1}\langle \boldsymbol{f}_{\psi_i}(\boldsymbol{X}_{t_l}^{(m)}), \boldsymbol{X}_{t_{l+1}}^{(m)} - \boldsymbol{X}_{t_l}^{(m)}\rangle. \end{aligned} \tag{2.6}$$

Hence, corresponding to $\nabla\mathcal{E}_{L,T,M} = 0$ for the error functional in (2.4), we solve the normal equations for a to obtain the solution $\widehat{a}_{L,T,M,\mathcal{H}}$:

$$\begin{aligned} &A_{M,L}\widehat{a}_{L,T,M,\mathcal{H}} = b_{M,L}, \text{ where} \\ &A_{M,L} := \frac{1}{M}\sum_{m=1}^{M} A^{(m)},\ b_{M,L} := \frac{1}{M}\sum_{m=1}^{M} b^{(m)} \end{aligned} \tag{2.7}$$

and corresponding desired MLE for the interaction kernel:

$$\widehat{\phi}_{L,T,M,\mathcal{H}} = \sum_{i=1}^{n}\widehat{a}_{L,T,M,\mathcal{H}}(i)\psi_i. \tag{2.8}$$

The normal equations (2.7) are solved by least squares, so the solution always exists. We will show in Sect. 4 that assuming a coercivity condition, the matrix $A_{M,L} \in \mathbb{R}^{n\times n}$ is invertible with high probability when M and L are large, so the least squares estimator is the unique solution to the normal equations, and the regularized MLE is the unique minimizer of the empirical error functional over $\mathcal{H}$.

2.2 Dynamics-Adapted Measures and Function Spaces

We will assess the estimation error in a suitable function space: $L^2(\mathbb{R}^+, \rho_T)$. Here ρ_T is the distribution of pairwise distances between all particles:

$$\rho_T(\mathrm{d}r) := \frac{1}{\binom{N}{2}T}\int_{t=0}^{T}\left[\sum_{i,i'=1,i<i'}^{N}\mathbb{E}[\delta_{r_{ii'}(t)}(\mathrm{d}r)]\,\mathrm{d}t\right], \tag{2.9}$$

where δ is the Dirac δ-distribution, so that $\mathbb{E}[\delta_{r_{ii'}(t)}(\mathrm{d}r)]$ is the distribution of the random variable $r_{ii'}(t) = ||\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i',t}||$, with $\boldsymbol{x}_{i,t}$ being the position of particle i at time t.

Here the expectation is taken over the distribution μ_0 of initial conditions and realizations of the system. The probability measure ρ_T depends on both μ_0 and the measure determining the random noise on the path space, while it is independent of the observed data. The measure ρ_T encodes the information about the dynamics marginalized to pairwise distances; regions with large ρ_T-measure correspond to pairwise distances between particles that are often encountered during the dynamics.

With observations of M trajectories at L discrete-times each, we introduce a corresponding measure

$$\rho_T^{L,M}(\mathrm{d}r) := \frac{1}{\binom{N}{2}LM} \sum_{l=0,m=1}^{L-1,M} \left[\sum_{i,i'=1,i<i'}^{N} \delta_{r_{ii'}^{(m)}(t_l)}(\mathrm{d}r) \right], \tag{2.10}$$

where $r_{ii'}^{(m)}(t) = ||\boldsymbol{x}_{i,t}^{(m)} - \boldsymbol{x}_{i',t}^{(m)}||$ is from the m-th observed trajectory. We think of this as an approximation to ρ_T, in two significantly different aspects. In L, because as $L \to +\infty$ our observations tend to be continuous in time, and in M, as $\rho_T^{L,M}$ can be thought of, after letting $M \to +\infty$, as an empirical approximation to ρ_T performed from data on the M independent trajectories.

Accuracy of the estimator. We measure the accuracy of our estimator $\widehat{\phi}_{L,T,M,\mathcal{H}}$ by the quantity

$$\|(\widehat{\phi}_{L,T,M,\mathcal{H}} - \phi)(\cdot)\cdot\|_{L^2(\mathbb{R}^+,\rho_T)}.$$

The function $\phi(\cdot)\cdot$, instead of ϕ, which at $r \in \mathbb{R}_+$ takes value $\phi(r)r$, appears naturally in our learning theory in Sect. 3, fundamentally because it is the derivative of the pairwise distance potential Φ in (1.2). We define

$$|||\varphi||| := \|\varphi(\cdot)\cdot\|_{L^2(\rho_T)} = \left(\int_{\mathbb{R}^+} |\varphi(r)r|^2 \rho_T(\mathrm{d}r) \right)^{1/2}. \tag{2.11}$$

Then, the mean squared error (MSE) of the estimator is

$$\mathbb{E}\left|\left|\left|\widehat{\phi}_{L,T,M,\mathcal{H}} - \phi\right|\right|\right|^2. \tag{2.12}$$

2.3 Hypothesis Spaces and Nonparametric Estimators

As standard in nonparametric estimation, we let the hypothesis space $\mathcal{H}$ grow in dimension with the number of observations, avoiding under- or over-parameterization, and leading to consistent estimators, that in fact reach an optimal min–max rate of convergence (see, e.g., [20,21,23]).

Similar to [40,41], we set the basis functions $\{\psi_i\}_{i=1}^n$ to be piecewise polynomials on a partition of the support of the density function of the empirical probability measure $\rho_T^{L,M}$.

Guided by the optimal rate convergence results in Sect. 3, we will set the dimension of the hypothesis space $\mathcal{H}$ to be

$$n = C(M/\log M)^{1/(2s+1)}, \tag{2.13}$$

Algorithm 1 Learning interaction kernels from many trajectories

1: **Input:** Data consisting of M independent trajectories $\{\boldsymbol{X}^m_{t_0:t_L}\}_{m=1}^M$; Hölder regularity s of the true kernel.
2: **Output:** An estimator $\widehat{\phi}_{L,T,M,\mathcal{H}_{n_M}}$ for the interaction kernel.
3: Compute the pairwise distances and the empirical measure $\rho_T^{L,M}$ in (2.10).
4: Construct the basis $\{\psi_p\}_{p=1}^{n_M}$ with adaptive partition based on $\rho_T^{L,M}$, and with n_M given by (2.13).
5: Assemble the normal equations (2.7) (in parallel).
6: Solve the normal equation and return $\widehat{\phi}_{L,T,M,\mathcal{H}_{n_M}}$ as in (2.8).

where the number s is the Hölder index of continuity of the basis functions, and it is to be chosen according the regularity of the true kernel. When T is large and when the relative position system is ergodic, we set

$$n = C(N_{ess}/\log N_{ess})^{1/(2s+1)},$$

where $N_{ess} := M\frac{T}{\tau}$, with τ denoting the auto-correlation time of the system, is the effective sample size of the data. Here the auto-correlation time τ is the equivalent of the mixing time for a reversible ergodic Markov chain [35].

We estimate the auto-correlation time by the sum of the temporal auto-correlation function of a pairwise distance $r_{i,j}$. (We refer to [51] for detailed discussion on the estimation of auto-correlation time, which is a whole subject by itself.)

We will prove bounds, that hold with high probability, on the mean squared error (MSE) of the MLE $\phi_{L,T,M,\mathcal{H}_{n_M}}$ in (2.8) for fixed and large M, for a fixed time T and for suitable hypothesis spaces $\mathcal{H}_{n_M}$ with dimension n_M as in (2.13). When continuous-time trajectories are observed, the MSE is of the order $(\frac{\log M}{M})^{\frac{2s}{2s+1}}$ with high probability, according to Theorem 3.2, and so is its expectation. In particular, this avoids the curse of dimensionality of the state space (dN). When the observations are discrete-time trajectories with observation gap Δt, the error is of the order $(\frac{\log M}{M})^{\frac{2s}{2s+1}} + \Delta t^{1/2}$ with a high probability according to Theorem 4.2.

2.4 Algorithmic and Computational Considerations

The algorithm is summarized in Algorithm 1. Note that the normal matrices $\{A^{(m)}\}$ and vectors $\{b^{(m)}\}$ are defined trajectory-wise and therefore may be computed in parallel. When the size of the system is large (i.e., dN is large), this allows one to accelerate the computation of the estimator, by assembling these normal matrices and vectors for each trajectory in parallel, and updating the normal matrix $A_{M,L}$ and vector $b_{M,L}$. The total computational cost of constructing our estimator, given P CPU's, is $O(L\frac{N^2d}{P}Mn^2 + n^3)$. This becomes $O(L\frac{N^2d}{P}M^{1+\frac{1}{2s+1}} + CM^{\frac{3}{2s+1}})$ when n is chosen optimally according to Theorem 3.2 and ϕ is at least in $C^{1,1}$ (corresponding to the index of regularity $s \geq 2$ in the theorem).

2.5 Accuracy of Trajectory Prediction

One application of estimating the interaction kernel from data is to perform predictions of the dynamics. Given an estimator, the following proposition bounds its accuracy in predicting the trajectories of the system driven by the true interaction kernel:

Proposition 2.1 *Let $\widehat{\phi} \in \mathcal{K}_{R,S}$ be an estimator of the true kernel ϕ , where $\mathcal{K}_{R,S}$ is the admissible set defined in (1.7). Denote by $\widehat{\boldsymbol{X}}_t$ and $\boldsymbol{X}_t$ the solutions of the systems with kernels $\widehat{\phi}$ and ϕ, respectively, starting from the same initial condition and with the same random noise. Then, we have*

$$\sup_{t\in[0,T]} \mathbb{E}\left[\frac{1}{N}\|\widehat{\boldsymbol{X}}_t - \boldsymbol{X}_t\|^2\right] \le 2T^2 e^{8T^2(R+1)^2S^2} \left|\left|\left|\widehat{\phi} - \phi\right|\right|\right|^2$$

where the measure ρ_T is defined by (2.9).

We postpone the proof to Sect. A.3.

3 Learning Theory: Continuous-Time Observations

We analyze first the regularized MLE in the case of continuous-time observations $\{\boldsymbol{X}^{(m)}_{[0,T]}\}_{m=1}^M$. We show that under a coercivity condition, the regularized MLE is consistent, and that with proper choice of the hypothesis spaces, we can achieve an optimal learning rate $(\frac{\log M}{M})^{\frac{2s}{2s+1}}$.

Recall from Eq.(2.1) that

$$\mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi) := \frac{1}{2\sigma^2 TN}\int_0^T \left(\|\boldsymbol{f}_\varphi(\boldsymbol{X}_t)\|^2 - 2\langle \boldsymbol{f}_\varphi(\boldsymbol{X}_t), d\boldsymbol{X}_t\rangle \mathrm{d}t\right)$$

is the negative log-likelihood of a trajectory $\boldsymbol{X}_{[0,T]}$, with respect to the measure induced by the system with interaction kernel φ. Then, the negative log-likelihood of independent trajectories $\{\boldsymbol{X}^{(m)}_{[0,T]}\}_{m=1}^M$ is

$$\mathcal{E}_{T,M}(\varphi) := \frac{1}{M}\sum_{m=1}^M \mathcal{E}_{\boldsymbol{X}^{(m)}_{[0,T]}}(\varphi)\,, \tag{3.1}$$

and the regularized MLE over a hypothesis space $\mathcal{H}$ is

$$\widehat{\phi}_{T,M,\mathcal{H}} \in \arg\min_{\varphi\in\mathcal{H}} \mathcal{E}_{T,M}(\varphi)\,. \tag{3.2}$$

The existence of the minimizer follows from the fact that the error functional $\mathcal{E}_{T,M}(\varphi)$ is quadratic in φ, which in turn is a consequence of the linearity of $\boldsymbol{f}_\varphi$ in φ. The uniqueness of the minimizer, however, requires a coercivity condition and is related to the learnability of the kernel, which we discuss in the next section.

3.1 Identifiability and Learnability: A Coercivity Condition

The uniqueness of the minimizer of the error functional $\mathcal{E}_{T,M}(\varphi)$ over the hypothesis space ensures that the kernel is identifiable. This is not granted, even when the number of observed trajectories is infinite: denote

$$\mathcal{E}_{T,\infty}(\varphi) := \mathbb{E}\mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi) = \lim_{M\to\infty} \mathcal{E}_{T,M}(\varphi) \quad a.s.\,, \tag{3.3}$$

where $\mathbb{E}$ here, and in all that follows unless otherwise indicated, is the expectation over initial conditions, independently sampled from μ_0, and over the Wiener measure underlying the random noise, and observe that

$$\mathcal{E}_{T,\infty}(\varphi) - \mathcal{E}_{T,\infty}(\phi) = \mathbb{E}\Big[\mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi) - \mathcal{E}_{T,\infty}(\phi)\Big] = \frac{1}{2\sigma^2 NT}\mathbb{E}\int_0^T \|\mathbf{f}_{\varphi-\phi}(\boldsymbol{X}_t)\|^2 \mathrm{d}t\,. \tag{3.4}$$

Only when $\mathbb{E}\int_0^T \|\mathbf{f}_{\varphi-\phi}(\boldsymbol{X}_t)\|^2 dt > 0$ for any $\varphi - \phi \neq 0$ can one ensure the uniqueness of minimizer. This motivates us to propose the following coercivity condition, introduced in [9] in the case of non-stochastic systems:

Definition 3.1 *(Coercivity condition)* We say that the stochastic system defined in (1.1) satisfies a coercivity condition on a set $\mathcal{H}$ of functions on $\mathbb{R}_+$, with a constant $0 < c_{\mathcal{H}}$, if

$$c_{\mathcal{H}}|||\varphi|||^2 \leq \frac{1}{2\sigma^2 NT}\int_0^T \sum_{i=1}^N \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i'=1}^N \varphi(r_{ii'}(t))\boldsymbol{r}_{ii'}(t)\Big\|^2\Big]\mathrm{d}t \tag{3.5}$$

for all $\varphi \in \mathcal{H}$ such that $\varphi(\cdot)\cdot \in L^2(\rho_T)$. Here $|||\cdot|||$ denotes the norm defined in (2.11). We will denote by $c_{\mathcal{H}}$ the largest constant for which the inequality holds and call it the coercivity constant.

The coercivity condition ensures identifiability of the kernel. We emphasize that the kernel is latent, in the sense that its values at $\{r_{ii'} = \|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|\}$ are undeterminable from data. In fact, to recover $(\phi(r_{ii'})) \in \mathbb{R}^{\frac{N(N-1)}{2}}$ from the observed trajectories, even if we ignore the stochastic noise in the system and assume to have access to $\mathbf{f}_\phi(\boldsymbol{x}) \in \mathbb{R}^{dN}$, which consists of a linear combination of $(\phi(r_{ii'}))$ with coefficients $\boldsymbol{r}_{ii'} = \boldsymbol{x}_{i'} - \boldsymbol{x}_i$, we face a linear system that is underdetermined as soon as dN (=number of known quantities) $\leq \frac{N(N-1)}{2}$ (=number of unknowns), i.e., for $d < (N-1)/2$. Thus, in general the exact values of ϕ at locations $\{r_{ii'}\}_{i,i'}$ cannot be determined. Furthermore, we have stochastic noise in the system. This suggests that the inverse problem of estimating the interaction kernel in a space of continuous functions is ill-posed. We will see that the coercivity condition ensures well-posedness in $L^2(\rho_T)$, both in the sense of uniqueness and in the sense of stability.

The coercivity condition plays a key role in the learning of the kernel. Beyond ensuring learnability of kernels by ensuring the uniqueness of minimizer over any compact

convex sets, it also enables us to control the error of the estimator by the discrepancy between the expectation of error functionals, as is shown in Proposition 3.1. We will use this property to establish the convergence of the estimators in later sections.

To simplify notation, we define a bilinear functional product over $\mathcal{H}$ by

$$\langle\!\langle \varphi_1, \varphi_2 \rangle\!\rangle := \frac{1}{2\sigma^2 NT} \mathbb{E} \int_0^T \langle \mathbf{f}_{\varphi_1}(\boldsymbol{X}_t), \mathbf{f}_{\varphi_2}(\boldsymbol{X}_t) \rangle \mathrm{d}t, \quad \forall \varphi_1, \varphi_2 \in \mathcal{H}. \tag{3.6}$$

Proposition 3.1 *Let $\mathcal{H}$ be a compact convex subset of $L^2(\mathbb{R}^+, \rho_T)$ and assume the coercivity condition (3.5) holds true on $\mathcal{H}$ with constant $c_{\mathcal{H}}$. Then, the error functional $\mathcal{E}_{T,\infty}$ defined in (3.3) has a unique minimizer over $\mathcal{H}$ in $L^2(\rho_T)$:*

$$\widehat{\phi}_{T,\infty,\mathcal{H}} = \operatorname*{arg\,min}_{\varphi \in \mathcal{H}} \mathcal{E}_{T,\infty}(\varphi)\,. \tag{3.7}$$

Moreover, for all $\varphi \in \mathcal{H}$,

$$\mathcal{E}_{T,\infty}(\varphi) - \mathcal{E}_{T,\infty}(\widehat{\phi}_{T,\infty,\mathcal{H}}) \geq c_{\mathcal{H}} \left|\!\left|\!\left| \varphi - \widehat{\phi}_{T,\infty,\mathcal{H}} \right|\!\right|\!\right|^2. \tag{3.8}$$

Proof From Eq. (3.4), we have $\mathcal{E}_{T,\infty}(\varphi) - \mathcal{E}_{T,\infty}(\phi) = \langle\!\langle \varphi - \phi, \varphi - \phi \rangle\!\rangle$. Then,

$$\begin{aligned}
&\mathcal{E}_{T,\infty}(\varphi) - \mathcal{E}_{T,\infty}(\widehat{\phi}_{T,\infty,\mathcal{H}}) \\
&= \langle\!\langle \varphi - \phi, \varphi - \phi \rangle\!\rangle - \langle\!\langle \widehat{\phi}_{T,\infty,\mathcal{H}} - \phi, \widehat{\phi}_{T,\infty,\mathcal{H}} - \phi \rangle\!\rangle \\
&= \langle\!\langle \varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}, \varphi + \widehat{\phi}_{T,\infty,\mathcal{H}} - 2\phi \rangle\!\rangle \\
&= \langle\!\langle \varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}, \varphi - \widehat{\phi}_{T,\infty,\mathcal{H}} \rangle\!\rangle + 2\langle\!\langle \varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}, \widehat{\phi}_{T,\infty,\mathcal{H}} - \phi \rangle\!\rangle \\
&\geq c_{\mathcal{H}} \left|\!\left|\!\left| \varphi - \widehat{\phi}_{T,\infty,\mathcal{H}} \right|\!\right|\!\right|^2 + 2\langle\!\langle \varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}, \widehat{\phi}_{T,\infty,\mathcal{H}} - \phi \rangle\!\rangle,
\end{aligned}$$

where the inequality follows from the coercivity condition. Then, Eq.(3.8) follows once we notice that $\langle\!\langle \varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}, \widehat{\phi}_{T,\infty,\mathcal{H}} - \phi \rangle\!\rangle \geq 0$ by the convexity of $\mathcal{H}$. In fact, since $t\varphi + (1-t)\widehat{\phi}_{L,\infty,\mathcal{H}} \in \mathcal{H}$, $\forall t \in [0,1]$, we have $\mathcal{E}_{T,\infty}(t\varphi + (1-t)\widehat{\phi}_{T,\infty,\mathcal{H}}) - \mathcal{E}_{T,\infty}(\widehat{\phi}_{T,\infty,\mathcal{H}}) \geq 0$ since $\widehat{\phi}_{T,\infty,\mathcal{H}}$ is a minimizer, and so, equivalently,

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

Sending $t \to 0^+$, we obtain $\langle\!\langle \varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}, 2\widehat{\phi}_{T,\infty,\mathcal{H}} - 2\phi \rangle\!\rangle \geq 0$. □

Well-conditioning from coercivity. When the hypothesis space $\mathcal{H}$ is a finite-dimensional linear space, the coercivity constant provides a lower bound for the smallest eigenvalue of the limit of the normal equations matrix $A_{M,L}$ in Eq.(2.7) as $M, L \to +\infty$. Therefore, when the sample size M is large and when the observation frequency L is high, the matrix $A_{M,L}$ is invertible with a high probability (see Corollary 4.2 for details), and thus, the coercivity condition ensures the uniqueness of the regularized MLE in Eq.(2.8):

Proposition 3.2 *Suppose that the coercivity condition holds on* $\mathcal{H} = \text{span}\{\psi_1, \dots, \psi_n\}$, *where the basis functions satisfy* $\langle \psi_p(\cdot)\cdot, \psi_{p'}(\cdot)\cdot\rangle_{L^2(\rho_T)} = \delta_{p,p'}$. *Let* $A_\infty = \big(\langle\!\langle \psi_p, \psi_{p'} \rangle\!\rangle\big)_{p,p'} \in \mathbb{R}^{n\times n}$ *with the bilinear functional* $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ *defined in (3.6). Then the smallest singular value of* A_∞ *is* $\lambda_{\min}(A_\infty) = c_{\mathcal{H}}$.

Proof For an arbitrary $a \in \mathbb{R}^n$, denoting $\psi = \sum_{p=1}^n a_p \psi_p$, we have

$$a^T A_\infty a = \langle\!\langle \psi, \psi \rangle\!\rangle \geq c_{\mathcal{H}} |||\psi|||^2 = c_{\mathcal{H}} \|a\|^2 \tag{3.9}$$

where the first equality follows from that the functional $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ is bilinear, and the inequality follows from the coercivity condition. Note that by the definition of the coercivity constant in (3.5), we have

$$c_{\mathcal{H}} = \sup_{\psi\in\mathcal{H}} \frac{\langle\!\langle \psi, \psi \rangle\!\rangle}{|||\psi|||^2} = \sup_{\psi\in\mathcal{H}, |||\psi|||=1} \langle\!\langle \psi, \psi \rangle\!\rangle,$$

which is attained at some $\psi_* \in \mathcal{H}$ since $\mathcal{H}$ is finite dimensional. Hence, the inequality in (3.9) becomes inequality for ψ_* and the smallest eigenvalue of A_∞ is $c_{\mathcal{H}}$. □

Proposition 3.2 suggests that for the hypothesis space $\mathcal{H}$, it is important to choose a basis that is orthonormal in $L^2(\rho_T)$, so as to make the matrix in the normal equations as well-conditioned as possible given the dynamics. In practice, the unknown ρ_T is approximated by the empirical density $\rho_T^{L,M}$. Therefore, when using local basis functions, it is natural to use a partition of the support of ρ_T^M.

The coercivity condition and positive integral operators. The coercivity condition introduces constraints on the hypothesis spaces and on the distribution of the solutions of the system, and it is therefore natural that it depends on the distribution μ_0 of the initial condition $\boldsymbol{X}_0$, the true interaction kernel ϕ, and the random noise. We review below briefly the recent developments in [37,38], where the coercivity condition is proved to hold on any compact sets of $L^2(\rho_T)$ for special classes of systems, such as linear systems and nonlinear systems with a stationary distribution. As discussed in [9,40,41] for the deterministic cases, we believe that the coercivity condition is "generally" satisfied for "relevant" hypothesis spaces, with a constant independent of the number of particles N, thanks to the exchangeability of both the distribution of the initial conditions and that of the particles at any time t.

The coercivity condition is ensured by the positiveness of integral operators that arise in the expectation in Eq.(3.5). More precisely, recall that the drift of the SDE is cyclic in the indexes. Thus, the distribution of $\boldsymbol{X}_t$ is exchangeable and

for $i \neq i'$, one has

$$\mathbb{E}\big[\varphi(r_{ii'}^t)\varphi(r_{ii''}^t)\langle \boldsymbol{r}_{ii'}^t, \boldsymbol{r}_{ii''}^t\rangle\big] = \mathbb{E}[\varphi(\|\boldsymbol{r}_{12}^t\|)\varphi(\|\boldsymbol{r}_{13}^t\|)\langle \boldsymbol{r}_{12}^t, \boldsymbol{r}_{13}^t\rangle].$$

One can rewrite Eq.(3.5) as

$$\begin{aligned}
&(c_{\mathcal{H}} - \frac{N-1}{N^2})\|\varphi(\cdot)\cdot\|^2_{L^2(\rho_T)} \\
&\quad \le \frac{(N-1)(N-2)}{N^2 T}\int_0^T \mathbb{E}[\varphi(\|\boldsymbol{r}_{12}^t\|)\varphi(\|\boldsymbol{r}_{13}^t\|)\langle \boldsymbol{r}_{12}^t, \boldsymbol{r}_{13}^t\rangle]\mathrm{d}t \\
&\quad = \int_0^\infty \int_0^\infty \varphi(r)\varphi(s)\overline{K}_T(r,s)\mathrm{d}r\mathrm{d}s,
\end{aligned}$$

where the integral kernel $\overline{K}_T : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ is defined as

$$\overline{K}_T(r,s) := (rs)^d \int_{S^{d-1}}\int_{S^{d-1}} \langle \xi, \eta\rangle \frac{1}{T}\int_0^T p_t(r\xi, s\eta)\mathrm{d}t\mathrm{d}\xi\mathrm{d}\eta, \tag{3.10}$$

with $p_t(u,v)$ denoting the joint density function of the random vector $(\boldsymbol{r}_{12}^t, \boldsymbol{r}_{13}^t)$ and S^{d-1} denoting the unit sphere in $\mathbb{R}^d$. It is shown in [37,38] that the associated integral operator defined by $\overline{K}_T$ is strictly positive definite, and therefore, the coercivity condition holds, for a large class of systems with interaction kernels in form of $\phi(r) = (a + r^\theta)^{\gamma-1} r^{\theta-2}$ with $a \ge 0$ and $\{(\theta, \gamma) \in (0,1] \times (1,2] : \theta\gamma > 1\}$.

3.2 Consistency and Rate of Convergence of the Estimator

In this section, we consider using a family of finite-dimensional linear spaces $\{\mathcal{L}_n : n \in \mathbb{N}^+\} \subset C^{1,1}[0,R]$ as hypothesis spaces and establish the consistency and rate of convergence of our estimators, which are our main theorems for continuous-time observations. We assume the spaces $\{\mathcal{L}_n : n \in \mathbb{N}^+\} \subset C^{1,1}[0,R]$ satisfying Markov–Bernstein-type inequality: there exist $c_1, \gamma > 0$ s.t. for all $\varphi \in \mathcal{L}_n$

$$\|\varphi'\|_\infty \le c_1 \dim(\mathcal{L}_n)^\gamma \|\varphi\|_\infty . \tag{3.11}$$

This condition has a long history and rich literature in classical approximation theory, where it is studied when function spaces satisfy (3.11) (e.g., see the survey paper [53]), which is an important step in establishing inverse approximation theorems. This kind of inequality holds true on many function spaces that are commonly used as approximation spaces in practice, including:

- $\mathcal{L}_n$: trigonometric polynomials of degree n on $[0, 2\pi]$ (similarly on $[0, R]$), for which $\|\phi'\|_\infty \le \frac{1}{2}(\dim(\mathcal{L}_n) - 1)\|\phi\|_\infty$. This result dates back to Bernstein [5].
- $\mathcal{L}_n$: the polynomial space consisting of all polynomials with degree less than $n-1$ on $[0, R]$ (see Theorem 3.3 in [48]), for which $\|\varphi'\|_\infty \le \frac{2}{R}(\dim(\mathcal{L}_n) + 1)^2\|\varphi\|_\infty$. As a result, (3.11) also holds true for polynomial splines; other extensions include rational functions. We refer to the reader to [30] for details.

If we choose a compact convex hypothesis set $\mathcal{H}_M$ contained in some $\mathcal{L}_n$, with a suitable correspondence between n and M, such that the distance between $\mathcal{H}_M$ and the true kernel ϕ vanishes as M increases, the following consistency result holds:

Theorem 3.1 (Strong consistency of estimators) *Suppose $\phi \in \mathcal{K}_{R,S}$, the admissible set defined in (1.7). Let $\{\mathcal{L}_n : n \in \mathbb{N}^+\} \subset C^{1,1}[0, R]$ satisfying (3.11) and*

$$\inf_{\varphi \in \mathcal{L}_n} \|\varphi - \phi\|_\infty \xrightarrow{n\to\infty} 0.$$

Let $S_0 \geq S$ and $\mathcal{H}_M = \mathcal{B}^\infty_{2S_0}(\mathcal{L}_{n_M}) = \{\varphi \in \mathcal{L}_{n_M} : ||\varphi||_\infty < 2S_0\}$ with $\dim(\mathcal{L}_{n_M}) = n_M$ and $\lim_{M\to\infty} \frac{n_M \log n_M}{M} = 0$. Finally, suppose the coercivity condition holds true on $\cup_n \mathcal{L}_n$. Then, we have

$$\lim_{M\to\infty} \left|\left|\left| \widehat{\phi}_{T,M,\mathcal{H}_M} - \phi \right|\right|\right| = 0 \quad \textit{with probability one.}$$

If we know the explicit approximation rate of the family $\{\mathcal{L}_n : n \in \mathbb{N}^+\}$, then by carefully choosing the dimension of hypothesis spaces as a function of M, we can obtain a near-optimal rate of convergence of our estimators.

Theorem 3.2 (Convergence rate of estimators) *Suppose $\phi \in \mathcal{K}_{R,S}$, the admissible set defined in (1.7). Assume that there exits a sequence of linear spaces $\{\mathcal{L}_n : n \in \mathbb{N}^+\} \subset C^{1,1}[0, R]$ satisfying (3.11) with the properties*

(i) $\dim(\mathcal{L}_n) \leq c_0 n$ *for* $n \in \mathbb{N}^+$,
(ii) $\inf_{\varphi\in\mathcal{L}_n} \|\varphi - \phi\|_\infty \leq c_2 n^{-s}$.

For example, when $\phi \in C^{k,\alpha}$ with $s = k + \alpha \geq 2$, we may choose $\mathcal{L}_n$ to consist of polynomial splines of degree $\lfloor s - 1 \rfloor$ with uniform knots on $[0, R]$. Let $\mathcal{H}_n = \mathcal{B}^\infty_{S_0}(\mathcal{L}_n)$ with $S_0 = c_2 + S$ and $n \asymp (M/\log M)^{1/(2s+1)}$, and assume that the coercivity condition holds on $\mathcal{L} := \cup_n \mathcal{L}_n$ with a constant $c_{\mathcal{L}} > 0$. Then, we have

$$\mathbb{E}\left[\left|\left|\left| \widehat{\phi}_{T,M,\mathcal{H}_n} - \phi \right|\right|\right|^2 \right] \leq \frac{C}{c^2_{\mathcal{L}}} \left(\frac{\log M}{M} \right)^{\frac{2s}{2s+1}},$$

where C is a constant depending only on σ, N, T, R, S_0.

It is fruitful to compare (up to log terms) the rate $2s/(2s + 1)$ to that for nonparametric 1-dimensional regression, where one can observe directly noisy values of the target function ϕ at sample points drawn i.i.d from ρ_T. For the function space $C^{k,\alpha}$, this rate is min–max optimal. Our numerical examples in Sect. 5 empirically validate the desired convergence rate for $s = 1, 2$ where we use piecewise constant and linear polynomials. Note that in our setting, learning ϕ is an inverse problem, as we do not directly observe the values $\{\phi(\|\boldsymbol{x}^{(m)}_{i',t} - \boldsymbol{x}^{(m)}_{i,t}\|)\}^{N,N,M}_{i,i'=1,\, m=1}$. While our result is stated in such a way that a knowledge of s is required, in fact an upper bound $\tilde{s}$ is sufficient, as choosing sufficiently regular splines, of degree $\lfloor \tilde{s} - 1 \rfloor$ would give the optimal s-dependent rate, at the cost of possibly larger constants. We also remark that we do not require that the underlying stochastic process satisfies certain mixing properties nor starts from a stationary distribution. Obtaining this optimal convergence rate in M for short-time trajectory observations is therefore satisfactory. For long trajectories

and under ergodicity assumptions, rates in terms of MT are likely to be obtainable: In Sect. 5, we present numerical evidence that suggests that the error does decrease with MT at a near-optimal rate.

3.3 Proof of the Main Theorems

In the following part, we present the proof for Theorem 3.2, which also yields the proof for Theorem 3.1. The main techniques include the Itô formula, concentration inequalities of unbounded random variables, and a generalization of a novel covering argument in [52] that enables us to deal efficiently with the fluctuations in the data due to the stochastic noise in the dynamics of the system.

One major obstacle in the non-asymptotic analysis of our regularized MLE estimators is the unboundedness of stochastic integral of the form $\frac{1}{T}\int_0^T \langle \mathbf{f}_\varphi(\boldsymbol{X}_t), d\boldsymbol{X}_t\rangle \mathrm{d}t$ appearing in the empirical error functional. Unlike the deterministic case $\sigma = 0$, our empirical error functional $\mathcal{E}_{T,M}(\cdot)$ is in general not continuous over $\mathcal{H}$ with respect to the $\|\cdot\|_\infty$ norm. In the following, we first leverage the general Itô formula described in Theorem A.3, to obtain a form of the empirical error functional that does not involve a stochastic integral and is amenable to analysis; we then show that it is continuous on $C^{1,1}([0, R])$ with respect to the $\|\cdot\|_{1,1}$ norm. Therefore, in the following preliminary results for the proofs of the main theorems, we consider the following generic hypothesis space:

Assumption 3.3 The hypothesis space $\mathcal{H}$ is a compact convex subset of $C^{1,1}([0, R])$ with respect to the uniform norm $\|\cdot\|_\infty$ and bounded above by $S_0 \geq S$.

Lemma 3.1 *Suppose $\varphi \in \mathcal{K}_{R,S}$, the admissible set defined in (1.7). Let*

$$V_\varphi(\boldsymbol{X}_t) = \frac{1}{2N}\sum_{i,i'} \Psi(\|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i',t}\|) \ \textit{with} \ \Psi'(r) = \varphi(r)r\,; \tag{3.12}$$

then, we have, almost surely

$$-(\mathrm{d}V_\varphi)(\boldsymbol{X}_t) = \langle \mathbf{f}_\varphi(\boldsymbol{X}_t), \mathrm{d}\boldsymbol{X}_t\rangle + \frac{\sigma^2}{2N}\sum_{i=1}^N \sum_{i'\neq i} \left(\varphi'(\|\boldsymbol{x}_{ii'}\|)\|\boldsymbol{x}_{ii'}\| + \varphi(\|\boldsymbol{x}_{ii'}\|)d\right)\mathrm{d}t$$

Proof Let $g(\boldsymbol{X}) = V_\varphi(\boldsymbol{X})$. Note that g is C^2, with derivatives

$$\frac{\partial g(\boldsymbol{X})}{\partial \boldsymbol{x}_i} = \frac{\partial V_\varphi(\boldsymbol{X})}{\partial \boldsymbol{x}_i} = \frac{1}{N}\sum_{i'\neq i} \varphi(\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|)(\boldsymbol{x}_i - \boldsymbol{x}_{i'})$$

$$\frac{\partial^2 g(\boldsymbol{X})}{\partial \boldsymbol{x}_i \partial \boldsymbol{x}_k} = -\delta_{ki}\frac{1}{N} \left(\sum_{i'\neq i} \varphi(\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|)\mathbf{I}_d + \frac{\varphi'(\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|)}{\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|}(\boldsymbol{x}_i - \boldsymbol{x}_{i'}) \otimes (\boldsymbol{x}_i - \boldsymbol{x}_{i'})\right)$$

$$- \delta_{k\neq i}\frac{1}{N} \left(\varphi(\|\boldsymbol{x}_i-\boldsymbol{x}_k\|)\mathbf{I}_d + \frac{\varphi'(\|\boldsymbol{x}_i-\boldsymbol{x}_k\|)}{\|\boldsymbol{x}_i-\boldsymbol{x}_k\|}(\boldsymbol{x}_i-\boldsymbol{x}_k)\otimes(\boldsymbol{x}_i-\boldsymbol{x}_k)\right).$$

Using Itô's formula (Theorem A.3) for the Itô process $g(\boldsymbol{X}_t)$, the conclusion follows. □

Proposition 3.3 *Suppose $\varphi_1, \varphi_2 \in \mathcal{H}$, then it holds almost surely that*

$$|\mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi_1) - \mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi_2)| \leq C_1\|\varphi_1-\varphi_2\|_\infty + C_2\|\varphi_1'-\varphi_2'\|_\infty,$$

where $C_1 = \frac{R^2S_0}{\sigma^2} + \frac{R^2}{2\sigma^2T} + \frac{d}{2}$ and $C_2 = \frac{R}{2}$.

Proof Note that

$$\mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi_1) - \mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi_2) = \frac{1}{2\sigma^2NT}\int_0^T \underbrace{\langle \mathbf{f}_{\varphi_1-\varphi_2}(\boldsymbol{X}_t), \mathbf{f}_{\varphi_1+\varphi_2}(\boldsymbol{X}_t)\rangle}_{I_1}\mathrm{d}t - 2\underbrace{\langle \mathbf{f}_{\varphi_1-\varphi_2}(\boldsymbol{X}_t), \mathrm{d}\boldsymbol{X}_t\rangle}_{I_2}.$$

I_1 satisfies

$$|I_1| \leq \|\mathbf{f}_{\varphi_1+\varphi_2}(\boldsymbol{X}_t)\|\|\mathbf{f}_{\varphi_1-\varphi_2}(\boldsymbol{X}_t)\| \leq N\|\varphi_1-\varphi_2\|_\infty\|\varphi_1+\varphi_2\|_\infty R^2,$$

since $\|\mathbf{f}_\varphi(\boldsymbol{X}_t)\| \leq \sqrt{N}R\|\varphi\|_\infty$. For I_2, Lemma 3.1 yields

$$\begin{aligned}|I_2| &\leq |V_{\varphi_1-\varphi_2}(\boldsymbol{X}_0) - V_{\varphi_1-\varphi_2}(\boldsymbol{X}_t)| \\ &\quad + \frac{\sigma^2}{2N}\left|\int_0^T \sum_{i=1}^N\sum_{i'\neq i}((\varphi_1-\varphi_2)'(r_{ii'})r_{ii'} + (\varphi_1-\varphi_2)(r_{ii'})d)\mathrm{d}t\right| \\ &\leq \frac{(N-1)}{2}\|\varphi_1-\varphi_2\|_\infty R^2 + \frac{(N-1)\sigma^2T}{2}(\|\varphi_1'-\varphi_2'\|_\infty R + \|\varphi_1-\varphi_2\|_\infty d),\end{aligned}$$

where we used

$$|V_\varphi(\boldsymbol{X}_t) - V_\varphi(\boldsymbol{X}_0)| \leq \frac{1}{2N}\sum_{ii'}\left|\int_{r_{ii',0}}^{r_{ii',T}}\varphi(r)r\mathrm{d}r\right| \leq \frac{(N-1)\|\varphi\|_\infty R^2}{2},$$

which follows from its definition in (3.12). Combining the estimates for I_1 and I_2, and using $\|\varphi_1+\varphi_2\|_\infty \leq 2S_0$, we obtain

$$\begin{aligned}&|\mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi_1) - \mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi_2)| \\ &\quad\leq \frac{R^2}{2\sigma^2}\|\varphi_1-\varphi_2\|_\infty\left(\|\varphi_1+\varphi_2\|_\infty + \frac{1}{T}\right)\end{aligned}$$

$$+\frac{1}{2}(d\|\varphi_1-\varphi_2\|_\infty+\|\varphi_1'-\varphi_2'\|_\infty R)$$
$$\le C_1\|\varphi_1-\varphi_2\|_\infty+C_2\|\varphi_1'-\varphi_2'\|_\infty, \tag{3.13}$$

where $C_1=\frac{R^2S_0}{\sigma^2}+\frac{R^2}{2\sigma^2T}+\frac{d}{2}$ and $C_2=\frac{R}{2}$. □

When $M=\infty$, i.e., we observe infinitely many trajectories, the expectation of our error functional $\mathcal{E}_{T,\infty}$, as in (3.3), does not involve the stochastic integral term. From the proof of Proposition 3.3, we see that it is continuous over $\mathcal{H}$ with respect to the $\|\cdot\|_\infty$ norm:

Corollary 3.1 *Suppose $\varphi_1,\varphi_2\in\mathcal{H}$, then, with $C_1=\frac{2R^2S_0}{\sigma^2}$, we have*

$$|\mathcal{E}_{T,\infty}(\varphi_1)-\mathcal{E}_{T,\infty}(\varphi_2)|\le C_1\|\varphi_1-\varphi_2\|_\infty.$$

Proof Using (3.4), we obtain that

$$\begin{aligned}|\mathcal{E}_{T,\infty}(\varphi_1)-\mathcal{E}_{T,\infty}(\varphi_2)| &= \frac{1}{2\sigma^2NT}\mathbb{E}\int_0^T\left|\langle\mathbf{f}_{\varphi_1-\varphi_2}(\boldsymbol{X}_t),\mathbf{f}_{\varphi_1+\varphi_2-2\phi}(\boldsymbol{X}_t)\rangle\right|\mathrm{d}t\\ &\le \frac{R^2}{2\sigma^2}\|\varphi_1-\varphi_2\|_\infty\|\varphi_1+\varphi_2-2\phi\|_\infty\\ &\le \frac{2R^2S_0}{\sigma^2}\|\varphi_1-\varphi_2\|_\infty\end{aligned}$$

□

Recall the definition

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

We now analyze the discrepancy between the empirical minimizer $\widehat{\phi}_{T,M,\mathcal{H}}$ and $\widehat{\phi}_{T,\infty,\mathcal{H}}$, which we called sampling error (SE) in the diagram in Fig. 1. We introduce a measurable function on the path space by

$$D_\varphi:=\mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\varphi)-\mathcal{E}_{\boldsymbol{X}_{[0,T]}}(\widehat{\phi}_{T,\infty,\mathcal{H}}) \tag{3.14}$$

for any $\varphi\in\mathcal{H}$. From Proposition 3.1, we have

$$\mathbb{E}D_\varphi=\mathcal{E}_{T,\infty}(\varphi)-\mathcal{E}_{T,\infty}(\widehat{\phi}_{T,\infty,\mathcal{H}})\ge c_{\mathcal{H}}\left|\!\left|\!\left|\varphi-\widehat{\phi}_{T,\infty,\mathcal{H}}\right|\!\right|\!\right|, \tag{3.15}$$

so D_φ in fact bounds (in expectation) the distance between φ and $\widehat{\phi}_{T,\infty,\mathcal{H}}$ w.r.t. the $|||\cdot|||$-norm. We now perform a non-asymptotic analysis of D_φ. We shall show that the random variable D_φ satisfies moment conditions, sufficient to guarantee strong concentration about its expectation (Proposition 3.4). To do this, we decompose D_φ as the sum of a bounded component only involving time integrals and an unbounded component involving stochastic integrals:

$$D_\varphi:=D_\varphi^{\mathrm{bd}}-D_\varphi^{\mathrm{ubd}}$$

$$D_\varphi^{\mathrm{bd}} := \frac{1}{2\sigma^2 NT}\int_0^T \langle \mathbf{f}_{\varphi-\widehat{\phi}_{T,\infty,\mathcal{H}}}(\boldsymbol{X}_t), \mathbf{f}_{\varphi+\widehat{\phi}_{T,\infty,\mathcal{H}}-2\phi}(\boldsymbol{X}_t)\rangle \mathrm{d}t$$
$$D_\varphi^{\mathrm{ubd}} := \frac{1}{\sigma^2 NT}\int_0^T \langle \mathbf{f}_{\varphi-\widehat{\phi}_{T,\infty,\mathcal{H}}}(\boldsymbol{X}_t), \mathrm{d}\boldsymbol{B}(t)\rangle$$

We prove moment conditions independently for each of these components in the next two Lemmata.

Lemma 3.2 (Bounds on D_φ^{ubd}) *For $\varphi \in \mathcal{H}$ and $p = 2, 3, 4, \ldots,$*

$$\mathbb{E}\big|D_\varphi^{\mathrm{ubd}}\big|^p \le C\big(\big\|\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}\big\|_\infty\big)^{p-2}\big|\!\big|\!\big|\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}\big|\!\big|\!\big|^2$$

where $C = (\frac{p(p-1)}{2})^{\frac{p}{2}}\frac{R^{p-2}}{\sigma^{2p}(NT)^{\frac{p+2}{2}}}$.

Proof First of all, note that

$$\big\|\mathbf{f}_{\varphi-\widehat{\phi}_{T,\infty,\mathcal{H}}}(\boldsymbol{X}_t)\big\| \le \sqrt{N}R\big\|\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}\big\|_\infty. \tag{3.16}$$

Therefore $\mathbf{f}_{\varphi-\widehat{\phi}_{T,\infty,\mathcal{H}}}(\boldsymbol{X}_t)$ is a L^2-integrable process. Applying Theorem A.5, we obtain

$$\begin{aligned}
&\mathbb{E}\left|\int_0^T \langle \mathbf{f}_{\varphi-\widehat{\phi}_{T,\infty,\mathcal{H}}}(\boldsymbol{X}_t), \mathrm{d}\boldsymbol{B}(t)\rangle\right|^p \\
&\le C_{p,T}\mathbb{E}\int_0^T \big\|\mathbf{f}_{\varphi-\widehat{\phi}_{T,\infty,\mathcal{H}}}(\boldsymbol{X}_t)\big\|^p \mathrm{d}t \\
&\le C_{p,T}\big(\sqrt{N}R\|\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}\|_\infty\big)^{p-2}\mathbb{E}\int_0^T \big\|\mathbf{f}_{\varphi-\widehat{\phi}_{T,\infty,\mathcal{H}}}(\boldsymbol{X}_t)\big\|^2 \mathrm{d}t \\
&\le C_{p,T}\big(\sqrt{N}R\|\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}\|_\infty\big)^{p-2} NT\big|\!\big|\!\big|\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}\big|\!\big|\!\big|^2,
\end{aligned}$$

with $C_{p,T} = \big(\frac{p(p-1)}{2}\big)^{\frac{p}{2}} T^{\frac{p-2}{2}}$. The conclusion then follows by adding in the scaling factor $\frac{1}{\sigma^2 NT}$. □

Lemma 3.3 (Bounds on D_φ^{bd}) *For $\varphi \in \mathcal{H}$ and $p = 2, 3, 4, \ldots,$*

$$\mathbb{E}\big|D_\varphi^{\mathrm{bd}}\big|^p \le \left(\frac{2S_0}{\sigma^2}\right)^p R^{2p-2}\big\|\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}\big\|_\infty^{p-2}\big|\!\big|\!\big|\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}\big|\!\big|\!\big|^2.$$

Proof From inequality (3.16) and the linear dependence of $\mathbf{f}_\varphi$ on φ, we have

$$\begin{aligned}
\big|D_\varphi^{\mathrm{bd}}\big| &\le \frac{2S_0 R}{\sigma^2\sqrt{N}T}\int_0^T \big\|\mathbf{f}_{\varphi-\widehat{\phi}_{T,\infty,\mathcal{H}}}(\boldsymbol{X}_t)\big\| \mathrm{d}t \\
&\le \frac{2S_0 R}{\sigma^2}\sqrt{\frac{1}{NT}\int_0^T \big\|\mathbf{f}_{\varphi-\widehat{\phi}_{T,\infty,\mathcal{H}}}(\boldsymbol{X}_t)\big\|^2 \mathrm{d}t}.
\end{aligned}$$

Therefore,

$$\mathbb{E}|D_\varphi^{\mathrm{bd}}|^p \le \left(\frac{2S_0}{\sigma^2}\right)^p R^{2p-2}\|\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}\|_\infty^{p-2}|||\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}|||^2.$$

□

Now we combine Lemmas 3.2 and 3.3 to prove the moment condition for D_φ.

Lemma 3.4 (Moment conditions) *For $\varphi \in \mathcal{H}$, and every $p = 2, 3, \ldots$, we have*

$$\mathbb{E}\left[\left|D_\varphi - \mathbb{E}D_\varphi\right|^p\right] \le \frac{1}{2}p!K_{\varphi,\mathcal{H}}^{p-2}C_{\varphi,\mathcal{H}},$$

where

$$K_{\varphi,\mathcal{H}} := C_0\|\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}\|_\infty, \quad C_{\varphi,\mathcal{H}} := C_0^2C_1|||\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}|||^2 \tag{3.17}$$

with $C_0 = \sqrt{\frac{2e^2R^2}{\sigma^4NT}}$, $C_1 = \max_{p\ge2}\frac{1}{\sqrt{2\pi p}NTR^2}\left(1 + \frac{c_{\sigma,S_0,R}}{C_0^p}\right)$, *and* $c_{\sigma,S_0,R} = \max_{p\ge2}(\frac{8S_0}{\sigma^2e})^p\frac{R^{2p-2}}{\sqrt{2\pi}p^{p+\frac{1}{2}}}$.

Proof The proof is based on the Jensen's inequality, Lemma 3.2 and Lemma 3.3.

$$\begin{aligned}
\mathbb{E}\left|D_\varphi - \mathbb{E}D_\varphi\right|^p &\le 2^{p-1}\mathbb{E}\left|D_\varphi^{\mathrm{bd}} - \mathbb{E}D_\varphi^{\mathrm{bd}}\right|^p + 2^{p-1}\mathbb{E}\left|D_\varphi^{\mathrm{ubd}}\right|^p\\
&\le 2^{2p-1}\mathbb{E}\left|D_\varphi^{\mathrm{bd}}\right|^p + 2^{p-1}\mathbb{E}[\left|D_\varphi^{\mathrm{ubd}}\right|^p\\
&\le \frac{1}{2}C_0\|\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}\|_\infty^{p-2}|||\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}|||^2\\
&\le \frac{1}{2}p!(C_1\|\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}\|_\infty)^{p-2}C_1^2C_2|||\varphi - \widehat{\phi}_{T,\infty,\mathcal{H}}|||^2,
\end{aligned}$$

where the constants are

$$\begin{aligned}
C_0 &:= \left(\frac{8S_0}{\sigma^2}\right)^p R^{2p-2} + (2p(p-1))^{\frac{p}{2}}\frac{R^{p-2}}{\sigma^{2p}(NT)^{\frac{p+2}{2}}},\\
C_1 &:= \sqrt{\frac{2e^2R^2}{\sigma^4NT}}, \qquad C_2 := \max_{p\ge2}\left(\frac{1}{\sqrt{2\pi p}NTR^2} + \frac{c_{\sigma,S_0,R}}{C_1^p}\right),\\
c_{\sigma,S_0,R} &:= \max_{p\ge2}\left(\frac{8S_0}{\sigma^2e}\right)^p\frac{R^{2p-2}}{\sqrt{2\pi}p^{p+\frac{1}{2}}},
\end{aligned}$$

and the last inequality is derived from the Stirling's lemma. □

We now tie the discrepancy functionals for finite and infinite M:

Proposition 3.4 (Sampling Error bound) *Let $0 < \delta < 1$ and $\{\phi_j\}_{j=1}^{\mathcal{N}}$ be an η-net of functions in a compact convex hypothesis space $\mathcal{H} \subset Ball_{S_0}(L^\infty[0, R])$, that is for any function $\varphi \in \mathcal{H}$, there exists some j such that $\|\varphi - \phi_j\|_\infty \leq \eta$. Denote*

$$\mathcal{D}_{\varphi_j,\infty} := \mathbb{E}D_{\varphi_j} = \mathcal{E}_{T,\infty}(\varphi_j) - \mathcal{E}_{T,\infty}(\widehat{\phi}_{T,\infty,\mathcal{H}})$$
$$\mathcal{D}_{\varphi_j,M} := \mathcal{E}_{T,M}(\varphi_j) - \mathcal{E}_{T,M}(\widehat{\phi}_{T,\infty,\mathcal{H}}).$$

Then with probability at least $1 - \frac{\delta}{2}$, we have

$$\mathcal{D}_{\varphi_j,\infty} - \mathcal{D}_{\varphi_j,M} \leq \epsilon_{M,\delta,\mathcal{N}} + \frac{1}{2}\mathcal{D}_{\varphi_j,\infty} \tag{3.18}$$

for all j, where $\epsilon_{M,\delta,\mathcal{N}} = \frac{C}{M}\log(\frac{2\mathcal{N}}{\delta})$, $C = 2\frac{C_0^2 C_1}{c_{\mathcal{H}}} + 4C_0S_0$, and C_0, C_1 as in (3.17), with $c_{\mathcal{H}}$ the coercivity constant defined in (3.5).

Proof For each $\varphi_j \in \mathcal{H}$, recall that in Eq.(3.15), the coercivity condition on $\mathcal{H}$ implies that

$$\mathcal{D}_{\varphi_j,\infty} \geq c_{\mathcal{H}}\left|\left|\left|\varphi_j - \widehat{\phi}_{T,\infty,\mathcal{H}}\right|\right|\right|^2.$$

Then, Eq.(3.17) in Lemma 3.4 yields that

$$\mathbb{E}|D_{\varphi_j} - \mathbb{E}D_{\varphi_j}|^p \leq \frac{1}{2}p!K_{\varphi_j,\mathcal{H}}^{p-2}\frac{C_0^2C_1}{c_{\mathcal{H}}}\mathcal{D}_{\varphi_j,\infty}. \tag{3.19}$$

Therefore, the random variable D_{φ_j} satisfies the moment condition in Corollary A.2, and so $\forall\epsilon > 0$

$$\mathbb{P}\left\{\mathcal{D}_{\varphi_j,\infty} - \mathcal{D}_{\varphi_j,M} \geq \sqrt{\epsilon(\epsilon + \mathcal{D}_{\varphi_j,\infty})}\right\} \leq \exp\left(\frac{-M\epsilon}{\frac{2C_0^2C_1}{c_{\mathcal{H}}} + 2K_{\varphi_j,\mathcal{H}}}\right).$$

We have $K_{\varphi_j,\mathcal{H}} \leq 2C_0S_0$ where C_0 is defined in (3.17). Taking a union bound on all these events, over $j \in \{1, 2, \ldots, \mathcal{N}\}$, we obtain that

$$\mathbb{P}\left\{\max_{1\leq j\leq\mathcal{N}}\frac{\mathcal{D}_{\varphi_j,\infty} - \mathcal{D}_{\varphi_j,M}}{\sqrt{\mathcal{D}_{\varphi_j,\infty} + \epsilon}} \geq \sqrt{\epsilon}\right\} \leq \mathcal{N}\exp\left(\frac{-M\epsilon}{\frac{2C_0^2C_1}{c_{\mathcal{H}}} + 4C_0S_0}\right). \tag{3.20}$$

Setting the right-hand side to be $\frac{\delta}{2}$, we get $\epsilon_{M,\delta,\mathcal{N}} = \frac{C}{M}\log(\frac{2\mathcal{N}}{\delta})$, where $C := 2\frac{C_0^2C_1}{c_{\mathcal{H}}} + 4C_0S_0$. Using the inequality $\sqrt{\epsilon_{M,\delta,\mathcal{N}}(\epsilon_{M,\delta,\mathcal{N}} + D_{\varphi_j,\infty})} \leq \epsilon_{M,\delta,\mathcal{N}} + \frac{1}{2}D_{\varphi_j,\infty}$, we conclude that with probability at least $1 - \frac{\delta}{2}$

$$\mathcal{D}_{\varphi_j,\infty} - \mathcal{D}_{\varphi_j,M} \leq \epsilon_{M,\delta,\mathcal{N}} + \frac{1}{2}\mathcal{D}_{\varphi_j,\infty}.$$

□

Proof of Theorem 3.2 For $\mathcal{H}_n = \mathcal{B}_{S_0}^{\infty}(\mathcal{L}_n)$, let $\{\varphi_j : j = 1, \ldots, \mathcal{N}\}$ be an η-net of $\mathcal{H}_n$. Let

$$\widehat{\phi}_{T,M,\mathcal{H}_n} = \operatorname{argmin}_{\varphi\in\mathcal{H}_n} \mathcal{E}_{T,M}(\varphi).$$

Then there exists φ_{j_M} in the net such that $\|\varphi_{j_M} - \widehat{\phi}_{T,M,\mathcal{H}_n}\|_\infty \le \eta$; by Corollary 3.1

$$\left|\mathcal{D}_{\varphi_{j_M},\infty} - \mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_n},\infty}\right| = \left|\mathcal{E}_{T,\infty}(\varphi_{j_M}) - \mathcal{E}_{T,\infty}(\widehat{\phi}_{T,M,\mathcal{H}_n})\right| \le \eta\frac{2S_0R^2}{\sigma^2}. \tag{3.21}$$

On the other hand, since $\mathcal{H}_n \subset \mathcal{L}_n \subset C^{1,1}([0, R])$, thanks to the almost sure bound in Proposition 3.3 and the uniformly bound $\sup_{\varphi\in\mathcal{L}_n} \frac{\|\varphi'\|_\infty}{\|\varphi\|_\infty} \le c_1(\dim(\mathcal{L}_n))^\gamma$ from assumption (3.11), we have, almost surely,

$$\begin{aligned}\left|\mathcal{D}_{\varphi_{j_M},M} - \mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_n},M}\right| &= \left|\mathcal{E}_{T,M}(\varphi_{j_M}) - \mathcal{E}_{T,M}(\widehat{\phi}_{T,M,\mathcal{H}_n})\right| \\ &\le \eta(C_1 + c_1C_2\dim(\mathcal{L}_n)^\gamma),\end{aligned} \tag{3.22}$$

where $C_1 = \frac{R^2S_0}{\sigma^2} + \frac{R^2}{2\sigma^2T} + \frac{d}{2}$, $C_2 = \frac{R}{2}$.

By Lemma 3.4, for each $\eta > 0$, with probability at least $1 - \frac{\delta}{2}$, (3.18) holds for this η-net $\{\varphi_j : j = 1, \ldots, \mathcal{N}\}$. Combining (3.18) with (3.21) and (3.22), we conclude that, with probability at least $1 - \frac{\delta}{2}$,

$$\begin{aligned}&\mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_n},\infty} - \mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_n},M} \\ &\le |\mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_n},\infty} - \mathcal{D}_{\varphi_{j_M},\infty}| + |\mathcal{D}_{\varphi_{j_M},\infty} - \mathcal{D}_{\phi_{j_M},M}| \\ &\quad + |\mathcal{D}_{\phi_{j_M},M} - \mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_n},M}| \\ &\le \eta(C_0 + C_1 + c_1C_2\dim(\mathcal{L}_n)^\gamma) + \mathcal{D}_{\phi_{j_M},\infty} - \mathcal{D}_{\phi_{j_M},M} \\ &\le \eta(C_0 + C_1 + c_1C_2\dim(\mathcal{L}_n)^\gamma) + \epsilon_{M,\delta,\mathcal{N}} + \frac{1}{2}\mathcal{D}_{\phi_{j_M},\infty} \\ &\le \eta(\frac{3C_0}{2} + C_1 + c_1C_2\dim(\mathcal{L}_n)^\gamma) + \epsilon_{M,\delta,\mathcal{N}} + \frac{1}{2}\mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_n},\infty}.\end{aligned}$$

Notice that $\mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_n},M} \le 0$, so the above inequality implies that

$$\mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_n},\infty} \le \eta(3C_0 + 2C_1 + 2c_1C_2\dim(\mathcal{L}_n)^\gamma) + 2\epsilon_{M,\delta,\mathcal{N}}. \tag{3.23}$$

Let $\mathcal{S}$ be a metric space and $\eta > 0$. We define the covering number $\mathcal{N}(\mathcal{S}, \eta)$ to be the minimal number of disks in $\mathcal{S}$ with radius η covering $\mathcal{S}$. The covering number of $\mathcal{H}_n$ satisfies $\mathcal{N}(\mathcal{H}_n, \eta) \le \left(\frac{4S_0}{\eta}\right)^{c_0n}$ (e.g., Proposition 5 in [20]). By the triangle inequality, we split the error we want to control into sampling error (SE) and approximation error (AE) (see Fig. 1)

$$\left|\left|\left|\widehat{\phi}_{T,M,\mathcal{H}_n} - \phi\right|\right|\right|^2 \le 2\left|\left|\left|\widehat{\phi}_{T,M,\mathcal{H}_n} - \widehat{\phi}_{T,\infty,\mathcal{H}_n}\right|\right|\right|^2 + 2\left|\left|\left|\widehat{\phi}_{T,\infty,\mathcal{H}_n} - \phi\right|\right|\right|^2. \tag{3.24}$$

From (3.23) and the coercivity condition (3.15), we obtain that, with probability at least $1-\frac{\delta}{2}$,

$$\begin{aligned}\left|\!\left|\!\left|\widehat{\phi}_{T,M,\mathcal{H}_n}-\widehat{\phi}_{T,\infty,\mathcal{H}_n}\right|\!\right|\!\right|^2 &\le \frac{1}{c_{\mathcal{H}_n}}\mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_n},\infty}\\ &\le \frac{\eta}{c_{\mathcal{H}_n}}(3C_0+2C_1+2c_1C_2\dim(\mathcal{L}_n)^{\gamma})\\ &\quad+\frac{2}{c_{\mathcal{H}_n}}\epsilon_{M,\delta,\mathcal{N}}. \end{aligned} \tag{3.25}$$

Let $\phi_{\mathcal{H}_n}:=\operatorname{argmin}_{\psi\in\mathcal{H}_n}\|\psi-\phi\|_\infty$. By coercivity condition (3.15), we have

$$\begin{aligned}\left|\!\left|\!\left|\widehat{\phi}_{T,\infty,\mathcal{H}_n}-\phi_{\mathcal{H}_n}\right|\!\right|\!\right|^2 &\le \frac{1}{c_{\mathcal{H}_n}}(\mathcal{E}_{T,\infty}(\phi_{\mathcal{H}_n})-\mathcal{E}_{T,\infty}(\widehat{\phi}_{T,\infty,\mathcal{H}_n}))\\ &\le \frac{1}{c_{\mathcal{H}_n}}(\mathcal{E}_{T,\infty}(\phi_{\mathcal{H}_n})-\mathcal{E}_{T,\infty}(\phi)) \le \frac{1}{c_{\mathcal{H}_n}}\left|\!\left|\!\left|\phi_{\mathcal{H}_n}-\phi\right|\!\right|\!\right|^2,\end{aligned}$$

where we used

$$\mathcal{E}_{T,\infty}(\phi)\le\mathcal{E}_{T,\infty}(\widehat{\phi}_{T,\infty,\mathcal{H}_n}),\quad \left|\mathcal{E}_{T,\infty}(\phi)-\mathcal{E}_{T,\infty}(\varphi)\right| \le \left|\!\left|\!\left|\phi-\varphi\right|\!\right|\!\right|^2,\quad \forall\varphi\in\mathcal{H}_n.$$

Therefore, we have

$$\left|\!\left|\!\left|\widehat{\phi}_{T,\infty,\mathcal{H}_n}-\phi\right|\!\right|\!\right|^2 \le (2+\frac{2}{c_{\mathcal{H}_n}})\left|\!\left|\!\left|\phi_{\mathcal{H}_n}-\phi\right|\!\right|\!\right|^2 \le \frac{4R^2}{c_{\mathcal{H}_n}}n^{-2s}. \tag{3.26}$$

Now we combine the estimates (3.24), (3.25) and (3.26) together, and let $\eta=n^{-2s-\gamma}$ and $n=(\frac{M}{\log M})^{\frac{1}{2s+1}}$, and note that $c_{\mathcal{L}}=c_{\cup_n\mathcal{L}_n}=c_{\cup_n\mathcal{H}_n}\le c_{\mathcal{H}_n}=c_{\mathcal{L}_n}\le 1$ for all n. We obtain that, with probability at least $1-\frac{\delta}{2}$, the following estimate holds true:

$$\begin{aligned}\left|\!\left|\!\left|\widehat{\phi}_{T,M,\mathcal{H}_n}-\phi\right|\!\right|\!\right|^2 &\le \frac{2\eta}{c_{\mathcal{L}}}(3C_0+2C_1+2c_1C_2\dim(\mathcal{L}_n)^{\gamma})+\frac{8R^2}{c_{\mathcal{L}}}n^{-2s}+\frac{4}{c_{\mathcal{L}}}\epsilon_{M,\delta,N}\\ &\le \frac{C_3}{c_{\mathcal{L}}}n^{-2s}+\frac{C_4}{c_{\mathcal{L}}}\frac{n\log n}{M}+\frac{4C}{c_{\mathcal{L}}M}\log(\frac{2}{\delta})\\ &\le \frac{C_5}{c_{\mathcal{L}}}\left(\frac{\log M}{M}\right)^{\frac{2s}{2s+1}}+\frac{4C}{c_{\mathcal{L}}M}\log(2/\delta),\end{aligned} \tag{3.27}$$

where we used (3.18) to get $\epsilon_{M,\delta,N}$, C, and $\{C_i\}_{i=0}^2$ is defined in (3.18), (3.21), and (3.22) respectively, and

$$\begin{aligned}C_3&=6C_0+4C_1+4c_0^{\gamma}c_1C_2+8R^2\\ C_4&=4c_0C|\log(4S_0)|+4c_0(2s+\gamma)C\end{aligned}$$

$$C_5 = C_3 + \frac{C_4}{2s+1}.$$

The bound in expectation is obtained by standard techniques, writing

$$\mathbb{E}\left|\left|\left|\widehat{\phi}_{T,M,\mathcal{H}_n} - \phi\right|\right|\right|^2 = \int_0^\infty \mathbb{P}\left\{\left|\left|\left|\widehat{\phi}_{T,M,\mathcal{H}_n} - \phi\right|\right|\right|^2 > \epsilon\right\} \mathrm{d}\epsilon\,,$$

and splitting the integration interval into $[0, \frac{C_5}{c_{\mathcal{L}}}(\frac{\log M}{M})^{\frac{2s}{2s+1}}]$ and $[\frac{C_5}{c_{\mathcal{L}}}(\frac{\log M}{M})^{\frac{2s}{2s+1}}, \infty]$. On the first interval, we use $\mathbb{P}\left\{\left|\left|\left|\widehat{\phi}_{T,M,\mathcal{H}_n} - \phi\right|\right|\right|^2 > \epsilon\right\} \leq 1$. On the second one, we use a change of variables and the probability estimate (3.27). We obtain

$$\begin{aligned}\mathbb{E}\left|\left|\left|\widehat{\phi}_{T,M,\mathcal{H}_n} - \phi\right|\right|\right|^2 &\leq \frac{C_5}{c_{\cup_n \mathcal{H}_n}}\left(\frac{\log M}{M}\right)^{\frac{2s}{2s+1}} + \frac{4C}{c_{\cup_n \mathcal{H}_n}}\frac{1}{M} \\ &\leq \frac{C_6}{c^2_{\cup_n \mathcal{H}_n}}\left(\frac{\log M}{M}\right)^{\frac{2s}{2s+1}} \end{aligned} \tag{3.28}$$

where C_6 is an absolute constant only depending on σ, N, T, S_0, R. □

Proof of Theorem 3.1 In this proof, $a \lesssim b$ means that there exists a constant c such that $a \leq cb$. For any $\epsilon > 0$, we claim

$$\sum_{M=1}^{\infty} \mathbb{P}\left\{\left|\left|\left|\widehat{\phi}_{T,M,\mathcal{H}_M} - \phi\right|\right|\right|^2 \geq \epsilon\right\} < \infty.$$

Strong consistency will then follow from the Borel–Cantelli lemma. Notice that

$$\begin{aligned}\mathbb{P}\left\{\left|\left|\left|\widehat{\phi}_{T,M,\mathcal{H}_M} - \phi\right|\right|\right|^2 \geq \epsilon\right\} \leq\ & \mathbb{P}\left\{\left|\left|\left|\widehat{\phi}_{T,M,\mathcal{H}_M} - \widehat{\phi}_{T,\infty,\mathcal{H}_M}\right|\right|\right|^2 \geq \frac{\epsilon}{2}\right\} \\ &+ \mathbb{P}\left\{\left|\left|\left|\widehat{\phi}_{T,\infty,\mathcal{H}_M} - \phi\right|\right|\right|^2 \geq \frac{\epsilon}{2}\right\},\end{aligned}$$

and $\mathbb{P}\left\{\left|\left|\left|\widehat{\phi}_{T,\infty,\mathcal{H}_M} - \phi\right|\right|\right|^2 \geq \frac{\epsilon}{2}\right\} = 0$ when M is large enough (see (3.26)). It suffices to prove

$$\sum_{M=1}^{\infty} \mathbb{P}\left\{\left|\left|\left|\widehat{\phi}_{T,M,\mathcal{H}_M} - \widehat{\phi}_{T,\infty,\mathcal{H}_M}\right|\right|\right|^2 \geq \epsilon\right\} < \infty\,.$$

Let $\{\varphi_j\}_{j=1}^{\mathcal{N}}$ be an η net for $\mathcal{H}_M$. Consider the event

$$\Lambda_{\eta,M,\epsilon} = \{\max_{1\leq j\leq \mathcal{N}} \frac{1}{2}\mathcal{D}_{\varphi_j,\infty} - \mathcal{D}_{\varphi_j,M} \geq \epsilon\}\,.$$

The bound (3.20) in Proposition 3.4 yields

$$\mathbb{P}\left\{\Lambda_{\eta,M,\epsilon}\right\} \lesssim \mathcal{N}\exp\left(-c_{\mathcal{H}_M} M\epsilon\right). \tag{3.29}$$

Using the fact that there exists $j_M \in \{1, 2 \ldots, \mathcal{N}\}$ such that $\|\phi - \varphi_{j_M}\|_\infty \le \eta$, and following the same argument as in (3.21) and (3.22), we obtain,

$$\mathbb{P}\left\{\frac{1}{2}\mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_M},\infty} - \mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_M},M} \gtrsim \eta n_M^\gamma + \epsilon\right\} \le \mathbb{P}\left\{\Lambda_{\eta,M,\epsilon}\right\} \lesssim \mathcal{N}\exp\left(-c_{\mathcal{H}_M} M\epsilon\right).$$

Notice that $\mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_M},M} \le 0$ and $\mathcal{D}_{\widehat{\phi}_{T,M,\mathcal{H}_M},\infty} \ge c_{\mathcal{H}_M}\left|\!\left|\!\left|\widehat{\phi}_{T,M,\mathcal{H}_M} - \widehat{\phi}_{T,\infty,\mathcal{H}_M}\right|\!\right|\!\right|^2$, so that we have

$$\mathbb{P}\left\{c_{\mathcal{H}_M}\left|\!\left|\!\left|\widehat{\phi}_{T,M,\mathcal{H}_M} - \widehat{\phi}_{T,\infty,\mathcal{H}_M}\right|\!\right|\!\right|^2 \gtrsim \eta n_M^\gamma + \epsilon\right\} \lesssim \mathcal{N}\exp\left(-c_{\mathcal{H}_M} M\epsilon\right).$$

Let $\eta n_M^\gamma = \epsilon$, i.e., $\eta = n_M^{-\gamma}\epsilon$, by assumption, we have $c_{\mathcal{H}_M} \ge c_{\cup_M \mathcal{H}_M} > 0$ and $\lim_{M\to\infty} \frac{n_M \log n_M}{M} = 0$, we have

$$\begin{aligned}\mathbb{P}\left\{\left|\!\left|\!\left|\widehat{\phi}_{T,M,\mathcal{H}_M} - \widehat{\phi}_{T,\infty,\mathcal{H}_M}\right|\!\right|\!\right|^2 \gtrsim \epsilon\right\} &\lesssim \exp\left(n_M \log\frac{n_M}{\epsilon} - c_{\cup_M\mathcal{H}_M} M\epsilon\right)\\ &= \exp\left(-c_{\cup_M\mathcal{H}_M} M\epsilon\left(1 - \frac{n_M}{M\epsilon}\log\frac{n_M}{\epsilon}\right)\right)\\ &\lesssim \exp\left(-\frac{1}{2}c_{\cup_M\mathcal{H}_M} M\epsilon\right)\end{aligned}$$

when M is large enough. By the comparison theorem, the series

$$\sum_{M=1}^{\infty}\mathbb{P}\left\{\left|\!\left|\!\left|\widehat{\phi}_{T,M,\mathcal{H}_M} - \widehat{\phi}_{T,\infty,\mathcal{H}_M}\right|\!\right|\!\right|^2 \gtrsim \epsilon\right\}$$

converges. The claim is proved. □

4 Learning Theory: Discrete-Time Observations

In this section, we analyze the estimation error of the (regularized) MLE $\widehat{\phi}_{L,T,M,\mathcal{H}}$, defined in (2.5) for finite-dimensional linear space $\mathcal{H}$ and for discrete-time observations. We show that it is of order $\sqrt{\frac{n}{M}} + \Delta t^{1/2}$ with high probability, where n is the dimension of $\mathcal{H}$ and Δt is the observation gap. As a consequence, the MLE is consistent when $M \to \infty$ and $\Delta t \to 0$; and the MLE converges at an optimal rate as when n is optimally chosen as in (2.13).

This section is organized as follows: we shall first prove the main theorem, Theorem 4.2, on the error bounds of the MLE in Sect. 4.1. We postpone the technical details, including concentration inequalities and discretization error bounds, to Sects. 4.2–4.3.

Recall that we denote $\boldsymbol{X}_{[0,T]}$ the solution to system (1.1) with the true interaction kernel ϕ and denote $\{\boldsymbol{X}_{t_0:t_L}^{(m)}\}_{m=1}^M$ independent trajectories observed at discrete times $t_l = l\Delta t$ with $\Delta t = T/L$. Recall that when $\mathcal{H} = \text{span}\{\psi_p\}_{p=1}^n$, the MLE $\widehat{\phi}_{L,T,M,\mathcal{H}}$ is given in (2.8).

Throughout this section, we assume

Assumption 4.1 (Basis functions) Assume that the basis functions $\{\psi_p\}_{p=1}^n \subset C_b^1(\mathbb{R}^+, \mathbb{R})$ satisfy the following conditions:

(a) $\{\psi_p(\cdot)(\cdot)\}_{p=1}^n$ are orthonormal in $L^2(\rho_T)$;
(b) $\max_k \|\psi_p\|_\infty \le b_0$, $\max_k \|\psi_k'(\cdot)(\cdot)\|_\infty \le b_1$;
(c) there exists a constant c_{ρ_T} such that $n \le c_{\rho_T} \min(b_0^2 R, b_1 R^{3/2})$.

Item (a) aims to make the normal equations matrix nonsingular, as discussed in Proposition 3.2. In item (b), the uniform bound for the derivatives aims to control the discretization error due to discrete-time observations; the uniform boundness on the functions will be used for concentration inequalities. Item (c) states that the number of such orthonormal basis functions is bounded by the measure ρ_T and the uniform bounds of the functions and their derivatives. Item (c) often follows from $(a)-(b)$, with an intuition from examples including polynomials and trigonometric polynomials, and smoothed piecewise polynomials, if $r^2\rho_T(\mathrm{d}r)$ is equivalent to the Lebesgue measure on an interval $[R_0, R] \subset \mathbb{R}^+$. Such an interval is where pairwise distances explore with a noticeable probability (see, for example, in Figs. 2 and 6). It exists in general when the initial distribution spreads out the pairwise distances or when the relative position system is ergodic, since the density of ρ_T is smooth and nonnegative on $\mathbb{R}^+$.

4.1 Error Bounds for the MLE

We show first that the $L^2(\rho_T)$ error of the estimator $\widehat{\phi}_{L,T,M,\mathcal{H}}$ in (2.8) converges as $M \to \infty$ and $\Delta t := T/L \to 0$, with high probability.

Theorem 4.2 (Error bounds for the MLE) *Let the hypothesis space be* $\mathcal{H} = \mathrm{span}\{\psi_i\}_{i=1}^n$, *where the set of functions* $\{\psi_i\}_{i=1}^n$, *satisfying Assumption 4.1, are orthonormal in* $L^2(\rho_T)$ *with respect to the norm* $|||\cdot|||$ *defined in Eq. (2.11). Suppose that the coercivity condition holds on* $\mathcal{H}$ *with a constant* $c_{\mathcal{H}} > 0$. *Then, with a probability at least* $1 - (4n + 2n^2)\exp\left(-\frac{\epsilon^2}{8c_1^2}\right)$, *the error of the estimator* $\widehat{\phi}_{L,T,M,\mathcal{H}}$ *in (2.8) satisfies*

$$|||\widehat{\phi}_{L,T,M,\mathcal{H}} - \phi||| \le |||\widehat{\phi}_{T,\infty,\mathcal{H}} - \phi||| + c_2\left(\sqrt{\frac{n}{M}}\epsilon + c_3 \Delta t^{\frac{1}{2}}\right), \tag{4.1}$$

where $\widehat{\phi}_{T,\infty,\mathcal{H}}$ *is the projection of the true kernel to* $\mathcal{H}$, $\Delta t = T/L \le c_{\mathcal{H}}/(2c_3)$, *and the constants are*

$$\begin{aligned} c_1 &= Rb_0(Rb_0 + 2\sigma/\sqrt{T}), \\ c_2 &= 4c_{\mathcal{H}}^{-1}(1 + |||\phi_{T,\infty,\mathcal{H}}|||), \\ c_3 &= dN(b_1 + b_0)Rb_0(Rb_0\Delta t^{\frac{1}{2}} + \sqrt{d}\sigma)\sqrt{c_{\rho_T}\min(b_0^2 R, b_1 R^{3/2})}. \end{aligned} \tag{4.2}$$

Remark 1 *(The discretization error may dominate the statistical error)* When the observation gap $\Delta t = 0$, we recover the min–max learning rate $M^{-\frac{s}{2s+1}}$ proved in the previous section, if we choose the optimal dimension $n = C(M/\log M)^{1/(2s+1)}$ for the hypothesis space. However, when $\Delta t > 0$, once $M^{-\frac{s}{2s+1}}(\log M)^{-\frac{1}{2(2s+1)}} \lesssim \Delta t^{\frac{1}{2}}$, the discretization error will dominate the error of the estimator, preventing us from observing the min–max learning rate. This phenomenon is well-illustrated by the left plots in Figs. 5 and 9 in our numerical experiments.

Remark 2 We assumed C_b^1 regularity for the basis functions $\{\psi_p\}$ for the above numerical error analysis, stronger than that of piecewise polynomials (which may be discontinuous) used in the numerical tests. Such a difference between the regularity requirements stems from the numerical representation, and we can view the piecewise polynomials as numerical approximations of regular functions. This view is supported by the numerical experiments: the estimator has only small jumps at the discontinuities in the high probability region.

Remark 3 A smaller coercivity constant $c_{\mathcal{H}}$ corresponds to a worse conditioned problem (Proposition 3.2), and so the condition $L \gtrsim 1/c_{\mathcal{H}}$ that requires L to be larger for small $c_{\mathcal{H}}$ makes sense.

We shall prove Theorem 4.2 as follows: we first outline the main idea and introduce the key elements, such as the normal matrices and vectors and the empirical error functionals in their entries; then, we provide a proof with key but technical estimations, including the concentration inequalities and discretization error bounds, postponed to Sects. 4.2–4.3.

The error of the MLE $\widehat{\phi}_{L,T,M,\mathcal{H}}$ consists of three parts: approximation error, discretization error and sampling error:

$$\begin{aligned}\left|\!\left|\!\left|\widehat{\phi}_{L,T,M,\mathcal{H}} - \phi\right|\!\right|\!\right| \le & \underbrace{\left|\!\left|\!\left|\widehat{\phi}_{T,\infty,\mathcal{H}} - \phi\right|\!\right|\!\right|}_{\text{Approximation Error}} + \underbrace{\left|\!\left|\!\left|\widehat{\phi}_{L,T,\infty,\mathcal{H}} - \widehat{\phi}_{T,\infty,\mathcal{H}}\right|\!\right|\!\right|}_{\text{Discretization Error}} \\ & + \underbrace{\left|\!\left|\!\left|\widehat{\phi}_{L,T,M,\mathcal{H}} - \widehat{\phi}_{L,T,\infty,\mathcal{H}}\right|\!\right|\!\right|}_{\text{Sampling Error}}, \end{aligned} \tag{4.3}$$

where $\widehat{\phi}_{L,T,\infty,\mathcal{H}}$ is the infinite-data estimator. We shall study the discretization error and the sampling error by analyzing the differences between their coefficient vectors.

All these coefficients are solutions to the corresponding normal equations (e.g., Eq. (2.7)). To facilitate the study of these normal matrices and vectors, we introduce the following notions. For any $f, g \in C_b^1(\mathbb{R}^{Nd}, \mathbb{R}^{Nd})$, we define the following functionals of the observation paths:

$$\begin{aligned}
\xi(f, \boldsymbol{X}_{t_0:t_L}) &= \frac{1}{TN}\sum_{l=1}^{L-1}\langle f(\boldsymbol{X}_{t_l}), \boldsymbol{X}_{t_{l+1}} - \boldsymbol{X}_{t_l}\rangle,\\
\eta(f, g, X_{t_0:t_L}) &= \frac{1}{LN}\sum_{l=1}^{L-1}\langle f(\boldsymbol{X}_{t_l}), g(\boldsymbol{X}_{t_l})\rangle,\\
\xi(f, \boldsymbol{X}_{[0,T]}) &= \frac{1}{TN}\int_0^T \langle f(\boldsymbol{X}_t), \mathrm{d}\boldsymbol{X}_t\rangle,\\
\eta(f, g, \boldsymbol{X}_{[0,T]}) &= \frac{1}{TN}\int_0^T \langle f(\boldsymbol{X}_t), g(\boldsymbol{X}_t)\rangle \mathrm{d}t.
\end{aligned} \tag{4.4}$$

Correspondingly, we define the empirical functionals

$$\begin{aligned}
\xi_{M,L}(f) &= \frac{1}{M}\sum_{m=0}^{M}\xi(f, \boldsymbol{X}^{(m)}_{t_0:t_L}), \qquad \eta_{M,L}(f, g) = \frac{1}{M}\sum_{m=0}^{M}\eta(f, g, \boldsymbol{X}^{(m)}_{t_0:t_L}),\\
\xi_{M,\infty}(f) &= \frac{1}{M}\sum_{m=0}^{M}\xi(f, \boldsymbol{X}^{(m)}_{[0,T]}), \qquad \eta_{M,\infty}(f, g) = \frac{1}{M}\sum_{m=0}^{M}\eta(f, g, \boldsymbol{X}^{(m)}_{[0,T]}),
\end{aligned} \tag{4.5}$$

Using the notation of empirical functional introduced in (4.4)–(4.5), we consider the following normal matrixes and vectors:

$$\begin{aligned}
b_{M,L}(k) &= \xi_{M,L}(f_{\psi_p}), & A_{M,L}(k,k') &= \eta_{M,L}(f_{\psi_p}, f_{\psi_{k'}}),\\
b_{\infty,L}(k) &= \mathbb{E}[\xi(f_{\psi_p}, \boldsymbol{X}_{t_0:t_L})], & A_{\infty,L}(k,k') &= \mathbb{E}[\eta(f_{\psi_p}, f_{\psi_{k'}}, \boldsymbol{X}_{t_0:t_L})],\\
b_{\infty}(k) &= \mathbb{E}[\xi(f_{\psi_p}, \boldsymbol{X}_{[0,T]})], & A_{\infty}(k,k') &= \mathbb{E}[\eta(f_{\psi_p}, f_{\psi_{k'}}, \boldsymbol{X}_{[0,T]})].
\end{aligned} \tag{4.6}$$

It is clear that (here, to ease the notation, we denote the coefficient $\widehat{a}_{L,T,M,\mathcal{H}}$ in Fig. 1 as $a_{M,L}$, and similarly for others)

$$\begin{aligned}
\widehat{\phi}_{L,T,M,\mathcal{H}} &= \sum_{i=1}^{n} a_{M,L}(i)\psi_i, \quad \text{with } a_{M,L} = A_{M,L}^{-1} b_{M,L},\\
\widehat{\phi}_{L,T,\infty,\mathcal{H}} &= \sum_{i=1}^{n} a_{\infty,L}(i)\psi_i, \quad \text{with } a_{\infty,L} = A_{\infty,L}^{-1} b_{\infty,L},\\
\widehat{\phi}_{T,\infty,\mathcal{H}} &= \sum_{i=1}^{n} a_{\infty}(i)\psi_i, \quad \text{with } a_{\infty} = A_{\infty}^{-1} b_{\infty}.
\end{aligned} \tag{4.7}$$

Here the matrix A_∞ is invertible due the coercivity condition: its smallest eigenvalue is the coercivity constant $c_{\mathcal{H}}$ (see Proposition 3.2). The matrix $A_{\infty,L}$ is invertible when $L = T/(\Delta t)$ is large, with its smallest eigenvalue bounded below by $c_{\mathcal{H}} - c_3\Delta t^{1/2}$, see Corollary 4.1. Furthermore, Corollary 4.2 in Sect. 4.3 shows that, with probability at $1-\delta$, the matrix $A_{M,L}$ is invertible with its smallest eigenvalue bounded blow by $c_{\mathcal{H}} - \left(\sqrt{\frac{n}{M}}\epsilon + c_3\Delta t^{\frac{1}{2}}\right)$.

Proof of Theorem 4.2 By Eq. (4.3), it suffices to prove upper bounds for the discretization error and the sampling error separately:

$$\begin{aligned} &\text{discretization error:} && \left|\!\left|\!\left|\widehat{\phi}_{L,T,\infty,\mathcal{H}} - \widehat{\phi}_{T,\infty,\mathcal{H}}\right|\!\right|\!\right| \le c_2 c_3 \Delta t^{1/2}, \\ &\text{sampling error:} && \left|\!\left|\!\left|\widehat{\phi}_{L,T,M,\mathcal{H}} - \widehat{\phi}_{L,T,\infty,\mathcal{H}}\right|\!\right|\!\right| \le c_2 \sqrt{\frac{n}{M}}\epsilon, \end{aligned}$$

where the second inequality holds with probability at least $1-\delta$.

For the discretization error, since $\{\psi_i(\cdot)(\cdot)\}$ are orthonormal in $L^2(\rho_T)$, we have, by Eq. (4.7):

$$\begin{aligned} \left|\!\left|\!\left|\widehat{\phi}_{L,T,\infty,\mathcal{H}} - \widehat{\phi}_{T,\infty,\mathcal{H}}\right|\!\right|\!\right|^2 &= \left\| \sum_{i=1}^{n} [a_{\infty,L}(i) - a_\infty(i)]\psi_i \right\|^2 \\ &= \left\| a_{\infty,L} - a_\infty \right\|^2 = \| A_{\infty,L}^{-1} b_{\infty,L} - A_\infty^{-1} b_\infty \|^2. \end{aligned}$$

Using the formula $A_{\infty,L}^{-1} - A_\infty^{-1} = A_{\infty,L}^{-1}(A_\infty - A_{\infty,L})A_\infty^{-1}$ (see, e.g., [25, Appendix B9]), we have

$$\begin{aligned} \left\| A_{\infty,L}^{-1} b_{\infty,L} - A_\infty^{-1} b_\infty \right\| &= \left\| A_{\infty,L}^{-1}(b_{\infty,L} - b_\infty) + (A_{\infty,L}^{-1} - A_\infty^{-1}) b_\infty \right\| \\ &\le \left\| A_{\infty,L}^{-1} \right\| \left(\left\| b_{\infty,L} - b_\infty \right\| + \left\| A_\infty - A_{\infty,L} \right\| \left\| A_\infty^{-1} b_\infty \right\| \right). \end{aligned}$$

Note that (i) $\left\| A_{\infty,L}^{-1} \right\| \le 2c_{\mathcal{H}}^{-1}$, since $c_3 \Delta t^{1/2} < \frac{1}{2} c_{\mathcal{H}}$; (ii) by Proposition 4.1 (in Sect. 4.3) in combination of Assumption 4.1,

$$\left\| b_{\infty,L} - b_\infty \right\| \le c_3 \Delta t^{1/2}, \quad \left\| A_{\infty,L} - A_\infty \right\| \le c_3 \Delta t^{1/2};$$

and (iii) $\left\| A_\infty^{-1} b_\infty \right\| = \left|\!\left|\!\left|\widehat{\phi}_{T,\infty,\mathcal{H}}\right|\!\right|\!\right|$. Then,

$$\left\| A_{\infty,L}^{-1} b_{\infty,L} - A_\infty^{-1} b_\infty \right\| \le 2c_{\mathcal{H}}^{-1} \left(1 + \left|\!\left|\!\left|\widehat{\phi}_{T,\infty,\mathcal{H}}\right|\!\right|\!\right|\right) c_3 \Delta t^{1/2}, \tag{4.8}$$

and the inequality for the discretization error follows.

Similarly, for the sampling error, we have

$$\begin{aligned} &\left|\!\left|\!\left|\widehat{\phi}_{L,T,M,\mathcal{H}} - \widehat{\phi}_{L,T,\infty,\mathcal{H}}\right|\!\right|\!\right| \\ &= \left\| a_{M,L} - a_{\infty,L} \right\| = \| A_{M,L}^{-1} b_{M,L} - A_{\infty,L}^{-1} b_{\infty,L} \| \\ &= \left\| A_{M,L}^{-1}(b_{M,L} - b_{\infty,L}) + (A_{M,L}^{-1} - A_{\infty,L}^{-1}) b_{\infty,L} \right\| \\ &\le \left\| A_{M,L}^{-1} \right\| \left(\left\| b_{M,L} - b_{\infty,L} \right\| + \left\| A_{M,L}^{-1} - A_{\infty,L}^{-1} \right\| \left\| A_{\infty,L}^{-1} b_{\infty,L} \right\| \right). \end{aligned}$$

Note that (i) $\left\|A_{M,L}^{-1}\right\| \leq 2c_{\mathcal{H}}^{-1}$ when M and L are large enough such that $\sqrt{\frac{n}{M}}\epsilon + c_3\Delta t^{\frac{1}{2}} < \frac{1}{2}c_{\mathcal{H}}$; (ii) we have, by Proposition 4.2 (in Sect. 4.3), that

$$\left\|b_{M,L} - b_{\infty,L}\right\| \leq \sqrt{\frac{n}{M}}\epsilon, \quad \left\|A_{M,L} - A_{\infty,L}\right\| \leq \sqrt{\frac{n}{M}}\epsilon$$

hold with probability at least $1-\delta$; (iii) since $c_3\Delta t^{\frac{1}{2}} < \frac{1}{2}c_{\mathcal{H}}$ and $\left\|A_{\infty}^{-1}b_{\infty}\right\| = \left|\left|\left|\widehat{\phi}_{T,\infty,\mathcal{H}}\right|\right|\right|$, we have

$$\left\|A_{\infty,L}^{-1}b_{\infty,L}\right\| \leq 2\left(1 + \left|\left|\left|\widehat{\phi}_{T,\infty,\mathcal{H}}\right|\right|\right|\right).$$

Then, the inequality for the sampling error follows. □

In in the next two subsections, we prove Proposition 4.1–4.2 that we just used in the above proof. Section 4.2 studies the concentration inequalities and discretization error bounds for the empirical error functionals in (4.5), which are the entries of the normal matrices and vectors. Based on them, Sect. 4.3 provides error bounds for the normal matrices and the normal vectors in Proposition 4.1–4.2.

4.2 Concentration and Discretization Error of Empirical Functionals

We introduce concentration inequalities for the above empirical functionals on the path space of diffusion processes and a bound on the discretization error of the estimator due to discrete-time approximations. Our first lemma studies concentration of the discrete-time empirical functionals $\xi_{M,L}$ and $\eta_{M,L}$.

Lemma 4.1 (Concentration of empirical functionals) *Let $\{X_{t_0:t_L}^{(m)}\}_{m=1}^{M}$ be discrete-time observations, with $t_l = l\Delta t$ and $L = T/\Delta t$, of the system (1.1) with ϕ. Then for any $f, g \in C_b(\mathbb{R}^{dN}, \mathbb{R}^{dN})$, the error functionals defined in (4.5) satisfy the concentration inequalities:*

$$\begin{aligned}\mathbb{P}\left\{|\xi_{M,L}(f) - \mathbb{E}[\xi_{M,L}(f)]| > \epsilon\right\} &\leq 4e^{-\frac{M\epsilon^2}{8C_1^2}}\\ \mathbb{P}\left\{|\eta_{M,L}(f,g) - \mathbb{E}[\eta_{M,L}(f,g)]| > \epsilon\right\} &\leq 2e^{-\frac{M\epsilon^2}{2C_2^2}},\end{aligned} \tag{4.9}$$

for any $\epsilon > 0$, where $C_1 = \frac{1}{N}\|f\|_\infty \max(\frac{2\sigma\sqrt{N}}{\sqrt{T}}, \|f_\phi\|_\infty)$, and $C_2 = \frac{1}{N}\|f\|_\infty\|g\|_\infty$. Furthermore,

$$\mathbb{P}\left\{|\xi_{M,L}(f) - \mathbb{E}[\xi_{M,L}(f)]| < \epsilon,\ |\eta_{M,L}(f,g) - \mathbb{E}[\eta_{M,L}(f,g)]| < \epsilon\right\} \geq 1 - 8e^{-\frac{M\epsilon^2}{8C^2}}, \tag{4.10}$$

where $C = \frac{1}{N}\|f\|_\infty \max(\frac{2\sigma}{\sqrt{T/N}}, \|f_\phi\|_\infty, \|g\|_\infty)$.

Proof Note that $\left|\eta(f, g, \boldsymbol{X}_{[t_0:t_L]})\right| \leq \|f\|_\infty \|g\|_\infty$. Then, the exponential inequality for η_{ML} follows from the Hoeffding inequality, which states that, for i.i.d. random variables $\{Z_i\}$ bounded above by K, one has $\mathbb{P}\left\{\left|\frac{1}{M}\sum_{m=1}^{M}(Z_i - \mathbb{E}Z_i)\right| > \epsilon\right\} \leq 2\exp\left(-\frac{M\epsilon^2}{2K^2}\right)$.

To study ξ_{ML}, we decompose $\xi(f, \boldsymbol{X}_{[t_0:t_L]})$ into two parts, a bounded part and a martingale part:

$$\begin{aligned}\xi(f, \boldsymbol{X}_{[t_0:t_L]}) =& \frac{1}{TN}\sum_{l=0}^{L-1}\langle f(\boldsymbol{X}_{t_l})\mathbf{1}_{[t_l,t_{l+1}]}(s), \boldsymbol{X}_{t_{l+1}} - \boldsymbol{X}_{t_l}\rangle \\ =& \frac{1}{TN}\int_0^T \langle f^L(s), f_\phi(\boldsymbol{X}_s)\rangle \mathrm{d}s + \frac{1}{TN}\int_0^T \langle f^L(s), \sigma \mathrm{d}\boldsymbol{B}_s\rangle =: Z_T + Y_T,\end{aligned}$$

where we denote $f^L(s) := \sum_{l=0}^{L-1} f(\boldsymbol{X}_{t_l})\mathbf{1}_{[t_l,t_{l+1}]}(s)$. We call Z_T a bounded part because

$$|Z_T| = \left|\frac{1}{TN}\int_0^T \langle f^L(s), f_\phi(\boldsymbol{X}_s)\rangle \mathrm{d}s\right| \leq \frac{1}{N}\|f\|_\infty \|f\|_\infty.$$

We call Y_T a martingale part since $TNY_T = \int_0^T \langle f^L(s), \sigma \mathrm{d}\boldsymbol{B}_s\rangle$ is a martingale. Correspondingly, we can write

$$\xi_{ML} = \frac{1}{M}\sum_{m=1} Z_T^{(m)} + Y_T^{(m)}.$$

Then, denoting $K_1 := \frac{1}{N}\|f\|_\infty \|f_\phi\|_\infty$ and $K_2 = 2\sigma\|f\|_\infty$, and noticing that $C_1 = K_1 + K_2/\sqrt{TN}$, we can conclude the first concentration inequality in (4.9) from

$$\begin{aligned}&\mathbb{P}\left\{|\xi_{M,L}(f) - \mathbb{E}[\xi_{M,L}(f)]| > \epsilon\right\} \\ &\leq \mathbb{P}\left\{\left|\frac{1}{M}\sum_{m=1} Z_T^{(m)} - \mathbb{E}Z_T\right| \geq \frac{\epsilon}{2}\right\} + \mathbb{P}\left\{\left|\frac{1}{M}\sum_{m=1} Y_T^{(m)}\right| \geq \frac{\epsilon}{2}\right\} \\ &\leq 2e^{-\frac{M\epsilon^2}{2K_1^2}} + e^{-\frac{TNM\epsilon^2}{8K_2^2}},\end{aligned}$$

where the first exponential bound follows directly from Hoeffding inequality applied to $\{Z_T^{(m)}\}$, and the second exponential bound $\mathbb{P}\left\{\left|\frac{1}{M}\sum_{m=1} Y_T^{(m)}\right| \geq \frac{\epsilon}{2}\right\} \leq e^{-\frac{M\epsilon^2}{8K_2^2}}$ is proved as follows.

Note that $\mathbb{E}Y_T = 0$ and $TNY_T = \int_0^T \langle f^L(s), \sigma \mathrm{d}\boldsymbol{B}_s\rangle$ is a martingale satisfying $\mathbb{E}[e^{\sigma^2\int_0^T |f^L(s)|^2 ds}] < \infty$ because $\left|f^L(s)\right|^2 \leq \|f\|_\infty$. By the Novikov theorem, the process $(e^{\alpha TNY_T - \frac{\alpha^2}{2}\sigma^2\int_0^T |f^L(s)|^2 ds}, T \geq 0)$ is a martingale for any $\alpha \in \mathbb{R}$ (see, e.g., [31, Corollary 5.13]). Therefore, with $\alpha = \frac{\lambda}{MTN}$, we have

$$\mathbb{E}\left[e^{\frac{\lambda}{M}Y_T}\right] = \mathbb{E}\left[e^{\sigma^2\frac{\lambda^2}{(MTN)^2}\int_0^T |f^L(s)|^2 ds}\right] \le e^{\sigma^2\frac{\lambda^2}{M^2TN}\|f\|_\infty^2}$$

for any $\lambda > 0$. As a consequence, we have

$$\mathbb{P}\left\{\left|\frac{1}{M}\sum_{m=1} Y_T^{(m)}\right| \ge \frac{\epsilon}{2}\right\} \le \inf_{\lambda>0} e^{-\frac{\lambda\epsilon}{2}}\mathbb{E}\left[e^{\frac{\lambda}{M}Y_T}\right] \le \inf_{\lambda>0} e^{-\frac{\lambda\epsilon}{2}+\lambda^2\frac{\sigma^2\|f\|_\infty}{TNM}} \le e^{-\frac{TNM\epsilon^2}{8K_2^2}}.$$

Lastly, Eq. (4.10) follows directly from Eq. (4.9). □

We remark that here we focus on the case $M \to \infty$ with finite time T. If the relative position system is ergodic, one can extend the concentration inequalities to the case when $T \to \infty$.

The next lemma shows that the discretization error of the empirical functionals, as discrete-time approximation of the integrals, is at order $\Delta t^{\frac{1}{2}}$.

Lemma 4.2 (Discretization error of empirical functionals) *Let $f, g \in C_b^1(\mathbb{R}^{dN}, \mathbb{R}^{dN})$. Let $\boldsymbol{X}_{t_0:t_L}$ be a discrete-time trajectory, with $t_l = l\Delta t$ and $L = T/\Delta t$, of the system (1.1) with ϕ. Then, the error functionals defined in (4.4) satisfy*

$$\begin{aligned}\left|\mathbb{E}[\xi(f, \boldsymbol{X}_{t_0:t_L})] - \mathbb{E}[\xi(f, \boldsymbol{X}_{[0,T]})]\right| &\le C_1\Delta t^{\frac{1}{2}};\\ \left|\mathbb{E}[\eta(f, g, \boldsymbol{X}_{t_0:t_L})] - \mathbb{E}[\eta(f, g, \boldsymbol{X}_{[0,T]})]\right| &\le C_2\Delta t^{\frac{1}{2}},\end{aligned} \tag{4.11}$$

where the constants are

$$\begin{aligned}C_1 &= \|\nabla f\|_\infty\|f_\phi\|_\infty\left(\|f_\phi\|_\infty\Delta t^{\frac{1}{2}}/N + \sqrt{d/N}\sigma\right),\\ C_2 &= (\|\nabla f\|_\infty\|g\|_\infty + \|\nabla g\|_\infty\|f\|_\infty)\left(\|f_\phi\|_\infty\Delta t^{\frac{1}{2}}/N + \sqrt{d/N}\sigma\right).\end{aligned}$$

Proof Note that since $\boldsymbol{X}_{[0,T]}$ is a solution to system (1.1), we have for each l,

$$\begin{aligned}\mathbb{E}\int_{t_l}^{t_{l+1}}\langle f(\boldsymbol{X}_{t_l}) - f(\boldsymbol{X}_s), \mathrm{d}\boldsymbol{X}_s\rangle &= \mathbb{E}\int_{t_l}^{t_{l+1}}\langle f(\boldsymbol{X}_{t_l}) - f(\boldsymbol{X}_s), f_\phi(\boldsymbol{X}_s)\rangle \mathrm{d}s\\ &\le \|\nabla f\|_\infty\|f_\phi\|_\infty\mathbb{E}\int_{t_l}^{t_{l+1}}|\boldsymbol{X}_{t_l} - \boldsymbol{X}_s|\mathrm{d}s\\ &\le \|\nabla f\|_\infty\|f_\phi\|_\infty\left(\|f_\phi\|_\infty\Delta t^2 + \sqrt{dN}\sigma\Delta t^{3/2}\right),\end{aligned}$$

where in the first inequality we have applied the mean value theorem to bound $f(\boldsymbol{X}_{t_l}) - f(\boldsymbol{X}_s)$:

$$|f(\boldsymbol{X}_{t_l}) - f(\boldsymbol{X}_s)| \le \|\nabla f\|_\infty|\boldsymbol{X}_{t_l} - \boldsymbol{X}_s|,$$

and in the second inequality, we used the fact that

$$\mathbb{E}|\boldsymbol{X}_{t_l} - \boldsymbol{X}_s| = \mathbb{E}\left|\int_{t_l}^{s} f_\phi(\boldsymbol{X}_r))\mathrm{d}r + \sigma(\boldsymbol{B}_s - \boldsymbol{B}_{t_l})\right| \leq \|f_\phi\|_\infty(s - t_l) + \sqrt{dN}\sigma(s - t_l)^{1/2}.$$

Thus, we obtain the bound in (4.11) by a summation over l:

$$\mathbb{E}[\xi(f, \boldsymbol{X}_{t_0:t_L}) - \xi(f, \boldsymbol{X}_{[0,T]})] = \frac{1}{TN}\sum_{l=1}^{L-1}\mathbb{E}\int_{t_l}^{t_{l+1}}\langle f(\boldsymbol{X}_{t_l}) - f(\boldsymbol{X}_s), \mathrm{d}\boldsymbol{X}_s\rangle, \leq \|\nabla f\|_\infty\|f_\phi\|_\infty\left(\|f_\phi\|_\infty\Delta t/N + \sqrt{d/N}\sigma\Delta t^{1/2}\right).$$

Similarly, we have

$$\left|\langle f(\boldsymbol{X}_{t_l}), g(\boldsymbol{X}_{t_l})\rangle - \langle f(\boldsymbol{X}_s), g(\boldsymbol{X}_s)\rangle\right| \leq (\|\nabla f\|_\infty\|g\|_\infty + \|\nabla g\|_\infty\|f\|_\infty)\left|\boldsymbol{X}_s - \boldsymbol{X}_{t_l}\right|,$$

and the bound for η follows from the fact that

$$\mathbb{E}[\eta(f, g, \boldsymbol{X}_{t_0:t_L}) - \eta(f, g, \boldsymbol{X}_{[0,T]})] = \frac{1}{TN}\sum_{l=1}^{L-1}\mathbb{E}\int_{t_l}^{t_{l+1}}\left|\langle f(\boldsymbol{X}_{t_l}), g(\boldsymbol{X}_{t_l})\rangle - \langle f(\boldsymbol{X}_s), g(\boldsymbol{X}_s)\rangle\right|\mathrm{d}s.$$

□

4.3 Error Bounds for the Normal Matrix and Vector

Proposition 4.1 (Discretization error) *For the normal matrix $A_{\infty,L}$ and vector $b_{\infty,L}$ defined in (4.6) with $\{\psi_p\}_{p=1}^n$ satisfying Assumption 4.1, we have*

$$\|b_{\infty,L} - b_\infty\| \leq \sqrt{n}C\Delta t^{\frac{1}{2}}, \quad \|A_{\infty,L} - A_\infty\| \leq \sqrt{n}C\Delta t^{\frac{1}{2}},$$

where the constant C is $C = dN(b_1 + b_0)Rb_0(Rb_0\Delta t^{\frac{1}{2}} + \sqrt{\mathrm{d}}\sigma)$.

Proof Applying Lemma 4.2, in combination of the basic fact that $\|b\| \leq \sqrt{n}\max_{k=1,\ldots,n}|b(k)|$ for any $b \in \mathbb{R}^n$, and $\|A\| \leq \sqrt{n}\max_{k,k'=1,\ldots,n}|A(k,k')|$ for any $A \in \mathbb{R}^{n\times n}$, we obtain

$$\|b_{\infty,L} - b_\infty\| \leq \sqrt{n}C_1\Delta t^{\frac{1}{2}}, \quad \|A_{\infty,L} - A_\infty\| \leq \sqrt{n}C_2\Delta t^{\frac{1}{2}},$$

with constants C_1 and C_2 in the form of

$$C_1 = \|f_\phi\|_\infty\left(\|f_\phi\|_\infty\Delta t^{\frac{1}{2}}/N + \sqrt{d/N}\sigma\right)\max_{k=1,\ldots,n}\|\nabla f_{\psi_p}\|_\infty,$$

$$C_2 = \left(\|f_\phi\|_\infty \Delta t^{\frac{1}{2}}/N + \sqrt{d/N}\sigma \right) \max_{k,k'=1,\ldots,n} (\|\nabla f_{\psi_p}\|_\infty \|f_{\psi'_p}\|_\infty + \|\nabla f_{\psi_{k'}}\|_\infty \|f_{\psi_p}\|_\infty).$$

To complete the proof, we are left to estimate $\left\| f_{\psi_p} \right\|_\infty$ and $\left\| \nabla f_{\psi_p} \right\|_\infty$. From the definition of $f_\cdot$, we have

$$\left\| f_{\psi_p} \right\|_\infty^2 = \sup_x \sum_{i=1}^N \left| \frac{1}{N} \sum_{j=1}^N \psi_p(|\boldsymbol{X}_j - \boldsymbol{X}_i|)(\boldsymbol{X}_j - \boldsymbol{X}_i) \right|^2 \leq R^2 b_0^2 N, \quad (4.12)$$

and $\left\| f_\phi \right\|_\infty \leq R b_0 \sqrt{N}$ as well. Note that for each $i, i' \in \{1, \ldots, N\}$, with notation $\boldsymbol{r}_{ji} = \boldsymbol{x}_j - \boldsymbol{x}_i$ and $r_{ji} = |\boldsymbol{r}_{ji}|$, we have,

$$\begin{aligned} \nabla_{\boldsymbol{x}'_i} \left(f_{\psi_p}(\boldsymbol{x}) \right)_i = \delta_{ii'} \frac{1}{N} \sum_{j=1, j\neq i}^N \left(\psi_p(r_{ji}) \mathbf{I}_d + \psi'_p(r_{ji}) \frac{\boldsymbol{r}_{ji} \otimes \boldsymbol{r}_{ji}}{r_{ji}} \right) \\ + \delta_{i'\neq i} \frac{1}{N} \left(\psi_p(r_{ii'}) \mathbf{I}_d + \psi'_p(r_{ii'}) \frac{\boldsymbol{r}_{ii'} \otimes \boldsymbol{r}_{ii'}}{r_{ii'}} \right). \end{aligned}$$

Thus, the norm of this $d \times d$ matrix is uniformly bounded,

$$\sup_x \left\| \nabla_{\boldsymbol{x}'_i} \left(f_{\psi_p}(\boldsymbol{x}) \right)_i \right\| \leq d(b_1 + b_0),$$

and as a result, the norm of the $dN \times dN$ matrix is uniformly bounded,

$$\left\| \nabla f_{\psi_p} \right\|_\infty \leq dN(b_1 + b_0).$$

Combining the above estimates with $\|f_\phi\|_\infty \leq R b_0 N$ (the same as $\|f_{\psi_p}\|_\infty$), we obtain that C_1 and C_2 are both bounded by C. □

It follows directly that the matrix $A_{\infty,L}$ is invertible:

Corollary 4.1 *The smallest eigenvalue of the matrix $A_{\infty,L}$ defined in (4.6) is bounded below by $c_{\mathcal{H}} - c_3 \Delta t^{1/2}$ when $c_3 \Delta t^{1/2} < c_{\mathcal{H}}$, with c_3 defined in (4.2).*

Proof Recall that from Proposition 3.2, we have $a^T A_\infty a \geq c_{\mathcal{H}} |a|^2$ for an arbitrary $a \in \mathbb{R}^n$. Then,

$$a^T A_{\infty,L} a = a^T (A_{\infty,L} - A_\infty) a + a^T A_\infty a \geq (c_{\mathcal{H}} - c_2 \Delta t^{1/2}) \|a\|^2$$

by Proposition 4.1 with the bound of $\sqrt{n}$ in Assumption 4.1. □

We prove next that the matrix $A_{M,L}$ is invertible and concentrates around $A_{\infty,L}$.

Proposition 4.2 (Concentration of the normal matrix and vector) *Suppose that the coercivity condition holds on $\mathcal{H} = \text{span}\{\psi_i\}_{i=1}^n$ with a constant $c_{\mathcal{H}} > 0$, where*

$\{\psi_p\}_{p=1}^n$ satisfying Assumption 4.1. Then, the normal matrix $A_{M,L}$ and vector $b_{M,L}$ defined in (4.6) satisfy concentration inequalities in the sense that for any $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left\{\left\|A_{M,L} - A_{\infty,L}\right\| > \epsilon\right\} &\le 2n^2 e^{-\frac{M\epsilon^2}{2nC^2}} \\
\mathbb{P}\left\{\left\|b_{M,L} - b_{\infty,L}\right\| > \epsilon\right\} &\le 4ne^{-\frac{M\epsilon^2}{8nC^2}}, \\
\mathbb{P}\left\{\left\|A_{M,L} - A_{\infty,L}\right\| < \epsilon,\ \left\|b_{M,L} < b_{\infty,L}\right\| < \epsilon\right\} &\ge 1 - (4n + 2n^2)e^{-\frac{M\epsilon^2}{8nC^2}},
\end{aligned} \tag{4.13}$$

where the constant C is $C = Rb_0(RS_0 + 2\sigma/\sqrt{T})$.

Proof Recall that by definition in (4.6), $b_{M,L}(k) = \xi_{M,L}(f_{\psi_p})$ with $\mathbb{E}[b_{M,L}] = b_{\infty,L}$ and $A_{M,L}(k,k') = \eta_{M,L}(f_{\psi_p}, f_{\psi_{k'}})$ with $\mathbb{E}[A_{M,L}] = A_{\infty,L}$. Lemma 4.1 implies that each of these entries concentrates around their mean:

$$\begin{aligned}
\mathbb{P}\left\{|\xi_{M,L}(f_{\psi_p}) - b_{\infty,L}(k)| > \frac{\epsilon}{\sqrt{n}}\right\} &\le 4e^{-\frac{M\epsilon^2}{8nC^2}}, \\
\mathbb{P}\left\{|\eta_{M,L}(f_{\psi_p}, f_{\psi'_p}) - A_{M,L}(k,k')| > \frac{\epsilon}{\sqrt{n}}\right\} &\le 2e^{-\frac{M\epsilon^2}{2nC^2}}.
\end{aligned}$$

where the constant C is obtained from (4.12). In combination of the basic fact that $\|b\| \le \sqrt{n}\max_{k=1,\dots,n}|b(k)|$ for any $b \in \mathbb{R}^n$, and $\|A\| \le \sqrt{n}\max_{k,k'=1,\dots,n}|A(k,k')|$ for any $A \in \mathbb{R}^{n\times n}$, we obtain

$$\begin{aligned}
\mathbb{P}\left\{\left\|b_{M,L} - b_{\infty,L}\right\| > \epsilon\right\} &\le \sum_k \mathbb{P}\left\{|\xi_{M,L}(f_{\psi_p}) - b_{\infty,L}(k)| > \frac{\epsilon}{\sqrt{n}}\right\} \le 4ne^{-\frac{M\epsilon^2}{8nC^2}}, \\
\mathbb{P}\left\{\left\|A_{M,L} - A_{\infty,L}\right\| > \epsilon\right\} &\le \sum_{k,k'} \mathbb{P}\left\{|\eta_{M,L}(f_{\psi_p}, f_{\psi'_p}) - A_{\infty,L}(k,k')| > \frac{\epsilon}{\sqrt{n}}\right\} \\
&\le 2n^2 e^{-\frac{M\epsilon^2}{2nC^2}}.
\end{aligned}$$

The third exponential inequality follows directly by combining the first two. □

Corollary 4.2 *Denote $\lambda_{\min}(A_{ML})$ the smallest eigenvalue of the normal matrix A_{ML} defined in (4.6). We have*

$$\mathbb{P}\{\lambda_{\min}(A_{ML}) > c_{\mathcal{H}} - \epsilon\} > 1 - \delta$$

with $\delta = 2n^2\exp\left(-\frac{M\epsilon_1^2}{2nc_1^2}\right)$, for any $\epsilon_1 > 0$ and any $\Delta t = T/L$ such that $\epsilon_1 + c_3\Delta t^{\frac{1}{2}} = \epsilon < c_{\mathcal{H}}$, where c_1 and c_3 are defined in (4.2).

Proof Note that for any $a \in \mathbb{R}^n$ such that $\|a\| = 1$, we have, by Corollary 4.1, $a^T A_{\infty,L} a \ge c_{\mathcal{H}} - c_3\Delta t^{\frac{1}{2}}$. Meanwhile, Proposition 4.2 implies that

$$\|a^T A_{M,L} a - a^T A_{\infty,L} a\| \le \|A_{M,L} - A_{\infty,L}\| \le \epsilon_1$$

with probability at least $1-\delta$. Thus,

$$\|a^T A_{M,L} a\| \geq \|a^T A_{\infty,L} a\| - \epsilon_1 \geq c_{\mathcal{H}} - c_2 \Delta t^{\frac{1}{2}} - \epsilon_1,$$

and the corollary follows. □

Remark 4 The above corollary requires $\epsilon > c_3 \Delta t^{\frac{1}{2}}$. This condition can be removed if the coercivity holds for the discrete-time observations on $\mathcal{H}$ with a constant $c_{\mathcal{H},T,L}$, which can be tested numerically from a data set with a large M. In fact, we obtain directly from the above proof that $\mathbb{P}\left\{\lambda_{\min}(A_{ML}) > c_{\mathcal{H},T,L} - \epsilon\right\} > 1-\delta$ with $\delta = 2n^2 \exp\left(-\frac{M\epsilon_1^2}{2nc_1^2}\right)$, for any $\epsilon > 0$.

Remark 5 In practice, the minimum eigenvalue of A_∞ may be small due to the redundancy of the local basis functions or due to the coercivity constant on $\mathcal{H}$ being small. Thus, the smallest eigenvalue of $A_{M,L}$ may be zero. On the other hand, these matrices are always symmetric and nonnegative, so it is advisable to regularize the matrix by pseudo-inverse.

5 Examples and Numerical Simulation Results

In this section, we performed numerical experiment to validate that our estimator defined in (2.5) and implemented by Algorithm 1, behaves in practice as predicted by the theory. We consider two examples: a stochastic opinion dynamical system and a stochastic Lennard-Jones system, using observations from simulated data.

The setup for the numerical simulations is as follows. We simulate sample paths on the time interval $[0, T]$ with the standard Euler–Maruyama scheme (see (2.3)), with a sufficiently small time step length $\mathrm{d}t$. When observations are made at every time step, i.e., $\Delta t = t_{l+1} - t_l = \mathrm{d}t$ for each l, we view $\boldsymbol{X}_{\text{train},M} := \{\boldsymbol{X}_{t_0:t_L}^{(m)}\}_{m=1}^{M}$ as continuous-time trajectories. When observations occur spaced in time with observation gap Δt equal to an integer multiple of $\mathrm{d}t$, we refer to them here as discrete-time observations.

From the observations, we construct the empirical probability measure $\rho_T^{L,M}$ (defined in (2.10)), and let $[R_{\min}, R_{\max}]$ be its support. We choose the hypothesis spaces $\mathcal{H}$ consisting of piecewise constant or piecewise linear polynomials on interval-based partitions of $[R_{\min}, R_{\max}]$. This choice is dictated by the ease of obtaining an orthonormal basis for $\mathcal{H}$, ease and efficiency of computation, and ability to capture local features of the interaction kernel. To avoid discontinuities at the extremes of the intervals in the partition and to reduce stiffness of the equations of the system with the estimated interaction kernels, we interpolate the estimator linearly on a fine grid and extrapolate it with a constant to the left of $R_{\min}$ and the right of $R_{\max}$. This post-processing procedure ensures the Lipschitz continuity of the estimators. We use the post-processed estimators to predict and generate the dynamics with the estimated interaction kernels.

We mainly focus on the case where T is small and report on the results as follows:

- **Interaction kernel estimation.** We compare ϕ and $\widehat{\phi}_{T,M,\mathcal{H}}$, the true and estimated interaction kernels (after smoothing), by plotting them side by side, superimposed with an estimate of ρ_T, obtained as in (2.10) by using M_{ρ_T} ($M_{\rho_T} \gg M$) independent trajectories. The estimated kernel is plotted in terms of its mean and standard deviation, computed over 10 independent learning trials. To demonstrate the dependence of the estimator on the sample size and the scale of the random noise, we report the above for different values of M and σ.
- **Trajectory prediction.** In the spirit of Proposition 2.1, we compare the discrepancy between the true trajectories (evolved using ϕ) and predicted trajectories (evolved using $\widehat{\phi}_{T,M,\mathcal{H}}$) on both the training time interval $[0, T]$ and a future time interval $[T, T_f]$, over two different sets of initial conditions—one taken from the training data, and one consisting of new samples from μ_0. When simulating the trajectories for the systems driven by $\widehat{\phi}_{T,M,\mathcal{H}}$ using the EM scheme, we use the same initial conditions and the same realization of the random noise as in the trajectory of the system driven by ϕ. The mean trajectory error is estimated using M test trajectories (the same number as in the training data).
- **Rate of convergence.** We report the convergence rate of $\widehat{\phi}_{T,M,\mathcal{H}}$ to ϕ in the $|||\cdot|||$ norm on $L^2(\rho_T)$ as the sample size M increases, with the dimension of $\mathcal{H}$ growing with M according to Theorem 3.2, for different scales σ of the random noise. We also investigate numerically the convergence rate when both T and M increase, with the dimension of the hypothesis space $\mathcal{H}$ set according to the effective sample size as discussed in Sect. 2.2.
- **Discretization errors from discrete-time observations.** To study the discretization error due to discrete-time observations, we report the convergence rate (in M) of estimators $\widehat{\phi}_{L,T,M,\mathcal{H}}$ obtained from data with different observation gaps $\Delta t = T/L$. We also verify numerically that the $|||\cdot|||$ error of the estimators increases with Δt as predicted by Theorem 4.2. These experiments are carried out for different values of the square root of the diffusion constant σ.

We will report the conclusions of our experiments in Sect. 5.3

5.1 Example 1: Stochastic Opinion Dynamics

We first consider a 1D system of stochastic opinion dynamics with interaction kernel

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

It is straightforward to see that ϕ is in $C_c^{1,1}([0, 2])$ and nonnegative. Systems of this form are motivated in various applications, from biology to in social science, where ϕ models how the opinions of people influence each other (see [7,11,18,33,43] and references therein), with one or a multiplicity of consensuses may be eventually reached.

Table 3 (OD) Parameters for the system

d	N	M_{ρ_T}	$\mathrm{d}t$	$[0; T; T_f]$	μ_0	$\deg(\psi)$	n
1	10	$5 \cdot 10^4$	0.01	[0; 5; 50]	$\mathcal{U}([0, 8])^{\otimes N}$	0	$40(\frac{M}{\log M})^{\frac{1}{3}}$

In the system we consider, each agent tries to align its opinions more with its farther neighbors than with its closer neighbors: such interactions are called heterophilious. For deterministic systems of this type, [43] shows that the opinions of agents merge into clusters, with the number of clusters significantly smaller than the number of agents. This is natural, as increased alignment with farther neighbors increases mixing and consensus. In our stochastic setting, the random noise prevents the opinions from converging to single opinions. Instead, soft clusters form at large time, that are metastable states for the dynamics, i.e., states where agents dwell for long times, rarely switching between them.

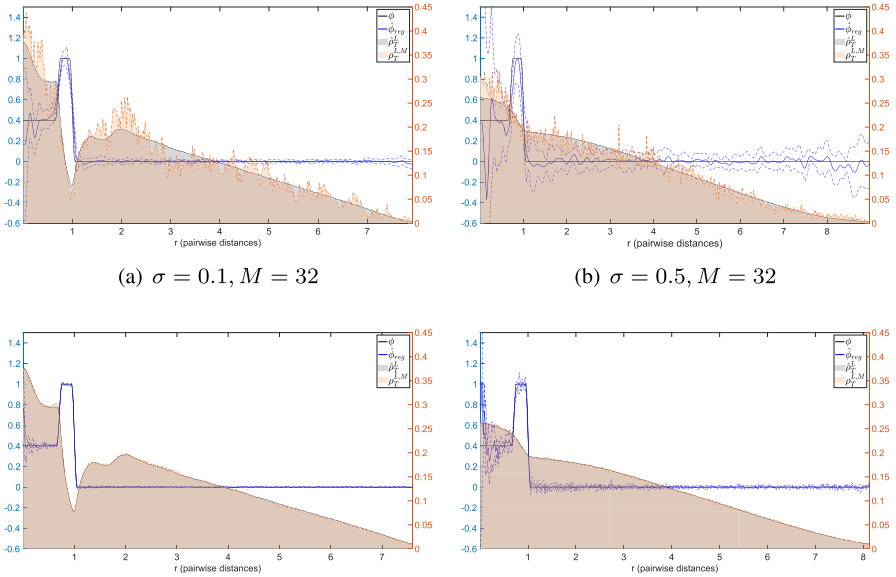
We study the performance of our estimators of the interaction kernel, from trajectory data. Table 3 summarizes parameters of the setup. In this example, we choose $\mathcal{H}_{n_M}$ to be the function space consisting of piecewise constant functions on n_M uniform partitions of the interval [0, 10].

Figure 2 shows that, as the number of trajectories increases, we obtain increasingly accurate approximations to the true interaction kernel, including at locations with sharp transitions of ϕ. The lack of artifacts at these locations is an advantage provided by the use of local basis. The estimators oscillate near 0, with amplitudes scaling with the level of noise. We believe that the reason for this phenomenon is that due to the structure of the equations, we have terms of the form $\phi(0)\mathbf{0} = \mathbf{0}$ at, and near, 0, with subsequent loss of information about the interaction kernel about 0.

We then use the learned interaction kernels $\widehat{\phi}$ in Fig. 2 to predict the dynamics and summarize the results in Fig. 3 and Table 4. Even with $M = 32$, our estimator produces very accurate approximations of the true trajectories both in the training time interval [0, 5] and the future time interval [5, 50], including number and location of clusters, and the time of their formation. As M increases to 4096, we have more accurate predictions on the locations of clusters. We impute this improvement to the better reconstruction of estimators at locations near 0.

Next we investigate the convergence rate of estimators. It is well known in approximation theory (see Theorem 6.1 in [48]) that $\inf_{\varphi \in \mathcal{H}_n} \|\varphi - \phi\|_\infty \leq \mathrm{Lip}[\phi] n^{-1}$. With the dimension n being proportional to $(\frac{M}{\log M})^{\frac{1}{3}}$, Fig. 4 shows that the learning rate in terms of M is around $M^{-0.34}$, which matches the optimal min–max rate $M^{-\frac{1}{3}}$ stated in Theorem 3.2 with $s = 1$.

We also study the convergence of the estimator as the length of the trajectory T increases, for the estimator $\widehat{\phi}_{T,M,\mathcal{H}}$ from continuous-time trajectories (i.e., without gaps between observations). The auto-correlation time for this system is estimated to be about $\tau = 10$ time units. Therefore, we use relatively long trajectories up to $T = 1500$ time units to test the convergence, contributing up to about 150 effective samples. We set the dimension of the hypothesis space to be $n = 4(\frac{MT/\mathrm{d}t}{\log(MT/\mathrm{d}t)})^{\frac{1}{3}}$ for each pair (M, T), where $\mathrm{d}t$ is the time step size of the Euler–Maruyama scheme.



(a) $\sigma = 0.1, M = 32$ (b) $\sigma = 0.5, M = 32$

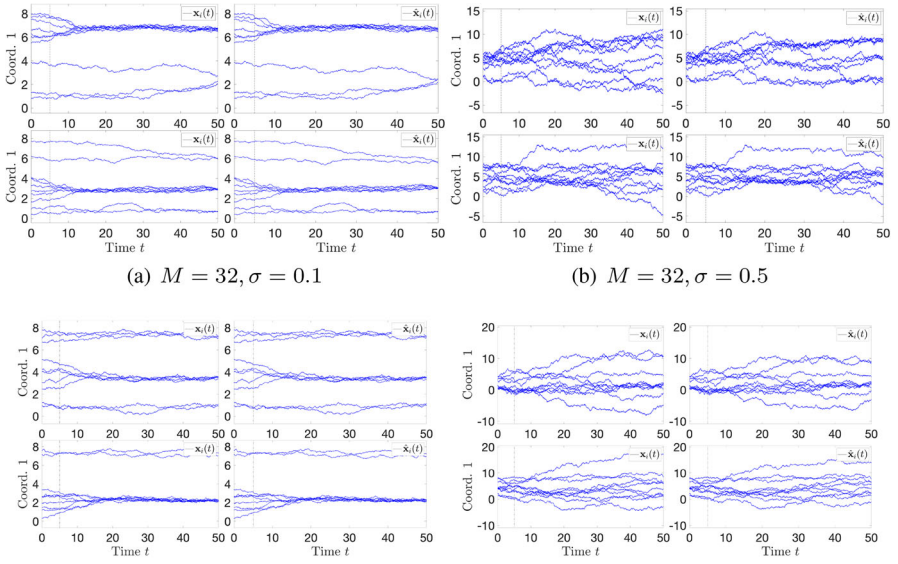
(c) $\sigma = 0.1, M = 4096$ (d) $\sigma = 0.5, M = 4096$

Fig. 2 Stochastic opinion dynamics: comparison between true and estimated interaction kernels $\widehat{\phi}_{T,M,\mathcal{H}}$ for different values of M and σ, together with histograms (shaded regions) for ρ_T and ρ_T^M. In black: the true interaction kernel. In blue: the mean of estimators in 10 independent trials, with dash lines representing the standard deviation. From top to bottom: learning from $M = 2^5, 2^{12}$ trajectories for kernels in systems with $\sigma = 0.1$ (left) and $\sigma = 0.5$ (right). The standard deviation bars on the estimated interaction kernels become smaller if M increases and σ decreases. The mean of the estimation error can be found in Fig. 4a (color figure online)

The convergence rate of the estimators in terms of MT is about 0.33, showing the equivalence of learning from a single long trajectory with multiple short trajectories when the underlying process is ergodic.

We also investigate the effects of the scale of the random noise, which is represented by the standard deviation σ. Figure 2 shows that the estimators for the system with $\sigma = 0.5$ have much large oscillations than those with $\sigma = 0.1$. The left plot in Fig. 4 shows that the scale of the random noise does not affect the learning rate, matching our theory. We also see that the absolute $L^2(\rho_T)$ error of estimators increases as the system noise increases; this may indicate that the coercivity constant decreases as the level of noise in the system increases. The left plot in Fig. 5 shows that the scale of the errors increase linearly in σ (in particular, when the observation gap is 1).

Finally, we study the discretization error due to approximation of the integral in the likelihood using discrete-time observations. In the left plot of Fig. 5, as the observation gap k increases, the learning rate curves become flat, due to the error induced by discretization of the likelihood function (2.1). The right plot shows that the absolute error of the estimator is dominated by $\sigma\, O((\Delta t)^{1/2})$.



(a) $M = 32, \sigma = 0.1$ (b) $M = 32, \sigma = 0.5$

(c) $M = 4096, \sigma = 0.1$ (d) $M = 4096, \sigma = 0.5$

Fig. 3 Stochastic opinion dynamics: trajectory prediction. In each panel: $\boldsymbol{X}_t$ (left column) and $\widehat{\boldsymbol{X}}_t$ (right column) obtained with the true kernel ϕ and the estimated interaction kernel $\widehat{\phi}_{T,M,\mathcal{H}}$ from $M = 32$ (top panel) and 4096 (bottom panel) trajectories, for an initial condition in the training data (top row in each panel) and a (new) initial condition randomly chosen from μ_0 (bottom row in each panel). The black dashed vertical line at $t = T = 5$ divides the "training" interval $[0, T]$ from the "prediction" interval [5,50]. As M increases, our estimators achieve better approximation of the true kernel overall, and at regions near 0 (see Fig. 2). As a result, they produced more faithful prediction of the number and location of clusters for large time. Statistics of trajectory prediction errors are reported in Table 4

Table 4 Stochastic opinion dynamics: means and standard deviations of trajectory prediction errors

	$[0, 5]$	$[5, 50]$
$M = 32, \sigma = 0.1$, $\text{mean}_{\text{traj}}$: Training ICs	$2.0 \cdot 10^{-1} \pm 1.4 \cdot 10^{-1}$	$6.3 \cdot 10^{-1} \pm 5.7 \cdot 10^{-1}$
$M = 32, \sigma = 0.1$, $\text{mean}_{\text{traj}}$: Random ICs	$1.7 \cdot 10^{-1} \pm 1.2 \cdot 10^{-1}$	$5.7 \cdot 10^{-1} \pm 3.9 \cdot 10^{-1}$
$M = 32, \sigma = 0.5$, $\text{mean}_{\text{traj}}$: Training ICs	$3.8 \cdot 10^{-1} \pm 1.7 \cdot 10^{-1}$	$4.0 \cdot 10^{0} \pm 2.3 \cdot 10^{0}$
$M = 32, \sigma = 0.5$, $\text{mean}_{\text{traj}}$: Random ICs	$3.6 \cdot 10^{-1} \pm 1.1 \cdot 10^{-1}$	$3.5 \cdot 10^{0} \pm 1.4 \cdot 10^{0}$
$M = 4096, \sigma = 0.1$, $\text{mean}_{\text{traj}}$: Training ICs	$2.1 \cdot 10^{-2} \pm 2.0 \cdot 10^{-2}$	$9.3 \cdot 10^{-2} \pm 1.6 \cdot 10^{-1}$
$M = 4096, \sigma = 0.1$, $\text{mean}_{\text{traj}}$: Random ICs	$2.1 \cdot 10^{-2} \pm 2.3 \cdot 10^{-2}$	$9.8 \cdot 10^{-2} \pm 1.7 \cdot 10^{-1}$
$M = 4096, \sigma = 0.5$, $\text{mean}_{\text{traj}}$: Training ICs	$5.2 \cdot 10^{-2} \pm 3.5 \cdot 10^{-2}$	$3.8 \cdot 10^{-1} \pm 3.0 \cdot 10^{-1}$
$M = 4096, \sigma = 0.5$, $\text{mean}_{\text{traj}}$: Random ICs	$5.2 \cdot 10^{-2} \pm 3.5 \cdot 10^{-2}$	$3.8 \cdot 10^{-1} \pm 3.0 \cdot 10^{-1}$

The tests with "Training ICs" use initial conditions from the training data set. The tests with "Random ICs" use initial conditions that are randomly drawn from μ_0. Means are taken over M trajectories. There is little difference between errors on training and test ICs, indicating the prediction of trajectories generalizes perfectly to new ICs

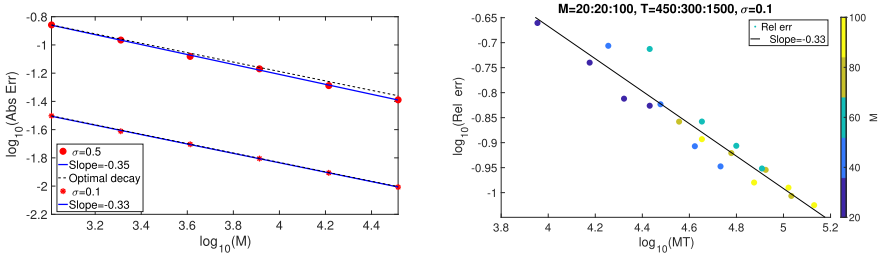


Fig. 4 Stochastic opinion dynamics: learning rates for continuous-time observations. Left: the convergence rate of the estimators in terms of M is 0.35 for $\sigma = 0.5$ and is 0.33 for $\sigma = 0.1$, close to the theoretical optimal min–max rate $1/3$ (shown in the black dot line). Right: the convergence rate of the estimators in terms of MT, when both M and T increase, is about 0.33. The colors of points are assigned according to M. The learning rate is still close to the theoretical optimal min–max rate $1/3$, showing the equivalence of learning from a single long trajectory with multiple short trajectories when the underlying process is ergodic (color figure online)

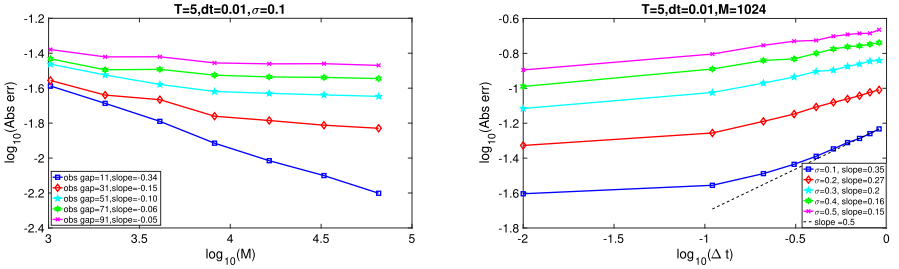


Fig. 5 Stochastic opinion dynamics: discretization error due to discrete-time observation. Left: the learning rates of estimators $\widehat{\phi}_{L,T,M,\mathcal{H}}$ obtained from data with different observation gaps $\Delta t = k\mathrm{d}t$ for k ranging from 11 to 100. Recall that $L = T/\Delta t$. As k increases, the learning rate curves become flat, due to the bias induced by discretization of the likelihood function (2.1) on coarse time grids. Right: the log–log plot of the absolute error of the estimator in terms of observation gap $\Delta t = k\mathrm{d}t$ for k ranging from 1 to 100, for systems with different levels of random noise in terms of σ, computed with $M = 1024$, $T = 5$ and $\mathrm{d}t = 0.01$ fixed. The orders of the absolute error in both σ and Δt are bounded by the theoretical order $\sigma\, O((\Delta t)^{1/2})$, dominating the statistical error due to sampling, finite-dimensional approximation, and noise. The slopes of the lines are calculated using points whose x coordinate fall in the range $[-1, 0]$

5.2 Example 2: Stochastic Lennard Jones Dynamics

In this example, we consider the Lennard-Jones-type kernel $\phi(r) = \frac{\Phi'(r)}{r}$, with

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

for some $p > q \in \mathbb{N}$. The system of particles is assumed to be associated with a potential energy function only depending on the pairwise distance and Φ, and the evolution is driven by minimization of the energy function. In particular, ϵ represents the depth of the potential well, r is the distance between the particles, and r_m is the distance at which the potential reaches its minimum. At r_m, the potential function has the value $-\epsilon$. The r^{-p} term, which is the repulsive term, describes Pauli repulsion

Table 5 (Stochastic LJ) Parameters for the Lennard-Jones kernel

p	q	ϵ	r_m	r_{trunc}
8	2	1	1	0.95

Table 6 (Stochastic LJ) Parameters for the system

d	N	M_{ρ_T}	dt	$[0; T; T_f]$	μ_0	deg(ψ)	n
2	10	$5 \cdot 10^4$	0.001	[0;0.5;20]	$\mathcal{N}(0, I)$	1	$30(\frac{M}{\log M})^{\frac{1}{5}}$

at short ranges due to overlapping electron orbitals, and the r^{-q} term, which is the attractive long-range term. The corresponding system has wide applications in molecular dynamics and material sciences where ϕ models atom–atom interactions. Note that ϕ is singular at $r = 0$: we truncate it at r_{trunc} by connecting it with an exponential function of the form $a \exp(-br^{12})$ so that it has a continuous derivative on $\mathbb{R}^+$.

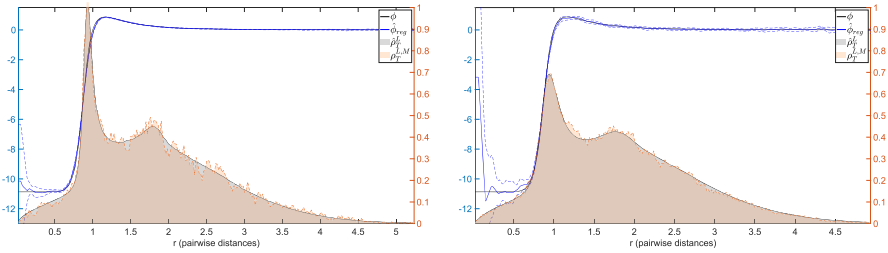
In this system, the particle–particle interactions are all short-range repulsions and long-range attractions. The short-range repulsion force prevents the particles to collide and long-range attractions keep the particles in the flock. In the deterministic setting, the system evolves to equilibrium configurations very quickly, which are crystal-like structure, whose pairwise distance corresponds to the local minimizers of the associated energy function. Tables 5 and 6 summarize the system and learning parameters.

Note that the true kernel ϕ is not compactly supported. But in our simulations, we observe the dynamics up to a time T which is a fraction of the equilibrium time. Since the particles only explore a bounded region due to the large-range attraction, ρ_T is essentially compactly supported on a bounded region (see the histogram background of Fig. 6), on which ϕ is in our admission space.

We use piecewise linear functions on n uniform partitions of the learning interval to approximate the true kernel ϕ. With $M = 32$, Fig. 6 shows that we have already obtained faithful approximations to the true interaction kernel, except for on regions are close 0. Increasing number of observations improves the accuracy of estimators at locations near 0, which seems to be very helpful for the system with larger noise level.

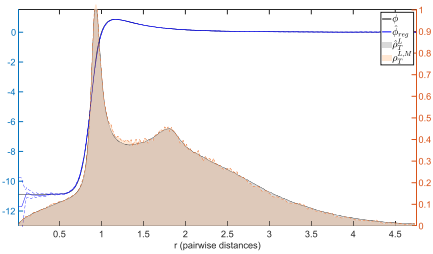
In terms of the trajectory prediction, we use the learned interaction kernels $\widehat{\phi}$ in Fig. 2. We summarize the results in Fig. 7 and Table 7. In the experiments, we study two cases, one with small random noise where the particles still form an equilibrium configuration, and then, this configuration has small fluctuation in the space; the other one with medium level of random noise, where the random noise begins to break the formation of a fixed equilibrium configuration and we see the transition between different configurations. We see that in both cases, our estimators produce good prediction of the true dynamics in both training and future time interval.

We plot the convergence rate of estimators in terms of M in the right plot of Fig. 8. In this case, we have $\inf_{\varphi \in \mathcal{H}_n} \|\varphi - \phi\|_\infty \leq \text{Lip}[\phi'] n^{-2}$. We choose a choice of dimension n proportional to $(\frac{M}{\log M})^{\frac{1}{5}}$, our numerical results show that the learning rate is around $M^{-0.39}$, which matches the optimal min–max rate $M^{-\frac{2}{5}}$ stated in Theorem 3.2.

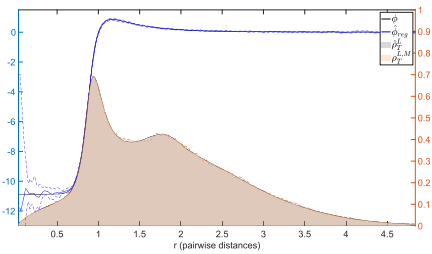


(a) $\sigma = 0.05, M = 128$

(b) $\sigma = 0.25, M = 128$



(c) $\sigma = 0.05, M = 1024$



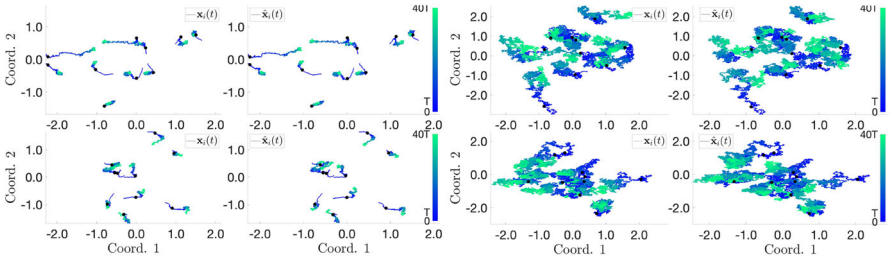
(d) $\sigma = 0.25, M = 1024$

Fig. 6 Stochastic Lennard-Jones dynamics: comparison between true and estimated interaction kernels with different values of M, together with histograms (shaded regions) for ρ_T and ρ_T^M. In black: the true interaction kernel. In blue: the mean of estimators in 10 independent trials, with dash lines representing the standard deviation. From top to bottom: learning from $M = 2^7, 2^{10}$ trajectories for kernels in systems with $\sigma = 0.05$ (left) and $\sigma = 0.25$ (right). The standard deviation bars on the estimated interaction kernels become smaller if M increases and σ decreases. More details of the estimation errors can be found in Fig. 8 a) (color figure online)

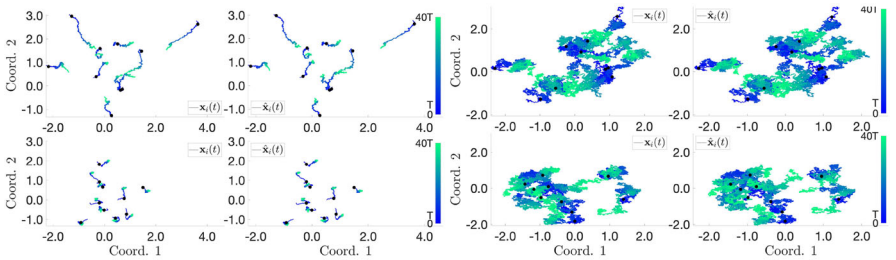
We also study the convergence of the estimators as the length of the trajectory T increases. In this example with $\sigma = 0.35$, the estimated auto-correlation time is about $\tau = 10$ time units. Therefore, we use relatively long trajectories up to $T = 1200$ time units, contributing up to about 120 effective samples. We set the dimension of the hypothesis space to be $n = 4(\frac{MT/\mathrm{d}t}{\log(MT/\mathrm{d}t)})^{\frac{1}{5}}$ for each pair (M, T), where $\mathrm{d}t$ is the time step size of the Euler–Maruyama scheme. The right plot of Fig. 8 shows that the rate is 0.39, indicating the equivalence between a single long trajectory and multiple short trajectories for inference.

Next, we investigate the effects of the scale of the random noise on learning. We observe phenomenon similar to those in Example 1. Figure 6 shows that the estimators for the system with $\sigma = 0.25$ oscillate more than the one with $\sigma = 0.05$ at locations near 0. The random noise also did not affect the learning rates, suggested by the left plot of Fig. 8. As the random noise increases, absolute $L^2(\rho_T)$ error of estimators also increase, suggesting that coercivity constant is getting smaller.

At last, we study the effects of discretization error induced by discrete observations. As the observation gap increases, the discretization errors flatten the learning rate curve of M, see left plot of Fig. 8. Similar to Example 1, the right plot of Fig. 8 shows that the absolute error of the estimator is of order close to the theoretical order $\sigma\, O((\Delta t)^{1/2})$ (Fig. 9).



(a) $\sigma = 0.05, M = 128$ (b) $\sigma = 0.25, M = 128$



(c) $\sigma = 0.05, M = 1024$ (d) $\sigma = 0.25, M = 1024$

Fig. 7 (Stochastic Lennard-Jones dynamics) In each panel: true trajectory $\boldsymbol{X}_t$ (left column) and learned trajectory $\widehat{\boldsymbol{X}}_t$ (right column) obtained with the true kernel ϕ and the estimated kernel $\widehat{\phi}$ from $M = 128$ and 1024 trajectories, for an initial condition in the training data (top row) and an initial condition randomly chosen (bottom row). The black dot at $t = 0.5$ divides the "training" interval [0, 0.5] from the "prediction" interval [0.5,20]. The trajectory prediction errors are small in all cases. The statistics of the errors are presented in Table 7

Table 7 (Stochastic Lennard-Jones dynamics) trajectory errors: ICs used in the training set (first two rows), new ICs randomly drawn from μ_0 (second set of two rows)

	[0, 0.5]	[0.5, 20]
$M = 128, \sigma = 0.05$, $\text{mean}_{\text{traj}}$: Training ICs	$3.1 \cdot 10^{-2} \pm 8.3 \cdot 10^{-3}$	$3.0 \cdot 10^{-1} \pm 3.9 \cdot 10^{-1}$
$M = 128, \sigma = 0.05$, $\text{mean}_{\text{traj}}$: Random ICs	$3.1 \cdot 10^{-2} \pm 9.3 \cdot 10^{-3}$	$3.1 \cdot 10^{-1} \pm 4.2 \cdot 10^{-1}$
$M = 128, \sigma = 0.25$, $\text{mean}_{\text{traj}}$: Training ICs	$5.5 \cdot 10^{-1} \pm 2.4 \cdot 10^{-2}$	$1.3 \cdot 10^{0} \pm 7.5 \cdot 10^{-1}$
$M = 128, \sigma = 0.25$, $\text{mean}_{\text{traj}}$: Random ICs	$5.8 \cdot 10^{-2} \pm 2.3 \cdot 10^{-2}$	$1.3 \cdot 10^{0} \pm 7.3 \cdot 10^{-1}$
$M = 1024, \sigma = 0.05$, $\text{mean}_{\text{traj}}$: Training ICs	$1.2 \cdot 10^{-2} \pm 3.4 \cdot 10^{-3}$	$1.7 \cdot 10^{-1} \pm 2.7 \cdot 10^{-1}$
$M = 1024, \sigma = 0.05$, $\text{mean}_{\text{traj}}$: Random ICs	$1.2 \cdot 10^{-2} \pm 3.6 \cdot 10^{-3}$	$1.5 \cdot 10^{-1} \pm 2.5 \cdot 10^{-1}$
$M = 1024, \sigma = 0.25$, $\text{mean}_{\text{traj}}$: Training ICs	$2.2 \cdot 10^{-2} \pm 6.2 \cdot 10^{-3}$	$3.2 \cdot 10^{-1} \pm 3.7 \cdot 10^{-1}$
$M = 1024, \sigma = 0.25$, $\text{mean}_{\text{traj}}$: Random ICs	$2.2 \cdot 10^{-2} \pm 6.4 \cdot 10^{-3}$	$3.2 \cdot 10^{-1} \pm 3.5 \cdot 10^{-1}$

Means are taken over the same number of trajectories as in the training data set

5.3 Conclusions from the Numerical Experiments

Numerical results show that in case of continuous-time observations, the algorithm effectively estimates the interaction kernel, achieves the near-optimal learning rate in M, is robust to different magnitudes of the random noise, and the system with the

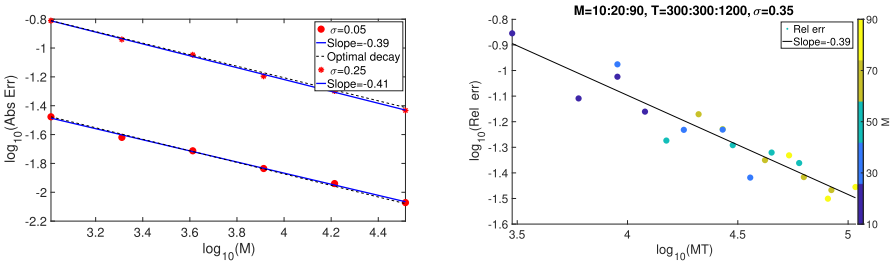


Fig. 8 Stochastic Lennard–Jones: learning rates for continuous-time observations. Left: the learning rate of the estimators in terms of M is 0.39 when $\sigma = 0.05$ and is 0.41 when $\sigma = 0.25$, close to the theoretical optimal min–max rate $2/5$ (shown in the black dot line). Right: the convergence rate of the estimators in terms of MT, when both M and T increase. The colors of points are assigned according to M. The rate is still close to the optimal min–max rate $2/5$, showing the equivalence of learning from a single long trajectory with multiple short trajectories when the underlying process is ergodic (color figure online)

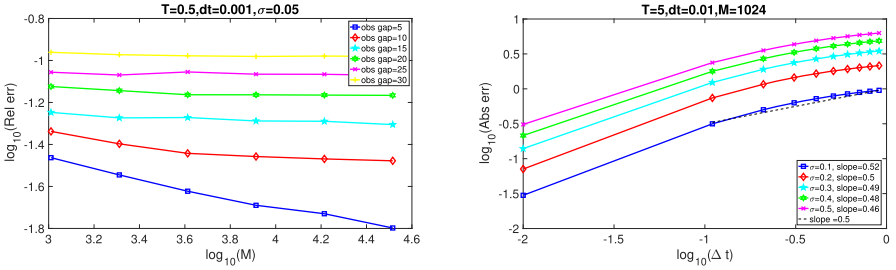


Fig. 9 Stochastic Lennard-Jones: discretization error due to discrete-time observation. Left: The learning rate of estimators in terms of different observation gap $\Delta t = k\mathrm{d}t$ for $k = 5 : 5 : 30$. The learning rate becomes flat, due to the bias induced by discretization of the likelihood function on coarse time grids. Right: the log–log plot of the absolute error of the estimator in terms of observation gap $\Delta t = k\mathrm{d}t$ for $k = 5 : 5 : 45$, for systems with different levels of random noises in terms of σ, computed with $M = 1024$, $T = 0.5$ and $\mathrm{d}t = 0.001$ fixed. The orders of the absolute error in both σ and Δt are close to the theoretical order $\sigma\, O((\Delta t)^{1/2})$. The slopes of the lines are calculated using points whose x coordinate fall in the range $[-1, 0]$

estimated kernels accurately predicts trajectories. In case of discrete-time observations, the estimator has an estimation error of order $\Delta t^{1/2}$, due to the discretization error in the approximation of the likelihood ratio. These numerical results are in full agreement with the learning theory in Sects. 3–4:

- In case of continuous-time observations, the estimators in 10 trials are faithful approximations of the true interaction kernels, with a mean close to the truth. The standard deviation of the estimators decreases as the sample size increases and gets larger as the diffusion constant increases.
- The estimator from data achieves the min–max learning rate $(\log M/M)^{s/(2s+1)}$ in Theorem 3.2 by the appropriate choice of the hypothesis spaces and their dimension as a function of M. For ϕ in $C^{k+\alpha}$ with $k + \alpha \geq 2$, the learning rate is around $M^{-\frac{1}{3}}$ when using piecewise constant estimators ($s = 1$) and the learning rate is around $M^{-\frac{2}{5}}$ using the piecewise linear estimators ($s = 2$), which is the mini–max optimal rate for the case $k + \alpha = 2$.

– The estimators predict transient dynamics well in the training time interval, and the results validate Proposition 2.1: the trajectory discrepancy is controlled by $L^2(\rho_T)$ error of estimators, demonstrating the effectiveness of distances in $L^2(\rho_T)$ in quantifying the performance of estimators. In addition, the estimators even predict in a remarkably accurate fashion the collective behavior of particles in larger future time intervals, indicating that the bound in Proposition 2.1 may be overly pessimistic in some cases. Our intuition is that this benign phenomenon benefits from the large support of ρ_T, encouraged by the randomness of the initial conditions and presence of stochastic noise.
– In case of discrete-time observations with observation gap Δt, the estimation error of the estimator is of order $\Delta t^{1/2}$ and depends linearly on σ, the square root of the diffusion constant. Therefore, as Δt increases, the discretization error dominates the estimation error, consistently with the learning theory in Sect. 4, which leads to bounding the estimation error of the estimator by $M^{-\frac{s}{2s+1}} + \sigma\mathcal{O}(\Delta t^{1/2})$.
– When the length T of the trajectories increases, the optimal learning rate (in M) is still achieved. The estimation errors of the estimator exhibit a convergence rate around $(\frac{\log(MT)}{MT})^{s/(2s+1)}$ with $s = 1, 2$ respectively, demonstrating an equivalence of "information" between few long trajectories and many short trajectories initiated at suitably random initial conditions, as discussed above in Sect. 2.3.

6 Final Remarks and Future Work

There are many venues in which the present work could be extended.

The first notable extension is to heterogeneous particle systems with multiple types of particles, which arise in many applications. In this case, one assumes that there are different interaction kernels, modeling the non-symmetric interactions between different types of particles. Examples of these systems are considered in [41] in the deterministic case, with the theoretical analysis achieved in [40], where the coercivity condition is generalized to the multiple-particle-types setting, and (near-)optimal convergence rates of the estimators were established. We believe a similar extension is possible in the stochastic case, combining the ideas of this work and [40].

Another notable extension is to second-order differential systems of interacting particles or systems with possible external potentials, where interaction kernels of more general forms than those considered here arise. In the deterministic case, [41] considers examples of such systems, with a forthcoming theoretical analysis. In the stochastic case, the extension would require significant effort, especially if important cases of systems with degenerate diffusion (e.g., stochastic Langevin) were considered. We also remark again that in this work we do not observe velocities, as done in the works just cited in the case of deterministic systems: here we fully take into account the discretization (in time) error, and if we let $\sigma \to 0$, the results here would imply similar results in the deterministic case. Extending these considerations to second-order systems would be valuable.

Further work is also needed to formalize the considerations we put forward in Sect. 2.3 regarding ergodic systems, and design robust and optimal algorithms in the regimes of observation a long trajectory or many independent trajectories.

We assume in this work that all particles are observed. A desirable extension is to the case of partial observations of a subset of particles or macroscopic observations of the population density, which is a practical concern when the system is large with millions of particles in high dimension. Since it is an ill-posed inverse problem to recover the missing trajectories of unobserved particles [54], a new formulation based on the corresponding mean field equations [27,28,43] is under investigation.

In this work, we assume that the noise coefficient is a known constant: there has been of course significant work in estimating the noise coefficient, for example in the case of interacting particle systems see the recent work [26] and references therein, and for the case of model reduction for Langevin equations with state-dependent diffusion coefficient [19].

Acknowledgements We thank the anonymous referee for their helpful comments. FL and MM are grateful for partial support from NSF-1913243, FL for NSF-1821211; MM for NSF-1837991, NSF-1546392, AFOSR-FA9550-17-1-0280 and the Simons Fellowship; ST for AMS Simons travel grant and a start-up fund sponsored by University of California Santa Barbara.

A Appendix

A.1 Preliminaries for SDEs

Let $(\boldsymbol{X}_t, t \geq 0)$ be a stochastic process on $\mathbb{R}^n$ satisfying

$$\mathrm{d}\boldsymbol{X}_t = V(\boldsymbol{X}_t, t)\mathrm{d}t + \sigma(\boldsymbol{X}_t, t)\mathrm{d}\boldsymbol{B}_t. \tag{A.1}$$

We first review the existence and uniqueness of strong solutions for SDEs (see Theorem 5.4 in [32])

Theorem A.1 (Existence and Uniqueness) *If the following conditions are satisfied*

– *The coefficients V and σ are locally Lipschitz in $\boldsymbol{x}$ uniformly in t, that is for every T and K, there is a constant C depend only on T and K such that for all $\|\boldsymbol{x}\|, \|\boldsymbol{x}\| \leq K$ and all $0 \leq t \leq T$*

$$\|V(\boldsymbol{x}, t) - V(\boldsymbol{x}, t)\| + \|\sigma(\boldsymbol{x}, t) - \sigma(\boldsymbol{Y}, t)\| < C\|\boldsymbol{x} - \boldsymbol{x}\|. \tag{A.2}$$

– *Coefficients satisfy the linear growth condition*

$$\|V(\boldsymbol{x},t)\| + \|\sigma(\boldsymbol{x},t)\| < C(1+\|\boldsymbol{x}\|). \tag{A.3}$$

– $\boldsymbol{X}_0$ *is independent of* $(\boldsymbol{B}_t, 0 \le t \le T)$, *and* $\mathbb{E}\|\boldsymbol{X}_0\|^2 < \infty$.

Then, there exists a unique strong solution $\boldsymbol{X}_t$ *of the SDEs (A.1).* $\boldsymbol{X}_t$ *has continuous paths; moreover,*

$$\mathbb{E}[\sup_{0\le t\le T} \|\boldsymbol{X}_t\|^2] < C_1(1+\mathbb{E}[\|\boldsymbol{X}_0\|^2]),$$

where constants C_1 *depend only on* C *and* T.

It is straightforward to show that $\mathbf{f}_\phi$ satisfy (A.2) and (A.3). Therefore, suppose μ_0 is independent of the underlying Brownian motion and has finite second moment; then, there exists a unique strong solution up to time T for system (1.1) for any $\boldsymbol{X}_0$ drawn from μ_0.

Theorem A.2 (Girsonov theorem) *Let* P_σ *be the probability measure induced by the solution of the SDEs (A.1) for* $t \in [T_0, T]$ *and a fixed starting value at time* T_0, *and let* W_σ *be the law of the respective driftless process. Suppose that* $\Sigma = \sigma\sigma'$ *is invertible and* V *fulfills the Novikov condition*

$$\mathbb{E}_{P_\sigma}\left[\exp\left(\frac{1}{2}\int_{T_0}^{T} \|V(\boldsymbol{X}_t,t)\|^2 \mathrm{d}t\right)\right] < \infty.$$

Then P_σ *and* W_σ *are equivalent measures with Radon–Nikodym derivative given by Girsonov's formula*

$$\frac{\mathrm{d}P_\sigma}{\mathrm{d}W_\sigma}(\boldsymbol{X}_{[T_0,s]}) = \exp\left(\int_{T_0}^{s} V^T\Sigma^{-1}\mathrm{d}\boldsymbol{X}_t - \frac{1}{2}\int_{T_0}^{s} V^T\Sigma^{-1}V\mathrm{d}t\right)$$

for all $s \in [T_0, t]$ *and* $\boldsymbol{X}_{[T_0,s]} = (\boldsymbol{X}_t)_{t\in[T_0,s]}$.

The proof of Theorem A.2 can be found in [31, Chapter 3.5], [45, Chapter 8.6].

Theorem A.3 (The Itô formula, see Theorem 4.1.2 in [45]) *Let* $g : \mathbb{R}^n \to \mathbb{R}$ *be a* C^2 *map and* (X_t) *be a solution to (A.1) with* σ *being a constant. Then, the process* $Y(t) = g(\boldsymbol{X}_t)$ *is an Itô process satisfying*

$$\mathrm{d}Y = \sum_{i=1}^{n} \frac{\partial g}{\partial \boldsymbol{x}_i}(\boldsymbol{X}_t)\mathrm{d}\boldsymbol{X}_i + \frac{1}{2}\sum_{i,j} \frac{\partial^2 g}{\partial \boldsymbol{x}_i \partial \boldsymbol{x}_j}(\boldsymbol{X}_t)\sigma^2\mathrm{d}t.$$

A.2 Useful Inequalities

Theorem A.4 (Bernstein inequality for unbounded random variables) *Let* $X_1, X_2, \ldots,$ X_M *be independent random variables with* $\mathbb{E}(X_i) = 0$. *If for some constants* $K_1, v_1 >$

0, *the bound* $\mathbb{E}|X_i|^p \leq \frac{1}{2}p!K_1^{p-2}v_1$ *holds for every* $2 \leq p \in \mathbb{N}$, *then*

$$\mathbb{P}\left\{\sum_{i=1}^{M} X_i \geq \epsilon\right\} \leq e^{-\frac{\epsilon^2}{2}(Mv_1+K_1\epsilon)^{-1}}. \tag{A.4}$$

For the proof of Theorem A.4 , we refer to [4] and David Pollard's book notes [47](page 14).

Corollary A.1 *Denote* $\mathcal{E}_M(g) = \frac{1}{M}\sum_{m=1}^{M} g(X_m)$ *for a measurable function* g. *If for some* $K_2, v_2 > 0$, *the bound*

$$\mathbb{E}|g - \mathbb{E}g|^p \leq \frac{1}{2}p!K_2^{p-2}v_2$$

holds for $2 \leq p \in \mathbb{N}$, *then there holds*

$$\mathbb{P}\{\mathbb{E}g - \mathcal{E}_M(g) \geq \epsilon\} \leq e^{-\frac{M\epsilon^2}{2(v_2+K_2\epsilon)}}, \forall \epsilon > 0. \tag{A.5}$$

Proof Applying Theorem A.4 on the random variable $\mathbb{E}g - g$, we immediately obtain the desired bound. □

Corollary A.2 *If for some* $K_3, v_3 > 0$, *the bound*

$$\mathbb{E}|g - \mathbb{E}g|^p \leq \frac{1}{2}p!K_3^{p-2}v_3|\mathbb{E}g|$$

holds for $2 \leq p \in \mathbb{N}$, *then*

$$\mathbb{P}\left\{\mathbb{E}g - \mathbb{E}_M(g) \geq \sqrt{\epsilon}\sqrt{\epsilon + |\mathbb{E}g|}\right\} \leq e^{-\frac{M\epsilon}{2(v_3+K_3)}}, \forall \epsilon > 0$$

Proof If we replace ϵ with $\sqrt{\epsilon(\epsilon + |\mathbb{E}g|)}$ in (A.5), and let $K_2 = K_3$, $v_2 = v_3|\mathbb{E}g|$, the desired bound follows from the inequality

$$\begin{aligned} & e^{-\frac{M\epsilon(\epsilon+|\mathbb{E}g|)}{2(v_2+K_2\sqrt{\epsilon(\epsilon+|\mathbb{E}g|)})}} \leq e^{-\frac{M\epsilon}{2(v_3+K_3)}} \\ \Leftrightarrow\ & v_3\epsilon + K_3(\epsilon + |\mathbb{E}g|) \geq K_3\sqrt{\epsilon(\epsilon + |\mathbb{E}g|)}, \end{aligned}$$

where the last inequality is true since $\sqrt{\epsilon(\epsilon + |\mathbb{E}g|)} \leq \epsilon + |\mathbb{E}g|$ for all $\epsilon \geq 0$. □

We also refer to [52] (see its Lemma 3 and Lemma 5) for the analog of Corollary A.1 and A.2.

Theorem A.5 (Moment inequality for stochastic integrals, see Theorem 7.1 in [42]) *Let* $\mathcal{M}^2([0, T]; \mathbb{R}^{n\times m})$ *denote the family of all* $n \times m$*-matrix-valued measurable* $\{\mathcal{F}_t\}_{t\geq t_0}$

-adapted process $f = \{(f_{ij}(t))_{n\times m}\}_{0\le t\le T}$ *such that* $\mathbb{E}\int_0^T \|f(t)\|^2 \mathrm{d}t < \infty$. *If* $p \ge 2$, $f \in \mathcal{M}^2([0,T];\mathbb{R}^{n\times m})$ *such that*

$$\mathbb{E}\int_0^T \|f(t)\|^p \mathrm{d}t < \infty,$$

then

$$\mathbb{E}\left\| \int_0^T f(s)\mathrm{d}\boldsymbol{B}(s)\right\|^p \le \left(\frac{p(p-1)}{2}\right)^{\frac{p}{2}} T^{\frac{p-2}{2}} \mathbb{E}\int_0^T \|f(s)\|^p \mathrm{d}s$$

In particular, for $p = 2$, *there is equality.*

A.3 Proof of Proposition 2.1

Proof of Proposition 2.1 For ease of notation, in this proof we use $\mathbb{E}$ to represent $\mathbb{E}_{\mu_0,\boldsymbol{B}}$. For every $t \in [0,T]$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t\|^2\right] &= \mathbb{E}\left[\left\| \int_0^t \mathbf{f}_\phi(\boldsymbol{X}(s)) - \mathbf{f}_{\widehat{\phi}}(\widehat{\boldsymbol{X}}(s))\mathrm{d}s\right\|^2\right] \\
&\le t\mathbb{E}\left[\int_0^t \left\|\mathbf{f}_\phi(\boldsymbol{X}(s)) - \mathbf{f}_{\widehat{\phi}}(\widehat{\boldsymbol{X}}(s))\right\|^2 \mathrm{d}s\right] \\
&\le 2T\mathbb{E}\left[\int_0^t \left\|\mathbf{f}_\phi(\boldsymbol{X}(s)) - \mathbf{f}_{\widehat{\phi}}(\boldsymbol{X}(s))\right\|^2 \mathrm{d}s\right] \\
&\quad + 2T\mathbb{E}\left[\int_0^t \left\|\mathbf{f}_{\widehat{\phi}}(\boldsymbol{X}(s)) - \mathbf{f}_{\widehat{\phi}}(\widehat{\boldsymbol{X}}(s))\right\|^2 \mathrm{d}s\right].
\end{aligned}$$

Letting $\boldsymbol{x}_{ji}(s) := \boldsymbol{x}_j(s) - \boldsymbol{x}_i(s)$, $\widehat{\boldsymbol{x}}_{ji}(s) := \widehat{\boldsymbol{x}}_j(s) - \widehat{\boldsymbol{x}}_i(s)$, and $F_\varphi(\boldsymbol{x}) = \varphi(\|\boldsymbol{x}\|)\boldsymbol{x}$, for $\varphi \in \mathcal{K}_{R,S}$ and $\boldsymbol{x} \in \mathbb{R}^d$,

$$\begin{aligned}
\left\|\mathbf{f}_{\widehat{\phi}}(\boldsymbol{X}(s)) - \mathbf{f}_{\widehat{\phi}}(\widehat{\boldsymbol{X}}(s))\right\|^2 &= \sum_{i=1}^N \left\| \frac{1}{N}\sum_{j=1}^N \left(F_{[\widehat{\phi}]}(\boldsymbol{x}_{ji}(s)) - F_{[\widehat{\phi}]}(\widehat{\boldsymbol{x}}_{ji}(s))\right)\right\|^2 \\
&\le 4\mathrm{Lip}^2(F_{[\widehat{\phi}]}) \left\|\boldsymbol{X}(s) - \widehat{\boldsymbol{X}}(s)\right\|^2, \quad \text{almost surely.}
\end{aligned}$$

Then, an application of Gronwall's inequality yields the estimate

$$\mathbb{E}\left[\left\|\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t\right\|^2\right] \le 2Te^{8T^2\mathrm{Lip}^2(F_{[\widehat{\phi}]})}\mathbb{E}\left[\int_0^T \left\|\mathbf{f}_\phi(\boldsymbol{X}(s)) - \mathbf{f}_{\widehat{\phi}}(\boldsymbol{X}(s))\right\|^2 \mathrm{d}s\right].$$

Note that by Jensen's inequality,

$$\frac{1}{T}\int_0^T \mathbb{E}\left[\left\|\mathbf{f}_\phi(\boldsymbol{X}(s)) - \mathbf{f}_{\widehat{\phi}}(\boldsymbol{X}(s))\right\|^2\right]\mathrm{d}s < N\left|\left|\left|\widehat{\phi} - \phi\right|\right|\right|^2.$$

Then, the conclusion follows by combining with the estimate $\mathrm{Lip}(F_{[\widehat{\phi}]}) \leq (R+1)S$. □

References

1. A. S. Baumgarten and K. Kamrin. A general constitutive model for dense, fine-particle suspensions validated in many geometries. *Proc Natl Acad Sci USA*, 116(42):20828–20836, 2019.
2. N. Bell, Y. Yu, and P. J. Mucha. Particle-based simulation of granular materials. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation - SCA '05*, page 77, Los Angeles, California, 2005. ACM Press.
3. S. Benachour, B. Roynette, D. Talay, and P. Vallois. Nonlinear self-stabilizing processes – I Existence, invariant probability, propagation of chaos. *Stochastic Processes and their Applications*, 75(2):173–201, 1998.
4. G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.
5. S. Bernstein. *Sur l'ordre de la meilleure approximation des fonctions continues par des polynômes de degré donné*, volume 4. Hayez, imprimeur des académies royales, 1912.
6. P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov. Universal algorithms for learning theory part i: piecewise constant functions. *J. Mach. Learn. Res.*, 6(Sep):1297–1321, 2005.
7. V. D. Blodel, J. M. Hendricks, and J. N. Tsitsiklis. On Krause's multi-agent consensus model with state-dependent connectivity. *Automatic Control, IEEE Transactions on*, 54(11):2586 – 2597, 2009.
8. F. Bolley, I. Gentil, and A. Guillin. Uniform Convergence to Equilibrium for Granular Media. *Arch Rational Mech Anal*, 208(2):429–445, 2013.
9. 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(05):909–951, 2017.
10. D. R. Brillinger. Learning a potential function from a trajectory. In *Selected Works of David Brillinger*, pages 361–364. Springer, 2012.
11. C. Brugna and G. Toscani. Kinetic models of opinion formation in the presence of personal conviction. *Phys. Rev. E*, 92(5):052818, 2015.
12. J. Carrillo, R. McCann, and C. Villani. Kinetic equilibration rates for granular media and related equations: Entropy dissipation and mass transportation estimates. *Rev. Mat. Iberoamericana*, pages 971–1018, 2003.
13. P. Cattiaux, A. Guillin, and F. Malrieu. Probabilistic approach for granular media equations in the non-uniformly convex case. *Probab. Theory Relat. Fields*, 140(1-2):19–40, 2007.
14. D. Chen, Y. Wang, G. Wu, M. Kang, Y. Sun, and W. Yu. Inferring causal relationship in coordinated flight of pigeon flocks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(11):113118, 2019.
15. X. Chen. Maximum likelihood estimation of potential energy in interacting particle systems from single-trajectory data. arXiv preprint arXiv:2007.11048, 2020.
16. A. Cohen, M. A. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Foundations of computational mathematics*, 13(5):819–834, 2013.
17. F. Comte and V. Genon-Catalot. Nonparametric drift estimation for i.i.d. paths of stochastic differential equations. *accepted for publication in The Annals of Statistics*, 2019.
18. I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433(7025):513 – 516, 2005.
19. M. C. Crosskey and M. Maggioni. Atlas: A geometric approach to learning high-dimensional stochastic systems near manifolds. *Journal of Multiscale Modeling and Simulation*, 15(1):110–156, 2017. arxiv:1404.0667.
20. F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.
21. F. Cucker and D. X. Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.
22. M. R. D'Orsogna, Y.-L. Chuang, A. L. Bertozzi, and L. Chayes. Self-propelled particles with soft-core interactions: patterns, stability, and collapse. *Phys. Rev. Lett.*, 96:104 – 302, 2006.

23. L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer, New York, 2002.
24. R. Hegselmann and U. Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *JASSS*, 5(3):33, 2002.
25. N. J. Higham. *Functions of matrices: theory and computation*, volume 104. SIAM, 2008.
26. H. Huang, J.-G. Liu, and J. Lu. Learning interacting particle systems: Diffusion parameter estimation for aggregation equations. *Mathematical Models and Methods in Applied Sciences*, 29(01):1–29, 2019.
27. P.-E. Jabin and Z. Wang. Mean field limit and propagation of chaos for Vlasov systems with bounded forces. *Journal of Functional Analysis*, 271(12):3588–3627, 2016.
28. P.-E. Jabin and Z. Wang. Quantitative estimates of propagation of chaos for stochastic systems with $W^{-1,\infty}$ kernels. *Invent. math.*, 214(1):523–591, 2018.
29. J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Number v. 160 in Applied Mathematical Sciences. Springer, New York, 2005.
30. S. Kalmykov, B. Nagy, V. Totik, et al. Bernstein-and Markov-type inequalities for rational functions. *Acta Mathematica*, 219(1):21–63, 2017.
31. I. Karatzas and S. E. Shreve. Brownian motion. In *Brownian Motion and Stochastic Calculus*, pages 47–127. Springer, 1998.
32. F. C. Klebaner. *Introduction to stochastic calculus with applications*. World Scientific Publishing Company, 2005.
33. U. Krause. A discrete nonlinear and non-autonomous model of consensus formation. *Communications in difference equations*, 2000:227–236, 2000.
34. Y. A. Kutoyants. *Statistical Inference for Ergodic Diffusion Processes*. Springer London, 2004.
35. D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.
36. L. Li, Y. Li, J.-G. Liu, Z. Liu, and J. Lu. A stochastic version of Stein Variational Gradient Descent for efficient sampling. *Commun. Appl. Math. Comput. Sci.*, 15(1):37–63, 2020.
37. Z. Li and F. Lu. On the coercivity condition in the learning of interacting particle systems. arXiv preprint arXiv:2011.10480, 2020.
38. Z. Li, F. Lu, M. Maggioni, S. Tang, and C. Zhang. On the identifiability of interaction functions in systems of interacting particles. *Stochastic Processes and their Applications*, 132:135–163.
39. Q. Liu and D. Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. arXiv:1608.04471 Cs Stat, 2019.
40. F. Lu, M. Maggioni, and S. Tang. Learning interaction kernels in heterogeneous systems of agents from multiple trajectories. *Journal of Machine Learning Research*, 22(32):1–67, 2021.
41. 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(29):14424–14433, 2019.
42. X. Mao. *Stochastic differential equations and applications*. Elsevier, 2007.
43. S. Motsch and E. Tadmor. Heterophilious Dynamics Enhances Consensus. *SIAM Rev.*, 56(4):577 – 621, 2014.
44. R. Nickl, K. Ray, et al. Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *Annals of Statistics*, 48(3):1383–1408, 2020.
45. B. Øksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 6th edition, 2013.
46. R. Olfati-Saber and R. M. Murray. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on automatic control*, 49(9):1520–1533, 2004.
47. D. Pollard. *Mini Book notes*. 2000. http://www.stat.yale.edu/~pollard/Books/Mini/Basic.pdf.
48. L. Schumaker. *Spline functions: basic theory*. Cambridge University Press, 2007.
49. A. V. Skorokhod. On the regularity of many-particle dynamical systems perturbed by white noise. *Journal of Applied Mathematics and Stochastic Analysis*, 9(4):427–437, 1996.
50. R. H. Stefano Almi, Massimo Fornasier. Data-driven evolutions of critical points. *Foundations of Data Science*, 2(3):207–255, 2020.
51. M. B. Thompson. A comparison of methods for computing autocorrelation time. arXiv preprint arXiv:1011.0175, 2010.
52. C. Wang and D.-X. Zhou. Optimal learning rates for least squares regularized regression with unbounded sampling. *Journal of Complexity*, 27(1):55–67, 2011.
53. J. P. Ward. l^p Bernstein inequalities and inverse theorems for RBF approximation on r^d. *Journal of Approximation Theory*, 164(12):1577–1593, 2012.

54. Z. Zhang and F. Lu. Cluster prediction for opinion dynamics from partial observations. *IEEE Transactions on Signal and Information Processing over Networks*, 2020.
55. M. Zhong, J. Miller, and M. Maggioni. Data-driven discovery of emergent behaviors in collective dynamics. *Physica D: Nonlinear Phenomena*, 411:132542, 2020.

Authors and Affiliations

Fei Lu[1] · Mauro Maggioni[2] · Sui Tang[3]

✉ Sui Tang
suitang@math.ucsb.edu

Fei Lu
feilu@math.jhu.edu

Mauro Maggioni
mauromaggionijhu@icloud.com

1 Department of Mathematics, Johns Hopkins University, Baltimore, USA

2 Department of Mathematics, Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, USA

3 Department of Mathematics, University of California, Santa Barbara, Santa Barbara, USA