Faculty of Natural Sciences

Department of Mathematics

Institute of Mathematics

Chair of Scientific Computing and
Uncertainty Quantification

Nonparametric online kernel inference for interacting particle systems

— MATH-599 | Master project in Mathematics (Spring 2024)

Author:
Eliott Van Dieren (eliott.vandieren@epfl.ch)

Supervisors:
Prof. Grigorios A. Pavliotis (Imperial College London, g.pavliotis@imperial.ac.uk)
Prof. Fabio Nobile (EPFL, fabio.nobile@epfl.ch)

July 2024

Contents

Summary - Résumé

English This master's thesis addresses the problem of online kernel inference for interacting particle systems within a nonparametric framework. The statistical inference of such kernels and the mean field limit of these systems has been an area of growing interest in recent years [23; 30; 37; 40]. However, nonparametric methods have received less attention, despite their usefulness in generalising kernel inference techniques.

This work generalises the inference of any smooth interacting kernel by applying the diffusion process drift estimation technique from [41] combined with nonparametric function representations. The methodology is tested with different smooth and non-smooth interacting kernels, employing Fourier series on the circle and Hermite expansions on the real line. The results are compared to maximum likelihood estimators for interacting particle systems [21; 39].

The numerical results are promising with theoretical convergence rates achieved. The study also explores the mean field regime, i.e. systems with a very large number of particles, to understand how much data is necessary to estimate the kernel. Indeed, one can benefit from the propagation of chaos in the mean field limit to express particle dynamics with only a few representative particles. Notably, this work shows that only two trajectories are needed to efficiently infer the Curie-Weiss interacting kernel on the real line for a system of 200 particles, whose inference would have been computationally demanding if using all available trajectories.

This thesis also covers the method's limitations such as stability issues on the real line when the order of the Hermite expansions is overly high, and provides insights into the method's adaptiveness to the number of modes in the function representations. We also show that the method lacks robustness when measurement noise is added to the particle positions. Future research directions include filtering such noisy data, extending to multi-dimensional particles, and inferring the confinement potential in parallel with the interaction potential.

French Ce projet de master aborde le problème de l'inférence de noyaux d'interaction pour les systèmes de particules en interaction, et ce de manière non-paramétrique. L'inférence statistique de tels noyaux et la limite de champ moyen de ces systèmes ont suscité un intérêt croissant ces dernières années [23; 30; 37; 40]. Cependant, les méthodes d'inférence non-paramétriques ont reçu moins d'attention, malgré leur potentiel de généralisation des techniques d'inférence de noyaux d'interaction.

Ce travail généralise l'inférence de tout noyau d'interaction (suffisamment lisse) en appliquant la technique d'inférence décrite dans [41], combinée avec des représentations de fonctions non-paramétriques. La méthode est testée avec différents noyaux d'interaction lisses et non-lisses, utilisant les séries de Fourier sur le cercle et les expansions de Hermite sur la ligne réelle. Les résultats sont comparés aux estimateurs du maximum de vraisemblance pour les systèmes de particules en interaction [21; 39].

Les résultats numériques sont prometteurs avec des taux de convergence théoriques atteints. Le projet explore également le régime en champ moyen, c'est-à-dire des systèmes avec un très grand nombre de particules, pour comprendre la quantité de données nécessaire pour estimer le noyau. En effet, on peut bénéficier de la propagation du chaos dans la limite en champ moyen pour exprimer la dynamique des particules avec seulement quelques particules représentatives. Notamment, ce travail montre que seules deux trajectoires sont nécessaires pour inférer efficacement le noyau d'interaction de Curie-Weiss sur la ligne réelle pour un système de 200 particules, dont l'inférence serait computationnellement très coûteuse si toutes les trajectoires disponibles étaient utilisées.

Cette thèse couvre également les limitations de la méthode, telles que les problèmes de stabilité lorsque l'ordre des expansions de Hermite est trop important, et étudie l'adaptabilité de la méthode au nombre de modes dans les représentations de fonctions. Nous montrons également que la méthode manque de robustesse lorsque du bruit de mesure est ajouté aux positions des particules. Les directions de recherche futures incluent le filtrage de ces données bruitées, l'extension aux particules multidimensionnelles et l'inférence du potentiel de confinement, en parallèle du potentiel d'interaction.

Notation

In this paragraph, we illustrate all the relevant notation symbols that are used throughout the work.

- General notations:
  - n is used as an arbitrary dimension;
  - $V^{(i)}$ denotes the i-th element of a vector $V \in \mathbb{R}^n$.
  - $M_{i,j}$ denotes the element i, j of a matrix M;
  - T is the upper bound of the time interval;
  - d is the dimension of the dynamic for the studied model;
  - N is the number of particles in the studied interacting particle system;
  - J is the dimension of the vector of parameters for the studied model;
  - n_{steps} denotes the number of simulations in a given numerical experiment;
  - $\Delta t = T/K$ with K steps of time discretization for numerical experiments;
  - $(t_i)_{i=0,\dots,K}$ are the timestamps chosen for the Euler-Maruyama (EM) scheme.
- Mathematical objects:
  - I_n is the identity matrix of dimension $n \times n$;
  - $\mathbb{1}_\alpha$ is the indicator function with condition α;
  - e_i is the i-th element of a given basis;
  - $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t \geq 0})$ is a probability space with $(\mathcal{F}_t)_{t>=0}$ being its filtration;
  - $B_t := \left(B_t^1, \dots, B_t^n\right)^\top$ is a $\mathbb{R}^n$-valued standard Brownian motion;
  - dB_t is a $\mathbb{R}^n$-valued vector containing standard Brownian motion increments;
  - $X^{(n)}(t, \omega) : [0, T] \times \Omega \to \mathbb{R}^d$ is a given particle $n = 1, ..., N$ in the studied interacting particle system (the ω will be omitted and we often write X_t for $X(t)$). When $d = 1$, X_t can be rewritten as a N-dimensional stochastic process;
  - $X_0 \in \mathbb{R}^{nd}$ is the initial datum of the dynamic;
  - $Y^{(n)}(t, \omega) : [0, T] \times \Omega \to \mathbb{R}^d$ is the noisy observation of a given particle $X^{(n)}(t, \omega)$ (the ω will be omitted);
  - $\theta \in \mathbb{R}^J$ is a set of parameters;
  - $\theta^* \in \mathbb{R}^J$ is the optimal set of parameters for a given model;
  - $f(x, \theta)$ is the estimated function of the true drift $f^*(x)$;
  - $\nabla_\theta f(x, \theta)$ is the transposed Jacobian of the function $f(x, \theta)$ w.r.t. θ;
  - l_t is the learning rate of the SGDCT method;
  - $\bar{F}, \bar{h}, \bar{c}$ are quantities used in the MLE method;
  - $\tilde{\sigma}$ is the measurement noise standard deviation;

- $\xi_t^{(i)} \sim \mathcal{N}(0,1)$ is a random variable used for measurement noise for $i = 1, ..., N$.

- Norms and operators:
  - $\overline{(\cdot)}$ is the empirical mean operator;
  - $|\cdot|$ is the Euclidean norm;
  - $\langle \cdot, \cdot \rangle$ is the inner product.
  - $\|\cdot\|$ is a general norm induced by an inner product or the mean-squared error operator for numerical experiments;

1 Introduction

Interacting Particle Systems (IPSs) are widely used in science to model agent interactions. The modelling of such systems has numerous applications, e.g., opinion dynamics [33], financial systems [20], networks of spiking neurons [1] and animal behaviour [12]. However, these systems often involve unknown parameters, necessitating several methodologies to estimate them from data [37; 40].

The interaction kernel is the function or operator representing interactions between particles. We aim to infer its parameters from data, or more broadly, to determine its representation. Kernel inference problems fall into two categories: (i) for systems with a finite number of particles, it is treated as a statistical learning problem, where methods like regression or likelihood maximisation are employed on particle trajectories, and (ii) for systems with a very large number of particles, as an inverse Partial Differential Equation (PDE) problem where only a macroscopic view of the particle distribution is available [7; 30]. As the particle count tends to infinity, their individual impact diminishes, a phenomenon known as the "propagation of chaos". This results in a symmetry property where interactions between any single particle and the rest of the pool can be expressed via a single representative particle. This links to using mean field Stochastic Differential Equations (SDEs), also called McKean-Vlasov SDEs. These equations are known to be linked with a non-linear, non-local PDE on the space of probability measures, called the Fokker-Plank equation. Nonparametric kernel inference techniques are particularly interesting due to their practicability and generality compared to parametric methods. Recent research has made strides in providing nonparametric estimators for large-scale systems using, e.g., regularised least squares on an error functional based on likelihood [23].

Contributions This work aligns with this broader applicability goal. We focus on inferring kernels using a diffusion process drift estimation technique called the "Stochastic Gradient Descent in Continuous Time" (SGDCT) and introduced in [41]. This method is combined with nonparametric function representations, i.e. Fourier series and Hermite expansions to infer their coefficients. More specifically, we apply the SGDCT methodology to IPSs where each particle follows an SDE. Our study covers the inference of multiple interaction kernels, the analysis of the convergence and numerical errors of the technique and its comparison to Maximum Likelihood Estimators (MLEs). We further explore the method's adaptiveness to the number of basis components of the kernel representation, and whether we can detect how many modes are necessary for a proper function approximation. Moreover, we discuss the measurement scheme and assess how many particle observations are needed to infer kernels for systems in the mean field regime, where statistical inference becomes intractable. Lastly, we introduce measurement noise to the particle data to assess the robustness of the inference method.

Outline of the work We begin by reviewing the literature on McKean-Vlasov SDEs and their parameter estimation techniques in Section 2. Section 3 describes the relevant methodologies and concepts required for this thesis. We then properly define our problem and the general methodology in Section 4. Sections 5 and 6 present the main results and numerical experiments on the circle and real line, respectively. The adaptiveness of the SGDCT concerning the number of components in the nonparametric function representation is discussed in Section 7. The Python library created for this project is briefly covered in Section 8, followed by the conclusion in Section 9.

2 Literature review

In the last few decades, significant efforts have been dedicated to the study of McKean-Vlasov SDEs [32; 18]. More precisely, research has focused on their well-posedness [14], ergodicity [2; 8; 45], existence and uniqueness [3; 25] and propagation of chaos [5; 17] since the 1960s. Conversely, the statistical inference of parameters for these nonlinear diffusion processes has been relatively less explored but has been emerging since the 2000s (with some earlier exceptions [21; 28]). For instance, some results on the convergence and asymptotic consistency of offline MLEs have been published using a continuous data stream in [16; 46]. Recently, these results have been expanded to path-dependent cases [27]. Parameter inference using a log-likelihood approximation based on continuous observations of a non-linear diffusion process has been proposed in [19].

As IPSs are tightly linked to McKean-Vlasov SDEs, we also refer to works based on parameter estimation for IPSs with N particles as $N \to \infty$, which enters the mean field limit of such systems [6; 10]. Notably, the work [37] uses a moment approximation technique to infer parameters for IPSs and their mean field limit. As mentioned briefly in the introduction, some progress has been made in the nonparametric representation of drift functions, such as in [11] and [15]. Studies on identification for nonparametric frameworks have also been conducted, notably in [24; 26] and on learning in [29; 30].

Despite this recent interest, these techniques remain offline, and less attention has been given to online parameter estimation, except for [40] where the authors study an online parametric MLE, analyse its convergence and compare it to traditional MLE techniques. This online setting is of particular interest as it allows tracking changes in the parameters over time and is more efficient if new data is gathered alongside the parameter inference. We refer to [41; 42; 43] for online parameter estimation on classical diffusion processes.

3 Background

We briefly review key concepts related to SDEs and IPSs. We also cover the MLE for diffusion processes and extend it to IPSs. We then describe the technique from [41] that we will use to infer interaction kernels and prove the convergence of the obtained estimator. Lastly, we review important concepts regarding orthonormal bases, both on the circle and real line, which will be utilised throughout the work for the non-parametric representation of interacting potentials.

3.1 Stochastic differential equations

In this subsection, we review basic concepts of stochastic calculus to introduce stochastic differential equations. We loosely follow the excellent lecture notes from Dr. A. Blumenthal for the EPFL course "Numerical Integration of Stochastic Differential Equations". For further details on SDEs and stochastic processes, we refer to the book by Prof. G. A. Pavliotis [36].

We first start by defining a probability space, as this will be the environment in which we will define further concepts.

Definition 3.1 (Probability space)**.** *We call a triplet* $(\Omega, \mathcal{F}, \mathbb{P})$ *a probability space, where*

- Ω *is the sample space, i.e. the non-empty set of all possible outcomes.*
- *The* σ*-algebra* $\mathcal{F} \subseteq 2^{\Omega}$ *is the collection of all events.*

- *The probability measure $\mathbb{P}$ is a function that returns the probability of a given event to happen. We define it as $\mathbb{P} : \mathcal{F} \to [0,1]$. We also need $\mathbb{P}(\Omega) = 1$, and that given $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ a countable collection of disjoint events, one has*

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

We then define a random variable, an important building block for stochastic processes and SDEs.

Definition 3.2 (Random variable)**.** *Let the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A mapping $X : \Omega \to \mathbb{R}^n$ is called a random variable (or random vector), if for each $B \in \mathcal{B}$, $X^{-1}(B) \in \mathcal{F}$, where we denote by $\mathcal{B}$ the Borel σ-algebra on $\mathbb{R}^n$, and write*

$$X^{-1}(B) = \{\omega \in \Omega | X(\omega) \in \mathcal{B}\}.$$

We say equivalently that X if $\mathcal{F}$-measurable.

For example, if we define $\Omega = \{\text{heads}, \text{tails}\}$, then one can define the random variable X as the outcome of a cointoss with an event ω

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = \text{heads} \\ 0, & \text{if } \omega = \text{tails}. \end{cases}$$

Definition 3.3 (Stochastic Process)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A stochastic process is a time-dependent collection of random variables denoted as $(X_t)_{t \in \mathcal{T}}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and has values in $\mathbb{R}^n$. We denote by $\mathcal{T}$ the parameter space, often $\mathbb{R}^+$.*

If one fixes t, a stochastic process reverts to a single random variable $X_t : \Omega \to \mathbb{R}^n$. Conversely, if one fixes the event $\omega \in \Omega$, then $X(t, \omega) : \mathcal{T} \to \mathbb{R}^n$ becomes a sample path. Hereunder, We define a key type of stochastic process: the Brownian motion.

Definition 3.4 (Brownian motion)**.** *A real-valued stochastic process $(B_t)_{t \geq 0}$ is called a standard Brownian motion if it satisfies the three following properties:*

1. *$B_0 = 0$ a.s. (Initial condition)*
2. *$B_t - B_s \sim \mathcal{N}(0, t-s)$ for all $0 \leq s < t$. (Normality of increments)*
3. *$B_{t_4} - B_{t_3} \perp\!\!\!\perp B_{t_2} - B_{t_1}$ for all $t_1 \leq t_2 < t_3 \leq t_4$ (Indepence of increments)*

From this, we find

$$\mathbb{P}(a \leq B_t \leq b) = \int_a^b \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}} dx,$$

as $B_t = B_t - B_0 \sim \mathcal{N}(0, t)$.

Definition 3.5 (Filtration on a probability space)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_t)_{t \geq 0}$ a stochastic process. The family of increasing σ-algebras*

$$\{\mathcal{F}_t = \mathcal{F}(X(s), 0 \leq s \leq t)\}_{t \geq 0}$$

is a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$, sometimes called "the history of X_t".

We also need to define an adaptive process, as this will be needed for strong solutions of SDEs.

Definition 3.6 (Adaptiveness)**.** *We say that a stochastic process $(X_t)_{t\geq 0}$ is adapted with respect to the filtration $\mathcal{F}_t$ if for each $t \in [0, T]$, X_t is $\mathcal{F}_t$-measurable.*

We can now define SDEs as we covered all the necessary basic concepts.

Definition 3.7 (Stochastic Differential Equation)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(B_t)_{t\geq 0}$ be a $\mathbb{R}^m$-valued standard Brownian motion, and the filtration $(\mathcal{F}_t)_{t>=0}$ such that for all $t \in [0, T]$ and any component i, $B_t^{(i)}$ is a 1-dimensional $\mathcal{F}_t$-measurable standard Brownian Motion.*

If we further define some functions $f : [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$, and $g_j : [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$ for $j = 1, ..., m$, then one can write a stochastic differential equation as

$$dX_t = f(t, X_t)dt + \sum_{j=1}^{m} g_j(t, X_t)dB_t^{(j)}, \quad X_0 = x_0,$$

which can also be rewritten component-wise as

$$dX_t^{(i)} = f^{(i)}(t, X_t)dt + \sum_{j=1}^{m} g_j^{(i)}(t, X_t)dB_t^{(j)}, \quad X_0^{(i)} = x_0^{(i)}, \quad i = 1, ..., n,$$

which is a one-dimensional SDE.

Definition 3.8 (Strong solution of SDEs)**.** *Let the SDE*

$$dX_t = f(t, X_t)dt + g(t, X_t)dB_t, \quad X_0 = x_0, \quad 0 \leq t \leq T. \tag{$\star$}$$

A strong solution of the SDE ($\star$) is a stochastic process $(X_t)_{0\leq t\leq T}$ with continuous sample paths, such that

1. *X_t is $\mathcal{F}_t$-measurable (i.e. X_t is adapted).*
2. *$f(t, X_t), g(t, X_t) \in M^2(0, T)$ where*

$$M^2(0, T) = \{\{G_t\}_{t\geq 0} \text{ stoch. process, progressively measurable, } \mathbb{E}\left[\int_0^T G_t^2 dt\right] < \infty\}.$$

3. *The SDE ($\star$) holds a.s. for each $t \in [0, T]$.*

Lastly, we cover the Euler-Maruyama method to sample an SDE numerically.

Definition 3.9 (Euler-Maruyama method)**.** *Let the SDE*

$$dX_t = f(t, X_t)dt + g(t, X_t)dB_t, \quad X_0 = x_0, \quad 0 \leq t \leq T.$$

Then, the Euler-Maruyama method is given as

$$X_{k+1} = X_k + f(t_k, X_k)\Delta t + g(t_k, X_k)\Delta B_k, \quad X_0 = x_0.$$

where $\Delta t = \frac{T}{K}$, K is the number of steps in the numerical method, $\Delta B_k \sim \mathcal{N}(0, \Delta t)$. It can be shown that the strong convergence rate of the Euler-Maruyama method is $1/2$, while its weak order is 1.

3.2 Interacting particle systems

In this work, we focus on systems of particles whose dynamics are governed by stochastic differential equations. Namely, we denote by $X_t := (X_t^{(i)})_{i=1,...,N} \in \mathbb{R}^{N\times d}$ a collection of N particles interacting together. Each particle is an Itô process that follows the SDE

$$dX_t^{(i)} = v\left(X_t^{(i)}\right)dt + \frac{1}{N}\sum_{n=1}^{N}\phi\left(X_t^{(i)}, X_t^{(n)}\right)dt + \sigma(X_t^{(i)})dB_t^{(i)} \quad i = 1, ..., N, \tag{3.1}$$

with $X_0^{(i)} \sim \mu_0$ for all $i = 1, ..., N$. The function $v : \mathbb{R}^d \to \mathbb{R}^d$ is the drift function, $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is the interaction kernel, $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is the diffusion matrix and B_t is a set of N d-dimensional standard Brownian motion. One notes that we can often rewrite $\phi(x, y)$ as $\phi(z)$, with $z = x - y$. This notation will be used throughout the work. For some examples and the study of the mean field limit, we refer to the lecture notes from Prof. D. Lacker [22].

3.3 Maximum likelihood estimators

In this section, we cover the main concepts of MLEs for diffusion processes and interacting particle systems. The discussion about diffusion processes follows the work from [39], while the derivation for interacting particle systems is based on a reparametrisation of the IPS into a multi-dimensional diffusion process.

3.3.1 MLE for diffusion processes

We cover here the main findings for MLE for diffusion processes. Following [39], it is of interest to start with a specific case such that the process follows the SDE

$$dX_t = -\nabla V_0(X_t)dt + \sqrt{2\beta^{-1}}dB_t, \tag{3.2}$$

which is also known as the overdamped Langevin equation. Here, $X_t \in \mathbb{R}^d$, $dB_t \in \mathbb{R}^d$ is a d-dimensional Brownian motion increment, $\beta > 0$ is the inverse temperature and $V_0 : \mathbb{R}^d \to \mathbb{R}$ is the confining potential, which is assumed in C^2, bounded from below and with a growth condition at infinity such that $\exp(-\beta V_0)$ is integrable. From this, (3.2) is ergodic with the Boltzmann-Gibbs measure

$$\rho_0(x) = Z^{-1}e^{-\beta V_0(x)},$$

where Z is the normalising constant over $\mathbb{R}^d$. The MLE will be used to estimate the potential V_0, supposing that β is known and constant. If we let Z_t be a solution of (3.2) when $V_0 \equiv 0$, then it solves

$$dZ_t = \sqrt{2\beta^{-1}}dB_t. \tag{3.3}$$

If we let $\mathbb{P}$ and $\mathbb{Q}$ be the two path-space measures generated on $[0, T]$ respectively by (3.2) and (3.3), then one can show that they are absolutely continuous with the Radon-Nikodym derivative

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = \exp(-TI_T(X)), \tag{3.4}$$

with

$$I_T(X) = \frac{\beta}{4T}\int_0^T (|\nabla V_0(X_t)|^2 dt + 2\langle\nabla V_0(X_t), dX_t\rangle).$$

If we fix a single path $(X_t)_{t\in[0,T]}$, then we can rewrite $I_T(X)$ as a functional depending on the choice of confining potential V, such that one has

$$\mathcal{I}_T(V) = \frac{\beta}{4T}\int_0^T (|\nabla V(X_t)|^2 dt + 2\langle \nabla V(X_t), dX_t\rangle). \tag{3.5}$$

One also notes that this is equal to the negative log-likelihood function for V and that the MLE should therefore try to minimise this equation. Rewriting (3.5) using the definition of X_t from (3.2), we can obtain the following splitting

$$\mathcal{I}_T(V) = \frac{\beta}{4T}\int_0^T \left(|\nabla V(X_t)|^2 - 2\langle \nabla V(X_t), \nabla V_0(X_t)\rangle\right) dt + \frac{\sqrt{2\beta}}{2T}\int_0^T \langle \nabla V(X_t), dW_t\rangle. \tag{3.6}$$

As as $T\to\infty$, and using the ergodicity of X_t, we have that $\mathcal{I}_T(V)\to\mathcal{I}_\infty(V)$ a.s. where

$$\mathcal{I}_\infty(V) = \frac{\beta}{4}\int_{\mathbb{R}^d}\Big(|\nabla V(x)|^2 - 2\langle \nabla V(x), \nabla V_0(x)\rangle\Big)\,\rho_0(x)dx,$$

which is minimised when $\nabla V = \nabla V_0$, the drift from (3.2). However, we do not often have access to an infinite amount of data as $T\to\infty$. Hence, if we replace $\rho(x)dx$ by its finite time equivalent, i.e. the occupation measure

$$\mu_T(B) = \frac{1}{T}\int_0^T \mathbb{1}_{\{X_t\in B\}}dt,$$

then after some steps, one obtains

$$\mathcal{I}_T(V) = \frac{\beta}{2T}\left(V(X_T) - V(X_0)\right) + \frac{\beta}{4}\int_{\mathbb{R}^d}\Big(|\nabla V(x)|^2 - 2\beta^{-1}\Delta V(x)\Big)\,\mu_T(dx). \tag{3.7}$$

As described in [39], minimising this expression is not recommended, e.g., due to singularity issues of μ_T w.r.t. the Lebegue measure when $d>1$. Therefore, we need some regularisation of (3.7) to solve the likelihood maximisation problem in finite time. There are three main methods to achieve this:

1. **Bayesian approach**: this introduces a prior measure on V, which enables potential candidates sampling. Coupled to (3.5), and data $(X_t)_{t\in[0,T]}$, one can infer the posterior distribution of V. We refer to [38] for further details about this approach.

2. **Measure substitution**: this method replaces the occupation measure $\mu_T(dx)$ by a smooth probability density function $\rho_T(x)dx$. Indeed, one can rewrite $\mathcal{I}_T(V)$ as

   $$\mathcal{I}_T(V) = \frac{\beta}{2T}(V(X_T) - V(X_0)) + \tilde{\mathcal{I}}_T(V),$$

   where

   $$\tilde{\mathcal{I}}_T(V) = \frac{\beta}{4}\int_{\mathbb{R}^d}(|\nabla V(x)|^2 - 2\beta^{-1}\Delta V(x))\rho_T(x)dx.$$

   One can show that as T grows sufficiently enough, $\mathcal{I}_T(V)$ can be well approximated by $\tilde{\mathcal{I}}_T(V)$. Overall, this transforms the problem into a regularised form, but we now have to determine ρ_T. We refer to [39] for further developments about this approach.

3. **Parametrisation of V**: here, the function V is expressed as a linear combination of smooth basis functions $e_i(x)$ such that

   $$V(x,\theta) = \sum_{j=1}^{J}\theta_j e_j(x),$$

with parameters $(\theta_j)_{j=1,\ldots,J}$. From this, we can rewrite (3.5) as

$$\mathcal{I}_T(\theta) = \frac{\beta}{4T} \int_0^T \left(\left| \sum_{j=1}^{J} \theta_j \nabla e_j(X_t) \right|^2 dt + 2 \sum_{j=1}^{J} \theta_j \langle \nabla e_j(X_t), dX_t \rangle \right). \tag{3.8}$$

One can find a finite set of basis functions $(e_j)_j$ such that the quantity $\mathcal{I}_T(\theta)$ is convex in θ, and therefore has a unique minimum.

We will hereunder focus on the last method and detail the steps needed to obtain the estimators $(\hat{\theta}_j^{\text{MLE}})_{j=1,\ldots,J}$. We will suppose that we have access to the data $(X_t)_{t\in[0,T]}$ which is governed by the SDE

$$dX_t = -\nabla V(X_t)dt + \sigma(X_t)dB_t,$$

where $\nabla V : \mathbb{R}^d \to \mathbb{R}^d$ is the drift function that we want to estimate using the MLE and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is the diffusion matrix that we will assume constant for the remaining of this explanation. B_t is a d-dimensional standard Brownian motion. We then can write the functional $\mathcal{I}_T(V)$ as

$$\mathcal{I}_T(V) = \frac{1}{T} \int_0^T \left(\left\langle \nabla V(X_t), (\sigma\sigma^\top)^{-1} \nabla V(X_t) \right\rangle dt + 2 \left\langle \nabla V(X_t), (\sigma\sigma^\top)^{-1} dX_t \right\rangle \right). \tag{3.9}$$

We now parametrise $\nabla V(x)$ as

$$\nabla V(x) = \frac{1}{2}(\sigma\sigma^\top)\nabla V(x,\theta), \qquad \text{with } V(x,\theta) = \sum_{j=1}^{J} \theta_j e_j.$$

The main goal is therefore to find θ minimising (3.9). One can insert this parametrisation in (3.9) and finds that (up to multiplicative constants)

$$\mathcal{I}_T(\theta) = \theta^\top F \theta - 2\theta^\top h,$$

where $F = (f_{ij}) \in \mathbb{R}^{J\times J}$ has entries

$$f_{ij} = \frac{1}{T} \int_0^T \left\langle \nabla e_i(X_t), (\sigma\sigma^\top) \nabla e_j(X_t) \right\rangle dt, \quad i,j = 1,\ldots,J.$$

On the other hand, $h = (h^{(i)}) \in \mathbb{R}^J$ has entries

$$h^{(i)} = -\frac{2}{T} \int_0^T \langle \nabla e_i(X_t), dX_t \rangle, \quad i = 1,\ldots,J.$$

If F is positive-definite, one can show that the minimiser of $\mathcal{I}_T(\theta)$ is given as

$$\hat{\theta}^{\text{MLE}} = F^{-1}h. \tag{3.10}$$

3.3.2 MLE for interacting particle systems

We will derive an interacting particle system version of the parametric method from Section 3.3.1. We assume momentarily the drift function $v \equiv 0$ from (3.1) for simplicity, and we will try to infer the parameters θ to represent the interaction potential W such that $\phi = -W'$. The general idea is to set $d = N$ from Section 3.3.1 and perceive the N-particle system as a multi-dimensional diffusion process. This yields that $X_t = (X_t^{(i)})_{i=1,\dots,N}$ lies in $\mathbb{R}^N$ for one-dimensional particles. We first rewrite (3.1), assuming σ constant such that

$$dX_t = -\nabla V_0(X_t)dt + \sigma dB_t,$$

with $dX_t \in \mathbb{R}^N$, $\sigma \in \mathbb{R}^{N\times N}$ and lastly the drift $-\nabla V_0(x)$ given as

$$\nabla V_0(X_t) = \begin{pmatrix} \frac{1}{N}\sum_{n=1}^N W'\left(X_t^{(1)} - X_t^{(n)}\right) \\ \vdots \\ \frac{1}{N}\sum_{n=1}^N W'\left(X_t^{(N)} - X_t^{(n)}\right) \end{pmatrix}, \tag{3.11}$$

where $W : \mathbb{R} \to \mathbb{R}$ is the interacting potential that we want to infer. If we wish to parametrise $W(x) = \sum_{j=1}^J \theta_j e_j(x)$, then we can define $\nabla V(X_t, \theta)$ as

$$\nabla V(X_t, \theta) = \sum_{j=1}^J \theta_j \nabla f_j(X_t),$$

with entries $(\nabla V(X_t, \theta))^{(i)}$ given as

$$(\nabla V(X_t, \theta))^{(i)} = \sum_{j=1}^J \theta_j \underbrace{\left(\frac{1}{N}\sum_{n=1}^N e_j'\left(X_t^{(i)} - X_t^{(n)}\right)\right)}_{\nabla f_j^{(i)}(X_t)},$$

for $i = 1, ..., N$, such that

$$\nabla f_j(X_t) = \begin{pmatrix} \frac{1}{N}\sum_{n=1}^N e_j'\left(X_t^{(1)} - X_t^{(n)}\right) \\ \vdots \\ \frac{1}{N}\sum_{n=1}^N e_j'\left(X_t^{(N)} - X_t^{(n)}\right) \end{pmatrix}, \quad j = 1, ..., J. \tag{3.12}$$

By using the parameterization $\nabla V_0(X_t) = \frac{1}{2}(\sigma\sigma^\top)\nabla V(X_t, \theta)$, we can insert it in (3.9) and obtain the result up to a multiplicative constant:

$$\begin{aligned} \mathcal{I}_T(\theta) &= \frac{1}{T}\int_0^T \left(\left\langle \nabla V(X_t, \theta), (\sigma\sigma^\top)\nabla V(X_t, \theta)\right\rangle dt + 4\left\langle \nabla V(X_t, \theta), dX_t\right\rangle\right) \\ &= \frac{1}{T}\int_0^T \left(\left\langle \sum_{j=1}^J \theta_j \nabla f_j(X_t), (\sigma\sigma^\top)\sum_{j=1}^J \theta_j \nabla f_j(X_t)\right\rangle dt + 4\left\langle \sum_{j=1}^J \theta_j \nabla f_j(X_t), dX_t\right\rangle\right) \\ &= \theta^\top \bar{F}\theta - 2\theta^\top \bar{h}, \end{aligned} \tag{3.13}$$

where $\bar{F} = (\bar{f}_{ij})$ is a $\mathbb{R}^{J\times J}$-matrix with entries

$$\bar{f}_{ij} = \frac{1}{T}\int_0^T \left\langle \nabla f_i(X_t), (\sigma\sigma^\top)\nabla f_j(X_t)\right\rangle dt, \tag{3.14}$$

and $\bar{h} = (\bar{h}^{(i)})$ is a $\mathbb{R}^J$-vector with entries

$$\bar{h}^{(i)} = -\frac{2}{T}\int_0^T \left\langle \nabla f_i(X_t), dX_t\right\rangle. \tag{3.15}$$

As for the former case with diffusion processes, one finds the estimator $\hat{\theta}^{\mathrm{MLE}}$ by solving the linear system, which yields

$$\hat{\theta}^{\mathrm{MLE}} = \bar{F}^{-1}\bar{h}. \tag{3.16}$$

One also notes that this derivation for parametrised kernels matches the results obtained in the founding work [21] on MLE for large interacting particle systems.

If we extend this result with a non-zero drift v that we suppose known, then it yields the parameterisation

$$\nabla V(X_t, \theta) = -2(\sigma\sigma^\top)^{-1}v(X_t) + \sum_{j=1}^{J} \theta_j \nabla f_j(X_t).$$

We can then rewrite (3.13) as

$$\begin{aligned}\mathcal{I}_T(\theta) = \theta^\top \bar{F}\theta - 2\theta^\top \bar{h} + \frac{1}{T}\int_0^T \Bigg(\Bigg(4\left\langle v(X_t), (\sigma\sigma^\top)^{-1}v(X_t)\right\rangle \\ - 4\sum_{j=1}^{J} \theta_j \left\langle v(X_t), \nabla f_j(X_t)\right\rangle\Bigg)dt - 8\left\langle (\sigma\sigma^\top)^{-1}v(X_t), dX_t\right\rangle\Bigg).\end{aligned} \tag{3.17}$$

Then, we obtain the estimate $\hat{\theta}^{\mathrm{MLE}}$ by differentiating (3.17) w.r.t. θ such that

$$\hat{\theta}^{\mathrm{MLE}} = \bar{F}^{-1}(\bar{h} + \bar{c}), \tag{3.18}$$

where $\bar{c}$ is a $\mathbb{R}^J$-vector with entries

$$\bar{c}^{(i)} = \frac{2}{T}\int_0^T \left\langle v(X_t), \nabla f_i(X_t)\right\rangle dt. \tag{3.19}$$

3.4 Stochastic Gradient Descent in Continuous Time

This subsection covers the main inference method used throughout this work. Stochastic Gradient Descent in Continuous Time (SGDCT) is an online computationally efficient statistical inference methodology from [41]. Let $X_t \in \mathcal{X}(= \mathbb{R}^d)$ be a given diffusion process following the stochastic differential equation

$$dX_t = f^*(X_t)dt + \sigma dB_t \quad t \in [0, T], \tag{3.20}$$

where $B_t \in \mathbb{R}^d$ is a standard Brownian motion, $\sigma \in \mathbb{R}^{d\times d}$. The role of the SGDCT is to efficiently find a set of parameters θ that produces an estimate $f(x, \theta)$ approximating the unknown function $f^*(x)$ well enough. The algorithm can be summarised in a single SDE describing the dynamics of the parameter estimates $\theta_t \in \mathbb{R}^J$ over time

$$d\theta_t = l_t[\nabla_\theta f(X_t, \theta_t)(\sigma\sigma^\top)^{-1}dX_t - \nabla_\theta f(X_t, \theta_t)(\sigma\sigma^\top)^{-1}f(X_t, \theta_t)dt], \tag{3.21}$$

where l_t is the learning rate and $\nabla_\theta f(X_t, \theta_t) \in \mathbb{R}^{J\times d}$ is matrix-valued. Hereunder, we go through the main assumptions from [41] to guarantee the convergence of the inference method.

Let

$$g(x, \theta) = \frac{1}{2}\|f(x, \theta) - f^*(x)\|^2_{\sigma\sigma^\top} = \frac{1}{2}\left\langle f(x, \theta) - f^*(x), (\sigma\sigma^\top)^{-1}(f(x, \theta) - f^*(x))\right\rangle \tag{3.22}$$

be the function one would like to minimise to obtain a good estimator function $f(x, \theta)$. The issue with this minimization is that one has no information about f^*, which makes the computation of g and $\nabla_\theta g$ impossible. A workaround is to use $dX_t = f^*(X_t)dt + \sigma dB_t$ as a noisy observation of $f^*(X_t)dt$, which enables us to approximate $\nabla_\theta g$ and perform the gradient step in (3.21). If we denote by

$$\bar{h}(\theta) = \int_{\mathcal{X}} h(x,\theta)\pi(dx)$$

the average of a given $L^1(\mathcal{X}, \pi(x)dx)$ function for its invariant measure π, then one can show that θ_t will converge to the minimum of the function $\bar{g}(\theta)$. Several assumptions must be considered to prove the convergence of this method and are described below. We note that Assumption 3.2 is a sufficient condition for X_t to guarantee the uniqueness and existence of its invariant measure.

Assumption 3.1 (Learning rate assumptions). *Throughout this work, we assume* $\int_0^\infty l_t dt = \infty$, $\int_0^\infty l_t^2 dt < \infty$, $\int_0^\infty |l_t' dt| dt < \infty$ *and that* $\exists p$ *s.t.* $\lim_{t\to\infty} l_t^2 t^{\frac{1}{2}+2p} = 0$.

For example, $l_t = \frac{C}{C_0+t}$ *would satisfy these conditions for* $C_0, C \in \mathbb{R}^+$.

Assumption 3.2 (Diffusion matrix and target function assumptions). *The matrix* $\sigma\sigma^\top$ *must be non-degenerative and bounded. One also needs* $\lim_{|x|\to\infty} f^*(x)\cdot x = -\infty$.

Assumption 3.3 (Control of the ergodicity of X_t). *For* $\theta \in \mathbb{R}^J$, *one has*

- *The gradient* $\nabla_\theta g(x,\cdot) \in C^2(\mathbb{R}^J)$ *for all* $x \in \mathcal{X}$, $\partial^2(\nabla_\theta g)/\partial x^2 \in C(\mathcal{X}, \mathbb{R}^J)$, *as well as* $\nabla_\theta g(\cdot,\theta) \in C^\alpha(\mathcal{X})$ *for* $\alpha \in (0,1)$. *There must exist* K, q *such that*

$$\sum_{i=0}^{2} \left| \frac{\partial^i \nabla_\theta g}{\partial \theta^i}(x,\theta) \right| \leq K(1+|x|^q).$$

- *For every* $N > 0$, $\exists C(N)$ *such that* $\forall \theta_1, \theta_2 \in \mathbb{R}^J$ *and* $|x| \leq N$,

$$|\nabla_\theta f(x,\theta_1) - \nabla_\theta f(x,\theta_2)| \leq C(N)|\theta_1 - \theta_2|.$$

  Furthermore, $\exists K, q > 0$ *such that* $|\nabla_\theta f(x,\theta)| \leq K(1+|x|^q)$.

- *Lastly, we need* $f^* \in C_b^{2+\alpha}(\mathcal{X})$ *with* $\alpha \in (0,1)$, *such that is has two bounded derivatives w.r.t.* x, *and partial derivatives being Hölder continuous.*

We now cite the main theorem from [41].

Theorem 3.1 (Convergence theorem for SGDCT). *If the conditions 3.1, 3.2, and 3.3 are satisfied, then one has*

$$\lim_{t\to\infty} \|\nabla \bar{g}(\theta_t)\| = 0, \text{ almost surely.}$$

Proof : see [41, Section 3] that covers the proof of the theorem.

Furthermore, it has been shown in the subsequent work [42] that

$$\mathbb{E}[\|\theta_t - \theta^*\|^p] \leq \frac{\mathcal{K}}{(C_0+t)^{\frac{p}{2}}}. \tag{3.23}$$

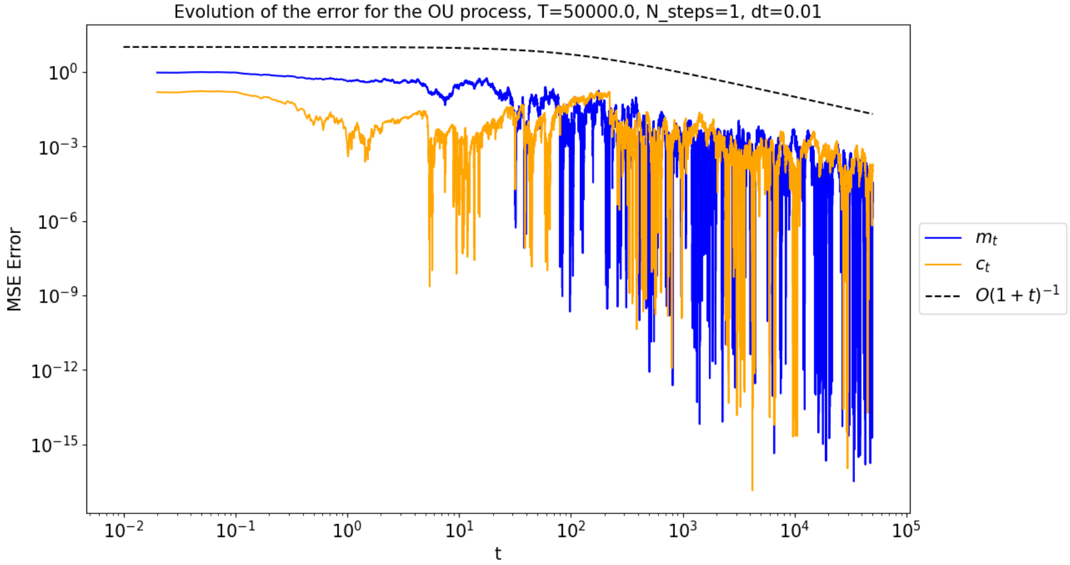


Figure 1: MSE error for c_t, m_t compared to (c^*, m^*) over time. We observe an error matching the theoretical convergence rate of $\mathcal{O}(1+t)^{-1}$.

Example 3.1 (SGDCT for the Ornstein-Ulhenbeck (OU) process)**.** *Let $X_t \in \mathbb{R}$ satisfy the SDE*

$$dX_t = c(m - X_t)dt + dW_t.$$

where we bound the domain of X_t in $[0, 2\pi]$, c is the speed of convergence, and m is the drift of the mean-reverting OU process. We want to estimate $\theta^* = (c^*, m^*)$. We compute $\nabla_\theta f$ that yields

$$\nabla f_\theta(X_t) = \begin{pmatrix} m_t - X_t \\ c_t \end{pmatrix},$$

and we obtain the SGDCT updates

$$\begin{aligned} & dc_t = \ell_t[(m_t - X_t)dX_t - c_t(m_t - X_t)^2 dt] \\ \iff & c_{t_{i+1}} = c_{t_i} + \ell_{t_i}(m_{t_i} - X_{t_i})(X_{t_{i+1}} - X_{t_i}) - \ell_{t_i} c_{t_i}(m_{t_i} - X_{t_i})^2 \Delta t, \end{aligned}$$

and

$$\begin{aligned} & dm_t = \ell_t[c_t dX_t - c_t^2(m_t - X_t)dt] \\ \iff & m_{t_{i+1}} = m_{t_i} + \ell_{t_i} c_{t_i}(X_{t_{i+1}} - X_{t_i}) - \ell_{t_i} c_{t_i}^2 (m_{t_i} - X_{t_i})\Delta t. \end{aligned}$$

For the learning rate we consider

$$\ell_t = \frac{10}{100 + t},$$

and define $\theta^* = (c^*, m^*) = (1, 1)^\top$, and $X_0 \sim U([0, 2\pi])$. We observe in Figure 1 that we successfully converge to the expected parameters as time grows.

3.5 Fourier series representation on the torus

We cover here the main concepts regarding the Fourier series and their implication on the circle $\mathbb{T}$. If we let f be a periodic function, then one can write its Fourier series as

$$f_N(x) = \sum_{n=-N}^{N} \hat{s}_n e^{i2\pi \frac{n}{P} x},$$

where N often goes to ∞, and P is the period of the Fourier series. Equivalently, one can rewrite f_N as

$$f_N(x) = A_0 + \sum_{n=1}^{N} \left(A_n \cos\left(2\pi \frac{n}{P} x\right) + B_n \sin\left(2\pi \frac{n}{P} x\right)\right).$$

This work will use Fourier series as an orthonormal basis on the circle. Indeed, the basis $\{e_n = e^{inx} : n \in \mathbb{Z}\}$ is an orthonormal basis in $L^2([\pi, \pi], dx)$. If we equip this space with the inner product

$$\langle f, g\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) g^*(x) dx,$$

where g^* is the complex conjugate of g, then this becomes a Hilbert space. For any $n, m \in \mathbb{Z}$, we have

$$\langle e_n, e_m\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{inx} e^{-imx} dx = \delta_{nm},$$

where δ is the Kronecker delta function. This proves the orthonormality of the proposed basis. One can therefore rewrite any function $f \in L^2([-\pi, \pi], dx)$ as

$$f(x) = \sum_{n=-\infty}^{\infty} \langle f, e_n\rangle e_n.$$

We will use the sine and cosine Fourier series notation to find the coefficients $(A_n, B_n)_n$. One has

$$A_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx) dx \text{ for } n \geq 0$$
$$B_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx) dx \text{ for } n \geq 1.$$

Example 3.2 (Fourier series representation)**.** *Let $f(x) = x^2$ periodic between $[-\pi, \pi]$. We will compute the Fourier series representation:*

$$A_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} x^2 dx = \frac{2\pi^2}{3},$$
$$A_n = \frac{1}{\pi} \int_{-\pi}^{\pi} x^2 \cos(nx) dx = \frac{2(\pi^2 n^2 - 2)\sin(\pi n) + 4\pi n \cos(\pi n)}{\pi n^3},$$
$$B_n = \frac{1}{\pi} \int_{-\pi}^{\pi} x^2 \sin(nx) dx = 0 \textit{ as the integrand is odd.}$$

where we used integration by parts twice for $(A_n)_{n\geq 1}$.

3.6 Hermite series representation on the real line

This section covers the main concepts regarding Hermite series representation in $\mathbb{R}$. We denote by H_n the physicist's Hermite polynomials such that

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2},$$

with a few examples given as

$$\begin{aligned} H_0(x) &= 1 \\ H_1(x) &= 2x \\ H_2(x) &= 4x^2 - 2 \\ &\vdots \end{aligned}$$

An interesting property of Hermite polynomials is their orthogonality. Indeed, if we introduce the inner product

$$\langle f, g \rangle_w = \int_{\mathbb{R}} f(x)g(x)w(x)dx, \quad w(x) = e^{-x^2},$$

and its related Hilbert space $L^2(\mathbb{R}, e^{-x^2}dx)$, the Hermite polynomials are orthogonal, i.e.,

$$\langle H_m(x), H_n(x) \rangle_w = \int_{-\infty}^{\infty} H_m(x)H_n(x)w(x)dx = \sqrt{\pi}2^n n! \delta_{nm}.$$

Hence, any function $f(x) \in L_2(\mathbb{R}, w(x)dx)$, i.e.

$$\int_{-\infty}^{\infty} |f(x)|^2 w(x)dx < \infty,$$

can be expressed using the Hermite polynomials series representation

$$f(x) = \sum_{n \geq 0} d_n H_n(x). \tag{3.24}$$

The coefficients d_n can be computed using orthogonality for $n \in \mathbb{N}$:

$$d_n = \frac{\langle f(x), H_n(x) \rangle}{\|H_n\|_w^2} = \frac{1}{2^n n! \sqrt{\pi}} \int_{-\infty}^{\infty} f(x)H_n(x)e^{-x^2}dx.$$

For example, we show below the Hermite expansion for the polynomial $p(x)$. We first need to check that $p(x) \in L^2(\mathbb{R}, e^{-x^2}dx)$, i.e.

$$\int_{\mathbb{R}} p(x)e^{-x^2}dx < \infty.$$

We find of interest to develop the general integral

$$I = \int_{\mathbb{R}} x^n e^{-x^2} dx$$

to show that any polynomial is indeed in $L^2(\mathbb{R}, e^{-x^2}dx)$ and for further computations. One observes that

$$I = ((-1)^n + 1) \int_0^{\infty} x^n e^{-x^2} dx = \begin{cases} 2\int_0^{\infty} x^n e^{-x^2} dx & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd.} \end{cases}$$

If we substitute x in I with the identity $u = x^2$, then we have for n even and finite:

$$I = \int_0^\infty u^{\frac{n-1}{2}} e^{-u} du = \Gamma\left(\frac{n+1}{2}\right) = \Gamma\left(\frac{n}{2} + \frac{1}{2}\right) = \frac{n!}{2^n(\frac{n}{2})!}\sqrt{\pi} < \infty,$$

which concludes that $p(x) \in L(\mathbb{R}, e^{-x^2}dx)$ as a linear combination of powers of x.

Example 3.3 (Polynomial Hermite expansion)**.** *Let $p(x) = x^2 + 2x + 1$. We will compute the Hermite polynomials coefficients d_0, d_1 and d_2:*

$$\begin{aligned} d_0 &= \frac{1}{\sqrt{\pi}}\int_{\mathbb{R}} p(x)H_0(x)e^{-x^2}dx = \frac{1}{\sqrt{\pi}}\left[\Gamma\left(\frac{3}{2}\right) + \Gamma\left(\frac{1}{2}\right)\right] = \frac{3}{2}, \\ d_1 &= \frac{1}{2\sqrt{\pi}}\int_{\mathbb{R}} p(x)H_1(x)e^{-x^2}dx = \frac{1}{\sqrt{\pi}}\int_{\mathbb{R}}(x^3 + 2x^2 + x)e^{-x^2}dx = \frac{1}{\sqrt{\pi}}\left[2\Gamma\left(\frac{3}{2}\right)\right] = 1, \\ d_2 &= \frac{1}{8\sqrt{\pi}}\int_{\mathbb{R}} p(x)H_2(x)e^{-x^2}dx = \frac{1}{8\sqrt{\pi}}\int_{\mathbb{R}}(x^2 + 2x + 1)(4x^2 - 2)e^{-x^2}dx \\ &= \frac{1}{4\sqrt{\pi}}\left[2\Gamma\left(\frac{5}{2}\right) + \Gamma\left(\frac{3}{2}\right) - \Gamma\left(\frac{1}{2}\right)\right] = \frac{1}{4}. \end{aligned}$$

One can also verify that $d_n = 0$ for all $n \geq 3$. From this, we can check $p(x) = d_0 + d_1H_1(x) + d_2H_2(x)$ as expected.

Practically, we will use the recursion formula

$$\begin{cases} H_0(x) = 1 \\ H_1(x) = 2x \\ H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x) \quad \text{for } n \geq 2 \end{cases}$$

to efficiently compute the Hermite polynomials in the following sections.

4 Problem statement and methodology

In this section, we describe the general problem statement and explain our methodology for the inference of interaction kernels. We further specify the method for two different domains: the circle and the real line.

4.1 Problem statement

Throughout this work, we will try to solve the following problem. Let $X_t = (X_t^i)_{i=1}^N \in \mathbb{R}^N$ be a set of N particles in an IPS as described in Section 3.2 with $d = 1$ and $t \in [0, T]$. We then recall (3.1), with the particular case where $\phi(x, y) = \phi(z)$ with $z = x - y$, such that

$$dX_t^{(i)} = v\left(X_t^{(i)}\right)dt + \frac{1}{N}\sum_{n=1}^N \phi\left(X_t^{(i)} - X_t^{(n)}\right)dt + \sigma dB_t^{(i)} \quad i = 1, ..., N, \tag{4.1}$$

where $X_0^{(i)} \sim \mu_0$ for all $i \in \{1, ..., N\}$. We assume $\sigma \in \mathbb{R}$ to be constant and define $(B_t^{(i)})_{t\geq 0}$ as a standard Brownian motion for each $i \in \{1, ..., N\}$. We can also write $v = -V'$ and $\phi = -W'$ with V, W being the confinement potential and the interaction potential, respectively. Using the general notation from Section 3.4, we can rewrite (4.1) as (3.20) by setting

$$f^*(X_t) = \begin{pmatrix} v\left(X_t^{(1)}\right) + \frac{1}{N}\sum_{n=1}^N \phi\left(X_t^{(1)} - X_t^{(n)}\right) \\ \vdots \\ v\left(X_t^{(N)}\right) + \frac{1}{N}\sum_{n=1}^N \phi\left(X_t^{(N)} - X_t^{(n)}\right) \end{pmatrix}. \tag{4.2}$$

4.2 Interaction kernel inference methodology

The general idea is to use the SGDCT method described in Section 3.4 to infer the interacting kernel in (4.1). We assume that we have access to discretized measurements of $X_k^{(i)}$ for all $k \in \{0, \Delta t, ..., T\}$ and $i \in \{1, ..., N\}$. In practice, we will use the Euler-Maruyama discretization method to simulate such data, i.e.

$$X_{k+1}^{(i)} = X_k^{(i)} + v(X_k^{(i)}) + \frac{1}{N}\sum_{n=1}^{N} \phi(X_k^{(i)} - X_k^{(n)})dt + \sigma dB_k^{(i)} \quad i = 1, ..., N. \tag{4.3}$$

This work will cover two distinct particle domains: the circle $\mathbb{T}$ and the real line $\mathbb{R}$. Every experiment will be run with $T = 5 * 10^4$, $\Delta t = 0.01$, $C = 10$ and $C_0 = 100$, such that

$$l_t = \frac{10}{100 + t}.$$

We assume that we want to infer a set of parameters $\theta \in \mathbb{R}^J$. As f^* is a vector-valued function in $\mathbb{R}^N$, we also have $f(X_t, \theta)$ in $\mathbb{R}^N$, and therefore $\nabla_\theta f(X_t, \theta) \in \mathbb{R}^{J \times N}$ defined as

$$\nabla_\theta f(X_t, \theta) = \begin{pmatrix} \frac{\partial f(X_t,\theta)^{(1)}}{\partial \theta^{(1)}} & \cdots & \frac{\partial f(X_t,\theta)^{(N)}}{\partial \theta^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(X_t,\theta)^{(1)}}{\partial \theta^{(J)}} & \cdots & \frac{\partial f(X_t,\theta)^{(N)}}{\partial \theta^{(J)}} \end{pmatrix}. \tag{4.4}$$

Algorithm 1 shows how to implement the SGDCT for IPSs. This would return the entire evolution of the J parameters, which is useful to analyse the behaviour of the inference method. If one is memory-constrained and does not want to keep an array of size $n_{\text{steps}} \times J \times K$, with $K = T/dt$, then it could be reduced to a $n_{\text{steps}} \times J \times 2$ array, with only the previous and current estimates of the J parameters.

Algorithm 1: SGDCT for Interacting Particle Systems

```
Input: $T$, $\Delta t$, $f(x, \theta)$, $((X_k)_{k=0}^{K})_{n=1}^{n_{\text{steps}}}$, $\mu_0^\theta$, $C$, $C_0$
Result: returns the estimates $\theta_t$ for $t \in \{0, \Delta t, ..., K\}$.
for $n = 1, ..., n_{steps}$ do
    $\theta_0^n \sim \mu_0^\theta$
    $X = ((X_k)_{k=0}^{K})_n$ // selecting the $n$-th trajectory of observations
    for $k = \Delta t, ..., T$ do
        $l_k = C/(C_0 + k)$
        $dX_k = X_k - X_{k-1}$
        $\theta_k^n = \theta_{k-1}^n + l_k \sigma^{-2} \nabla_\theta f(X_{k-1}, \theta_{k-1}^n) \left[dX_k - f(X_{k-1}, \theta_{k-1}^n)dt\right]$
    end
end
$\theta_t = \frac{1}{n_{\text{steps}}} \sum_{n=1}^{n_{\text{steps}}} \theta_t^n$
```

4.2.1 Methodology for the circle $\mathbb{T}$

SGDCT First, we discuss the SGDCT method for particles $(X_t^{(i)})_{i=1,...,N}$ periodically bounded between $[0, 2\pi]$, representing angles on the circle. This enables us to benefit from the Fourier series representation being an orthonormal basis as described in Section 3.5. Generally, we will infer ϕ using the truncated Fourier series representation of its related potential W, such that up to a constant K[1],

$$W(x) \approx K + \sum_{j=1}^{J} w_j \cos(jx),$$

where J is the number of modes considered and can be set from the start, or learned via an adaptive behaviour discussed in Section 7. Hence, we will use the SGDCT to infer the weights $\theta := \{w_j\}_{j=1}^{J}$. From this, we specify the function $f(X_t, \theta)$ and $\nabla_\theta f(X_t, \theta)$ as

$$f(X_t, \theta) = \begin{pmatrix} v(X_t^{(1)}) + \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{J} j w_j \sin\left(j\left(X_t^{(1)} - X_t^{(n)}\right)\right) \\ \vdots \\ v(X_t^{(N)}) + \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{J} j w_j \sin\left(j\left(X_t^{(N)} - X_t^{(n)}\right)\right) \end{pmatrix}, \tag{4.5}$$

and

$$\nabla_\theta f(X_t, \theta) = \begin{pmatrix} \frac{1}{N}\sum_{n=1}^{N} \sin\left(X_t^{(1)} - X_t^{(n)}\right) & \dots & \frac{1}{N}\sum_{n=1}^{N} \sin\left(X_t^{(N)} - X_t^{(n)}\right) \\ \vdots & \ddots & \vdots \\ \frac{1}{N}\sum_{n=1}^{N} J\sin\left(J\left(X_t^{(1)} - X_t^{(n)}\right)\right) & \dots & \frac{1}{N}\sum_{n=1}^{N} J\sin\left(J\left(X_t^{(N)} - X_t^{(n)}\right)\right) \end{pmatrix}. \tag{4.6}$$

MLE Secondly, we develop the likelihood maximisation method described in Section 3.3.2 on the circle. To match notations, we have the basis $\{e_i(x) := \cos(ix), i = 1, ..., J\}$, which yields

$$\nabla f_j(X_t) = -\begin{pmatrix} \frac{1}{N}\sum_{n=1}^{N} j\sin\left(j\left(X_t^{(1)} - X_t^{(n)}\right)\right) \\ \vdots \\ \frac{1}{N}\sum_{n=1}^{N} j\sin\left(j\left(X_t^{(N)} - X_t^{(n)}\right)\right) \end{pmatrix}, \quad j = 1, ..., J. \tag{4.7}$$

With $\sigma \in \mathbb{R}$, we write the matrix $\bar{F}$ as

$$\begin{aligned} \bar{f}_{ij} &= \frac{1}{T}\int_0^T \langle \nabla f_i(X_t), \sigma^2 \nabla f_j(X_t)\rangle \, dt \\ &= \frac{\sigma^2 ij}{TN^2}\sum_{k=1}^{N}\int_0^T \left(\sum_{n=1}^{N} \sin\left(i\left(X_t^{(k)} - X_t^{(n)}\right)\right)\right)\left(\sum_{n=1}^{N} \sin\left(j\left(X_t^{(k)} - X_t^{(n)}\right)\right)\right) dt, \end{aligned} \tag{4.8}$$

and the vector $\bar{h}$ as

$$\begin{aligned} \bar{h}^{(i)} &= -\frac{2}{T}\int_0^T \langle \nabla f_i(X_t), dX_t\rangle \\ &= \frac{2i}{TN}\sum_{k=1}^{N}\int_0^T \left(\sum_{n=1}^{N} \sin\left(i\left(X_t^{(k)} - X_t^{(n)}\right)\right)\right) dX_t^{(k)}. \end{aligned} \tag{4.9}$$

Computing the vector $\bar{c}$ as defined in Section 3.3.2 is straightforward, but as we will always have $v \equiv 0$ for the circle, we have $\bar{c} = 0$.

[1] We only consider even potentials, which allows us to use cosines for their Fourier representations.

4.2.2 Methodology for the real line $\mathbb{R}$

SGDCT Here, we discuss the SGDCT method for particles $X_t := (X_t^{(i)})_{i=1,\dots,N}$ with values lying in $\mathbb{R}$. Compared to the torus setting, the Fourier series is no longer an orthonormal basis for this space. As described in Section 3.6, if one can show that $W \in L^2(\mathbb{R}, e^{-x^2}dx)$ then we can express it as

$$W(x) = \sum_{j\geq 0} d_j H_j(x),$$

where $H_j(x)$ is the Physicist's Hermite polynomial of order j. Similarly to Section 4.2.1, we will truncate this representation to a finite number J of Hermite polynomials, i.e. to a polynomial of order J such that

$$W(x) \approx K + \sum_{j=1}^{J} w_j H_j(x).$$

Similarly to the former case, J can be set from the start or learned adaptively using a posteriori relative error. We will use the SGDCT to infer the weights $\theta := \{w_j\}_{j=1}^J$. From this, we can also specify $f(X_t, \theta)$ and $\nabla_\theta f(X_t, \theta)$ for the real line as

$$f(X_t, \theta) = \begin{pmatrix} v(X_t^{(1)}) + \frac{1}{N}\sum_{n=1}^N \sum_{j=1}^J 2jw_jH_{j-1}\left(X_t^{(1)} - X_t^{(n)}\right) \\ \vdots \\ v(X_t^{(N)}) + \frac{1}{N}\sum_{n=1}^N \sum_{j=1}^J 2jw_jH_{j-1}\left(X_t^{(N)} - X_t^{(n)}\right) \end{pmatrix}, \tag{4.10}$$

and

$$\nabla_\theta f(X_t, \theta) = \begin{pmatrix} \frac{1}{N}\sum_{n=1}^N 2H_0\left(X_t^{(1)} - X_t^{(n)}\right) & \dots & \frac{1}{N}\sum_{n=1}^N 2H_0\left(X_t^{(N)} - X_t^{(n)}\right) \\ \vdots & \ddots & \vdots \\ \frac{1}{N}\sum_{n=1}^N 2JH_{J-1}\left(X_t^{(1)} - X_t^{(n)}\right) & \dots & \frac{1}{N}\sum_{n=1}^N 2JH_{J-1}\left(X_t^{(N)} - X_t^{(n)}\right) \end{pmatrix}. \tag{4.11}$$

MLE We specify here the likelihood maximisation method described in Section 3.3.2 for the real line. To match notations, we have the basis $\{e_j(x) := H_j(x), j = 1, \dots, J\}$, which yields

$$\nabla f_j(X_t) = \begin{pmatrix} \frac{1}{N}\sum_{n=1}^N 2jH_{j-1}\left(X_t^{(1)} - X_t^{(n)}\right) \\ \vdots \\ \frac{1}{N}\sum_{n=1}^N 2jH_{j-1}\left(X_t^{(N)} - X_t^{(n)}\right) \end{pmatrix}, \quad j = 1, \dots, J. \tag{4.12}$$

With $\sigma \in \mathbb{R}$, we write the matrix $\bar{F}$ as

$$\begin{aligned} \bar{f}_{ij} &= \frac{1}{T}\int_0^T \langle \nabla f_i(X_t), \sigma^2 \nabla f_j(X_t)\rangle\, dt \\ &= \frac{4\sigma^2 ij}{TN^2}\sum_{k=1}^N \int_0^T \left(\sum_{n=1}^N H_{i-1}\left(X_t^{(k)} - X_t^{(n)}\right)\right)\left(\sum_{n=1}^N H_{j-1}\left(X_t^{(k)} - X_t^{(n)}\right)\right) dt, \end{aligned} \tag{4.13}$$

and the vector $\bar{h}$ as

$$\begin{aligned} \bar{h}^{(i)} &= -\frac{2}{T}\int_0^T \langle \nabla f_i(X_t), dX_t\rangle \\ &= -\frac{4i}{TN}\sum_{k=1}^N \int_0^T \left(\sum_{n=1}^N H_{i-1}\left(X_t^{(k)} - X_t^{(n)}\right)\right) dX_t^{(k)}. \end{aligned} \tag{4.14}$$

We will often need to add a confinement potential V such that $v = -V'$. Hence, the vector $\bar{c}$ as defined in Section 3.3.2 will be specified for each choice of confinement potential throughout the work on the real line.

4.2.3 Complexity comparison between SGDCT and MLE

We here study the complexity of both inference methods. We assume that the data is already generated and that $f(x, \theta)$ is a sum of J weighted basis elements, as this is the setting at hand, both for the real line and circle.

SGDCT We first show the cost to compute each quantity in the Algorithm 1, and then the different matrix-vector operation costs.

Table 1: Quantity computation complexity for SGDCT

Quantity	dX_t	$f(X_t, \theta_t)$	$\nabla_\theta f(X_t, \theta_t)$
Cost	$\mathcal{O}(N)$	$\mathcal{O}(N^2 J)$	$\mathcal{O}(N^2 J)$

As shown in Table 1, the cost to obtain $f(X_t, \theta_t)$ is $\mathcal{O}(N^2 J)$, as each element is the mean over every particle of a weighted sum of J elements of the chosen basis. This cost will be paid at every timestep. As $\nabla_\theta f(X_t, \theta) \in \mathbb{R}^{J \times N}$, we compute its product with dX_t and $f(X_t, \theta_t)$ for a cost of $\mathcal{O}(NJ)$. Based on this short cost analysis, the total cost of a parameter update for all timesteps is $\mathcal{O}(N^2 JK)$, where $K = T/dt$. If we further assume that this algorithm is run n_{steps} times, then it becomes

$$\text{cost}_{\text{SGDCT}} = \mathcal{O}\left(N^2 J K n_{\text{steps}}\right).$$

MLE We now have a look at the MLE for IPSs. Table 2 outlines the costs for computing the J gradients ∇f_j, the matrix $\bar{F}$, and vectors $\bar{h}$ and $\bar{c}$ for a single timestep. We discretize the integrals of (3.14), (3.15) and (3.19) as a sum of K elements, with $K = T/dt$. In our experiments, we derive θ_t from the MLE at each timestep, but we only add one term to the cumulative sum of the discretized integral, thereby avoiding the recomputation of all integrals at every timestep. Consequently, the total costs are multiplied by K, similar to the SGDCT. Finally, we need to invert $\bar{F}$ in $\mathcal{O}(J^3)$, and then an additional $\mathcal{O}(J^2)$ to compute $\bar{F}^{-1}(\bar{h} + \bar{c})$ at each step. Based on these computations and assuming n_{steps} repetitions, we obtain the total cost

$$\text{cost}_{\text{MLE}} = \mathcal{O}\left(K n_{\text{steps}}(N^2 J + N J^2 + J^3)\right).$$

Table 2: Quantity computation complexity for the MLE

Quantity	$\nabla f_j(X_t)$	$\bar{F}$	$\bar{h}$	$\bar{c}$
Cost	$\mathcal{O}(N^2)$	$\mathcal{O}(NJ^2)$	$\mathcal{O}(NJ)$	$\mathcal{O}(NJ)$

5 Main results on the circle $\mathbb{T}$

This section presents the main results of this work on the circle. Firstly, we apply the SGDCT to infer parameters with different interaction kernels. We also study its behaviour concerning the number of modes for the Fourier series and the number of particles in the system. We will assume $v \equiv 0$ for this entire section. We refer to [9] for further information on McKean-Vlasov equations on the torus.

5.1 SGDCT for the weighted cosines potential

This section explores the methodology described in Section 4.2.1 for an interacting potential defined as a weighted sum of cosines on the circle. Namely, we define

$$W(x) = \sum_{j=1}^{\tilde{J}} \gamma_j \cos(jx), \text{ which yields } \phi(x) = -W'(x) = \sum_{j=1}^{\tilde{J}} j\gamma_j \sin(jx).$$

We can freely choose γ_j for $j = 1, ..., \tilde{J}$ and assert whether the SGDCT finds every w_j close to γ_j when we set $J = \tilde{J}$. Choosing this toy kernel will help us analyse the behaviour of the methodology and assess whether it successfully finds the chosen weights.

Let $(\gamma_j)_{j=1}^{\tilde{J}} = [1, 1/2, ..., 1/\tilde{J}]$ such that the potential we try to infer is given as

$$W(x) = \sum_{j=1}^{\tilde{J}} \frac{1}{j} \cos(jx).$$

Then, this setup yields the IPS following the SDE

$$dX_t^{(i)} = \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{\tilde{J}} \sin\left(j\left(X_t^{(i)} - X_t^{(n)}\right)\right) dt + dB_t^{(i)} \quad i = 1, ..., N \tag{5.1}$$

from which we will simulate n_{steps} paths using the Euler-Maruyama method for the true dynamics.

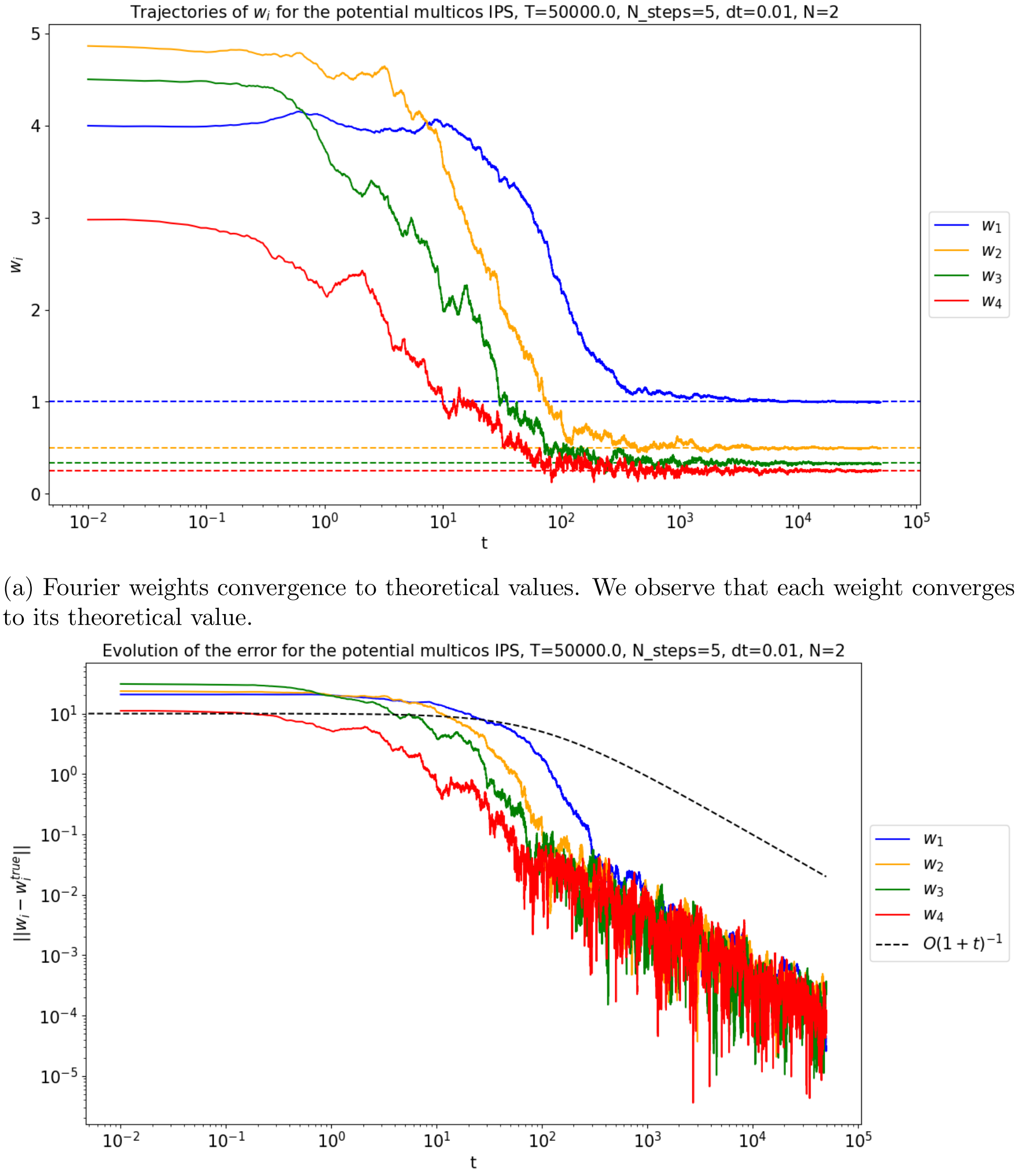
5.1.1 Study over the number of modes in truncated Fourier series

Case $J = \tilde{J}$ In this case, one would hope that every w_j converges to its corresponding true value γ_j for $j = 1, .., \tilde{J}$. For $N = 2$ particles and $\gamma_j = 1/j$ for $j = 1, ..., 4$, we obtain a convergence of w_j to γ_j for the J weights as shown in Figure 2a. Moreover, Figure 2b shows the mean squared error (MSE) evolution and its closeness to the theoretical rate of $\mathcal{O}(1+t)^{-1}$. For both figures, each line represents the average weight estimates over the n_{steps} trajectories.

Case $J < \tilde{J}$ Such as in the case of $J = \tilde{J}$, we would like the weights w_j to be near γ_j for $j = 1, ..., J$, but we might observe some discrepancies as this will be an approximation due to the series truncation to $J < \tilde{J}$ modes. This is indeed what Figure 3a shows, with the two weights close but not exactly converging to their theoretical values of 1 and 1/2. This can be explained as the two modes related to the two remaining true weights are not considered, which introduces an approximation. This approximation is non-negligible as the weights w_3 and w_4 have a magnitude of 1/3 and 1/4 respectively.

Case $J > \tilde{J}$ In this case, one would hope that every w_j converges to its corresponding true value γ_j for $j = 1, .., \tilde{J}$, and that every other mode for $j > \tilde{J}$ converges to zero. Figure 3b shows this is the case and that all weights w_j corresponding to non-existent modes in the original potential (namely w_5, w_6) converge to zero.

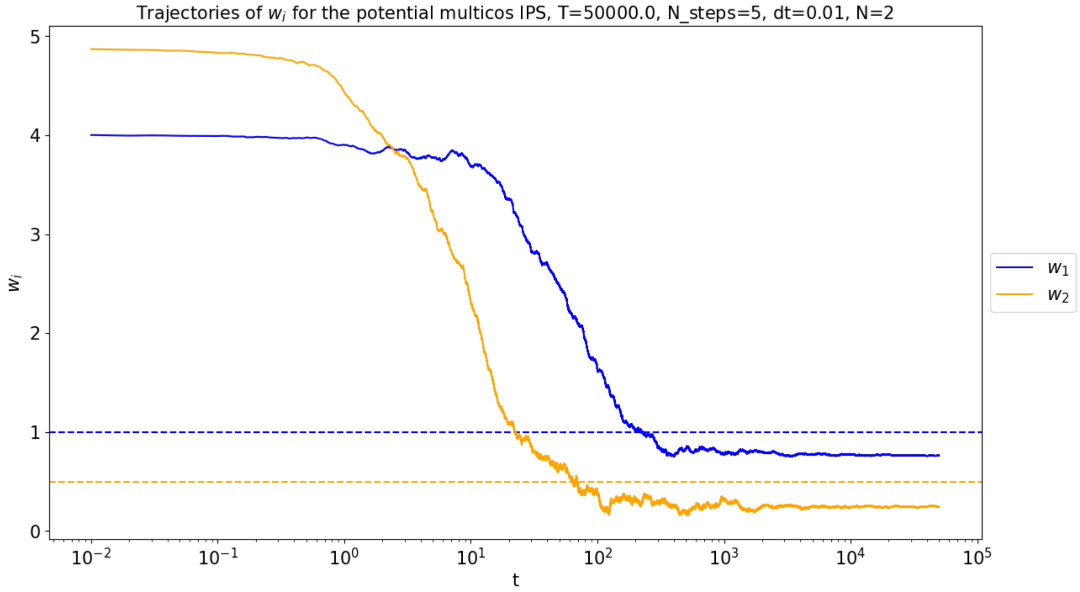
Based on these observations, generalising the method by not specifying the number of modes required in the truncated Fourier series representation is of interest. This study is done in Section 7 where we explore the adaptiveness of the method to efficiently find the number of weights required for a satisfactory kernel inference.



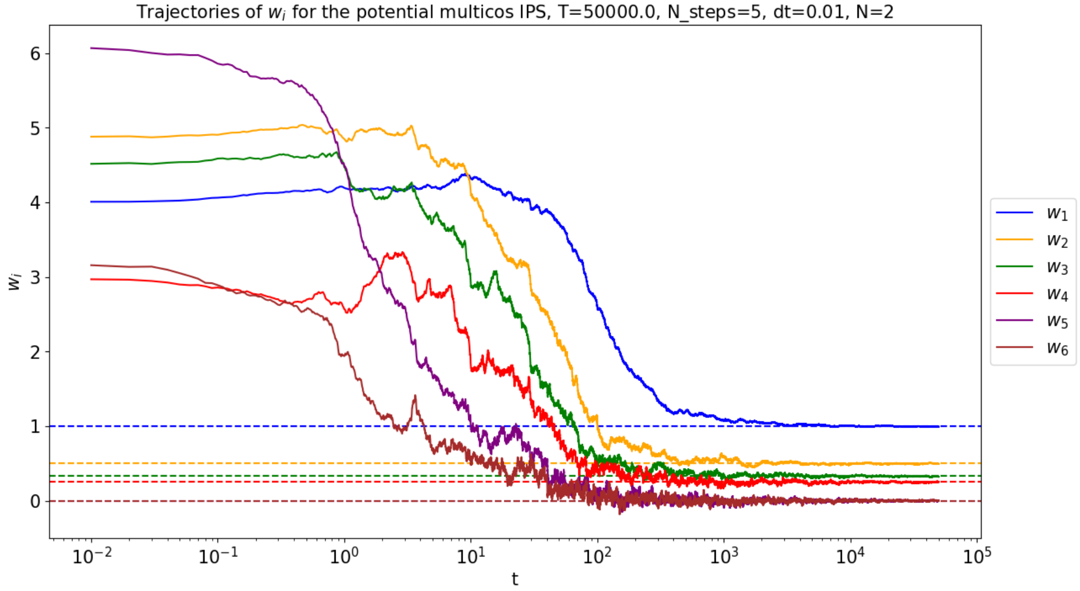
(a) Fourier weights convergence to theoretical values. We observe that each weight converges to its theoretical value.

(b) Fourier weights error rates compared to theoretical convergence rate. The theoretical rate of $\mathcal{O}(1+t)^{-1}$ is achieved.

Figure 2: Error rates and convergence plots for the weighted cosine sum interaction potential with $\tilde{J} = J$. Figure 2a shows that the estimated weights converge to their true value as time increases. On the other hand, we observe in Figure 2b that the error follows the theoretical rate of $\mathcal{O}(1+t)^{-1}$ (dotted line). For both figures, each line represents the average weight estimates over the number of trajectories. The theoretical weights are given as $\{1, 1/2, 1/3, 1/4\}$ and the error computation is done using the MSE formula.

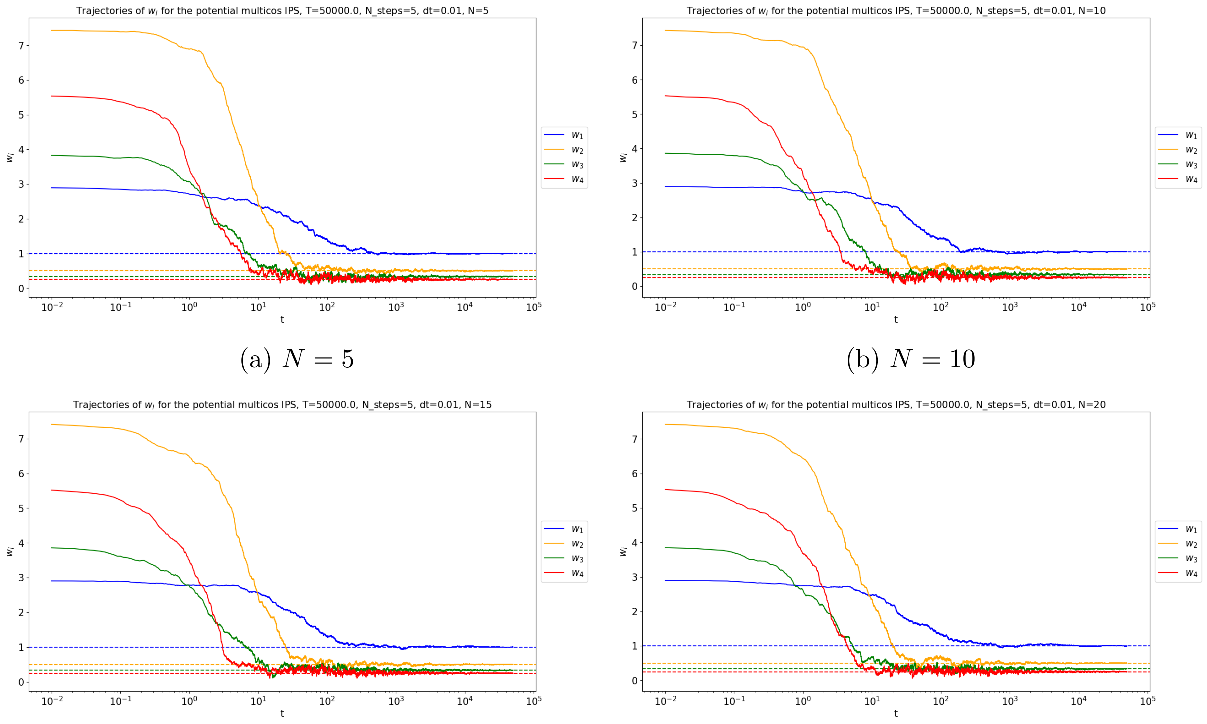


(a) $J = 2$



(b) $J = 6$

Figure 3: Study over the number of Fourier weights in the truncated Fourier series. Figure 3a shows that the weight estimates w_1, w_2 do not converge to γ_1, γ_2. This is caused by the unexplained variance from the two remaining weights that are not considered. Figure 3b shows that the weight estimates converge to their theoretical values. The first four weights converge to their four true values (blue,orange,green and red lines) while the two additional weights (purple and brown lines) converge to zero, as their respective modes do not appear in the studied interaction potential.



(a) $N = 5$

(b) $N = 10$

(c) $N = 15$

(d) $N = 20$

Figure 4: Study over the number of particles in the system, for $J = 4$. We observe that the convergence of the weights does not seem impacted by the number of particles.

5.1.2 Study over the number of interacting particles in the system

In this subsection, we study the behaviour of SGDCT based on the number of interacting particles in the system. Figure 4 illustrates the convergence of the weights for the truncated Fourier Series. We do not observe a significant difference in convergence to the theoretical weights as N grows. However, the more particles in the system, the higher the computation time, as shown in Section 4.2.3.

5.1.3 Comparison between SGDCT and MLE

We use here the MLE described in Section 3.3.2 and detailed for the circle in Section 4.2.1. Figure 5 shows the error rates for the multi-cosine interaction potential. The error follows the rate of $\mathcal{O}(t)^{-1}$, with some stability issues for the very first steps of the method. One can explain these stability issues by the low amount of data used for the first numerical integrals necessary for the construction of the matrix $\bar{F}$ and vector $\bar{h}$. Indeed, as time grows, this issue disappears as we integrate over a longer time frame. We note that as the true weight magnitude decreases, its related error also tends to be lower. When we compare this rate to the SGDCT rate of $\mathcal{O}(1+t)^{-1}$, and from the numerical experiments, we observe that the SGDCT is more stable in the first iterations and converges with approximately the same rate as time grows.

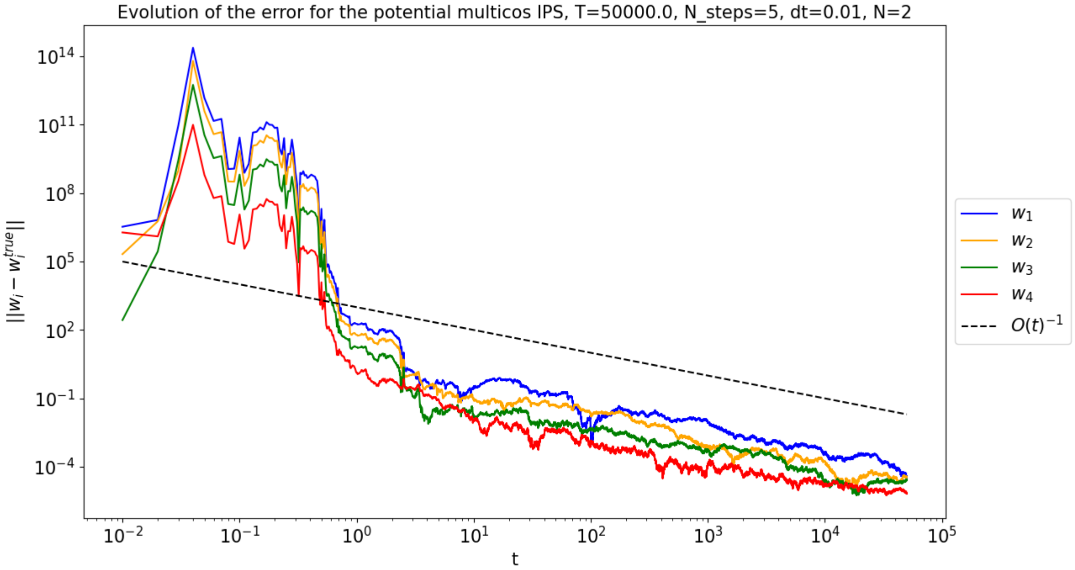


Figure 5: Error rates for the multi-cosine interaction potential for the MLE. The error follows the rate of $\mathcal{O}(t)^{-1}$, with some stability issues for the very first steps of the method. The stability issues might be due to the integral discretization. As time grows, this issue disappears as we integrate over a longer time frame. Each line represents the average of the weight estimates over the number of trajectories. The theoretical weights are given as $\{1, 1/2, 1/3, 1/4\}$ and the error computation is done using the MSE formula.

5.2 SGDCT for the von Mises potential

In this section, we explore the behaviour of the estimation methodology with the von Mises interaction potential. We can write the von Mises potential as

$$W(x) = \frac{\exp(\kappa \cos(x - \mu))}{2\pi I_0(\kappa)}, \tag{5.2}$$

where μ and κ are the location and concentration of the distribution, respectively. $I_n(x)$ is the modified Bessel function of the first kind of order n defined as

$$I_n(x) = \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(k+n+1)} \left(\frac{x}{2}\right)^{2k+n}, \tag{5.3}$$

where $\Gamma(x)$ is the Gamma function. A convenient property of the von Mises potential is that it can be expressed as a Fourier series with explicit weights, such that

$$W(x) = \frac{1}{2\pi} + \frac{1}{\pi I_0(\kappa)} \sum_{j=1}^{\infty} I_j(\kappa) \cos(jx). \tag{5.4}$$

such that we obtain by recalling $\phi = -W'$,

$$\phi(x) = \frac{1}{\pi I_0(\kappa)} \sum_{j=1}^{\infty} j I_j(\kappa) \sin(jx). \tag{5.5}$$

We are therefore working with the system of interacting particles governed by the SDEs

$$dX_t^{(i)} = \frac{1}{N\pi I_0(\kappa)} \sum_{n=1}^{N} \sum_{j=1}^{\infty} j I_j(\kappa) \sin\left(j\left(X_t^{(i)} - X_t^{(n)}\right)\right) dt + dB_t^{(i)} \quad i = 1, ..., N. \tag{5.6}$$

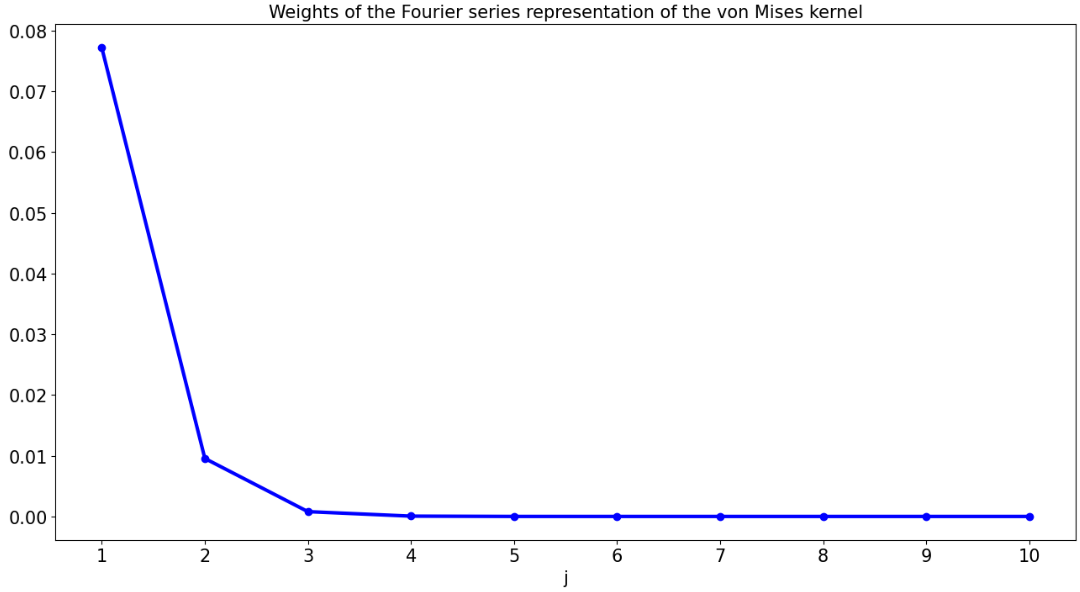


Figure 6: Weights of the Fourier series representation for the von Mises interaction potential (first 10 weights). We observe a steep decrease such that after $j = 4$, the weights are very close to zero. One can therefore argue that we can approximate this kernel with very few Fourier series modes.

We set $\mu = 0$ and $\kappa = 0.5$ for the numerical experiments. We also recall $N = 2$ particles, $J = 4$ Fourier modes and $n_{\text{steps}} = 5$ trajectories. As shown in Figure 7, we observe a convergence matching the theoretical rate of $\mathcal{O}(1+t)^{-1}$. Figure 6 illustrates the steep decay of the weights for the Fourier series representation of the von Mises distribution. Hence, only a few modes are required to reproduce the von Mises kernel.

5.2.1 Comparison between SGDCT and MLE

Similarly to the multi-cosine potential previously discussed, we observe some instability for the first steps of the MLE estimator for the von Mises potential, as shown in Figure 8. We also observe a similar error ranking related to the magnitude of the theoretical weights.

5.3 SGDCT for the Onsager potential

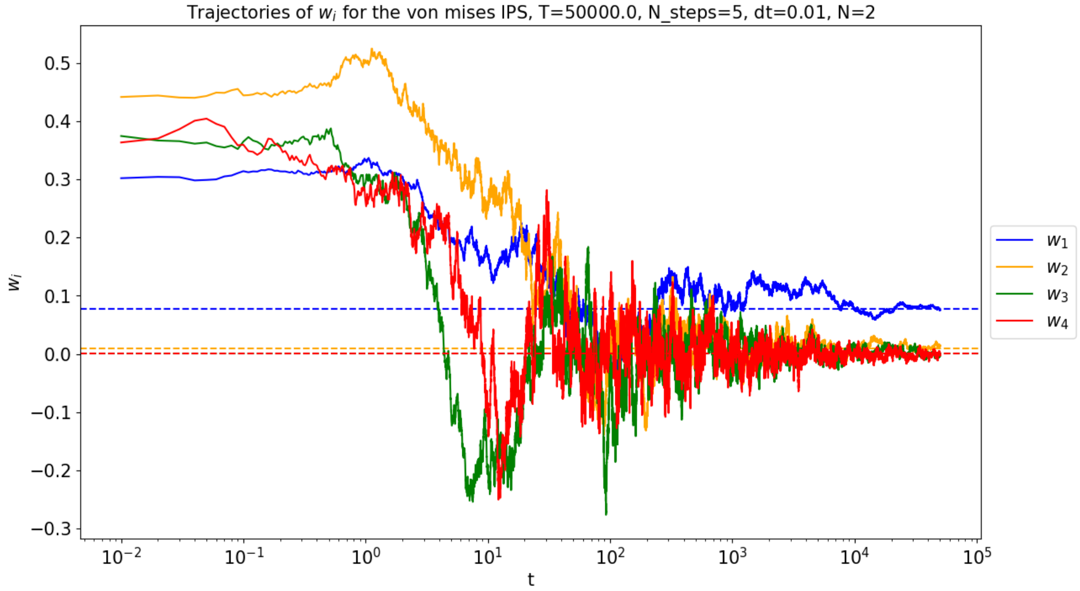
The Onsager Kernel can be described as

$$W(x) = |\sin(x)|, \text{ and } \quad \phi(x) = -W'(x) = -\operatorname{sign}(\sin(x))\cos(x).$$

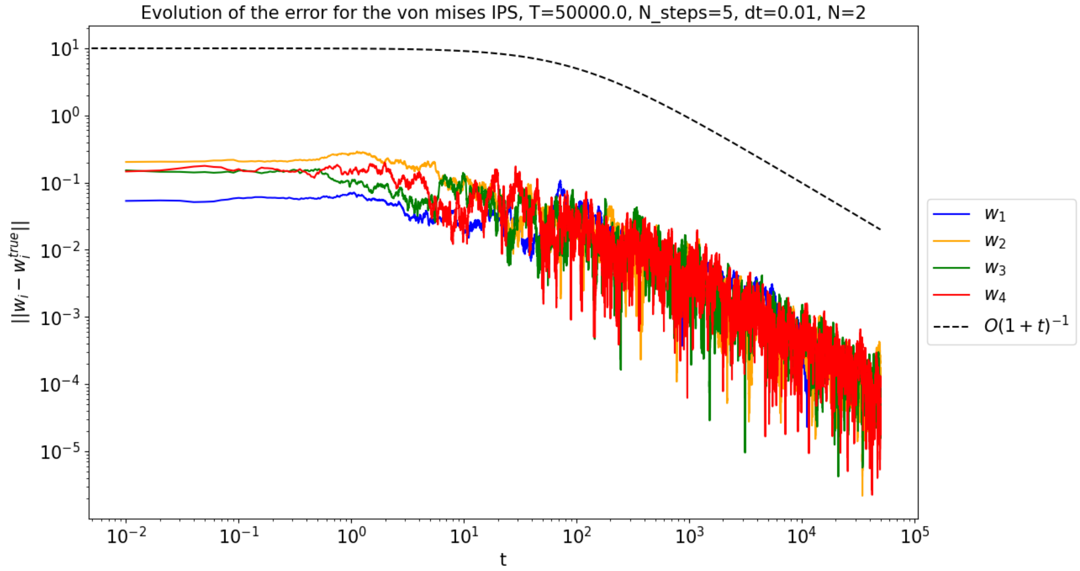
It was invented in 1949 by Lars Onsager to model the phase transitions of equilibria of dilute colloidal solutions of rod-like molecules, especially between the nematic and isotropic phases. In this model, x represents the difference between the angles of two rod-like molecules. We refer to [34; 35] for more information about this kernel in other contexts.

With the Onsager kernel, we have for X_t bounded between $[0, 2\pi]$ that

$$dX_t^{(i)} = -\frac{1}{N}\sum_{n=1}^{N} \operatorname{sign}\left(\sin\left(X_t^{(i)} - X_t^{(n)}\right)\right)\cos\left(X_t^{(i)} - X_t^{(n)}\right) dt + dB_t^{(i)} \quad i = 1, ..., N. \tag{5.7}$$



(a) Fourier weights convergence to theoretical values. We observe the weights corresponding to their true value as time grows.



(b) Fourier weights error rates compared to theoretical convergence rate. We observe a convergence rate matching the theoretical rate of $\mathcal{O}(1+t)^{-1}$.

Figure 7: Error rates and convergence plots for the von Mises interaction potential. Figure 7a shows that the estimated weights converge to their true value as time increases. On the other hand, we observe in Figure 7b that the error follows the theoretical rate of $\mathcal{O}(1+t)^{-1}$. For both figures, each line represents the average of the weight estimates over the number of trajectories. The theoretical weights are given as $\left(\frac{I_j(\kappa)}{\pi I_0(\kappa)}\right)_{j=1,\ldots,4}$ and the error computation is done using the MSE formula.

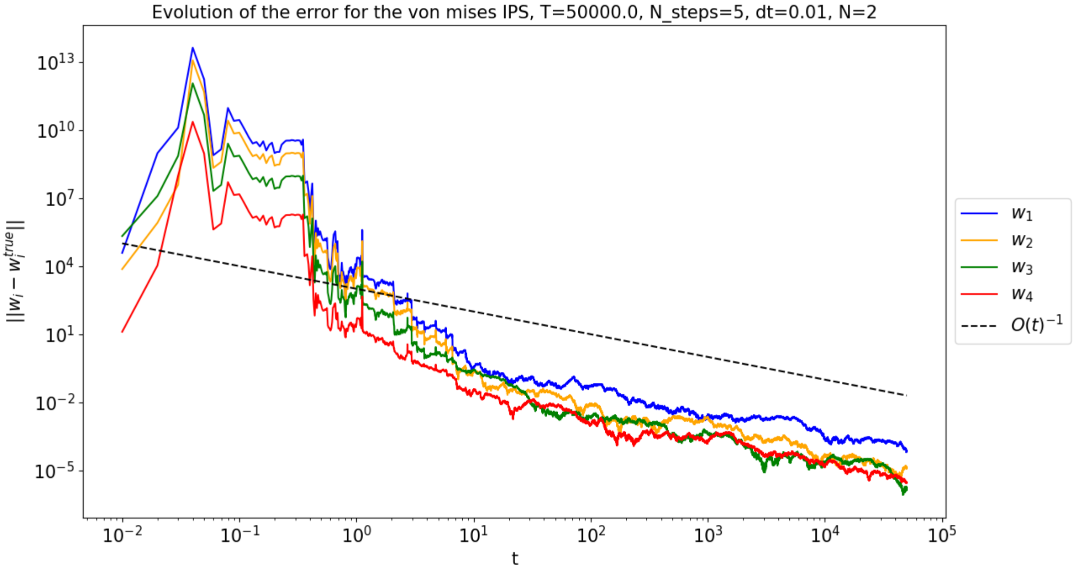


Figure 8: Error rates for the von Mises interaction potential for the MLE. The error follows the rate of $\mathcal{O}(t)^{-1}$, with some stability issues for the very first steps of the method. The stability issues might be due to the integral discretisation. As time grows, this issue disappears as we integrate over a longer time frame. Each line represents the average of the weight estimates over the number of trajectories. The theoretical weights are given as $\left(\frac{I_j(\kappa)}{\pi I_0(\kappa)}\right)_{j=1,\dots,4}$ and the error computation is done using the MSE formula.

One can express W with its Fourier series transform as shown in [31] such that

$$W(x) = \frac{2}{\pi} - \frac{4}{\pi}\sum_{j=1}^{\infty}\frac{1}{4j^2-1}\cos(2jx). \tag{5.8}$$

As the Onsager kernel is even, one can try to approximate it with its truncated Fourier series representation, such that

$$W(x) \approx K + \sum_{j=1}^{J} w_j \cos(jx).$$

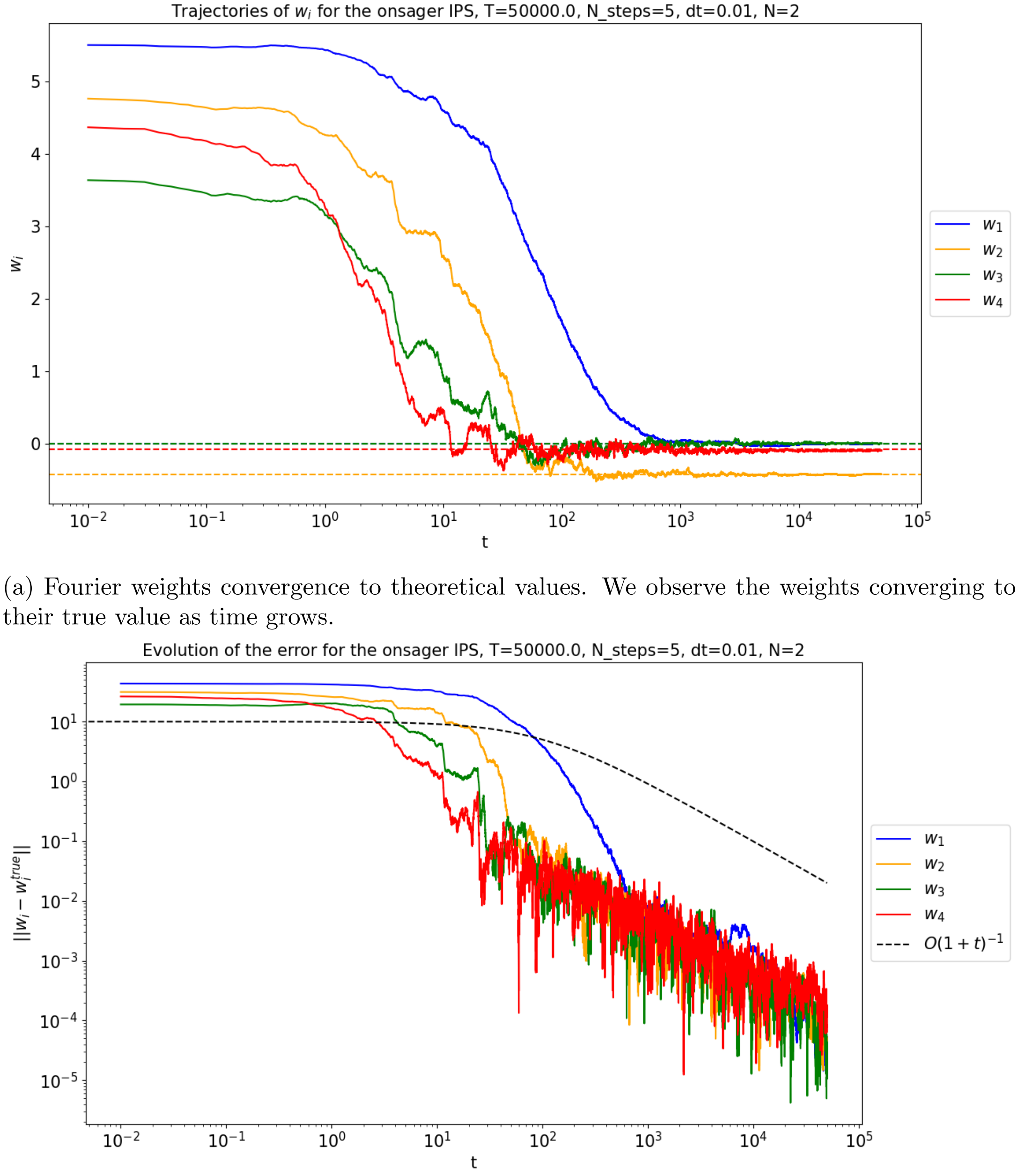
Figure 9 shows convergence for all the modes to their theoretical values and that only even modes are non-trivial. This is expected due to the multiplicative factor two in the cosine expression in (5.8). One notes that the first weight converges slower compared to the other weights.

5.3.1 Comparison between SGDCT and MLE

Similarly to other kernels on the torus, we observe some instability for the first steps of the MLE estimator, with errors rising to 10^{17} which is several order of magnitude higher than the two previous kernels, as shown in Figure 10. We also note that the fourth weight w_4 seems to reach some plateau after $t = 10^3$.

5.4 SGDCT for opinion dynamics

The last type of kernel that will be covered on the torus is linked to opinion dynamics. We consider the following system of interacting particles described in [40]. Namely, we study the



(a) Fourier weights convergence to theoretical values. We observe the weights converging to their true value as time grows.

(b) Fourier weights error rates compared to theoretical convergence rate. We observe a convergence rate matching the theoretical rate of $\mathcal{O}(1+t)^{-1}$. The blue line (w_1) shows a slower error convergence compared to other weights.

Figure 9: Error rates and convergence plots for the Onsager interaction potential. Figure 9a shows that the estimated weights converge to their true value as time increases. We note that only even modes are non-trivial, due to the kernel definition (5.8). Figure 9b outlines that the error follows the theoretical rate of $\mathcal{O}(1+t)^{-1}$. For both figures, each line represents the average of the weight estimates over the number of trajectories. The theoretical weights are given as $-\frac{4}{\pi(4j^2-1)}$ for j even and 0 otherwise. The error computation is done using the MSE formula.

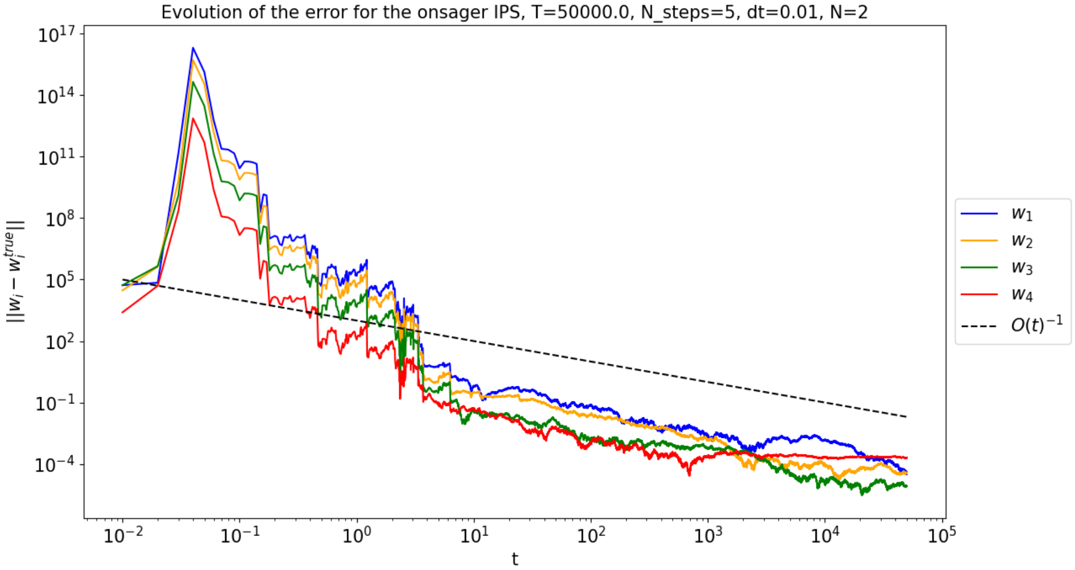


Figure 10: Error rates for the Onsager interaction potential for the MLE. The error follows the rate of $\mathcal{O}(t)^{-1}$, with some stability issues for the very first steps of the method. The stability issues might be due to the integral discretisation. As time grows, this issue disappears as we integrate over a longer time frame. Each line represents the average of the weight estimates over the number of trajectories. The theoretical weights are given as $-\frac{4}{\pi(4j^2-1)}$ for j even and 0 otherwise. The error computation is done using the MSE formula.

model

$$dX_t^{(i)} = -\frac{1}{N}\sum_{n=1}^{N}\phi_\theta(||X_t^{(i)} - X_t^{(n)}||)(X_t^{(n)} - X_t^{(i)})dt + dB_t^{(i)} \quad n = 1, ..., N, \tag{5.9}$$

with the function ϕ_θ as defined in [4]

$$\phi_\theta(r) = \begin{cases} \theta_1 & ||r|| \leq \theta_2 \\ 0 & \text{else.} \end{cases} \tag{5.10}$$

From this, we can define W with a given constant $K \in \mathbb{R}$ that satisfies $\phi = -W'$, such that

$$\phi(x) = -\phi_\theta(x)x, \quad \text{and } W(x) = \begin{cases} \frac{\theta_1 x^2}{2} + K & \|x\| \leq \theta_2 \\ K & \text{else.} \end{cases}$$

We can rewrite this opinion dynamics potential using its Fourier series expression

$$W(x) = \frac{1}{2}A_0 + \sum_{n=1}^{\infty} A_n \cos(nx)$$

where $A_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} W(x)dx$ and $A_n = \frac{1}{\pi}\int_{-\pi}^{\pi} W(x)\cos(nx)dx$ which yields (assuming $0 \leq \theta_2 \leq \pi$):

$$\begin{aligned} A_0 &= \frac{1}{\pi}\int_{-\pi}^{\pi}\left(K + \frac{\theta_1}{2}x^2\mathbb{1}_{\|x\|\leq\theta_2}\right)dx = 2K + \frac{1}{\pi}\int_{-\theta_2}^{\theta_2}\frac{\theta_1}{2}x^2dx = 2K + \frac{\theta_1\theta_2^3}{3\pi} \\ A_n &= \frac{1}{\pi}\int_{-\pi}^{\pi}\left(K + \frac{\theta_1}{2}x^2\mathbb{1}_{\|x\|\leq\theta_2}\right)\cos(nx)dx = \frac{1}{\pi}\int_{-\theta_2}^{\theta_2}\frac{\theta_1}{2}x^2\cos(nx)dx \\ &= \frac{\theta_1}{\pi n^3}\left[(n^2\theta_2^2 - 2)\sin(n\theta_2) + 2n\theta_2\cos(n\theta_2)\right]. \end{aligned} \tag{5.11}$$

We observe again that the resulting potential W is even, and we can therefore approximate it with its truncated Fourier series transform

$$W(x) \approx K + \sum_{j=1}^{J} w_j \cos(jx).$$

For the numerical experiments, we set $\theta_1 = 1$ and $\theta_2 = 3$ and try to infer γ_j for $j = 1, ..., \infty$ such that

$$W(x) = K + \frac{\theta_1\theta_2^3}{6\pi} + \sum_{j=1}^{\infty} \underbrace{\left(\frac{\theta_1}{\pi j^3}\left[(j^2\theta_2^2 - 2)\sin(j\theta_2) + 2j\theta_2\cos(j\theta_2)\right]\right)}_{\gamma_j} \cos(jx).$$

Unfortunately, we do not observe the convergence of the weights to their theoretical values here, as shown in Figure 11. This is due to the non-smoothness of ϕ and its two discontinuities at $\pm\theta_2$. We also note that there is no clear decay in the weights for the Fourier series representation, which means that we would need a high number of modes to correctly represent the interaction kernel (in hundreds), while we only use four modes here.

5.4.1 Comparison between SGDCT and MLE

Figure 12 shows similar results for the MLE compared to the SGDCT for the opinion dynamics kernel. Indeed, both errors stagnate around 10^{-1} after $t = 10^2$. However, we still observe an instability for the very first steps of the MLE, while this is not observed for the SGDCT.

6 Main results on the real line $\mathbb{R}$

We cover here the SGDCT methodology on the real line, with the Hermite polynomials expansion as discussed in 4.2.2.

6.1 SGDCT for the weighted Hermite potential

We define here the interacting potential as a weighted sum of Hermite polynomials in $\mathbb{R}$, similar to the circle case in Section 5.1. Namely, we define

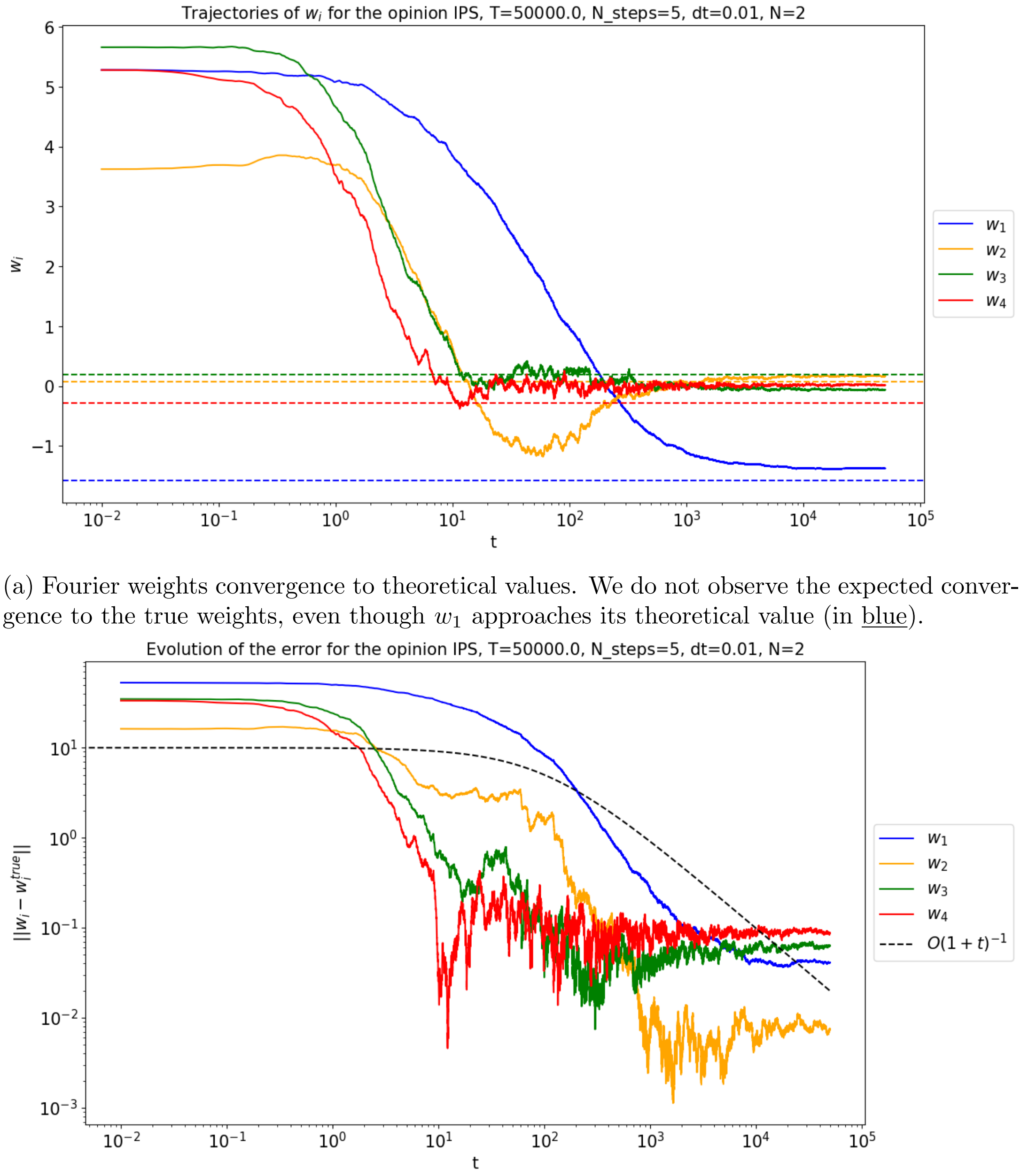
$$W(x) = \sum_{j=1}^{\tilde{J}} \gamma_j H_j(x), \text{ which yields } \phi(x) = -W' = -\sum_{j=1}^{\tilde{J}} 2j\gamma_j H_{j-1}(x).$$

As for the cosine potential discussed in Section 5.1, we can freely choose γ_j for $j = 1, ..., \tilde{J}$. Let $(\gamma_j)_{j=1}^{\tilde{J}} = [1, 1/2, ..., 1/\tilde{J}]$ such that the potential we try to infer is given as

$$W(x) = \sum_{j=1}^{\tilde{J}} \frac{1}{j} H_j(x).$$

We will also introduce a quadratic Ornstein-Uhlenbeck confining potential

$$V(x) = \frac{1}{2}x^2.$$



(a) Fourier weights convergence to theoretical values. We do not observe the expected convergence to the true weights, even though w_1 approaches its theoretical value (in blue).

(b) Fourier weights error rates compared to theoretical convergence rate. The error does not decrease according to its theoretical rate as time grows.

Figure 11: Error rates and convergence plots for the opinion dynamics interaction potential. Figure 11a shows that the estimated weights diverge from their true value. Figure 11b shows that the error does not follow the theoretical rate of $\mathcal{O}(1+t)^{-1}$. For both figures, each line represents the average of the weight estimates over the number of trajectories. The error computation is done using the MSE formula.

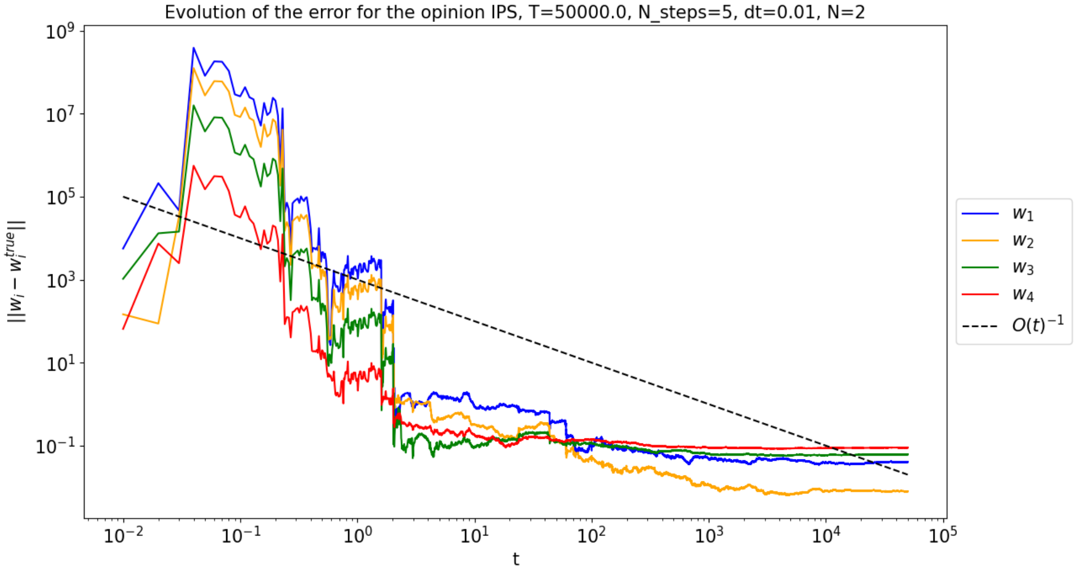


Figure 12: Error rates for the opinion dynamics interaction potential for the MLE. We do not observe convergence to the theoretical weights, as it stagnates after $t = 10^2$. Each line represents the average of the weight estimates over the number of trajectories. The error computation is done using the MSE formula.

Then, this setup yields the IPS following the SDE

$$dX_t^{(i)} = -X_t^{(i)} - \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{\tilde{J}} 2H_{j-1}\left(X_t^{(i)} - X_t^{(n)}\right) dt + dB_t^{(i)} \quad i = 1, ..., N \tag{6.1}$$

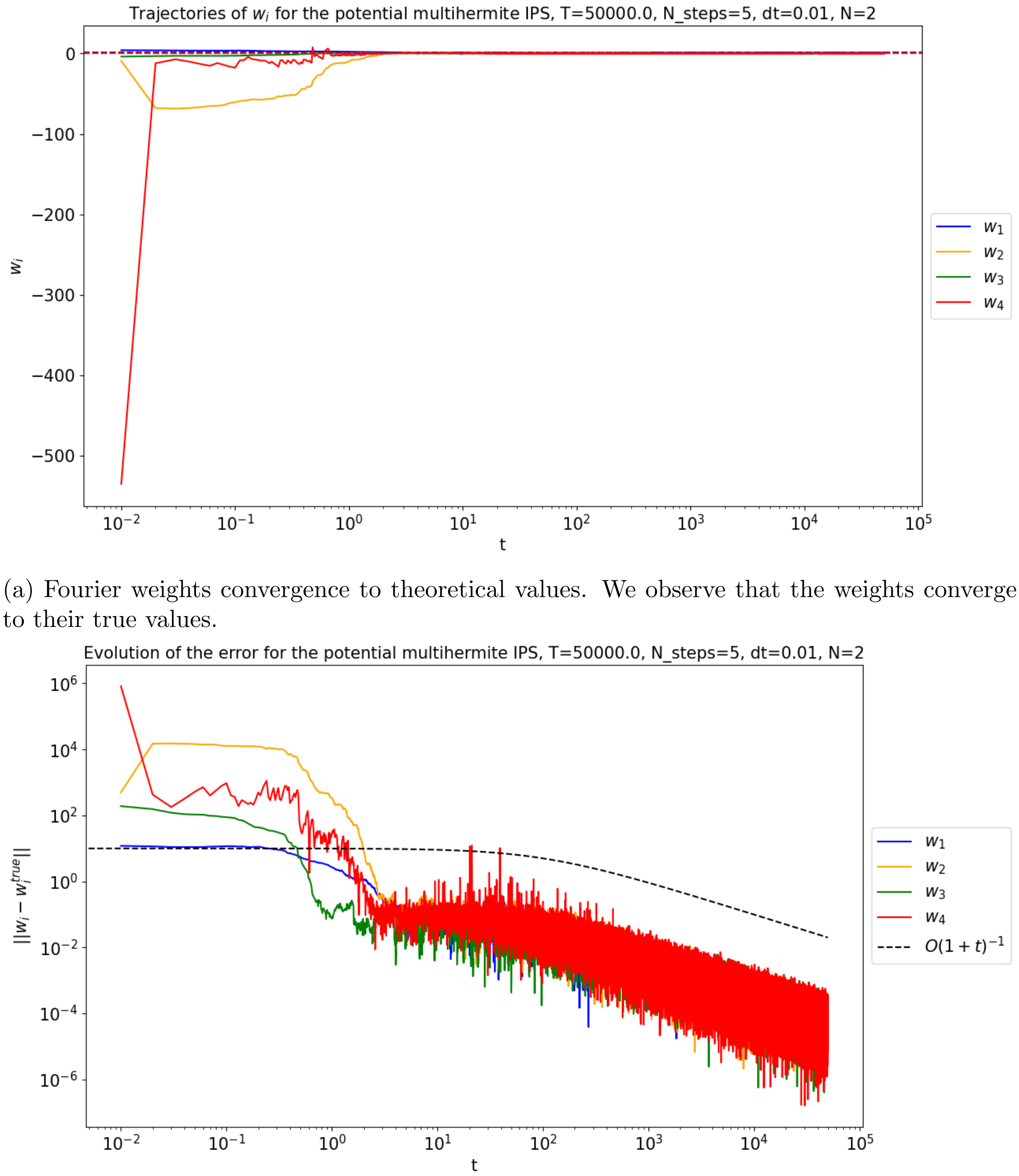
from which we will simulate n_{steps} paths using the Euler-Maruyama method for our true dynamics.

Figure 13 shows the convergence and error rates for the estimated weights of the Hermite expansion. The high variance up to $t = 10^2$ of w_4 is due to the high order of Hermite polynomials used in the nonparametric expansion of $f(x, \theta)$. Indeed, (3.21) shows that $f(x, \theta)$ takes dX_t as input, and for $J = 4$, the high order leading term is $16x^4$, which is quite high. This leads to high values for w_4, as shown in Figure 13a.

This shows that the SGDCT inference method using Hermite polynomials is of interest for potentials that can be expressed using low-order Hermite polynomials. Indeed, if we let J grow significantly, the values of $f(x, \theta)$ would increase as such as the parameter estimates would not converge properly. This can be seen as a major limitation of the technique on the real line, and more generally to unbounded domains where values of $f(x, \theta)$ can reach high levels. One notes that this was not the case on the torus, as the basis was bounded in $[-1, 1]$, which helps avoid stability issues. A workaround for this instability issue is to add a strong confinement potential such that deltas of particle positions dX_t remain low enough, as the particles would be contained in a smaller domain.

6.1.1 Comparison between SGDCT and MLE

Now, we explore the MLE on the real line. Here, we set $V(x) = \frac{1}{2}x^2$, such that $v(x) = -x$ for the vector $\bar{c}$ defined in Section 3.3.2. As shown in Figure 14 the MLE is non-conclusive for the first



(a) Fourier weights convergence to theoretical values. We observe that the weights converge to their true values.

(b) Fourier weights error rates compared to theoretical convergence rate. The convergence is met after $t = 5$.

Figure 13: Error rates and convergence plots for the weighted sum of Hermite polynomials interaction potential, coupled with a quadratic Ornstein-Uhlenbeck confining potential. The major observation is the scale of the weights (mainly the fourth weight, in red). They are indeed far from the expected theoretical values but as time grows, we still observe a convergence up to an error of 10^{-6}, which is satisfactory. The explanation for this very volatile behaviour is the high order of polynomials used in the computation of $f(x, \theta)$ in (3.21). For both figures, each line represents the average of the weight estimates over the number of trajectories. The theoretical weights are $\{1, 1/2, 1/3, 1/4\}$. The error computation is done using the MSE formula.

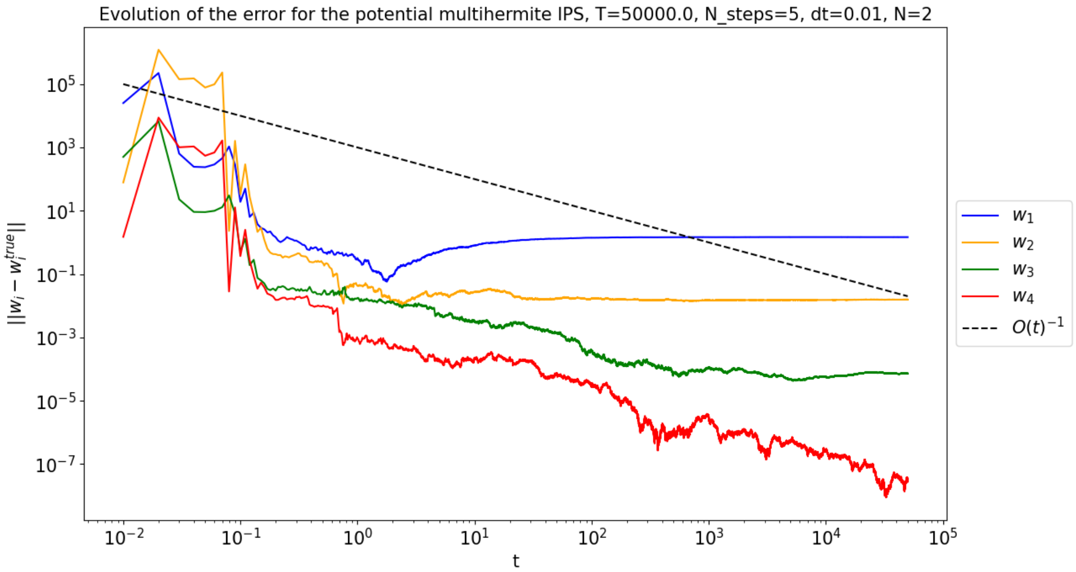


Figure 14: Error rates for the weighted sum of Hermite polynomials interaction potential for the MLE, coupled with a quadratic Ornstein-Uhlenbeck confining potential. The error does not follow the rate of $\mathcal{O}(t)^{-1}$, except for w_4. Each line represents the average of the weight estimates over the number of trajectories. The theoretical weights are $\{1, 1/2, 1/3, 1/4\}$. The error computation is done using the MSE formula.

two weights, while the last two weights are well enough approximated. The SGDCT performs significantly better, even with the small instabilities observed for the first few iterations.

6.2 SGDCT for the Curie-Weiss kernel

We will focus on applying the methodology described in Section 4.2.2 on the following problem from [37]. The studied interaction potential is known as the Curie-Weiss quadratic interaction kernel [13]

$$W(x, \kappa) = \frac{\kappa}{2} x^2.$$

If we add a quadratic Ornstein-Uhlenbeck confining potential, then this yields the following mean-reverting system of SDEs in $\mathbb{R}$

$$dX_t^{(i)} = -X_t^{(i)} dt - \kappa \left(X_t^{(i)} - \overline{X}_t \right) dt + dB_t^{(i)}, \quad i = 1, ..., N, \tag{6.2}$$

where $\overline{(\cdot)}$ represents the empirical mean operator. From this, we can observe that $W \in L^2(\mathbb{R}, e^{-x^2} dx)$ (as shown in Section 3.6 for any polynomial) which allows us to use the truncated Hermite expansion approximation

$$W(x) \approx K + \sum_{j=1}^{J} w_j H_j(x).$$

As W is a quadratic polynomial, we easily recover its three coefficients for the exact Hermite series expansion, and we will check whether SGDCT finds the same weights d_1, d_2. The theoretical coefficients are

$$\begin{aligned}
d_0 &= \frac{1}{\sqrt{\pi}} \int_{\mathbb{R}} \frac{1}{2}\kappa x^2 H_0(x) e^{-x^2} dx = \frac{\kappa}{2\sqrt{\pi}} \Gamma\left(\frac{3}{2}\right) = \frac{\kappa}{4} \\
d_1 &= \frac{1}{2\sqrt{\pi}} \int_{\mathbb{R}} \frac{1}{2}\kappa x^2 H_1(x) e^{-x^2} dx = 0 \\
d_2 &= \frac{1}{8\sqrt{\pi}} \int_{\mathbb{R}} \frac{1}{2}\kappa x^2 H_2(x) e^{-x^2} dx = \frac{\kappa}{4\sqrt{\pi}} \Gamma\left(\frac{5}{2}\right) - \frac{\kappa}{8\sqrt{\pi}} \Gamma\left(\frac{3}{2}\right) = \frac{3\kappa}{16} - \frac{\kappa}{16} = \frac{\kappa}{8}.
\end{aligned}$$

We set $\kappa = 0.5$ for the numerical experiments and recall $N = 2$. We also set $J = 2$ to recover d_1 and d_2. Figure 15 shows that the method manages to recover the two weights d_1, d_2. It is however more volatile than previously studied kernel inferences, e.g., on the circle.

6.2.1 Reaching the mean field limit

We now enter the mean field regime for the Curie-Weiss kernel. We consider $N = 200$ interacting particles, with the same initial condition $X_0^{(i)} = 0$ for $i = 1, ..., N$, to ensure symmetry. If we refer back to the complexity analysis of the SGDCT in Section 4.2.3, we observe a quadratic relation to the number of particles N. Hence, as N grows, we might face some computational difficulties.

To tackle this computation issue, one can benefit from the propagation of chaos from the mean field regime. We refer to [44] for further developments about this symmetry property. To infer the kernel with SGDCT, we practically need at least two particles to infer the kernel. Indeed, this is necessary in our setting to approximate the interacting component in $f(X_t, \theta)$ and $\nabla_\theta f(X_t, \theta)$. Here, we study the system defined with (6.2), but only feed the SGDCT the first two particles. The results are promising and shown in Figure 16. We observe that we recover the two theoretical values with only 1% of the original system data, which shows the method's practicality. Interestingly, increasing the number of observed particles does not drastically improve the inference behaviour, but inherently increases its computation time.

6.2.2 Comparison between SGDCT and MLE

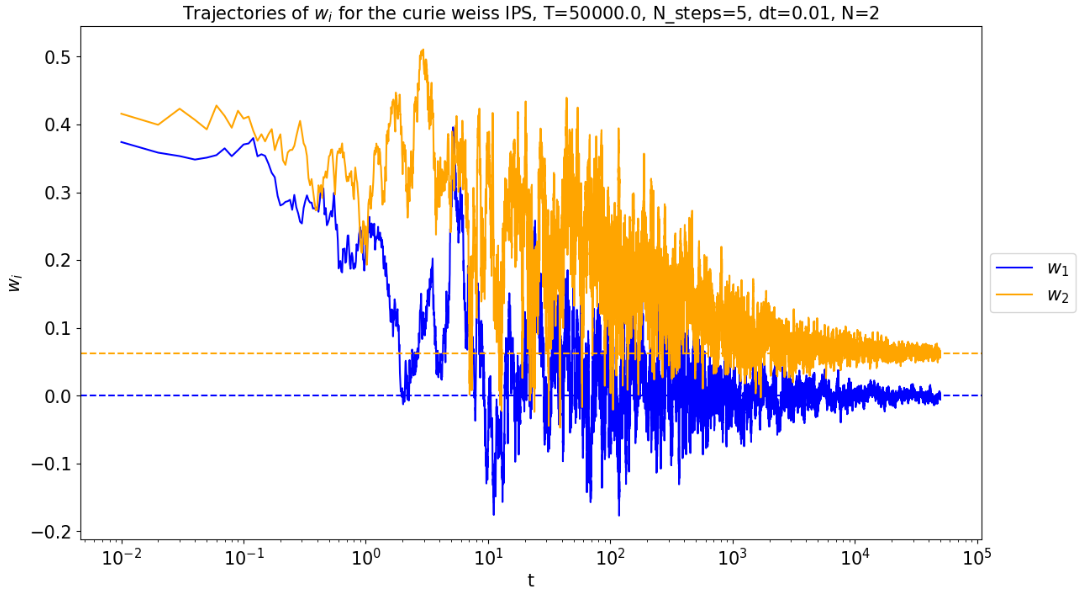
As for the previous kernel, we compare here the SGDCT and the MLE methods. The obtained convergence follows closely the rate $\mathcal{O}(t)^{-1}$ and reaches similar levels compared to the SGDCT. The computation time is roughly equivalent for both methods.

6.2.3 Adding measurement noise to observations

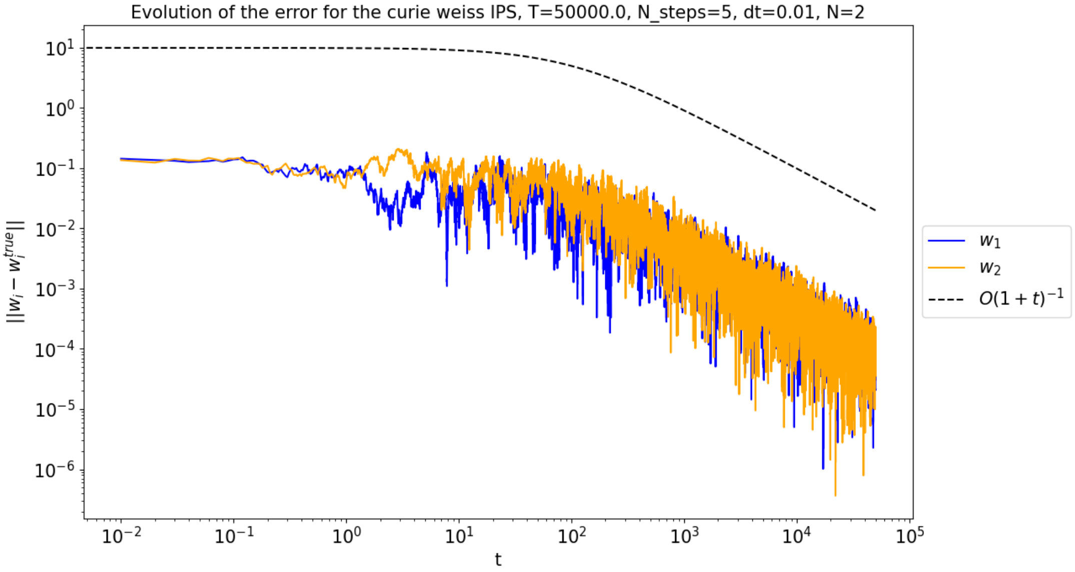
Here, we study the SGDCT behaviour when we feed noisy data. We define the new data as

$$Y_t^{(i)} = X_t^{(i)} + \tilde{\sigma}\xi_t^{(i)},$$

where $\xi_t^{(i)} \sim \mathcal{N}(0, 1)$. We will study the error between the obtained weight estimates and their theoretical values as $\tilde{\sigma}$ varies. For this, we use the Curie-Weiss model defined in (6.2), with $N = 2$, $T = 5 * 10^4$, $\Delta t = 10^{-2}$ and $n_{\text{steps}} = 1$. Figure 18 shows that the two errors vary differently with respect to the measurement noise $\tilde{\sigma}$. Indeed, the first weight inference is more stable compared to the second one, as its error remains flat as $\tilde{\sigma}$ increases compared to the other one that explodes. This can be due to w_1 being linked to a constant function H_0, while w_2 is linked to the linear Hermite polynomial H_1.

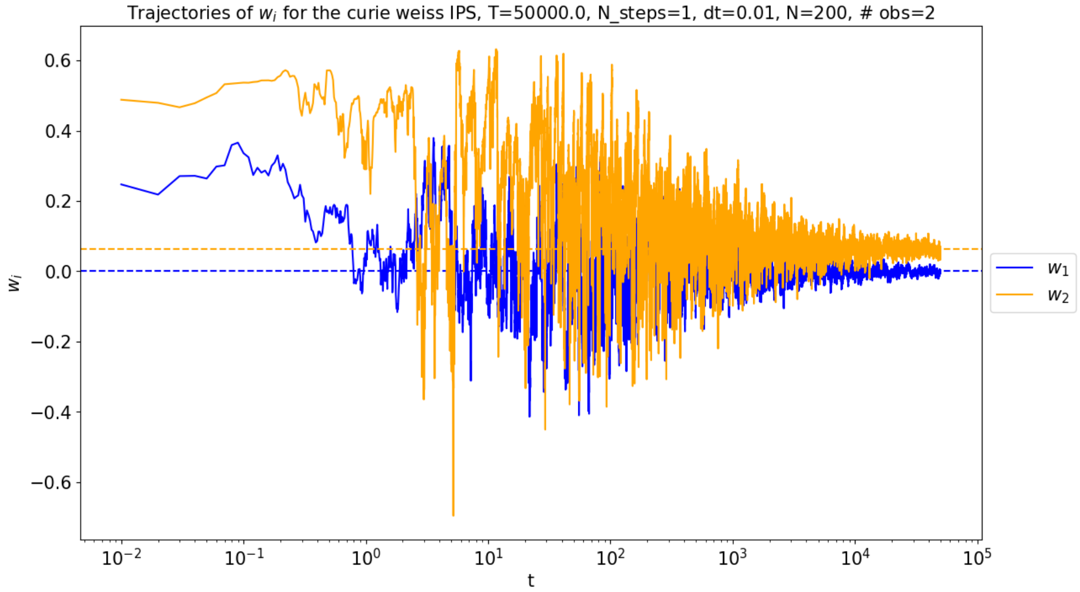


(a) Fourier weights convergence to theoretical values. We observe that the weights converge to their true values.

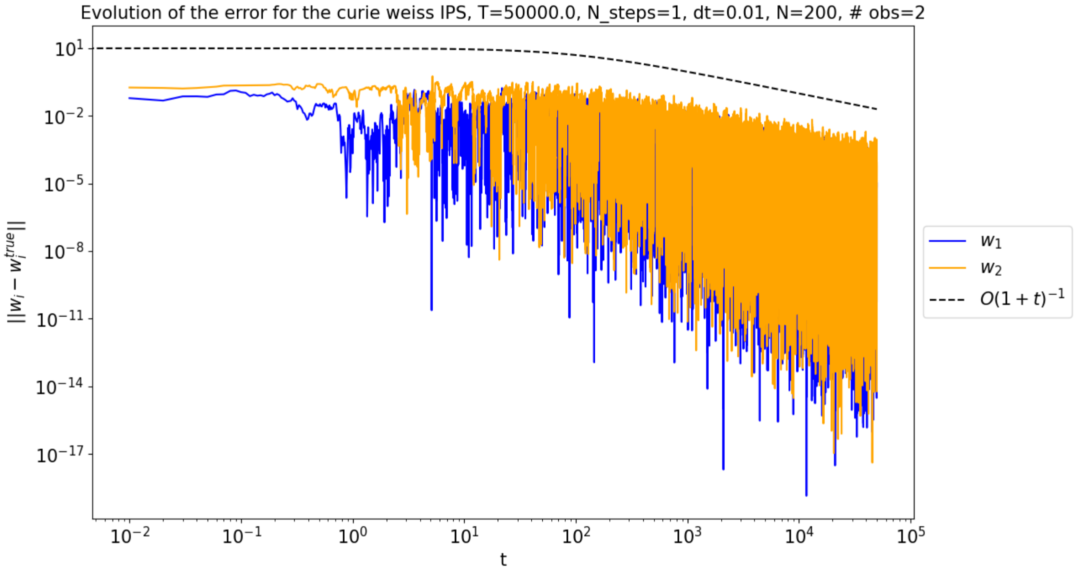


(b) Fourier weights error rates compared to theoretical convergence rate. Convergence is observed and follows the theoretical rate of $\mathcal{O}(1+t)^{-1}$.

Figure 15: Error rates and convergence plots for the Curie-Weiss interaction potential and Ornstein-Uhlenbeck confining potential. Figure 15a shows that the estimated weights converge to their true value as time increases. Figure 15b that the error follows the theoretical rate of $\mathcal{O}(1+t)^{-1}$. For both figures, each line represents the average of the weight estimates over the number of trajectories. The theoretical weights are $\{0, \kappa/8\}$. The error computation is done using the MSE formula.



(a) Fourier weights convergence to theoretical values. We observe that the weights converge to their true values.



(b) Fourier weights error rates compared to theoretical convergence rate. Convergence is observed and follows the theoretical rate of $\mathcal{O}(1+t)^{-1}$.

Figure 16: Error rates and convergence plots for the Curie-Weiss interaction potential and Ornstein-Uhlenbeck confining potential in the mean-field limit. Figure 16a shows that the estimated weights converge to their true value as time increases. Figure 16b that the error follows the theoretical rate of $\mathcal{O}(1+t)^{-1}$. For both figures, only one trajectory is observed and we assume that only two particles are in the system, instead of 200. The theoretical weights are $\{0, \kappa/8\}$. The error computation is done using the MSE formula.

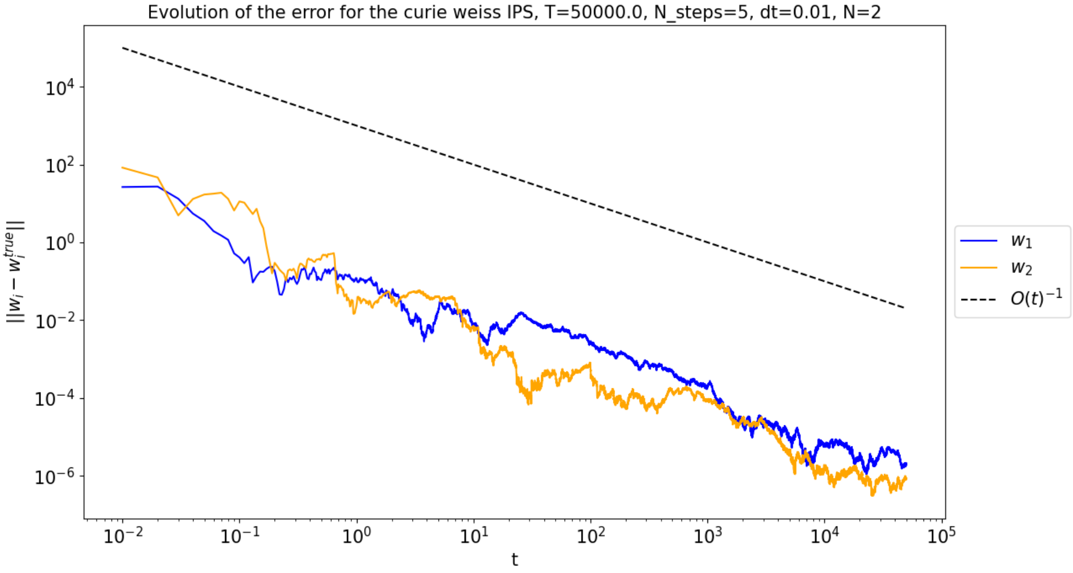


Figure 17: Error rates for the Curie-Weiss interaction potential for the MLE, coupled with a quadratic Ornstein-Uhlenbeck confining potential. The error follows the rate of $\mathcal{O}(t)^{-1}$. Each line represents the average of the weight estimates over the number of trajectories. The theoretical weights are $\{0, \kappa/8\}$. The error computation is done using the MSE formula.

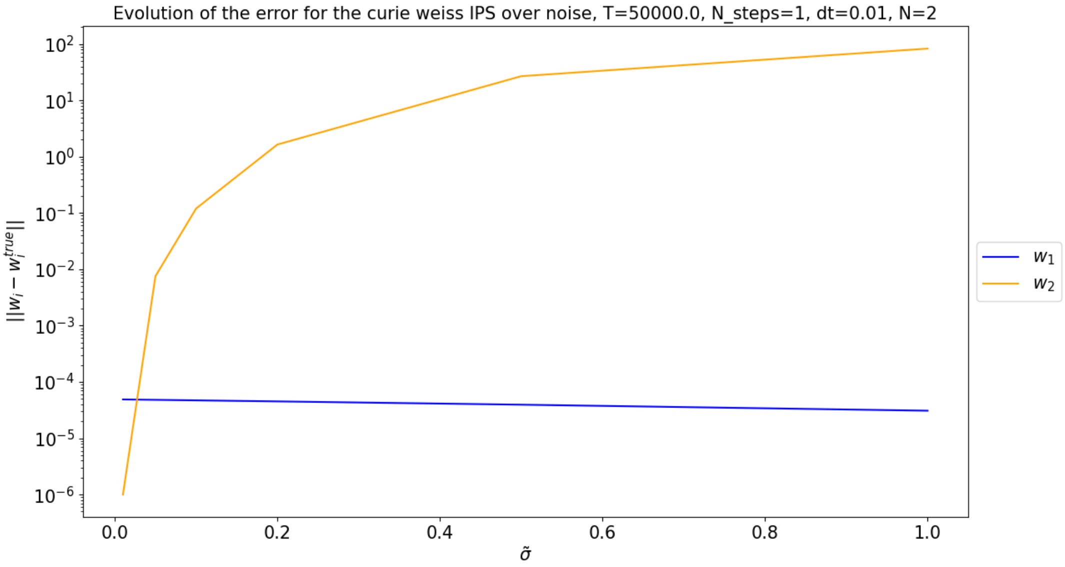


Figure 18: Evolution of the error between the weight estimates and their theoretical values as the measurement grows. The blue error line shows no direct link to $\bar{\sigma}$, while the orange line grows rapidly as $\bar{\sigma}$ increases. This can be due to w_1 being linked to a constant function H_0, while w_2 is linked to the linear Hermite polynomial H_1.

7 SGDCT adaptiveness for the truncation number

We discuss the adaptive algorithm to find the correct number of truncated modes required for a satisfactory kernel approximation based on a-posteriori errors. This is valid for the Hermite basis on the real line and the Fourier series basis on the circle.

As a reminder, the main goal is to infer an interaction kernel estimate based on several trajectories of observed data from the studied interacting particle system. To solve this problem, we try to find $\theta \in \mathbb{R}^J$ such that

$$W(x) \approx K + \sum_{j=1}^{J} w_j \cos(jx) \text{ or } W(x) \approx K + \sum_{j=1}^{J} w_j H_j(x) \tag{7.1}$$

for the circle and real line respectively. However, one might not know how many modes in the approximation are required to find a sufficiently close approximation of the true interacting potential. Therefore, we describe a straightforward algorithm that tries to find the smallest number of modes required to approximate the interaction potential function. The general idea is to start with a given truncation number J and increase it until the relative L^2 error between the two approximations is small enough, which would imply that we reached some plateau of optimality. Several extensions such as needing multiple low relative errors to avoid local minima can also be considered. Indeed, if we refer back to the study of the number of modes in Section 5.1.1, we see that after reaching $J = 4$, the relative error would be close to zero, and hence the algorithm would return $J^* = 4$.

Algorithm 2: Adaptive SGDCT algorithm

```
Input: $((X_k)_{k=0}^{T/\Delta t})_{n=1}^{n_{\text{steps}}}$, tol, $J$
Result: returns the estimates $\theta_t^{J^*}$ for $t \in \{0, \Delta t, ..., T/\Delta t\}$
Compute $W^J$ and related $\theta_t^J$ with SGDCT
$err = \infty$
while $err \geq$ tol do
    $J = J + 1$
    Compute $W^J$ and related $\theta_t^J$ with SGDCT
    $err = \|W^J - W^{J-1}\|_2$
end
$J^* \leftarrow J - 1$
returns $\theta^{J^*}$
```

8 Python library

This section covers the Python code that has been created for this project. A major focus of this thesis was to write the codebase as general and efficient as possible. This makes it possible to infer any smooth interacting kernel satisfying convergence conditions using the SGDCT and the MLE method, by only providing data for particle trajectories, and the function used as confinement potential if any. The codebase is available upon request at eliott.vandieren@epfl.ch.

8.1 Defining a problem setting

Several quantities need to be defined for a given problem. Firstly, we need to know how many particles are in the system (N). One can also reference the number of weights they wish to use in the function representation (J). The learning rate needs to be specified, by giving the quantities C_0 and C. For the inference method, we need information about the initial parameters. Hence, the initial distribution of θ_0, or θ_0 as an array must be given. Lastly, one needs to refer whether the particles will evolve on the circle or real line, as this will dictate which basis is used for the kernel inference.

8.2 Kernel inference

For the kernel inference, the code can either use the MLE described in Section 3.3.2 or the SGDCT described in Section 3.4. For the inference methods to run, we need to feed them particle trajectories X_t, the initial values θ_0, the particle increments if on the circle[2] and the number of trajectories of X_t that will be used (n_{steps}). For both methods, the output will be an array of dimension $n_{\text{steps}} \times K \times J$, with $K = T/dt$ and where the mean can be taken over the first dimension to obtain θ_t as defined in Algorithm 1.

8.3 Error measurements and plotting

There is a single function that plots the weight estimates over time on one hand and on the other the error compared to theoretical weights. Examples of outputs can be seen for example in Figure 2. It takes the output of the inference method, the true weights, whether the method was MLE or SGDCT, and some filename and directory to save the figures.

9 Final remarks

This master's thesis has demonstrated the effective application of the diffusion process drift estimation technique SGDCT using nonparametric function representations to model interacting kernels in IPSs, relying solely on particle position data. Specifically, coupling the SGDCT with Fourier series on the circle and Hermite expansions on the real line yielded excellent results for inferring multiple interacting kernels, both with and without a confinement potential. Various interaction kernels have been studied and some limitations have been shown such as some stability issues for Hermite expansions. The SGDCT method has been compared to the MLE for every kernel. The convergence and error rates matched the theoretical expectations for smooth kernels, and some insights regarding the adaptiveness of the method with an unknown number of basis elements have been presented. In the mean field regime, only 1% of the original data was required due to the propagation of chaos. Lastly, we have shown how the SGDCT reacts to measurement noise, which has unfortunately not been satisfactory.

Future research directions include filtering of noisy data, extending the inference framework to higher dimension particles ($d \gg 1$), and parallelising the method to infer both the interacting kernel and the confinement potential, which was assumed to be known in this work.

References

[1] Javier Baladron, Diego Fasoli, Olivier Faugeras, and Jonathan Touboul. Mean-field description and propagation of chaos in networks of hodgkin-huxley and fitzhugh-nagumo neurons. *The Journal of Mathematical Neuroscience*, 2:1–50, 2012.

[2] This avoids stability issues when the data is periodic in $[0, 2\pi]$. Indeed, suppose we only keep values between $[0, 2\pi]$, and we wrap around from 2π to 0. In that case, it is impossible to differentiate whether the particle made a significant negative jump or a smaller positive jump. Hence, an additional array of position increments is needed.

[2] Kaveh Bashiri. On the long-time behaviour of mckean-vlasov paths. 2020.

[3] Martin Bauer, Thilo Meyer-Brandis, and Frank Proske. Strong solutions of mean-field stochastic differential equations with irregular drift. 2018.

[4] Eli Ben-Naim, Paul L Krapivsky, and Sidney Redner. Bifurcations and patterns in compromise processes. *Physica D: nonlinear phenomena*, 183(3-4):190–204, 2003.

[5] Saıd Benachour, Bernard Roynette, Denis Talay, and Pierre Vallois. Nonlinear self-stabilizing processes–i existence, invariant probability, propagation of chaos. *Stochastic processes and their applications*, 75(2):173–201, 1998.

[6] Jaya Prakash Narayan Bishwal et al. Estimation in interacting diffusions: Continuous and discrete sampling. *Applied Mathematics*, 2(9):1154–1158, 2011.

[7] Mattia Bongini, Massimo Fornasier, Markus Hansen, and Mauro Maggioni. Inferring interaction rules from observations of evolutive systems i: The variational approach. *Mathematical Models and Methods in Applied Sciences*, 27(05):909–951, 2017.

[8] Oleg A Butkovsky. On ergodic properties of nonlinear markov chains and stochastic mckean–vlasov equations. *Theory of Probability & Its Applications*, 58(4):661–674, 2014.

[9] José A Carrillo, RS Gvalani, GA Pavliotis, and A Schlichting. Long-time behaviour and phase transitions for the mckean–vlasov equation on the torus. *Archive for Rational Mechanics and Analysis*, 235(1):635–690, 2020.

[10] Xiaohui Chen. Maximum likelihood estimation of potential energy in interacting particle systems from single-trajectory data. *Electronic Communications in Probability*, 26:1–13, 2021.

[11] Fabienne Comte, Valentine Genon-Catalot, and Catherine Larédo. Nonparametric moment method for mckean-vlasov stochastic differential equations. 2024.

[12] Felipe Cucker and Steve Smale. On the mathematics of emergence. *Japanese Journal of Mathematics*, 2:197–227, 2007.

[13] Donald A Dawson. Critical dynamics and fluctuations for a mean-field model of cooperative behavior. *Journal of Statistical Physics*, 31(1):29–85, 1983.

[14] PE Chaudru de Raynal. Strong well posedness of mckean–vlasov stochastic differential equations with hölder drift. *Stochastic Processes and their Applications*, 130(1):79–107, 2020.

[15] Laetitia Della Maestra and Marc Hoffmann. Nonparametric estimation for interacting particle systems: Mckean–vlasov models. *Probability Theory and Related Fields*, pages 1–63, 2022.

[16] Laetitia Della Maestra and Marc Hoffmann. The lan property for mckean–vlasov models in a mean-field regime. *Stochastic Processes and their Applications*, 155:109–146, 2023.

[17] Alain Durmus, Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. An elementary approach to uniform in time propagation of chaos. *Proceedings of the American Mathematical Society*, 148(12):5387–5398, 2020.

[18] Tadahisa Funaki. A certain class of diffusion processes associated with nonlinear parabolic equations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 67(3):331–348, 1984.

[19] Valentine Genon-Catalot and Catherine Larédo. Parametric inference for small variance and long time horizon mckean-vlasov diffusion models. *Electronic Journal of Statistics*, 15 (2):5811–5854, 2021.

[20] Kay Giesecke, Gustavo Schwenkler, and Justin A Sirignano. Inference for large financial systems. *Mathematical Finance*, 30(1):3–46, 2020.

[21] Raphael A Kasonga. Maximum likelihood theory for large interacting systems. *SIAM Journal on Applied Mathematics*, 50(3):865–875, 1990.

[22] Daniel Lacker. Mean field games and interacting particle systems. *preprint*, 2018.

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

[24] Quanjun Lang and Fei Lu. Identifiability of interaction kernels in mean-field equations of interacting particles. *arXiv preprint arXiv:2106.05565*, 2021.

[25] Juan Li and Hui Min. Weak solutions of mean-field stochastic differential equations and application to zero-sum stochastic differential games. *SIAM Journal on Control and Optimization*, 54(3):1826–1858, 2016.

[26] Zhongyang Li, Fei Lu, Mauro Maggioni, Sui Tang, and Cheng Zhang. On the identifiability of interaction functions in systems of interacting particles. *Stochastic Processes and their Applications*, 132:135–163, 2021.

[27] Meiqi Liu and Huijie Qiao. Parameter estimation of path-dependent mckean-vlasov stochastic differential equations. *Acta Mathematica Scientia*, 42(3):876–886, 2022.

[28] Eva Löcherbach. Lan and lamn for systems of interacting diffusions with branching and immigration. In *Annales de l'IHP Probabilités et statistiques*, volume 38, pages 59–90, 2002.

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

[30] Fei Lu, Mauro Maggioni, and Sui Tang. Learning interaction kernels in heterogeneous systems of agents from multiple trajectories. *The Journal of Machine Learning Research*, 22(1):1518–1584, 2021.

[31] Marcello Lucia and Jesenko Vukadinovic. Exact multiplicity of nematic states for an onsager model. *Nonlinearity*, 23(12):3157, 2010.

[32] Henry P McKean Jr. A class of markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences*, 56(6):1907–1911, 1966.

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

[34] Mohammad Ali Niksirat and Xinwei Yu. On stationary solutions of the 2d doi–onsager model. *Journal of Mathematical Analysis and Applications*, 430(1):152–165, 2015.

[35] Lars Onsager. The effects of shape on the interaction of colloidal particles. *Annals of the New York Academy of Sciences*, 51(4):627–659, 1949.

[36] Grigorios A Pavliotis. Stochastic processes and applications.

[37] Grigorios A Pavliotis and Andrea Zanoni. Eigenfunction martingale estimators for interacting particle systems and their mean field limit. *SIAM Journal on Applied Dynamical Systems*, 21(4):2338–2370, 2022.

[38] Yvo Pokern, Omiros Papaspiliopoulos, Gareth O Roberts, and AM Stuart. Non parametric bayesian drift estimation for one-dimensional diffusion processes. 2009.

[39] Yvo Pokern, Andrew M Stuart, and Eric Vanden-Eijnden. Remarks on drift estimation for diffusion processes. *Multiscale modeling & simulation*, 8(1):69–95, 2009.

[40] Louis Sharrock, Nikolas Kantas, Panos Parpas, and Grigorios A Pavliotis. Online parameter estimation for the mckean–vlasov stochastic differential equation. *Stochastic Processes and their Applications*, 162:481–546, 2023.

[41] Justin Sirignano and Konstantinos Spiliopoulos. Stochastic gradient descent in continuous time. *SIAM Journal on Financial Mathematics*, 8(1):933–961, 2017.

[42] Justin Sirignano and Konstantinos Spiliopoulos. Stochastic gradient descent in continuous time: A central limit theorem. *Stochastic Systems*, 10(2):124–151, 2020.

[43] Simone Carlo Surace and Jean-Pascal Pfister. Online maximum-likelihood estimation of the parameters of partially observed diffusion processes. *IEEE transactions on automatic control*, 64(7):2814–2829, 2018.

[44] Alain-Sol Sznitman. Topics in propagation of chaos. *Ecole d'été de probabilités de Saint-Flour XIX—1989*, 1464:165–251, 1991.

[45] Julian Tugaut. Convergence to the equilibria for self-stabilizing processes in double-well landscape. 2013.

[46] Jianghui Wen, Xiangjun Wang, Shuhua Mao, and Xinping Xiao. Maximum likelihood estimation of mckean–vlasov stochastic differential equation and its application. *Applied Mathematics and Computation*, 274:237–246, 2016.