European Journal of Applied Mathematics (2024), 1–26
doi:10.1017/S0956792524000706

CAMBRIDGE UNIVERSITY PRESS

PAPER

Recent kernel methods for interacting particle systems: first numerical results

Christian Fiedler[1], Michael Herty[2], Chiara Segala[2] and Sebastian Trimpe[1]

[1]Institute for Data Science in Mechanical Engineering, RWTH Aachen University, Aachen, Germany
[2]Institut für Geometrie und Praktische Mathematik, RWTH Aachen University, Aachen, Germany
Corresponding author: Chiara Segala; Email: segala@igpm.rwth-aachen.de

Received: 21 November 2023; **Revised:** 18 July 2024; **Accepted:** 07 August 2024

Keywords: kernel methods; interacting particle systems; mean-field limit; machine learning

2020 Mathematics Subject Classification: 46E22, 93A16 (Primary); 82B40, 82C40 (Secondary)

Abstract

Interacting particle systems (IPSs) are a very important class of dynamical systems, arising in different domains like biology, physics, sociology and engineering. In many applications, these systems can be very large, making their simulation and control, as well as related numerical tasks, very challenging. Kernel methods, a powerful tool in machine learning, offer promising approaches for analyzing and managing IPS. This paper provides a comprehensive study of applying kernel methods to IPS, including the development of numerical schemes and the exploration of mean-field limits. We present novel applications and numerical experiments demonstrating the effectiveness of kernel methods for surrogate modelling and state-dependent feature learning in IPS. Our findings highlight the potential of these methods for advancing the study and control of large-scale IPS.

1. Introduction

Interacting particle systems (IPSs), or synonymously multiagent systems (MASs), are dynamical systems that model the interaction of a group of homogenous particles or agents. These system classes have attracted an enormous amount of attention in the mathematical community, primarily because they exhibit emergent phenomena like flocking, swarming or consensus. For an overview and introduction to the vast literature on this subject, we refer to refs. [1, 7–9, 37, 38]. IPSs arise in a wide range of domains, from physical processes like gas dynamics [20], to biology, sociology and very recently even data science [21, 22, 26, 39, 43, 57, 59, 69]. In many applications of IPS, one has to work with a large number of particles, for example, in gas dynamics, the modelling of large animal populations, human crowds or large-scale traffic models. Analytical investigations (especially of emergent phenomena) and numerical methods (like simulation, optimisation and control) become very challenging or even impossible for such large-scale IPS.

Recently, a fruitful exchange between the fields of IPS and machine learning has gained momentum. On the one hand, theory and methods from IPS – in particular, large-scale IPS and mean-field limits – have been used to analyse and design methods in machine learning, for example, in the context of clustering problems [43], deep neural networks [44, 56] or ensemble-based optimisation methods [10, 19, 35, 45, 46, 61]. On the other hand, machine learning methods are increasingly used to investigate IPS, and enhance or even replace modelling with learning-based methods. In particular, while physical processes like gas dynamics have been very successfully treated using first-principles modelling [20], complex phenomena like animal motion, crowd dynamics or traffic flow are much more challenging to

model. For example, due to the increased availability of large data sets and computational resources, it is natural to try to learn the interaction rules of IPS from data. This question has received considerable attention lately, both from a theoretical as well as practical perspective, cf. [12, 53–55]. A particularly important class of machine learning approaches consists of kernel methods [65], which encompass, for example, Gaussian processes (GPs) [71] and support vector machines [68]. Kernel methods allow the systematic modelling of domain knowledge [66], are supported by a well-developed theory [48, 68] and lead to efficient and reliable numerical algorithms [65], capable of scaling up to very large data sets [51, 52]. All of this makes kernel methods natural candidates for machine learning on IPS. In this work, we consider two novel developments in this context.

First, surrogate models are a common approach to tackle large-scale and expensive modelling, simulation and optimisation tasks in scientific computing, statistics and machine learning [40, 63]. The basic idea is to approximate a computationally expensive function by a surrogate model that is cheap to evaluate. Kernel-based methods are also here a standard tool, in particular, GPs. We provide initial numerical investigations on this approach in the context of IPS by considering two prototypical tasks, for which we use a kernel method for the approximation.

Second, in kinetic theory, transitioning from microscopic to mesoscopic levels (considering distributions of particles rather than individual particles) is a common approach to managing the complexity of IPS. The mean-field limit, which involves the number of particles approaching infinity, is a well-established method for this transition. Extensive studies on mean-field limits have provided rigorous analytical frameworks for understanding the emergent behaviour in IPS [13, 14, 16, 18, 23, 24, 34, 41].

Building on these advances, recent theoretical work has explored the application of kernel methods in the mean-field limit [31, 33]. These methods enable learning state-dependent features of IPS from data, facilitating the analysis of large-scale systems. The primary motivation is to infer maps that measure or estimate specific aspects of the system's state, such as reaction to stimuli in swarming or susceptibility in opinion dynamics. In this work, we provide the first numerical experiments evaluating this approach.

We now provide an outline of the remaining paper. In Section 2, we introduce the necessary notation and provide background on reproducing kernel Hilbert spaces (RKHSs) and their application to interpolation and approximation problems. Section 3 delves into IPS models and their numerical treatments, highlighting well-known examples and the corresponding mean-field limits. In Section 4, we investigate the use of kernel methods in the context of IPS, starting with surrogate models in Section 4.1. In Section 4.2, we provide a self-contained exposition of mean-field limit of kernels and kernel methods, tailored to our needs. These developments are illustrated with a concrete class of kernels having a mean-field limit, which then forms the foundation of the numerical investigations. The paper concludes in Section 5, offering also an outlook on possible future applications.

2. Background on kernel methods

We now present necessary background material. First, in Section 2.1 we concisely give the main definitions and results on kernels and reproducing KHSs, since these function spaces will be used as candidate spaces for the interpolation and approximation problems considered later on. Next, in Section 2.2, we outline the standard approach to function interpolation in RKHSs. Finally, in Section 2.3, we recall some concepts and results related to learning and approximation with optimisation problems in RKHSs.

2.1. Kernels and RKHSs

In the following, we provide a concise overview of the necessary background on kernels and their reproducing RKHSs, following [68, Chapter 4]. We will work primarily with Hilbert spaces of functions of the following type.

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

Definition 2.1. *Let $\mathcal{X} \neq \emptyset$ be a non-empty set, $H \subseteq \mathbb{R}^{\mathcal{X}}$ a Hilbert space of functions, and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ some bivariate map. k is called a* reproducing kernel *for a Hilbert space $H \subseteq \mathbb{R}^{\mathcal{X}}$ of real-valued functions if*

1. $k(\cdot, x) \in H \quad \forall x \in \mathcal{X}$
2. $f(x) = \langle f, k(\cdot, x)\rangle_H \quad \forall f \in H, x \in \mathcal{X}.$

To effectively work with these function spaces, we need also the following concept.

Definition 2.2. *Let $\mathcal{X} \neq \emptyset$ be a non-empty set. A map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive semidefinite if for all $x, x' \in \mathcal{X}$ we have $k(x, x') = k(x', x)$, and for all $N \in \mathbb{N}_+$, $x_1, \ldots, x_N \in \mathcal{X}$, $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$ we have $\sum_{i,j=1}^{N} \alpha_i \alpha_j k(x_j, x_i) \geq 0$. If in the case of pairwise distinct inputs $x_1, \ldots, x_N$, equality in this latter condition holds only if $\alpha_1 = \ldots = \alpha_N = 0$, we call k positive definite. In the following, we call a positive semidefinite (or positive definite) map k a kernel (on $\mathcal{X}$).*

Remark 2.3.

1. *A map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is symmetric and positive (semi)definite if and only if all matrices $(k(x_j, x_i))_{i,j=1,\ldots,N}$ are symmetric and positive (semi)definite, for all $N \in \mathbb{N}_+$ and $x_1, \ldots, x_N \in \mathcal{X}$. This motivates the terminology of symmetric and positive (semi)definite.*
2. *Unfortunately, the terminology regarding kernels is highly non-uniform in the literature. What we call symmetric and positive semidefinite is often called of positive type or positive definite, and what we call symmetric and positive definite is often called strictly positive definite. Furthermore, the terminology kernel is often used for Mercer kernels, which are symmetric and positive semidefinite continuous bivariate functions, often on a compact metric space.*

A Hilbert space of functions has at most one reproducing kernel, and such a reproducing kernel is a kernel (i.e., symmetric and positive semidefinite), cf. [68, Lemma 4.19, Theorem 4.20]. Furthermore, a map k is a reproducing kernel for some Hilbert space of functions if and only if k is a kernel, and in this case this Hilbert space of functions is unique [68, Theorem 4.21]. We call this Hilbert space the reproducing kernel Hilbert space (RKHS) corresponding or associated to k, and denote it by $(H_k, \langle \cdot, \cdot \rangle_k)$ with correspondent induced norm $\| \cdot \|_k$. Finally, if k is a kernel, the linear space

$$H_k^{\text{pre}} = \left\{ \sum_{n=1}^{N} \alpha_{nk}(\cdot, x_n) \mid x_1, \ldots, x_N \in \mathcal{X},\ \alpha_1, \ldots, \alpha_N \in \mathbb{R}, N \in \mathbb{N} \right\}$$

is dense in H_k and is called the pre-RKHS associated with k.

Remark 2.4. *We tailored the exposition of kernels and RKHSs to our needs. In the machine learning literature, slightly different, but equivalent definitions are used.*

1. *A map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a* kernel *if there exists a Hilbert space $\mathcal{H}$ (called* feature space*) and a map $\Phi : \mathcal{X} \to \mathcal{H}$ (called* feature map*) such that*

$$k(x, x') = \left\langle \Phi(x'), \Phi(x) \right\rangle_{\mathcal{H}} \quad \forall x, x' \in \mathcal{X}.$$

   The motivation for this definition comes from the kernel trick*, which allows to use linear algorithms on inputs lifted (usually by a nonlinear map) to a high-dimensional (even infinite-dimensional) new space, as long as inner products of the transformed inputs can be efficiently computed. This is exactly the situation described in the preceding definition. It turns out that this definition is equivalent to our definition of a kernel, cf. [69, Theorem 4.16] for more details.*
2. *A Hilbert space of functions $H \subseteq \mathbb{R}^{\mathcal{X}}$ is called a* reproducing kernel Hilbert space *(RKHS) if $\delta_x : H \to \mathbb{R}$, $\delta_x(f) = f(x)$ is continuous for all $x \in \mathcal{X}$. This property holds if and only if H has a reproducing kernel [69, Lemma 4.19, Theorem 4.20], so this is again equivalent to our definition.*

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

Two features make kernels and their RKHSs particularly interesting in the context of function interpolation and approximation, as well as learning. On the one hand, they allow efficient algorithmic solutions of such problems, often amounting to solving a linear equation system or solving a finite-dimensional convex optimisation problem, cf. the next two sections. On the other hand, since a RKHS is generated from its associated reproducing kernel, the latter determines the properties of the functions from the RKHS. In particular, by designing appropriate kernels, one can systematically construct function spaces containing only functions with desirable properties. For example, if $\mathcal{X}$ is a topological space, then a bounded and continuous kernel k enforces that H_k contains only bounded and continuous functions [68, Lemma 4.28]. Similarly, if $\mathcal{X}$ is a metric space, then a Lipschitz- or more generally Hölder-continuous kernel enforces corresponding continuity properties for its RKHS functions, cf. [30] for an in-depth discussion. Another relevant property is invariance w.r.t. specified transformations of the input, as described in the following result. Please refer to Appendix A for the proof of Lemma 2.5.

Lemma 2.5. *Let $\mathcal{X} \neq \emptyset$ be a set, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a kernel on $\mathcal{X}$, and $T : \mathcal{X} \to \mathcal{X}$ some map. The following two statements are equivalent.*

1. *$k(T(x), x') = k(x, x')$ for all $x, x' \in \mathcal{X}$.*
2. *$f(T(x)) = f(x)$ for all $x \in \mathcal{X}$ and $f \in H_k$.*

Finally, a wide variety of kernels as well as construction techniques for kernels are available, see [68, Chapter 4], [66] and [62, Chapter 4]. For simplicity, we will use in the following one of the most popular choices, the Gaussian or Squared-Exponential (SE) kernel on $\mathbb{R}^d$, defined by

$$k(x, x') = k_\gamma(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\gamma^2}\right), \tag{2.1}$$

where $\gamma \in \mathbb{R}_{>0}$ is called the length scale of the kernel. The corresponding RKHS H_k contains very smooth functions, in particular, $H_k \subseteq C^\infty(\mathbb{R}^d, \mathbb{R})$, cf. [68, Section 4.4] for more details.

2.2. Kernel interpolation

Let us recall the basic setting of function interpolation with finite data. Consider two sets $\mathcal{X}, \mathcal{Y} \neq \emptyset$ (the input and output space), a collection $\mathcal{F} \subseteq \mathcal{Y}^\mathcal{X}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$ (the space of candidate functions), and $(x_1, y_1), \ldots, (x_N, y_N) \in \mathcal{X} \times \mathcal{Y}$ (the data). We say that a function $f \in \mathcal{F}$ interpolates the data if $f(x_n) = y_n$ for all $n = 1, \ldots, N$, and the interpolation problem consists in finding (if it exists) such a function f, potentially with additional properties. Consider now the case $\mathcal{Y} = \mathbb{R}$ and $\mathcal{F} = H_k$, where k is a kernel on $\mathcal{X}$. The resulting problem is usually called kernel interpolation, which is well-understood, cf. [60, Chapter 3]. The following result summarises the elements of the corresponding theory, which will be relevant for us.

Proposition 2.6. *Let $\mathcal{X} \neq \emptyset$ be some set, k a kernel on $\mathcal{X}$, and $(x_1, y_1), \ldots, (x_N, y_N) \in \mathcal{X} \times \mathbb{R}$. The kernel interpolation problem of finding $f \in H_k$ with $f(x_n) = y_n$ is solvable if and only if $\vec{y} \in \mathrm{im}(K)$, where we defined*

$$\vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad K = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{pmatrix}.$$

Furthermore, if the interpolation problem is solvable, then $f = \sum_{n=1}^N \alpha_n k(\cdot, x_n)$ is the unique solution to the optimization problem

$$\min_{\substack{f \in H_k \\ f(x_n) = y_n \ \forall n}} \|f\|_k$$

where $\vec{\alpha} \in \mathbb{R}^N$ is such that $\vec{y} = K\vec{\alpha}$.

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

In particular, if k is positive definite, then any interpolation problem with pairwise distinct $x_1, \dots, x_N$ is solvable, and the coefficients $\vec{\alpha}$ of the minimum norm interpolating function are given by $\vec{\alpha} = K^{-1}\vec{y}$.

Note that the matrix K defined above is usually called kernel matrix or Gram matrix. For the proof of Proposition 2.6, see [61, Theorem 3.4, Corollary 3.5]. This result states that the search for an interpolating function in the (in general infinite-dimensional) space H_k boils down to solving a finite-dimensional linear equation system, where the problem matrix is even positive semidefinite. Furthermore, we even get the minimum norm interpolating functions, i.e., the solution in closed form of a (in general infinite-dimensional) optimization problem over the function space H_k. Finally, if a kernel is positive definite, then *any* interpolation problem (with pairwise distinct input points) can be solved in the corresponding RKHS. This explains also the importance of positive definite kernels.

2.3. Approximation with kernels: kernel machines and kernel ridge regression

Next, we turn to function approximation with kernels, as motivated by supervised machine learning problems. Consider some unknown function $f_* : \mathcal{X} \to \mathbb{R}$, which is only accessible through noisy samples $y_n = f_*(x_n) + \eta_n, n = 1, \dots, N$, where the model additive noises $\eta_1, \dots, \eta_N$ are, for example, independent centred random variables with finite variances. The goal is to find a good approximation $\vec{f}$ of f_* from the data $(x_1, y_1), \dots, (x_N, y_N)$. To do so, we first have to fix a space of candidate functions, for which we choose here an RKHS, so the goal is to search for a good approximation $\vec{f} \in H_k$. One standard approach from machine learning is regularised empirical risk minimisation (on RKHSs), which amounts to the optimisation problem

$$\min_{f \in H_k} \frac{1}{N} \sum_{n=1}^{N} \ell\left(x_n, y_n, f(x_n)\right) + \lambda \|f\|_k^2, \tag{2.2}$$

where $\ell : \mathcal{X} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ is called loss function, and $\lambda \in \mathbb{R}_{>0}$ regularisation parameter. Intuitively, if at input $x \in \mathcal{X}$ the "true" output is $y \in \mathbb{R}$ and our approximation predicts output $t \in \mathbb{R}$, then we incur loss $\ell(x, y, t)$. The optimisation problem (2.2) contains two terms: The first term $\frac{1}{N} \sum_{n=1}^{N} \ell(x_n, y_n, f(x_n))$ is a data-fit term, measuring how good a given candidate function $f \in H_k$ performs on the data set, as measured by ℓ. However, since we only have noisy outputs (the y_n are not the real outputs $f_*(x_n)$), interpolating the data $(x_1, y_1), \dots, (x_N, y_N)$ is not meaningful in general. In fact, forcing $\vec{f}$ to exactly match the data might result in a bad prediction of f_*, a phenomenon called overfitting in machine learning. To avoid this, the second term $\lambda \|f\|_k^2$ acts as a regularisation. The RKHS norm $\|\cdot\|_k$ is used as a complexity measure, and the regularisation parameter $\lambda \in \mathbb{R}_{>0}$ determines the strength of the regularisation. An optimisation problem over an RKHS like (2.2) is often referred to as a kernel machine.

Finally, the loss function ℓ, the kernel k (inducing the RKHS H_k) and the regularisation parameter $\lambda \in \mathbb{R}_{>0}$ need to be chosen. The loss function is often determined by the problem setting, or chosen for theoretical and computational convenience, see [68, Chapters 2, 3] for many examples of loss functions and theoretical considerations regarding their choice in concrete learning problems. The kernel k is usually chosen based on properties of the associated RKHS H_k that are deemed appropriate for the problem at hand. For example, if it is known or suspected that the underlying function f_* is very smooth, then a kernel-inducing smooth RKHS function is chosen, like the SE kernel (2.1). In practice, one usually fixes a class of kernels up to some parameters, in this context called hyperparameters. For example, one might decide to use the SE kernel, and then the length scale is a hyperparameter that remains to be set. The regularisation parameter λ is also called a hyperparameter, and is usually chosen according to the noise level (the larger the noise magnitude, the larger the regularisation parameter). This intuitive notion will be made more precise towards the end of this section, when we discuss kernel ridge regression. In practice, the choice of hyperparameters is very important in machine learning problems, and different strategies like dataset-splitting, cross validation, or structural risk minimisation can be employed, cf. [58, Chapter 4].

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

In general, the candidate space H_k will be infinite-dimensional, so the question of existence and uniqueness of a solution of (2.2), and computational tractability of this optimisation problem, becomes very important. As another advantage of kernel methods and RKHSs, these questions actually do not pose a problem, as assured by the following result.

Proposition 2.7. *Let $\mathcal{X} \neq \emptyset$ be some set, k a kernel on $\mathcal{X}$, $L : \mathbb{R}^N \to \mathbb{R}_{\geq 0}$ a lower semicontinuous and strictly convex function, and $\lambda \in \mathbb{R}_{>0}$. For all $x_1, \dots, x_N$, the optimization problem*

$$\min_{f \in H_k} L\left(f(x_1), \dots, f(x_N)\right) + \lambda \|f\|_k^2$$

has a unique solution $\vec{f}$, which is of the form

$$\vec{f}(x) = \sum_{n=1}^{N} \alpha_n k(x, x_n),$$

where $\alpha_1, \dots, \alpha_N \in \mathbb{R}$.

This result is known as the Representer Theorem, with many variants and generalisations available, for example, [64] or [68, Section 5.1, 5.2] for more details and proofs. If ℓ is a benign loss function (so that $(t_1, \dots, t_N) \mapsto \frac{1}{N} \sum_{n=1}^{N} \ell(x_n, y_n, t_n)$ is lower semicontinuous and strictly continuous), then Proposition 2.7 assures that a unique solution of the (in general infinite-dimensional) optimisation problem (2.2) exists, and that it can be computed using a finite-dimensional optimisation problem.

Kernel ridge regression. In the following, we will focus on the ℓ_2- or least-squares loss function $\ell(x, y, t) = (y - t)^2$, so that (2.2) fulfils the conditions of Proposition 2.7. The resulting optimisation problem is

$$\min_{f \in H_k} \frac{1}{N} \sum_{n=1}^{N} \left(y_n - f(x_n)\right)^2 + \lambda \|f\|_k^2, \tag{2.3}$$

and finding a prediction $\vec{f}$ by solving this problem is called kernel ridge regression (KRR). It turns out that (2.3) has a closed-form solution, given as

$$\vec{f}(x) = \vec{k}(x)^\top (K + N\lambda I_N)^{-1} \vec{y},$$

where we defined

$$K = \left(k(x_j, x_i)\right)_{i,j=1,\dots,N}$$
$$\vec{k}(x)^\top = \begin{pmatrix} k(x, x_1) & \cdots & k(x, x_N) \end{pmatrix}$$
$$\vec{y} = \begin{pmatrix} y_1 & \cdots & y_N \end{pmatrix}.$$

Furthermore, in contrast to a more general kernel machine like (2.2), KRR can be given a concrete probabilistic interpretation. If we assume that the noise variables $\eta_1, \dots, \eta_N$ are independent and identically distributed centred Gaussian random variables, then the KRR solution is exactly the resulting posterior mean function occurring in GP regression, when using a zero prior mean function, and the kernel as the covariance function, cf. [48]. Similarly, the KRR solution can also be linked to the maximum a posteriori solution of Bayesian linear regression when using a Gaussian prior on the weights, cf. [58, Section 11.3]. In both cases, the regularisation parameter is linked to the noise level.

3. IPS models and their numerical treatment

In this section, we recall some well-known examples of interacting particle systems describing the agent dynamics on the microscopic level, as well as their corresponding mean-field limits. Furthermore, we describe established numerical methods used to simulate these systems, both on the microscopic and

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

mesoscopic levels. These example systems and the numerical methods will form the foundation for the numerical experiments in the next section.

3.1. Example systems and their mean-field limits

Various authors have introduced and studied multiagent models on the microscopic level that model social and political phenomena, aiming at understanding collective behaviours and self-organisation within society [4, 28, 42, 70]. Our first example, a well-known first-order model on the microscopic level, stems from this literature and has been used for example in opinion dynamics, cf. [69].

Let $x_i := x_i(t) \in \mathbb{R}^d$ be the state of agent $i = 1, \dots, M$ at time $t \geq 0$. In the case of opinion dynamics, x_i represents the opinion(s) of individual i. The interaction of the agents is modelled by a function $P : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, often called interaction function, and the dynamics are described by the first-order ordinary differential equation system

$$\dot{x}_i = \frac{1}{M} \sum_{j=1}^{M} P(x_i, x_j)\,(x_j - x_i)\,, \quad i = 1, \dots, M \tag{3.1}$$

starting from an initial condition $x_i(t_0) = x_i^0,\ i = 1, \dots, M$.

For our second example, we focus on second-order systems aiming at describing swarming or flocking behaviour. The latter refers to the aggregation behaviour of a group of similar entities, for example, animals of the same species. For concreteness, we focus on Cucker-Smale systems, which consider only the alignment behaviour of a group of agents [6, 22]. In the most common formulation of this model, the state of agent i is now given by its position $x_i \in \mathbb{R}^d$ and velocity $v_i \in \mathbb{R}^d$, evolving under the dynamics described by

$$\begin{aligned} \dot{x}_i &= v_i \\ \dot{v}_i &= \frac{1}{M} \sum_{j=1}^{M} H_\beta(x_i, x_j)\,(v_j - v_i)\,, \quad i = 1, \dots, M \end{aligned} \tag{3.2}$$

and initial conditions $x_i(t_0) = x_i^0,\ v_i(t_0) = v_i^0,\ i = 1, \dots, M$. The function H_β quantifies the intensity of interaction between individuals i and j, usually varying based on their mutual distance, with the underlying assumption that closer individuals have a greater influence compared to those farther apart. A common choice for the function H_β is

$$H_\beta(x_i, x_j) = \frac{1}{(\,1 + \|x_i - x_j\|^2\,)^\beta}\,, \tag{3.3}$$

where $\|\cdot\|$ is the usual Euclidean norm on $\mathbb{R}^d$. Under the assumption $\beta \geq 0$, it can be shown that the system is forward-complete and that mass and momentum are preserved.

Simulating dynamics of this nature for large systems of individuals requires significant computational resources, and for many interesting applications, such microscopic models can indeed involve very large populations of interacting individuals, ranging from several hundred thousand to millions. From a mathematical modelling perspective, these challenges have been addressed within the mean field research community, where deriving mean field equations serve as an initial step toward mitigating computational complexity, transitioning from a microscopic description, centred on phase-space particles, to a mesoscopic level, where the focus shifts to particle distributions. Let us briefly recall the formalisation of this.

Consider a continuous-time multiagent system with M agents, and suppose that the state space of an individual agent is Z. For example, for the first-order model (3.1), we have $Z = \mathbb{R}^d$, and for the second-order model (3.2), we have $Z = \mathbb{R}^d \times \mathbb{R}^d$. The state of the whole system at time $t \geq 0$ is then $(z_1(t), \dots, z_M(t)) \in Z^M$, and assuming indistinguishability of the agents, this corresponds to a time-varying empirical measure

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

$$\mu_t^M = \frac{1}{M}\sum_{i=1}^{M} \delta_{z_i(t)},$$

where δ_z is the Dirac distribution centred at atom $z \in Z$. The idea is now, using a weak formulation, to derive an evolution equation for the empirical measures, and go to the limit $M \to \infty$, which then leads to an evolution equation for probability measures over the state-space Z. Making this mean-field limit precise is a classic subject, and well-posedness results are available for a broad range of models, for example [14, 17].

Under regularity assumptions on the interaction function P and H_β, respectively, one can compute the mean-field limit of the microscopic models (3.1) and (3.2) introduced above. For the first-order model (3.1), one obtains the following strong form of the evolution equation for the agent distribution

$$\begin{aligned} &\partial_t \mu(t,x) + \nabla_x \cdot \left(\mu(t,x) \int P(x,y)(y-x) d\mu(t,y) \right) = 0\,, \\ &\mu(0,x) = \mu^0(x)\,. \end{aligned} \tag{3.4}$$

In the case of the Cucker-Smale model (3.2), the evolution equation for the agent distribution in strong form is

$$\begin{aligned} &\partial_t \mu(t,x,v) + v\nabla_x \cdot (\mu(t,x,v)) + \nabla_v \cdot \left(\mu(t,x,v) \int\int H_\beta(x,y)(w-v) d\mu(t,y,w) \right) = 0\,, \\ &\mu(0,x,v) = \mu^0(x,v)\,. \end{aligned} \tag{3.5}$$

Apart from allowing numerical tractability, the mean field Equations (3.4)–(3.5) can simplify the analysis of interacting particle systems, and allow to gain insights into the macroscopic properties of the model, such as its overall density, velocity, and direction, see, for example, [1, 9, 18]. This can be useful for studying the emergence of global behaviour and patterns, understanding phase transitions and analysing the stability of collective behaviours [2, 16, 29].

3.2. Numerics for the IPS models

We now turn to numerical approaches to approximate the first-order opinion dynamics model (3.1) and the second-order alignment model (3.2), as well as their mean field counterparts (3.4) and (3.5).

In the numerical experiments of Section 4, the dynamics on the microscopic level are discretized by a forward Euler scheme with time step Δt over the time horizon $[t_0,\ T]$, so for (3.1), we obtain the following discretization

$$x_i^{n+1} = x_i^n + \Delta t \left(\frac{1}{M} \sum_{j=1}^{M} P(x_i^n, x_j^n)(x_j^n - x_i^n) \right)\,,$$

while for (3.2) we have

$$\begin{aligned} x_i^{n+1} &= x_i^n + \Delta t\ v_i^n \\ v_i^{n+1} &= v_i^n + \Delta t \left(\frac{1}{M} \sum_{j=1}^{M} H_\beta(x_i^n, x_j^n)(v_j^n - v_i^n) \right)\,, \end{aligned}$$

where $x_i^n \approx x_i(t_n),\ v_i^n \approx v_i(t_n)$ with $t_n = n\ \Delta t \in [t_0,\ T]$.

We consider now the corresponding mean field counterparts, starting with the first-order model (3.1) and the associated mean field Equation (3.4). In order to approximate the latter, we use mean field Monte-Carlo (MFMC) methods as developed in ref. [3]. These methods fall in the class of fast algorithms for interacting particle systems such as direct simulation Monte-Carlo methods (DSMCs) [5, 11, 25], or most recently Random Batch Methods [47]. For the MFMC method, we consider $\hat{M}$ particles $x^0 \equiv \left\{x_i^0\right\}_i$ sampled from the initial distribution $\mu^0(x)$, and we introduce the following approximation for the mean

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

field dynamics (3.4)

$$x_i^{n+1} = (1 - \Delta t \vec{P}_i^n)\, x_i^n + \Delta t\, \vec{P}_i^n\, \vec{X}_i^n\,,$$

where the quantities $\vec{P}_i^n$ and $\vec{X}_i^n$ are computed from a sub-sample of $\hat{M}_s$ particles randomly selected from the whole ensemble of $\hat{M}$ sampled particles,

$$\vec{P}_i^n = \frac{1}{\hat{M}_s}\sum_{j=1}^{\hat{M}_s} P(x_i^n, x_{ij}^n)\,, \qquad \vec{X}_i^n = \frac{1}{\hat{M}_s}\sum_{j=1}^{\hat{M}_s} \frac{P(x_i^n, x_{ij}^n)}{\vec{P}_i^n} x_{ij}^n\,, \qquad i = 1, \ldots, \hat{M}\,.$$

Using this type of MC algorithm, we can reduce the cost due to the computation of the interaction term from $\mathcal{O}(\hat{M}^2)$ to $\mathcal{O}(\hat{M}_s\,\hat{M})$. Observe that for $\hat{M}_s = \hat{M}$, we obtain the explicit Euler scheme for the original particle system (3.1) with $\hat{M}$ particles.

Finally, for the implementation of the second-order mean field model (3.5), we use a two-dimensional Lax–Friedrichs scheme [49, 50], known to be a very stable scheme with much diffusion. It is a numerical method based on finite differences, forward in time and centred in space. For simplicity, we focus on the case $d = 1$, and consider a compact space-time domain $[t_0,\ T] \times [a_x,\ b_x] \times [a_v,\ b_v]$. Rewriting (3.5) in the compact form

$$\mu_t + (g(\mu))_x + (h(\mu))_v = 0\,,$$

where

$$\begin{aligned} \mu &:= \mu(t, x, v) \\ g(\mu) &:= v\,\mu(t, x, v) \\ h(\mu) &:= \mu(t, x, v) \int\int H_\beta(x, y)\,(w - v)\, d\mu(t, y, w)\,, \end{aligned}$$

the numerical scheme is given

$$\mu_{i,j}^{n+1} = \frac{1}{4}\left(\mu_{i+1,j}^n + \mu_{i-1,j}^n + \mu_{i,j+1}^n + \mu_{i,j-1}^n\right) - \frac{\Delta t}{2\Delta x}\left(g_{i+1,j}^n - g_{i-1,j}^n\right) - \frac{\Delta t}{2\Delta v}\left(h_{i,j+1}^n - h_{i,j-1}^n\right)\,,$$

where now $\mu_{i,j}^n \approx \mu(t_n, x_i, v_j)$, and the domain $[t_0,\ T] \times [a_x,\ b_x] \times [a_v,\ b_v]$ is discretized using equally spaced points with a spacing of Δt, Δx, Δv in the t, x, v direction, respectively, and $g_{i,j}^n$, $h_{i,j}^n$ are the numerical fluxes.

4. Kernel methods for IPS: Applications and numerical tests

We now present two novel applications of kernel methods to IPS, which we illustrate using numerical experiments based on the example models and associated numerical methods outlined in the preceding section. First, in Section 4.1, we describe the use of kernel methods for surrogate modelling in the context of IPS, which to the best of our knowledge is a novel use case. Section 4.2 is concerned with kernel-based learning of state-dependent features of IPS in the mean field setting, which naturally leads to mean-field limits of kernels. This scenario and the associated theory have been introduced in ref. [31, 33], and we provide a concise recap of the setting, the basic concepts and results from these references. We then consider a specific class of kernels, for which we can provide more concrete results than the general theory in the latter two references. These developments then form the basis for numerical experiments.

All experiments have been implemented in MATLAB®. For convenience, the experimental parameters are summarised in Table 1.

4.1. Surrogate modelling of IPS related properties

Consider an IPS with $M \in \mathbb{N}_+$ agents or particles and state-space X^M, where X is the state-space of an individual particle. Frequently one is not directly interested in a trajectory $\vec{x}_M : [0, T] \to X^M$, but

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

Table 1. *Simulation parameters for each test case.*

	t_0	T	Δt	M	s	$\hat{M}$	$\hat{M}_s$	Δx	Δv
Test A1	0	10	0.01	30	4	–	–	–	–
Test A2	0	10	0.01	30	5	–	–	–	–
Test B1	0	10	0.01	30	{2, 4, 8}	–	–	–	–
Test B2	0	10	0.01	{10, 100, 1000, ∞}	8	10,000	100	–	–
Test B3	0	10	0.01	30	4	–	–	–	–
Test B4	0	1	0.001	∞	4	–	–	0.05	0.05

rather in a functional of this trajectory. Such a functional is usually provided as a closed-form expression, or described implicitly by a numerical algorithm. In practice, one first computes the trajectory $\vec{x}_M$, using methods like the ones described in Section 3.2, and then applies the functional of interest to this trajectory. However, if the system is very large, so $M >> 1$, then the computation of $\vec{x}_M$ becomes very expensive. Similarly, for a very complex functional, even the second step can be very expensive. In this section, we propose to use kernel-based surrogate models to reduce the computational effort needed.

For illustrative purposes, we focus on simple but prototypical scenarios, which allow us to effectively evaluate the kernel-based techniques.

Test A1. Surrogate variance for the Cucker and Smale model. First, we consider the case of a pointwise-in-time functional of the state, i.e., we have a map $F_M : X^M \to \mathbb{R}$ that is applied on a state $\vec{x} \in X^M$. Given the trajectory $\vec{x}_M$ as above, this induces a corresponding trajectory of the functional, $t \mapsto F_M(\vec{x}_M(t))$. If the latter needs to be evaluated on a fine grid on $[0, T]$, for example, for visualisation purposes, this can become very expensive, in particular, if F requires complex computations. We therefore approximate $t \mapsto F_M(\vec{x}_M(t))$ by a kernel method as

$$\hat{F}_M(t) = \sum_{i=1}^{N} \alpha_i k(t, t_i) \tag{4.1}$$

from samples $(t_1, F_M(\vec{x}_M(t_1))), \ldots, (t_N, F_M(\vec{x}_M(t_M)))$, where $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$ are the coefficients determined by the chosen kernel method. Since the samples are the result of a computation and not a measurement, we do not incur measurement errors, and hence we have an interpolation problem, for which we use kernel interpolation, cf. Section 2.2.

Remark 4.1. A related problem is to learn the evolution of a functional from few measurements. In this case, the data will be noisy and kernel interpolation is inappropriate, and one could use KRR, for example. However, investigating this scenario in detail is beyond the present article.

As a concrete example, we use the microscopic Cucker-Smale model (3.2), so $X = (\mathbb{R}^d)^2$, for fixed initial conditions, and for ease of visualisation, we work with $d = 1$. As a functional of the state, we consider the (pointwise-in-time) variance of the velocities, which we denoted by $\mathcal{V}_M$. The goal is therefore to approximate $\mathcal{V}_M$ by

$$\hat{\mathcal{V}}_M(t) = \sum_{i=1}^{N} \alpha_i \, k_\gamma(t, t_i), \tag{4.2}$$

from data $(t_1, \mathcal{V}_M(t_1)), \ldots, (t_N, \mathcal{V}_M(t_N))$, where we chose for concreteness the SE kernel (2.1) with $\gamma = \frac{1}{\sqrt{2}}$. The underlying dynamics adhere to the second-order microscopic model (3.2) and it is discretized as explained in Section 3.2. It involves a swarm of $M = 30$ agents with $N = 4$ measurements in time of the true variance over a total number of 1000 time-steps. The initial input data are uniformly distributed, namely $x_i^0, v_i^0 \sim \mathcal{U}([1, 2])$, for every $i = 1, \ldots, M$. Figure 1 depicts the evolution of positions and velocities over time and shows a comparison between the exact function $\mathcal{V}_M$ and the approximated one $\hat{\mathcal{V}}_M$.

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

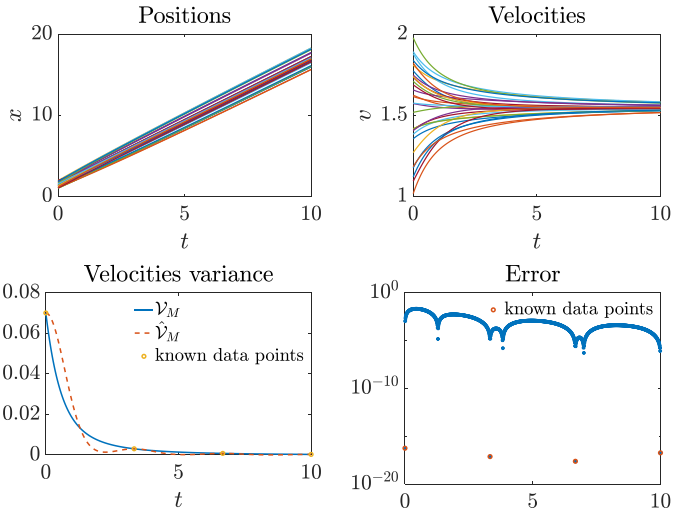

Figure 1. ***Test A1.*** *Top-left: position of particles evolving over time with dynamics (3.2). Top-right: evolution of velocities. Bottom-left: comparison true (blue) and approximated (red) variance from (4.2). Bottom-right: Error $t \to |\mathcal{V}_M - \hat{\mathcal{V}}_M|$. Dots underline the time-step where the true variance is accessible.*

We observe that there is a good agreement, in fact, the error $t \mapsto |\mathcal{V}_M - \hat{\mathcal{V}}_M|$ is in the interval $[10^{-18},\ 10^{-2}]$. Please refer to Table 1 for all the simulation parameters.

Test A2. Surrogate cost in a minimisation problem. We now consider the case of a functional, which depends on the whole trajectory, and not only pointwise-in-time. As an example, we consider an optimisation problem whose objective function is a functional of a trajectory of a (parameterized) IPS. For concreteness, consider the microscopic Cucker-Smale system (3.3) with $M \in \mathbb{N}_+$ agents, and we want to optimise the integrated (over time) variance of the velocities (w.r.t. a fixed initial condition) over the interaction parameter $\beta \in \mathbb{R}$ (appearing in the interaction function H_β). Furthermore, we consider the constraint $\beta \in K$, where $K \subseteq \mathbb{R}$ is compact. The corresponding minimisation problem can be formalised as

$$\min_{\beta \in K} \mathcal{J}(\beta) = \int_0^T \mathcal{V}_M(t)\, dt\,,$$
$$s.t. \quad \begin{cases} \dot{x}_i = v_i\,, \quad x_i(t_0) = x_i^0\,, \\ \dot{v}_i = \frac{1}{M} \sum_{j=1}^M H_\beta(x_i, x_j)\,(v_j - v_i)\,, \quad v_i(t_0) = v_i^0\,, \quad i = 1, \dots, M\,. \end{cases}$$

The previous problem is a one-dimensional minimisation problem and it could be solved without difficulty if the functional $\mathcal{J}$ is easily evaluated. However, numerical optimisation methods usually evaluate the underlying objective functions many times. Observe that in the present situation, the objective function involves the simulation of an IPS, and then the integral over (a functional of) the whole trajectory. For large T and M, this can become very expensive. Hence, it is reasonable to use a surrogate model for the objective function $\mathcal{J}$, which can be cheaply evaluated. We will use again kernel interpolation for this task, applying it to the data $(\beta_1, \mathcal{J}(\beta_1)), \dots, (\beta_N, \mathcal{J}(\beta_N))$. The corresponding surrogate function is hence given by

$$\hat{\mathcal{J}}(\beta) = \sum_{i=1}^N \alpha_i\, k_\gamma(\beta, \beta_i)\,,$$

where for concreteness, we chose again the SE kernel with $\gamma = \frac{1}{\sqrt{2}}$. For our experiment, the time integral in the definition of $\mathcal{J}$ is computed using the rectangular rule, and the underlying microscopic dynamics

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

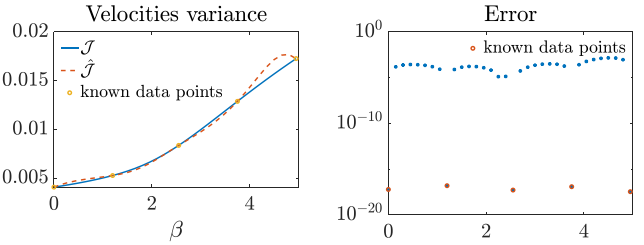

Figure 2. Test A2*. Left: comparison true (blue) and approximated (red) running cost. Right: Error* $\beta \to |\mathcal{J} - \hat{\mathcal{J}}|$*. Dots underline the points where the true variance is accessible, these correspond to the values of* $\beta \in \{0, 1.2, 2.55, 3.75, 4.95\}$.

are simulated as described in Section 3.2. The results for $N = 5$ are depicted in Figure 2 and we observe again a coincidence of the surrogate function $\hat{\mathcal{J}}$ and the true cost $\mathcal{J}$. This shows that kernel approximation could be used for an efficient minimisation using surrogate models. Additionally, one could easily seek the optimal value of β numerically by employing a gradient descent approach, since the gradient of $\hat{\mathcal{J}}$ can be easily computed. However, these considerations are beyond the scope of the present work.

4.2. Kernels in the mean-field limit

We now turn to kernels in the mean-field limit, which naturally arise when learning state-dependent functionals of large IPS. We first recall this latter learning problem as described in [33], and then describe some of the general theory of kernels in the mean-field limit, following [31, 33]. We then specialise the theory to our concrete setting, which forms the foundation for the numerical experiments in the remainder of this section.

Introduction. Consider a multiagent system consisting of $M \in \mathbb{N}_+$ agents or particles, and assume that the system state at each time instant $t \geq 0$ is completely described by all the individual agent states $x_i(t) \in \mathbb{R}^d$, $i = 1, \dots, M$, at time t, so the state of the whole system at time t is just $\vec{x}(t) = (x_i(t))_{i=1,\dots,M} \in (\mathbb{R}^d)^M$. We are interested in a certain state-dependent feature of the IPS. For example, in opinion dynamics, the individual state $x_i(t)$ corresponds to the opinion (in some potentially high-dimensional opinion space) of agent i, $i = 1, \dots, M$. A feature of interest could then be a measure of disagreement between the agents, or a measure of susceptibility to adversarial external influence. Since the IPS is completely described by its state $\vec{x}(t)$, it appears reasonable to model such a feature as a functional on the state space, i.e., if the system is in state $\vec{x}$, then the feature takes the value $f_M(\vec{x})$, where $f_M : (\mathbb{R}^d)^M \to \mathcal{Y}$, with $\mathcal{Y}$ some real vectorspace. Simple examples of such a feature are the mean and the variance of the agent states. These two cases are essentially trivial since the maps describing the features are given by analytical formulas. However, in modern applications of IPS like opinion dynamics, it can be very difficult to describe features using first-principles modelling. Instead, we can learn them from data. To simplify the following exposition, we consider only scalar-valued features (corresponding to $\mathcal{Y} = \mathbb{R}$ in the notation from above), since the vector-valued (or even matrix-valued) case can be covered by treating each component separately.

We assume that potentially noisy measurements of the feature of interest are available at certain time instances. More formally, let $0 \leq t_1 < \dots < t_N$, then we assume access to state snapshots $\vec{x}(t_1), \dots, \vec{x}(t_N)$ and noisy measurements of the feature at these times modelled as

$$y_n = f_M(\vec{x}(t_n)) + \eta_n, \quad n = 1, \dots, N,$$

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

where $\eta_1, \dots, \eta_N$ is additive noise. We can treat this as a standard supervised learning problem, with data set $(\vec{x}(t_1), y_1), \dots, (\vec{x}(t_N), y_N)$. If we want to use a kernel method like (2.2), we need a kernel of the form $k_M : (\mathbb{R}^d)^M \times (\mathbb{R}^d)^M \to \mathbb{R}$, and the resulting approximation of the map describing the feature then becomes

$$\vec{f}_M(\vec{x}) = \sum_{n=1}^{N} \alpha_n k_M(\vec{x}, \vec{x}(t_n)).$$

In particular, if we have at any time $t \geq 0$ a state snapshot $\vec{x}(t)$, then we can predict the feature at this time by $\vec{f}_M(\vec{x}(t))$. Moreover, if we have a model of the IPS dynamics, then we can predict the evolution of the feature by $t \mapsto \vec{f}_M(\vec{x}(t))$.

We consider now the case of a very large MAS, i.e., $M >> 1$. As described in Section 3 for concrete examples, the modelling, simulation and prediction of the dynamics in this setting can be made tractable by going to the kinetic level, for instance using the mean-field limit, corresponding to the limit $M \to \infty$. A key modelling assumption for this is the homogeneity of agents, i.e., the agents are indistinguishable. Under this condition, it appears reasonable to assume that also the feature of interest does not depend on the order of the agents, and standard results like [15, Lemma 1.2] suggest that also the maps f_M have a mean-field limit. But what happens to the learning problems in the limit $M \to \infty$? Recall that we need kernels of the form $k_M : (\mathbb{R}^d)^M \times (\mathbb{R}^d)^M \to \mathbb{R}$, so we have to consider the case of a sequence of kernels with inputs tuples of increasing length. It turns out that we can formulate an appropriate mean-field limit of kernels and their RKHSs, and that the resulting theory can be used for the learning problems.

Mean-field limit of functions and kernels. We first recall some preliminaries from measure theory. Let (X, d_X) be a compact metric space and denote by $\mathcal{P}(X)$ the set of Borel probability measures on X, which we endow with the topology of weak convergence. It is well-known that this topology can be metrized by the Kantorowich–Rubinstein metric

$$d_{\mathrm{KR}}(\mu_1, \mu_2) = \sup \left\{ \int_X \phi(x) \mathrm{d}(\mu_1 - \mu_2)(x) \mid \phi : X \to \mathbb{R} \text{ is 1-Lipschitz} \right\}. \tag{4.3}$$

Since X is separable as a compact metric space, this metric coincides with the 1-Wasserstein metric. Furthermore, since X is compact, also the metric space $(\mathcal{P}(X), d_{\mathrm{KR}})$ is compact. Given $\vec{x} \in X^M$, we denote the i-th component of $\vec{x}$ by x_i, and we define the *empirical measure* with atoms in $\vec{x}$ by $\hat{\mu}[\vec{x}] = \frac{1}{M} \sum_{i=1}^{M} \delta_{x_i}$, where δ_x denotes the Dirac measure centred at $x \in X$. It is well-known that the empirical measures are dense in $\mathcal{P}(X)$ w.r.t. the Kantorowich–Rubinstein metric. For more details and background, we refer to [28, Chapter 11].

The following definition makes precise the intuitive concept of a mean-field limit of a sequence of functions with an increasing limit of arguments.

Definition 4.2. *Consider functions $f_M : X^M \to \mathbb{R}$, $M \in \mathbb{N}_+$ and $f : \mathcal{P}(X) \to \mathbb{R}$. We say that f is the mean-field limit of $(f_M)_M$, or that $(f_M)_M$ converges in mean field to f, if*

$$\lim_{M \to \infty} \sup_{\vec{x} \in X^M} |f_M(\vec{x}) - f(\hat{\mu}[\vec{x}])| = 0.$$

This notion originated in the literature on mean field games [15], and is now common in the context of mean-field limits of IPS, cf. [36] and [32], for examples, in continuous and discrete-time, respectively.

The next well-known result, cf. [15, Lemma 1.2], ensures the existence of a (subsequential) mean-field limit of functions in the sense of Definition 4.2. For $M \in \mathbb{N}_+$, denote by $\mathcal{S}_M$ the set of permutations on $\{1, \dots, M\}$, and for a tuple $\vec{x} \in X^M$ and permutation $\sigma \in \mathcal{S}_M$, define $\sigma\vec{x} = (x_{\sigma(1)}, \dots, x_{\sigma(M)})$.

Proposition 4.3. *Let $f_M : X^M \to \mathbb{R}$, $M \in \mathbb{N}_+$, be a sequence of functions fulfilling the following conditions.*

1. *(Permutation-invariance) For all $M \in \mathbb{N}_+$, $\sigma \in \mathcal{S}_M$ and $\vec{x} \in X^M$, we have $f_M(\sigma\vec{x}) = f_M(\vec{x})$.*
2. *(Uniform boundedness) There exists $B_f \in \mathbb{R}_{\geq 0}$ such that for all $M \in \mathbb{N}_+$, $\vec{x} \in X^M$, we have $|f_M(\vec{x})| \leq B_f$.*

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

3. *(Uniform Lipschitz continuity) There exists* $L_f \in \mathbb{R}_{\geq 0}$ *such that for all* $M \in \mathbb{N}_+$, $\vec{x}, \vec{x}' \in X^M$ *we have*

$$|f_M(\vec{x}) - f_M(\vec{x}')| \leq L_f d_{KR}(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']).$$

Then there exists a Lipschitz continuous function $f : \mathcal{P}(X) \to \mathbb{R}$ *and a subsequence* $(f_{M_\ell})_\ell$ *such that*

$$\lim_{\ell \to \infty} \sup_{\vec{x} \in X^{M_\ell}} |f_{M_\ell}(\vec{x}) - f(\hat{\mu}[\vec{x}])| = 0.$$

This result can be used to justify the assumption that a mean-field limit map exists to model a state-dependent feature at the kinetic level. These developments can be extended to the case of kernels, as has been done first in [31].

Definition 4.4. *Consider bivariate functions* $\kappa_M : X^M \times X^M \to \mathbb{R}$, $M \in \mathbb{N}_+$ *and* $\kappa : X^M \times X^M \to \mathbb{R}$. *We say that* κ *is the mean-field limit of* $(\kappa_M)_M$, *or that* $(\kappa_M)_M$ *converges in mean field to* κ, *if*

$$\lim_{M \to \infty} \sup_{\vec{x}, \vec{x}' \in X^M} |\kappa_M(\vec{x}, \vec{x}') - \kappa(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}'])| = 0.$$

For the next result, define the metric

$$d^2_{\mathrm{KR}}((\mu, \nu), (\mu', \nu')) = d_{\mathrm{KR}}(\mu, \mu') + d_{\mathrm{KR}}(\nu, \nu')$$

on $\mathcal{P}(X) \times \mathcal{P}(X)$, which is one metric for the product topology, and hence $(\mathcal{P}(X) \times \mathcal{P}(X), d^2_{\mathrm{KR}})$ is again a compact metric space.

Proposition 4.5. *Let* $k_M : X^M \times X^M \to \mathbb{R}$, $M \in \mathbb{N}_+$, *be a sequence of kernels that fulfils the following conditions.*

1. *(Permutation-invariance) For all* $M \in \mathbb{N}_+$, $\sigma \in \mathcal{S}_M$ *and* $\vec{x}, \vec{x}' \in X^M$, *we have* $k_M(\sigma\vec{x}, \vec{x}') = k_M(\vec{x}, \vec{x}')$.
2. *(Uniform boundedness) There exists* $B_k \in \mathbb{R}_{\geq 0}$ *such that for all* $M \in \mathbb{N}_+$, $\vec{x}, \vec{x}' \in X^M$, *we have* $|k_M(\vec{x}, \vec{x}')| \leq B_f$.
3. *(Uniform Lipschitz continuity) There exists* $L_k \in \mathbb{R}_{\geq 0}$ *such that for all* $M \in \mathbb{N}_+$, $\vec{x}, \vec{x}', \vec{y}, \vec{y}' \in X^M$ *we have*

$$|k_M(\vec{x}, \vec{x}') - k_M(\vec{y}, \vec{y}')| \leq L_k d^2_{KR}\left((\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']), (\hat{\mu}[\vec{y}], \hat{\mu}[\vec{y}'])\right).$$

Then there exists a Lipschitz continuous kernel k *on* $\mathcal{P}(X)$ *and a subsequence* $(k_{M_\ell})_\ell$ *such that*

$$\lim_{\ell \to \infty} \sup_{\vec{x}, \vec{x}' \in X^{M_\ell}} |\kappa_{M_\ell}(\vec{x}, \vec{x}') - \kappa(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}'])| = 0.$$

In the preceding result, since k_{M_ℓ} and k are all kernels, they come with their unique RKHSs. Importantly, the corresponding RKHS functions inherit from their reproducing kernels properties that are relevant for the mean-field limit of functions, e.g., the permutation-invariance, cf. Lemma 2.5. Moreover, the mean-field limit of the kernels induces a certain limiting behaviour inside the RKHSs. Every function from the RKHS H_k arises as a mean-field limit of functions from the RKHSs corresponding to the kernels k_M, and conversely, every uniformly norm bounded sequence of functions from the RKHSs corresponding to the kernels k_M has a subsequence that converges in mean field to a function contained in H_k, cf. [33, Theorem 2.3] for details.

Mean-field limit of the kernel learning problem. All the preceding discussion suggests that we can use such kernels to learn feature functionals in the mean-field limit context. For this, we have to connect the learning problems for finite $M \in \mathbb{N}_+$ and for the mean-field limit. This can be done with the following mean field variant of the Representer Theorem, cf. [33, Theorem 3.3, Remark 3.4]. To simplify the exposition, assume from now on the situation of Proposition 4.5, relabel the convergent subsequence again by M, and define $H_M = H_{k_M}$ for all $M \in \mathbb{N}_+$.

Proposition 4.6. *Let* $N \in \mathbb{N}_+$, $\mu_1, \ldots, \mu_N \in \mathcal{P}(X)$, *and for* $n = 1, \ldots, N$ *consider* $\vec{x}_n^{[M]} \in X^M$, $M \in \mathbb{N}_+$, *such that* $\hat{\mu}[\vec{x}_n^{[M]}] \xrightarrow{d_{KR}} \mu_n$ *for* $M \to \infty$. *Let* $L : \mathbb{R}^N \to \mathbb{R}_{\geq 0}$ *be continuous and strictly convex and* $\lambda > 0$.

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

For each $M \in \mathbb{N}_+$, consider the problem

$$\min_{f \in H_M} L(f(\vec{x}_1^{[M]}), \ldots, f(\vec{x}_N^{[M]})) + \lambda \|f\|_M^2, \tag{4.4}$$

as well as the problem

$$\min_{f \in H_k} L\left(f(\mu_1), \ldots, f(\mu_N)\right) + \lambda \|f\|_k^2. \tag{4.5}$$

Then for each $M \in \mathbb{N}_+$, problem (4.4) has a unique solution $\vec{f}_M$, which is of the form

$$\vec{f}_M = \sum_{n=1}^{N} \alpha_n^{[M]} k_M(\cdot, \vec{x}_n^{[M]}) \in H_M$$

with $\alpha_1^{[M]}, \ldots, \alpha_N^{[M]} \in \mathbb{R}$, and problem (4.5) has a unique solution $\vec{f}$, which is of the form

$$\vec{f} = \sum_{n=1}^{N} \alpha_n k(\cdot, \mu_n) \in H_k$$

with $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$. Furthermore, there exists a subsequence $(\vec{f}_{M_j})_j$ such that $\vec{f}_{M_j} \to \vec{f}$ for $j \to \infty$ in mean field, and

$$L\left(\vec{f}_{M_j}(\vec{x}_1^{[M_j]}), \ldots, \vec{f}_{M_j}(\vec{x}_N^{[M_j]})\right) + \lambda \|\vec{f}_{M_j}\|_{M_j}^2 \to L\left(\vec{f}(\mu_1), \ldots, \vec{f}(\mu_N)\right) + \lambda \|\vec{f}\|_k^2.$$

for $j \to \infty$.

Kernel ridge regression. We can immediately specialise this result to the case of KRR. Let $(\mu_1, y_1), \ldots, (\mu_N, y_N) \in \mathcal{P}(X) \times \mathbb{R}$, $(\vec{x}_1^{[M]}, y_1^{[M]}), \ldots, (\vec{x}_N^{[M]}, y_N^{[M]}) \in X^M \times \mathbb{R}$, $M \in \mathbb{N}_+$, such that for all $n = 1, \ldots, N$, it holds that $\hat{\mu}[\vec{x}_n^{[M]}] \xrightarrow{d_{\mathrm{KR}}} \mu_n$ for $M \to \infty$. Consider the KRR problems

$$\min_{f \in H_M} \frac{1}{N} \sum_{n=1}^{N} \left(f(\vec{x}_n^{[M]}) - y_n^{[M]}\right)^2 + \lambda \|f\|_M^2, \quad M \in \mathbb{N}_+ \tag{4.6}$$

$$\min_{f \in H_k} \frac{1}{N} \sum_{n=1}^{N} \left(f(\mu_n) - y_n\right)^2 + \lambda \|f\|_k^2. \tag{4.7}$$

where $\lambda \in \mathbb{R}_{>0}$ is the regularisation parameter. The problems have unique solutions

$$\vec{f}_M(\vec{x}) = \vec{k}_M(\vec{x})^\top (K_M + N\lambda I_N)^{-1} \vec{y}_M, \quad M \in \mathbb{N}_+$$
$$\vec{f}(\mu) = \vec{k}(\mu)^\top (K + N\lambda I_N)^{-1} \vec{y},$$

where we defined for $M \in \mathbb{N}_+$

$$K_M = (k_M(\vec{x}_j^{[M]}, x_i^{[M]}))_{i,j=1,\ldots,N}$$
$$\vec{k}_M(x)^\top = \left(k_M(x, x_1^{[M]}) \cdots k_M(x, x_N^{[M]})\right)$$
$$\vec{y}_N^\top = \left(y_1^{[M]} \cdots y_N^{[M]}\right).$$

and

$$K = (k(\mu_j, \mu_i))_{i,j=1,\ldots,N}$$
$$\vec{k}(\mu)^\top = \left(k(\mu, \mu_1) \cdots k(\mu, \mu_N)\right)$$
$$\vec{y}^\top = \left(y_1 \cdots y_N\right).$$

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

According to Proposition 4.6, there exists a strictly increasing sequence $(M_j)_j$ such that $\vec{f}_{M_j} \to \vec{f}$ in mean field, and

$$\frac{1}{N}\sum_{n=1}^{N}\left(\vec{f}_{M_j}(\vec{x}_n^{[M_j]}) - y_n^{[M_j]}\right)^2 + \lambda\|\vec{f}_{M_j}\|_{M_j}^2 \to \frac{1}{N}\sum_{n=1}^{N}\left(\vec{f}(\mu_n) - y_n\right)^2 + \lambda\|\vec{f}\|_k^2$$

for $j \to \infty$.

A concrete example of mean field kernels. As a concrete example of kernels in the mean-field limit, we use the double-sum kernel, cf. [31, Section 5.2]. In contrast to the latter reference, we proceed with a more elementary approach. Let k_0 be a bounded kernel on X, so there exists $B_0 \in \mathbb{R}_{\geq 0}$ such that $|k_0(x, x')| \leq B_0$ for all $x, x' \in X$. Define now $k : \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R}$ by

$$k(\mu, \nu) = \int_X \int_X k_0(x, x')\mathrm{d}\mu(x)\mathrm{d}\nu(x'). \tag{4.8}$$

Since $\mu, \nu \in \mathcal{P}(X)$ are finite measures and k_0 is bounded, the double integral above is well-defined. Furthermore, we have

$$\begin{aligned} k(\mu, \nu) &= \int_X \int_X k_0(x, x')\mathrm{d}\mu(x)\mathrm{d}\nu(x') \\ &= \int_X \int_X \langle k_0(\cdot, x'), k_0(\cdot, x)\rangle_{k_0}\mathrm{d}\mu(x)\mathrm{d}\nu(x') \\ &= \left\langle \int_X k_0(\cdot, x')\mathrm{d}\nu(x), \int_X k_0(\cdot, x)\mathrm{d}\mu(x)\right\rangle_{k_0}, \end{aligned}$$

where the integrals in the last line are in the sense of Bochner, cf. [67, Theorem 1], and we used in the last step that the scalar product as a continuous linear functional commutes with the Bochner integral. The above equality shows that k is indeed a kernel on $\mathcal{P}(X)$, cf. Remark 2.3.

For $M \in \mathbb{N}_+$, define $k_M : X^M \times X^M \to \mathbb{R}$ by

$$k_M(\vec{x}, \vec{x}') = \frac{1}{M^2}\sum_{i,j=1}^{M} k_0(x_i, x_j'). \tag{4.9}$$

These bivariate maps are called double-sum kernels, and it is well-known that they are indeed kernels, and permutation-invariant. Furthermore, since for $M \in \mathbb{N}_+$ and $\vec{x}, \vec{x}' \in X^M$ we have

$$|k_M(\vec{x}, \vec{x}')| = \left|\frac{1}{M^2}\sum_{i,j=1}^{M} k_0(x_i, x_j')\right| \leq \frac{1}{M^2}\sum_{i,j=1}^{M}|k_0(x_i, x_j')| \leq B_0,$$

the kernels k_M are uniformly bounded in the sense of Proposition 4.5.

Observe now that for all $M \in \mathbb{N}_+$ and $\vec{x}, \vec{x}' \in X^M$, we have

$$\begin{aligned} k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) &= \int_X \int_X k_0(x, x')\mathrm{d}\hat{\mu}[\vec{x}](x)\mathrm{d}\hat{\mu}[\vec{x}'](x') \\ &= \frac{1}{M}\sum_{i=1}^{M}\frac{1}{M}\sum_{j=1}^{M} k(x_i, x_j') \\ &= k_M(\vec{x}, \vec{x}'), \end{aligned}$$

which implies that

$$\lim_{M\to\infty}\sup_{\vec{x},\vec{x}'\in X^M}|k_M(\vec{x}, \vec{x}') - k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}'])| = 0,$$

so the kernels k_M converge to k in mean field in the sense of Definition 4.4.

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

Remark 4.7. *The preceding developments work for any measurable space $(X, \mathcal{A}_X)$, where $\mathcal{P}(X)$ is now the set of probability measures defined on this measurable space, and any $\mathcal{A} \otimes \mathcal{A}$-$\mathcal{B}(\mathbb{R})$-measurable (here $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$) and bounded kernel $k_0 : X \times X \to \mathbb{R}$.*

Note that while we established the mean field convergence of k_M directly, without relying on Proposition 4.5, we still need all the assumptions of this latter result for the mean-field limit variant of the Representer Theorem stated as Proposition 4.6, cf. the corresponding proofs in [33]. The only missing property for the kernels k_M is uniform Lipschitz continuity. For a particular and broad class of kernels, the following result provides a sufficient condition for this property. Please refer to Appendix A for the proof of the following proposition. To the best of our knowledge, this result is new.

Proposition 4.8. *Let $\mathcal{X}$ be a normed vectorspace, $X \subseteq \mathcal{X}$ a nonempty Borel-measurable subset, $\phi : X \to \mathbb{R}$ a L-Lipschitz continuous function, define $\kappa_0 : X \times X \to \mathbb{R}$ by $\kappa_0(x, x') = \phi(\|x - x'\|)$, and for $M \in \mathbb{N}_+$ define $\kappa_M(\vec{x}, \vec{x}') = \frac{1}{M^2} \sum_{i,j=1}^{M} \kappa_0(\vec{x}_i, \vec{x}'_j)$. We then have for all $M \in \mathbb{N}_+$, $\vec{x}, \vec{x}', \vec{y}, \vec{y}' \in X^M$ that*

$$|\kappa_M(\vec{x}, \vec{x}') - \kappa_M(\vec{y}, \vec{y}')| \leq L\, d_{KR}^2 \left((\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']), (\hat{\mu}[\vec{y}], \hat{\mu}[\vec{y}']) \right).$$

Let's consider now $X \subseteq \mathcal{H}$ a nonempty subset of a Hilbert space. A kernel k_0 on X of the form $k_0(x, x') = \phi(\|x - x'\|)$ is called a *radial kernel*, or a *radial basis function (kernel)*. In the following, we consider $\mathcal{H} = \mathbb{R}^d$, $X \subseteq \mathbb{R}^d$ a nonempty compact subset, and choose k_0 as the Gaussian kernel, so in this case, $\phi(s) = \exp\left(-\frac{s^2}{2\gamma^2}\right)$. Observe that this ϕ is bounded, and (globally) Lipschitz continuous with Lipschitz bound given by $\max_{s \in \mathbb{R}} |\phi'(s)|$, so the resulting sequence of double-sum kernel fulfils all conditions from Proposition 4.5.

Remark 4.9. *We would like to point out the following delicate aspects of the preceding developments. By direct calculation, we have established the mean field convergence of the double-sum kernels k_M, as defined in (4.9), to k given by (4.8). Furthermore, the sequence of double-sum kernels based on the Gaussian kernel fulfils all the conditions of Proposition 4.5, so there exists a mean-field limit kernel that is bounded and Lipschitz continuous, and a subsequence of the double-sum kernel sequence, that converges in mean field to this latter kernel. However, we* did not prove that this kernel is (4.8). *If we had uniqueness of the mean-field limit kernel in Proposition 4.5, then this would trivially follow. Investigation of this uniqueness question is beyond the scope of the present work. However, it is clear that (4.8) is bounded, and by using mutatis mutandis the arguments from the proof of Proposition 4.8, one can verify that (4.8) is Lipschitz continuous. This means that (4.8) fulfils the properties from the limit kernel in Proposition 4.5.*

Experimental setup. Below, we conduct numerical experiments concerning mean field kernel methods, particularly emphasising learning tasks within large-scale IPS. These tests numerically validate the theoretical insights presented in this section. In particular, we focus on the kernel approximation of the variance $v_M : (\mathbb{R}^d)^M \to \mathbb{R}$ and the skewness $s_M : (\mathbb{R}^d)^M \to \mathbb{R}$ of the agent system $\vec{x} = (x_i)_{i=1,\dots,M} \in (\mathbb{R}^d)^M$, i.e.

$$v_M(\vec{x}) = \frac{1}{M} \sum_{i=1}^{M} \|x_i - m_M(\vec{x})\|^2,$$

$$s_M(\vec{x}) = \frac{1}{M} \sum_{i=1}^{M} \left(\frac{\|x_i - m_M(\vec{x})\|}{\sqrt{v_M(\vec{x})}} \right)^3,$$

where $m_M(\vec{x}) = \frac{1}{M} \sum_{i=1}^{M} x_i$ is the mean of the agents and $\|\cdot\|$ denotes the usual Euclidean norm in $\mathbb{R}^d$. The mean-field limit of those functions is $v_\infty, s_\infty : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$ given by

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

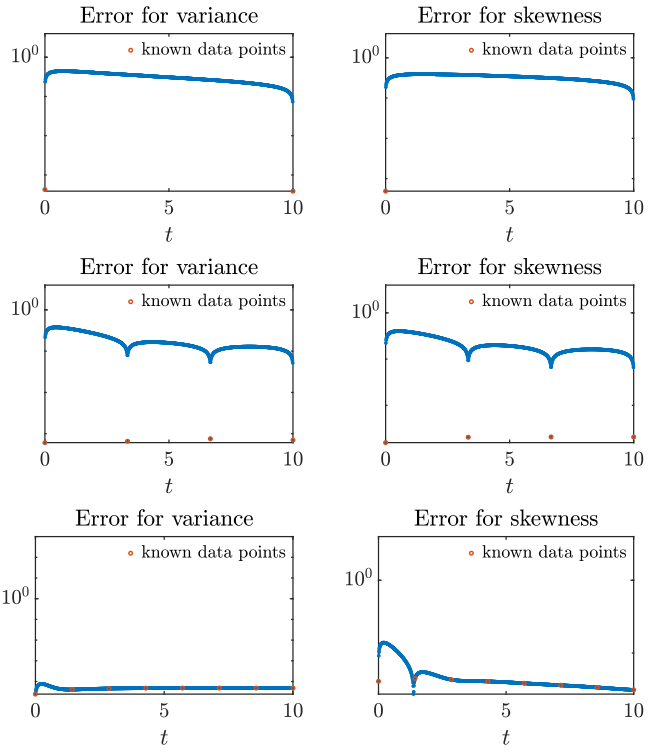

Figure 3. Test B1. *Inference problem for the microscopic dynamics (3.1). Evolution in time of the errors* $|v_M - \hat{v}_M|$ *(left column) and* $|s_M - \hat{s}_M|$ *(right column) for different number of measurements* $N \in \{2$ *(top row),* 4 *(middle row),* 8 *(bottom row)*$\}$*. The known data points are indicated as red dots.*

$$v_\infty(\mu) = \int_{\mathbb{R}^d} \|x - m_\infty(\mu)\|^2 \, \mu(x)dx \,,$$
$$s_\infty(\mu) = \int_{\mathbb{R}^d} \left(\frac{\|x - m_\infty(\mu)\|}{\sqrt{v_\infty(\mu)}} \right)^3 \mu(x)dx \,,$$

where the mean is $m_\infty(\mu) = \int_{\mathbb{R}^d} x \, d\mu(x)$. Please refer to Table 1 for simulation parameters of all the numerical tests, and to Section 3.2 for the models' discretization approaches.

Regarding the choice of the kernel, in the following numerical tests, we consider the double-sum kernel k_M in equation (4.9), and the correspondent mean-field limit kernel k in Equation (4.8). In both cases, we take $k_0 = k_\gamma$ given by (2.1).

Test B1. Microscopic first-order model. In this section, we present numerical tests for the opinion formation model (3.1). The interaction between the agents is described by P, which is given by

$$P(x_i, x_j) = \|x_i - x_j\|^2 \,. \tag{4.10}$$

P promotes mutual attraction among the agents as it consistently remains non-negative. The initial conditions x_i^0 for the agents $i = 1, \ldots, M$, are randomly chosen with uniform distribution in the interval $[1, 2]$, i.e., $\vec{x}(t_0) \sim \mathcal{U}([1, 2])^M$. For the given function P, it is known that the dynamics $x_i(t) \in [1, 2]$ for all $t \geq 0$ due to the non-negativity of the interaction rules. We consider first both functionals v_M and s_M in the noise-free case.

The numerical results presented in Figure 3 illustrate the inference of v_M and s_M under varying values of measurements $N \in \{2, 4, 8\}$, with $M = 30$ agents. These results underscore the high approximation

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

Table 2. Test B1. *L^∞ error over time for different number of measurements N.*

		Variance	Skewness
M	N	$\|v_M - \hat{v}_M\|_\infty$	$\|s_M - \hat{s}_M\|_\infty$
30	2	1.60e-2	5.42e-2
30	4	4.42e-3	4.17e-3
30	8	7.07e-5	1.37e-4

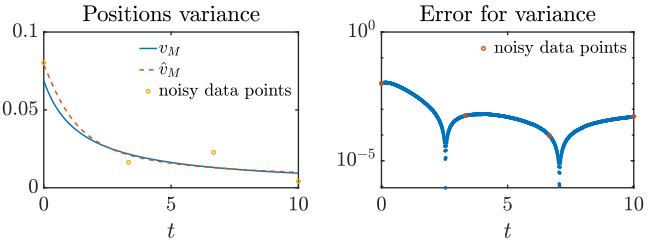

Figure 4. Test B1. *Inference problem for the microscopic dynamics (3.1) with noisy measurements. Left: comparison between true (blue) and approximated (red) variance. Right: error $|v_M - \hat{v}_M|$. Dots underline the values in time where the (noisy) variance is accessible.*

quality of the RKHS, even when the number of measurements is rather low. The graphical representation portrays the evolution in time of the errors $|v_M - \hat{v}_M|$ and $|s_M - \hat{s}_M|$, indicating a small discrepancy between the inferred functionals $\hat{v}_M(\vec{x}(t)), \hat{s}_M(\vec{x}(t))$, derived from solving the interpolation problem presented in Proposition 2.6, and the true solutions $v_M(\vec{x}(t)), s_M(\vec{x}(t))$ across all time points $t \geq 0$. This alignment is further substantiated by the numerical data in Table 2. Notably, as the number of known data points N increases, the error associated with approximated variance and skewness diminishes.

As an additional scenario, we introduce random noise perturbations η to the evaluations, as described in Section 2.3. Maintaining the same initial condition $\vec{x}(t_0) \sim \mathcal{U}([1,2])^M$, we now explore the approximation of v_M and s_M in the presence of noise. Specifically, we consider η_n that follows a normal distribution $\eta_n \sim \mathcal{N}(0, \sigma^2)$, for $i = 1, \ldots, N$, with $\sigma^2 = 0.01$. In this noisy scenario, we solve the minimisation problem (2.3) with $\lambda = \sqrt{\sigma^2}$.

Figure 4 illustrates the estimation of v_M and s_M under the condition of $N = 4$ measurements and a swarm size of $M = 30$ agents. Even with the presence of noise, these results demonstrate that the unknown functional can be successfully approximated. Notably, the approximation exhibits similar behaviour to the true function even in the presence of noise.

Test B2. Mean field first-order model. In this section, we explore the same opinion formation model, examining it first at the microscopic level while gradually increasing the number of particles, and subsequently, in the context of the mean-field limit described by Equation (3.4). The interaction between the agents P is still given by (4.10). As initial conditions of the agents, in the microscopic model (3.1), we consider a uniform distribution $\mathcal{U}([1,2])$. In other words, we set $\vec{x}(t_0) \sim \mathcal{U}([1,2])^M$, which means the parameter space is defined again as $[1,2] \subset \mathbb{R}$.

The numerical results presented in Figure 5 demonstrate the estimation error of $v_M(\vec{x}(t))$ and $s_M(\vec{x}(t))$ for $N = 8$ measurements and various values of M, where M is chosen from the set $\{10, 100, 1000\}$. We observe that the quality of approximation remains consistent across different values of the agent population, confirming the fact that there exists a well-defined mean-field limit. Consequently, the inference problem appears to be independent of the number of agents, and these findings are further substantiated by the numerical values provided in Table 3. The inference problem is addressed at the mean

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

Table 3. Test B2. *L^∞ error over time for increasing number of agents M, including the mean field limit (last row).*

		Variance	Skewness
M	N	$\|v_M - \hat{v}_M\|_\infty$	$\|s_M - \hat{s}_M\|_\infty$
10	8	1.26e-5	2.33e-4
100	8	1.43e-5	4.22e-4
1000	8	1.22e-5	6.71e-4
∞	8	3.21e-5	1.71e-4

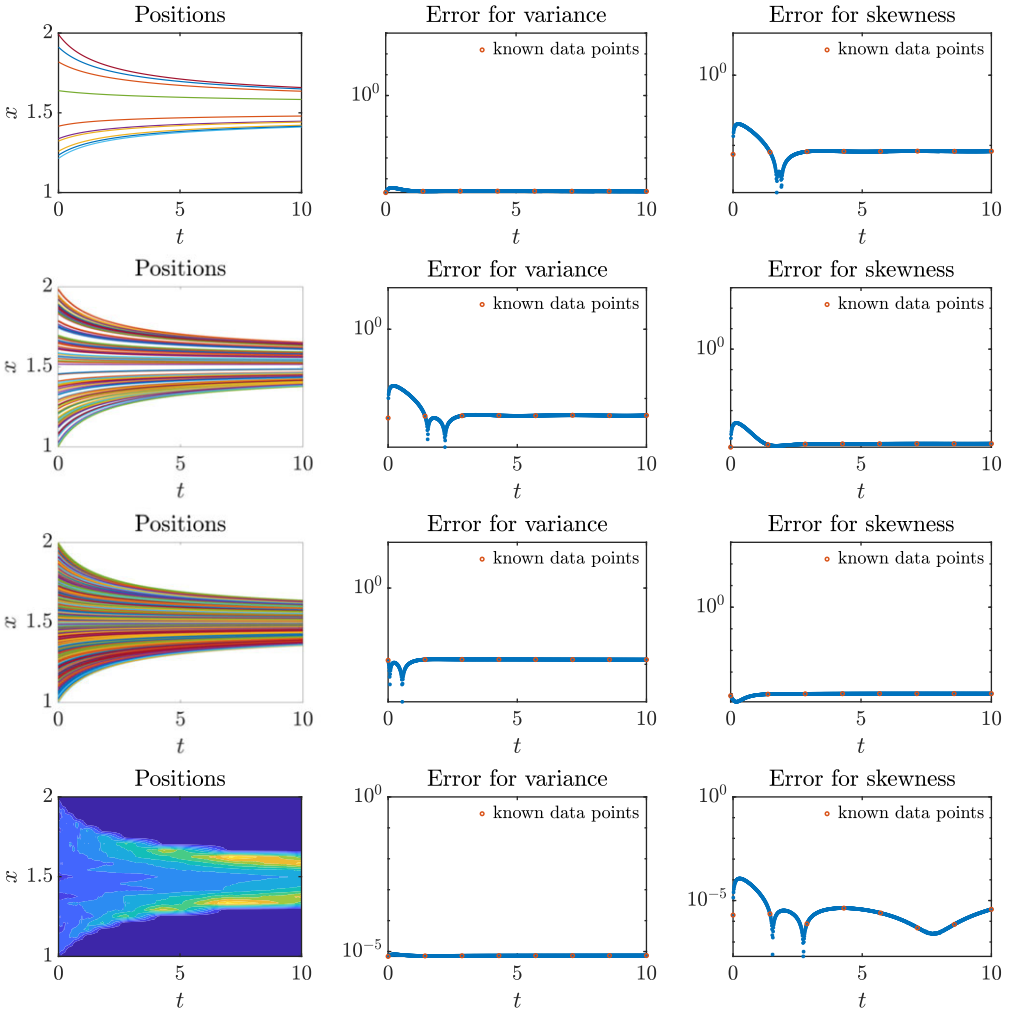

Figure 5. Test B2. *Evolution in time of particles and density (column 1) for the microscopic (3.1) (rows 1-2-3) and mean field (3.4) (row 4) dynamics. Evolution in time of the error $|v_M - \hat{v}_M|$ (column 2) and $|s_M - \hat{s}_M|$ (column 3) for different number of agents $M \in \{10$ (row 1), 100 (row 2), 1000 (row 3), ∞ (row 4)$\}$. The known data points are indicated as red dots.*

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

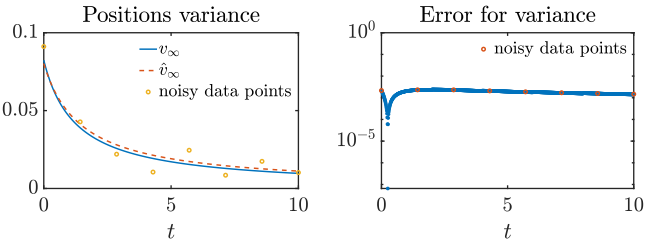

Figure 6. Test B2. *Approximation problem for the mean-field dynamics (3.4) with noisy measurements of variance evolution. Left: comparison between true (blue) and approximated (red) variance. Right: error* $|v_\infty - \hat{v}_\infty|$. *Dots underline the values in time where the (noisy) variance is accessible.*

field level, employing an MC simulation method as explained in the Section 3.2. We choose a subsample of $\hat{M}_s = 100$, and a sample of $\hat{M} = 10,000$ particles for the approximation of the density, we then reconstruct the evolution in the phase space by a histogram approach. For the initial condition, we adopt $\mu_0(x) = \chi_{[1,2]}(x)$ that is the indicator function of the interval $[a, b]$, i.e. the function that equals 1 when x is within the interval $[a, b]$ and equals 0 when x is outside that interval. In a manner consistent with our approach, we consider a set of $N = 8$ measurements, clearly marked as red dots in Figure 5. At these specific measurement times, the complete state $\mu(t_i, \cdot)$ is recorded. Both functionals for variance and skewness, denoted as v_∞ and s_∞, respectively, are assessed at these measurement times t_i for $i = 1, \ldots, N$. The error for the mean field case (see Figure 5, row 4) is of the same order as the scenario with a finite number of agents.

As for the microscopic model in Test B1, we investigate here the noisy scenario also for the mean field case. We introduce random noise perturbations η into the variance evaluations

$$y_n = v_\infty(\mu(t_n)) + \eta_n, \quad n = 1, \ldots, N,$$

and we solve the approximation problem as detailed in the mean-field kernel ridge regression paragraph in Section 4.2. We still consider $\mu_0(x) = \chi_{[1,2]}(x)$ as initial condition and we examine the approximation of v_∞ in the presence of noise. Here, η_n is assumed again to follow a normal distribution, specifically $\eta_n \sim \mathcal{N}(0, \sigma^2)$, for $i = 1, \ldots, N$, with σ^2 set to 0.01. In this noisy setting, we tackle the minimisation problem (4.7) using $\lambda = \sqrt{\sigma^2}$. Figure 6 showcases the estimation of v_∞ under the condition of having $N = 8$ noisy measurements of the system variance. Despite the presence of noise, the results indicate that the unknown functional can still be approximated effectively.

Test B3. Microscopic second-order model. In this section, we conduct numerical experiments to examine the second-order model (3.2) and the corresponding mean field PDE (3.5). The agent interactions, as represented by the parameter H_β, are governed by the Cucker–Smale function (3.3) with $\beta = 2$. In this model, the agents in the swarm align their velocities with the average velocity of their nearby neighbours, while they are also attracted to their neighbours, which helps to maintain group cohesion. In the context of this model, our primary focus is on approximating the velocity variance in a noise-free scenario, denoted as v_M. Both initial position and velocity conditions are randomly selected from a uniform distribution within the interval $[1, 2]$, specifically as $\vec{x}(t_0), \vec{v}(t_0) \sim \mathcal{U}([1, 2])^M$.

The numerical results displayed in Figure 7 demonstrate the inference process for a case with $N = 4$ measurements and a swarm consisting of $M = 30$ agents. These results highlight the remarkable accuracy of the RKHS approach, even when applied to a second-order microscopic system.

Test B4. Mean field second-order model. We now address the inference problem at the mean field level, using the two-dimensional Lax–Friedrichs scheme [50], as elaborated in Section 3.2. We consider Dirichlet's initial and Neumann's final boundary conditions. For the discretization of both x and y spaces,

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

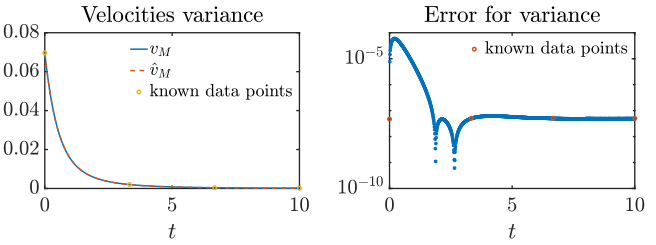

Figure 7. Test B3. *Inference problem for the second-order microscopic dynamics (3.2). Left: comparison true (blue) and approximated (red) variance. Right: error* $|v_M - \hat{v}_M|$*. Dots underline the values in time where the true variance is accessible.*

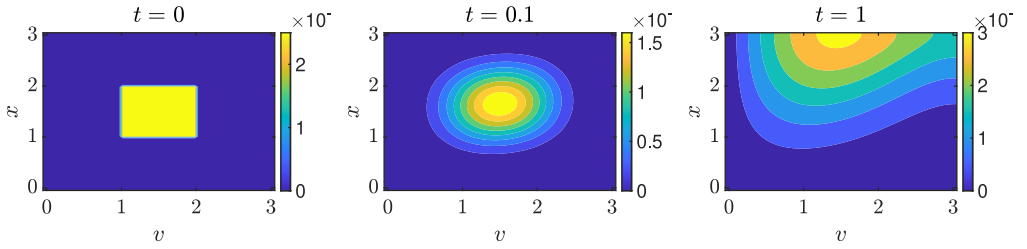

Figure 8. Test B4. *Three snapshot in time* $t \in \{0, 0.1, 1\}$ *of the density* $v(t, x, v)$ *evolution for the mean field model (3.5).*

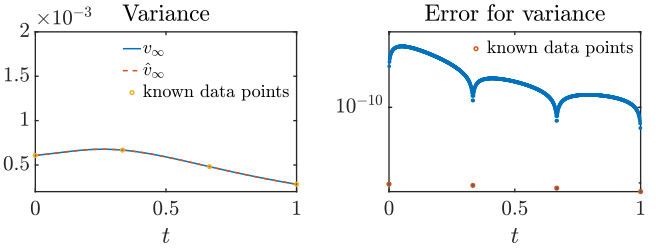

Figure 9. Test B4. *Inference problem for the second-order mean field dynamics (3.5). Left: comparison of true (blue) and approximated (red) variance. Right: error* $|v_\infty - \hat{v}_\infty|$*. Dots underline the values in time where the true variance is accessible.*

we take $\Delta x = \Delta v = 0.05$ in the interval $[0, 3]$. For time, in order to respect the Courant–Friedrichs–Lewy (CFL) stability condition, we take $\Delta t = 0.001$ in $[0\ 1]$.

In Figure 8, we provide three snapshots illustrating the density evolution over time. As initial condition, we opt for $\mu_0(x, v) = \chi_{[1,2]}(x) \times \chi_{[1,2]}(v)$, depicted in the first plot on the left. As expected from the model, the density is seen moving upwards in the x direction while concentrating in the v dimension, reflecting alignment behaviour. The observed diffusion is a result of the chosen numerical scheme.

Similar to the scenario with a finite number of agents, Figure 9 shows that no discernible differences exist between the kernel-based estimated variance and the actual variance.

5. Conclusion and outlook

In this paper, we have outlined recent and novel kernel-based approaches for numerical problems involving IPS. After providing a self-contained presentation of background on kernels and kernel methods, as

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

well as IPS and their numerical treatment, we presented interesting problem classes amenable to kernel methods. First, we investigated the usage of kernel methods for surrogate modelling in the context of IPS, which can provide a route to reducing the computational cost of properties of IPS. Our initial numerical results indicate that this could be a promising avenue for future research, in particular in the context of large-scale IPS. Second, we conducted the first numerical experiments on kernels in the mean-field limit, a recent development started in ref. [31, 33]. The numerical experiments show that this approach can indeed connect learning and approximation problems on the microscopic and mesoscopic levels. In future work, we plan to explore the learning rates of kernel approximations in a mean-field context and conduct a numerical error analysis.

Financial support. The authors thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for the financial support under Germany's Excellence Strategy EXC-2023 Internet of Production 390621612 and under the Excellence Strategy of the Federal Government and the Länder, 333849990/GRK2379 (IRTG Hierarchical and Hybrid Approaches in Modern Inverse Problems), 320021702/GRK2326, 442047500/SFB1481 within the projects B04, B05 and B06, through SPP 2410 Hyperbolic Balance Laws in Fluid Mechanics: Complexity, Scales, Randomness (CoScaRa) within the Project(s) HE5386/26-1 and HE5386/27-1, and through SPP 2298 Theoretical Foundations of Deep Learning within the Project(s) HE5386/23-1, mean field Theorie zur Analysis von Deep Learning Methoden (462234017). Support through the EU DATAHYKING No. 101072546 is also acknowledged. C. Segala is a member of the Italian National Group of Scientific Calculus (Indam GNCS).

Competing interests. The authors declare none.

References

[1] Albi, G., Bellomo, N., Fermo, L., et al. (2019) Vehicular traffic, crowds, and swarms: From kinetic theory and multiscale methods to applications and research perspectives. *Math. Models Methods Appl. Sci.* **29**(10), 1901–2005.
[2] Albi, G. & Ferrarese, F. (2023) Kinetic description of swarming dynamics with topological interaction and emergent leaders. *Multiscale Model. Simul.* **22**, 1169–1195.
[3] Albi, G. & Pareschi, L. (2013) Binary interaction algorithms for the simulation of flocking and swarming dynamics. *Multiscale Model. Simul.* **11**, 1–29.
[4] Albi, G., Pareschi, L. & Zanella, M. (2014) Boltzmann-type control of opinion consensus through leaders. *Philos. Ser. A Math. Phys. Eng. Sci.* **372**, 20140138,18.
[5] Babovsky, H. & Neunzert, H. (1986) On a simulation scheme for the boltzmann equation. *Math. Methods Appl. Sci.* **8**(1), 223–233.
[6] Bailo, R., Bongini, M., Carrillo, J. A. & Kalise, D. (2018) Optimal consensus control of the cucker-smale model. *IFAC-PapersOnLine* **51**(13), 1–6.
[7] Bellomo, N., Degond, P. & Tadmor, E. (2017) *Active Particles, Volume 1: Advances in Theory, Models, and Applications*, Switzerland, Birkhäuser.
[8] Bellomo, N., Degond, P. & Tadmor, E. (2019) *Active Particles, Volume 2: Advances in Theory, Models, and Applications*, Switzerland, Birkhäuser.
[9] Bellomo, N. & Soler, J. (2012) On the mathematical theory of the dynamics of swarms viewed as complex systems. *Math. Models Methods Appl. Sci.* **22**(supp01), 1140006,29.
[10] Benfenati, A., Borghi, G. & Pareschi, L. (2022) Binary interaction methods for high dimensional global optimization and machine learning. *Appl. Math. Optim.* **86**, Paper No. 9.
[11] Bobylev, A. & Nanbu, K. (2000) Theory of collision algorithms for gases and plasmas based on the Boltzmann equation and the Landau-Fokker-Planck equation. *Phys. Rev. E* **61**(4), 4576–4586.
[12] Bongini, M., Fornasier, M., Hansen, M. & Maggioni, M. (2017) Inferring interaction rules from observations of evolutive systems i: The variational approach. *Math. Models Methods Appl. Sci.* **27**(05), 909–951.
[13] Boudin, L. & Salvarani, F. (2009) A kinetic approach to the study of opinion formation. *M2AN Math. Model. Numer. Anal.* **43**(3), 507–522.
[14] Canizo, J. A., Carrillo, J. A. & Rosado, J. (2011) A well-posedness theory in measures for some kinetic models of collective motion. *Math. Models Methods Appl. Sci.* **21**(03), 515–539.
[15] Carmona, R. & Delarue, F. (2018) *Probabilistic Theory of Mean Field Games with Applications I-II*, Springer, Berlin.
[16] Carrillo, J. A., Choi, Y.-P. & Hauray, M. (2014) The derivation of swarming models: Mean-field limit and Wasserstein distances. In: *Collective Dynamics From Bacteria to Crowds*, Springer, Berlin, pp. 1–46.
[17] Carrillo, J. A., Fornasier, M., Rosado, J. & Toscani, G. (2010) Asymptotic flocking dynamics for the kinetic cucker–smale model. *SIAM J. Math. Anal.* **42**(1), 218–236.
[18] Carrillo, J. A., Fornasier, M., Toscani, G. & Vecil, F. (2010) Particle, kinetic, and hydrodynamic models of swarming. In: *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*, Springer, Berlin, pp. 297–336.
[19] Carrillo, J. A., Jin, S., Li, L. & Zhu, Y. (2021) A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM Control Optim. Calc. Var.* **27**(S5), PaperNo.S5,22.

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

[20] Cercignani, C., Illner, R. & Pulvirenti, M. (2013) *The Mathematical Theory of Dilute Gases*, Springer Science & Business Media, Berlin, Vol. **106**.
[21] Cristiani, E., Piccoli, B. & Tosin, A. (2014) *Multiscale Modeling of Pedestrian Dynamics*, Vol. **12**, Springer, Cham.
[22] Cucker, F. & Smale, S. (2007) Emergent behavior in flocks. *IEEE Trans. Autom. Control* **52**(5), 852–862.
[23] Degond, P., Herty, M. & Liu, J.-G. (2014) Flow on sweeping networks. *Multiscale model. Simul.* **12**, 538–565.
[24] Degond, P. & Motsch, S. (2008) Continuum limit of self-driven particles with orientation interaction. *Math. Models Methods Appl. Sci.* **18**(supp01), 1193–1215.
[25] Dimarco, G., Caflisch, R. & Pareschi, L. (2010) Direct simulation Monte Carlo schemes for coulomb interactions in plasmas. *Commun. Appl. Ind. Math.* **1**, 72–91.
[26] D'Orsogna, M. R., Chuang, Y.-L., Bertozzi, A. L. & Chayes, L. S. (2006) Self-propelled particles with soft-core interactions: Patterns, stability, and collapse. *Phys. Rev. Lett.* **96**(10), 104302.
[27] Dudley, R. M. (2002). *Real Analysis and Probability*, Cambridge University Press, Cambridge.
[28] Düring, B., Markowich, P., Pietschmann, J.-F. & Wolfram, M.-T. (2009) Boltzmann and Fokker-Planck equations modelling opinion formation in the presence of strong leaders. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **465**, 3687–3708.
[29] Estrada-Rodriguez, G. & Gimperlein, H. (2020) Interacting particles with Lévy strategies: Limits of transport equations for swarm robotic systems. *SIAM J. Appl. Math.* **80**(1), 476–498.
[30] Fiedler, C. (2023) Lipschitz and hölder continuity in reproducing kernel hilbert spaces, arXiv preprint arXiv: 2310.18078.
[31] Fiedler, C., Herty, M., Rom, M., Segala, C. & Trimpe, S. (2023) Reproducing kernel Hilbert spaces in the mean field limit. *Kinet. Relat. Models* **16**(6), 850–870.
[32] Fiedler, C., Herty, M. & Trimpe, S. (2023) Mean-field limits for discrete-time dynamical systems via kernel mean embeddings. *IEEE Control Syst. Lett.* **7**, 3914–3919.
[33] Fiedler, C., Herty, M. & Trimpe, S. (2023) On kernel-based statistical learning theory in the mean field limit. In: Thirty-seventh Conference on Neural Information Processing Systems.
[34] Fornasier, M., Haskovec, J. & Toscani, G. (2011) Fluid dynamic description of flocking via the Povzner-Boltzmann equation. *Phys. D* **240**(1), 21–31.
[35] Fornasier, M., Huang, H., Pareschi, L. & Sünnen, P. (2021) Consensus-based optimization on the sphere: Convergence to global minimizers and machine learning. *J. Mach. Learn. Res.* **22**, PaperNo.237,55.
[36] Fornasier, M., Lisini, S., Orrieri, C. & Savaré, G. (2019) Mean-field optimal control as gamma-limit of finite agent controls. *Eur. J. Appl. Math.* **30**(6), 1153–1186.
[37] Gibelli, L. (2020) *Crowd Dynamics, Volume 2: Theory, Models, and Applications*, Springer Nature, Cham.
[38] Gibelli, L. & Bellomo, N. (2019) *Crowd Dynamics, Volume 1: Theory, Models, and Safety Problems*, Springer, Cham.
[39] Gómez-Serrano, J., Graham, C. & Le Boudec, J.-Y. (2012) The bounded confidence model of opinion dynamics. *Math. Models Methods Appl. Sci.* **22**(02), 1150007,46.
[40] Gramacy, R. B. (2020) *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*, Chapman and Hall/CRC, Boca Raton, FL.
[41] Ha, S.-Y. & Tadmor, E. (2008) From particle to kinetic and hydrodynamic descriptions of flocking. *Kinet. Relat. Models* **1**(3), 415–435.
[42] Hegselmann, R., Krause, U., et al. (2002) Opinion dynamics and bounded confidence models, analysis, and simulation. *J. Artif. Soc. Soc. Simul.* **5**, 1–33.
[43] Herty, M., Pareschi, L. & Visconti, G. (2020) Mean field models for large data-clustering problems. *Netw. Heterog. Media* **15**(3), 463–487.
[44] Herty, M., Trimborn, T. & Visconti, G. (2022) Mean-field and kinetic descriptions of neural differential equations. *Found. Data Sci.* **4**(2), 271–298.
[45] Herty, M. & Visconti, G. (2020) Continuous limits for constrained ensemble Kalman filter. *Inverse Probl.* **36**(7), 075006,28.
[46] Herty, M. & Zanella, M. (2017) Performance bounds for the mean-field limit of constrained dynamics. *Discrete Contin. Dyn. Syst.* **37**(4), 2023–2043.
[47] Jin, S., Li, L. & Liu, J.-G. (2020) Random batch methods (RBM) for interacting particle systems. *J. Comput. Phys.* **400**, 108877.
[48] Kanagawa, M., Hennig, P., Sejdinovic, D. & Sriperumbudur, B. K. (2018) Gaussian processes and kernel methods: A review on connections and equivalences, arXiv preprint arXiv: 1807.02582.
[49] Lax, P. D. (1954) Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Commun. Pure Appl. Math.* **7**(1), 159–193.
[50] LeVeque, R. (2002) *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, Cambridge Texts in Applied Mathematics.
[51] Liu, F., Huang, X., Chen, Y. & Suykens, J. A. (2021) Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Trans. Pattern Anal.* **44**(10), 7128–7148.
[52] Liu, H., Ong, Y.-S., Shen, X. & Cai, J. (2020) When gaussian process meets big data: A review of scalable gps. *IEEE Trans. Neural Networks Learn.* **31**, 4405–4423.
[53] Lu, F., Maggioni, M. & Tang, S. (2021) Learning interaction kernels in heterogeneous systems of agents from multiple trajectories. *J. Mach. Learn. Res.* **22**, 1–67.
[54] Lu, F., Maggioni, M. & Tang, S. (2021) Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories. *Found. Comput. Math.* **22**, 1013–1067.
[55] Lu, F., Zhong, M., Tang, S. & Maggioni, M. (2019) Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proc. Natl. Acad. Sci.* **116**(29), 14424–14433.

[56] Mei, S., Misiakiewicz, T. & Montanari, A. (2019) Mean-field theory of two-layers neural networks: dimension-free bounds and Kernel limit. In: Conference on Learning Theory, PMLR, New York, pp. 2388–2464.
[57] Motsch, S. & Tadmor, E. (2014) Heterophilious dynamics enhances consensus. *SIAM Rev.* **56**(4), 577–621.
[58] Murphy, K. P. (2022) *Probabilistic Machine Learning: An Introduction*, MIT Press, Cambridge, MA.
[59] Pareschi, L. & Toscani, G. (2013) *Interacting Multi-Agent Systems. Kinetic Equations & Monte Carlo Methods*, Oxford University Press, USA.
[60] Paulsen, V. I. & Raghupathi, M. (2016) *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, Cambridge University Press, Cambridge, Vol. **152**.
[61] Pinnau, R., Totzeck, C., Tse, O. & Martin, S. (2017) A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.* **27**(01), 183–204.
[62] Rasmussen, C. E. & Williams, C. K. (2006) *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, Vol. **2**.
[63] Santner, T. J., Williams, B. J., Notz, W. I. & Williams, B. J. (2018) *the Design and Analysis of Computer Experiments*. 2nd ed., Springer, Cham.
[64] Schölkopf, B., Herbrich, R. & Smola, A. J. (2001) A generalized representer theorem. In: International Conference on Computational Learning Theory, Springer, Berlin, pp. 416–426.
[65] Scholkopf, B. & Smola, A. J. (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA.
[66] Shawe-Taylor, J. & Cristianini, N. (2004) *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge.
[67] Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B. & Lanckriet, G. R. (2010) Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.* **11**, 1517–1561.
[68] Steinwart, I. & Christmann, A. (2008) *Support Vector Machines*, Springer Science & Business Media, Berlin.
[69] Toscani, G. (2006) Kinetic models of opinion formation. *Commun. Math. Sci.* **4**, 481–496.
[70] Weidlich, W. (2003) Sociodynamics–a systematic approach to mathematical modelling in the social sciences. *Chaos, Solitons & Fractals*, **18**, 431–437. Complex Economic Phenomena in Time and Space in honour of Prof. Tonu Puu
[71] Williams, C. K. & Rasmussen, C. E. (2006) *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, Vol. **2**.

A. Additional material

Proof of Lemma 2.5

Let $x, x' \in \mathcal{X}$ be arbitrary, then $k(\cdot, x') \in H_k$, and hence the second statement implies that $k(T(x), x') = k(\cdot, x')(T(x)) = k(\cdot, x')(x) = k(x, x')$. Conversely, observe that by symmetry of k we have for all $x, x' \in \mathcal{X}$ that $k(x, T(x')) = k(T(x'), x) = k(x', x) = k(x, x')$. Since this holds for all $x, x' \in \mathcal{X}$, we find that $k(\cdot, T(x')) = k(\cdot, x')$ as an element of H_k, so we get for all $f \in H_k$ and $x \in \mathcal{X}$ that $f(T(x)) = \langle f, k(\cdot, T(x))\rangle_k = \langle f, k(\cdot, x)\rangle_k = f(x)$ by the reproducing property of k.

Proof of Proposition 4.8

Without loss of generality, we can assume that $L \in \mathbb{R}_{>0}$. Define for $x \in X$ the function $\varphi_x : X \to \mathbb{R}$ by $\varphi_x(x') = L^{-1}\phi(\|x' - x\|)$, and observe that since that for all $x', y' \in X$

$$\begin{aligned} |\varphi_x(x') - \varphi_x(y')| &= L^{-1}|\phi(\|x' - x\|) - \phi(\|y' - x\|)| \\ &\leq L^{-1}L|\,\|x' - x\| - \|y' - x\|\,| \\ &\leq \|(x' - x) - (y' - x)\| \\ &= \|x' - y'\| \end{aligned}$$

the function φ_x is 1-Lipschitz continuous.

https://doi.org/10.1017/S0956792524000706 Published online by Cambridge University Press

Let now $M \in \mathbb{N}_+$, $\vec{x}, \vec{x}', \vec{y}, \vec{y}' \in X^M$, then we get

$$\begin{aligned}
|\kappa_M(\vec{x},\vec{x}') - \kappa_M(\vec{y},\vec{y}')| &= \left| \frac{1}{M^2} \sum_{i,j=1}^{M} \kappa_0(x_i, x_j') - \frac{1}{M^2} \sum_{i,j=1}^{M} \kappa_0(y_i, y_j') \right| \\
&= \left| \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{M} \sum_{j=1}^{M} \phi(\|x_i - x_j'\|) - \frac{1}{M} \sum_{j=1}^{M} \phi(\|y_i - y_j'\|) \right) \right| \\
&\le \left| \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{M} \sum_{j=1}^{M} \phi(\|x_i - x_j'\|) - \frac{1}{M} \sum_{j=1}^{M} \phi(\|x_i - y_j'\|) \right) \right| \\
&\quad + \left| \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{M} \sum_{j=1}^{M} \phi(\|x_i - y_j'\|) - \frac{1}{M} \sum_{j=1}^{M} \phi(\|y_i - y_j'\|) \right) \right| \\
&= L \left| \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{M} \sum_{j=1}^{M} \varphi_{x_i}(x_j') - \frac{1}{M} \sum_{j=1}^{M} \varphi_{x_i}(y_j') \right) \right| \\
&\quad + L \left| \frac{1}{M} \sum_{j=1}^{M} \left(\frac{1}{M} \sum_{i=1}^{M} \varphi_{y_j'}(x_i) - \frac{1}{M} \sum_{i=1}^{M} \varphi_{y_j'}(y_i) \right) \right| \\
&\le L \frac{1}{M} \sum_{i=1}^{M} \left| \frac{1}{M} \sum_{j=1}^{M} \varphi_{x_i}(x_j') - \frac{1}{M} \sum_{j=1}^{M} \varphi_{x_i}(y_j') \right| \\
&\quad + L \frac{1}{M} \sum_{j=1}^{M} \left| \frac{1}{M} \sum_{i=1}^{M} \varphi_{y_j'}(x_i) - \frac{1}{M} \sum_{i=1}^{M} \varphi_{y_j'}(y_i) \right|.
\end{aligned}$$

Observe now that for all Borel-measurable $f : X \to \mathbb{R}$, we have

$$\frac{1}{M} \sum_{i=1}^{M} f(x_i) = \int_X f(x) \mathrm{d}\hat{\mu}[\vec{x}](x),$$

so we can continue with

$$\begin{aligned}
|\kappa_M(\vec{x},\vec{x}') - \kappa_M(\vec{y},\vec{y}')| &\le L \frac{1}{M} \sum_{i=1}^{M} \left| \int_X \varphi_{x_i}(x') \mathrm{d}\hat{\mu}[\vec{x}'](x') - \int_X \varphi_{x_i}(y') \mathrm{d}\hat{\mu}[\vec{y}'](y') \right| \\
&\quad + L \frac{1}{M} \sum_{j=1}^{M} \left| \int_X \varphi_{y_j'}(x) \mathrm{d}\hat{\mu}[\vec{x}](x) - \int_X \varphi_{y_j'}(y) \mathrm{d}\hat{\mu}[\vec{y}](y) \right| \\
&\le L \frac{1}{M} \sum_{i=1}^{M} \sup_{\substack{f:X\to\mathbb{R} \\ f \text{ 1-Lipschitz}}} \left| \int_X f(x') \mathrm{d}\hat{\mu}[\vec{x}'](x') - \int_X f(y') \mathrm{d}\hat{\mu}[\vec{y}'](y') \right| \\
&\quad + L \frac{1}{M} \sum_{j=1}^{M} \sup_{\substack{f:X\to\mathbb{R} \\ f \text{ 1-Lipschitz}}} \left| \int_X f(x) \mathrm{d}\hat{\mu}[\vec{x}](x) - \int_X f(y) \mathrm{d}\hat{\mu}[\vec{y}](y) \right| \\
&= L \frac{1}{M} \sum_{i=1}^{M} d_{\mathrm{KR}}(\hat{\mu}[\vec{x}'], \hat{\mu}[\vec{y}']) + L \frac{1}{M} \sum_{j=1}^{M} d_{\mathrm{KR}}(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{y}]) \\
&= L(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}'], \hat{\mu}[\vec{y}']) + d_{\mathrm{KR}}(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{y}])) \\
&= L d_{\mathrm{KR}}^2 \left((\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']), (\hat{\mu}[\vec{y}], \hat{\mu}[\vec{y}']) \right),
\end{aligned}$$

establishing the claim.

Cite this article: Fiedler C., Herty M., Segala C. and Trimpe S. Recent kernel methods for interacting particle systems: first numerical results. *European Journal of Applied Mathematics*, https://doi.org/10.1017/S0956792524000706