Inferring Interaction Rules from Observations of Evolutive Systems I: The Variational Approach

M. Bongini, M. Fornasier, M. Hansen, and M. Maggioni

Abstract

In this paper we are concerned with the learnability of nonlocal interaction kernels for first order systems modeling certain social interactions, from observations of realizations of the dynamics. This paper is the first of a series on learnability of nonlocal interaction kernels and presents a variational approach to the problem. In particular, we assume here that the kernel to be learned is bounded and locally Lipschitz continuous and the initial conditions of the systems are drawn identically and independently at random according to a given initial probability distribution. Then the minimization over a rather arbitrary sequence of (finite dimensional) subspaces of a least square functional measuring the discrepancy from observed trajectories produces uniform approximations to the kernel on compact sets. The convergence result is obtained by combining mean-field limits, transport methods, and a Γ-convergence argument. A crucial condition for the learnability is a certain coercivity property of the least square functional, majoring an L_2-norm discrepancy to the kernel with respect to a probability measure, depending on the given initial probability distribution by suitable push forwards and transport maps. We illustrate the convergence result by a few numerical experiments.

Keywords: interaction kernel learning, first order nonlocal systems, mean-field equations, Γ-convergence

1 Introduction

What are the instinctive individual reactions which make a group of animals forming coordinated movements, for instance a flock of migrating birds or a school of fish? Which biological interactions between cells produce the formation of complex structures, for instance organs? What are the mechanisms which induce certain significant changes in a large amount of players in the financial market? In this paper we are concerned with the "mathematization" of the problem of learning or inferring interaction rules from observations of evolutions. The framework we consider is the one of evolutions driven by gradient descents. The study of gradient flow evolutions to minimize certain energetic landscapes has been the subject of intensive research in the past years [2]. Some of the most recent models are aiming at describing time-dependent phenomena also in biology or even in social dynamics, borrowing a leaf from more established and classical models

in physics. For instance, starting with the seminal papers of Vicsek et. al. [20] and Cucker-Smale [10], there has been a flood of models describing consensus or opinion formation, modeling the exchange of information as long-range social interactions (forces) between active agents (particles). However, for the analysis, but even more crucially for the reliable and realistic numerical simulation of such phenomena, one presupposes a complete understanding and determination of the governing energies. Unfortunately, except for physical situations where the calibration of the model can be done by measuring the governing forces rather precisely, for some relevant macroscopical models in physics and most of the models in biology and social sciences the governing energies are far from being precisely determined. In fact, very often in these studies the governing energies are just predetermined to be able to reproduce, at least approximately or qualitatively, some of the macroscopical effects of the observed dynamics, such as the formation of certain patterns, but there has been little or no effort of matching data from real-life cases.

This attitude aiming just at a qualitative description tends however to reduce some of the investigations in this area to beautiful and mathematically interesting toy-cases, which have likely little to do with real-life scenarios. The aim of this paper is providing a mathematical framework for the reliable identification of the governing energies from data obtained by direct observations of corresponding time-dependent evolutions. This is a new kind of inverse problem, beyond more traditionally considered ones, as the forward map is a strongly nonlinear evolution, highly dependent on the probability measure generating the initial conditions. As we aim at a precise quantitative analysis, and to be very concrete, we will attack the learning of the energies for specific models in social dynamics governed by nonlocal interactions.

1.1 General abstract framework

Many time-dependent phenomena in physics, biology, and social sciences can be modeled by a function $x : [0,T] \to \mathcal{H}$, where $\mathcal{H}$ represents the space of states of the physical, biological or social system, which evolves from an initial configuration $x(0) = x_0$ towards a more convenient state or a new equilibrium. The space $\mathcal{H}$ can be a conveniently chosen Banach space or just a metric space; let $\mathrm{dist}_{\mathcal{H}}$ be the metric on $\mathcal{H}$. This implicitly assumes that x evolves driven by a minimization process of a potential energy $\mathcal{J} : \mathcal{H} \times [0,T] \to \mathbb{R}$. In this preliminary introduction we consciously avoid specific assumptions on $\mathcal{J}$, as we wish to keep a rather general view. We restrict the presentation to particular cases below.

Inspired by physics, for which conservative forces are the derivatives of the potential energies, one can describe the evolution as satisfying a gradient flow inclusion of the type

$$\dot{x}(t) \in -\partial_x \mathcal{J}(x(t), t), \tag{1}$$

where $\partial_x \mathcal{J}(x,t)$ is some notion of differential of $\mathcal{J}$ with respect to x, which might already take into consideration additional constraints which are binding the states to certain sets.

1.2 Example of gradient flow of nonlocal particle interactions

Assume that $x = (x_1, \dots, x_N) \in \mathcal{H} = \mathbb{R}^{d\times N}$ and that

$$\mathcal{J}_N(x) = \frac{1}{2N} \sum_{i,j=1}^{N} A(|x_i^{[a]} - x_j^{[a]}|),$$

where $A : \mathbb{R}_+ \to \mathbb{R}$ is a suitable nonlinear interaction kernel function, which, for simplicity we assume to be smooth, and $|\cdot|$ is the Euclidean norm in $\mathbb{R}^d$. Then, the formal unconstrained gradient flow (1) associated to this energy is written coordinatewise as

$$\dot{x}_i^{[a]}(t) = \frac{1}{N} \sum_{j\neq i} \frac{A'(|x_i^{[a]} - x_j^{[a]}|)}{|x_i^{[a]} - x_j^{[a]}|}(x_j^{[a]} - x_i^{[a]}), \quad i = 1, \dots, N. \tag{2}$$

Under suitable assumptions of local Lipschitz continuity and boundedness of

$$a(\cdot) := \frac{A'(|\cdot|)}{|\cdot|}, \tag{3}$$

this evolution is well-posed for any given $x(0) = x_0$ and it is expected to converge for $t \to \infty$ to configurations of the points whose mutual distances are close to local minimizers of the function A, representing steady states of the evolution as well as critical points of $\mathcal{J}_N$.
It is also well-known [2] (see also Proposition 2.2 below) that for $N \to \infty$ a mean-field approximation holds: if the initial conditions $x_i^{[a]}(0)$ are i.i.d. according to a compactly supported probability measure $\mu^0 \in \mathcal{P}_c(\mathbb{R}^d)$ for $i = 1, 2, 3, \dots$, the empirical measure $\mu_N(t) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i^{[a]}(t)}$ weakly converges for $N \to \infty$ to the probability measure-valued trajectory $t \to \mu(t)$ satisfying weakly the equation

$$\partial_t \mu(t) = -\nabla \cdot ((F^{[a]} * \mu(t))\mu(t)), \quad \mu(0) = \mu^0. \tag{4}$$

where $F^{[a]}(z) = -a(|z|)z = -A'(|z|)$, for $z \in \mathbb{R}^d$. In fact the differential equation (4) corresponds again to a gradient flow of the "energy"

$$\mathcal{J}(\mu) = \int_{\mathbb{R}^{d\times d}} A(|x - y|)d\mu(x)d\mu(y),$$

on the metric space $\mathcal{H} = \mathcal{P}_c(\mathbb{R}^d)$ endowed with the so-called Wasserstein distance. Continuity equations of the type (4) with nonlocal interaction kernels are currently the subject of intensive research towards the modeling of the biological and social behavior of microorganisms, animals, humans, etc. We refer to the articles [9, 8] for recent overviews on this subject. Despite the tremendous theoretical success of such research direction in terms of mathematical results, as we shall stress below in more detail, one of the issues which is so far scarcely addressed in the study of models of the type (2) or (4) is their actual applicability. Most of the results are addressing a purely *qualitative analysis* given

certain smoothness and asymptotic properties of the kernels A or a at the origin or at infinity, in terms of well-posedness or in terms of asymptotic behavior of the solution for $t \to \infty$. Certainly such results are of great importance, as such interaction functions, if ever they can really describe social dynamics, are likely to differ significantly from well-known models from physics and it is reasonable and legitimate to consider a large variety of classes of such functions. However, a solid mathematical framework which establishes the conditions of "learnability" of the interaction kernels from observations of the dynamics is currently not available and it will be the main subject of this paper.

1.3 Parametric energies and their identifications

Let us now consider an energy $\mathcal{J}[a]$ depending on a parameter function a. As in the example mentioned above, a may be defining a nonlocal interaction kernel as in (3). The parameter function a not only determines the energy, but also the corresponding evolutions driven according to (1), for fixed initial conditions $x(0) = x_0$. (Here we assume that the class of a is such that the evolutions exist and they are essentially well-posed.) The fundamental question to be addressed is: can we recover a with high accuracy given some observations of the realized evolutions? This question is prone to several specifications, for instance, we may want to assume that the initial conditions are generated according to a certain probability distribution or they are chosen deterministically ad hoc to determine at best a, that the observations are complete or incomplete, etc. As one quickly realizes, this is a very broad field to explore with many possible developments. Surprisingly, there are no results in this direction at this level of generality, and very little is done in the specific directions we mentioned in the example above. We refer for instance to [18, 15] for studies on the inference of social rules in collective behavior.

1.4 The optimal control approach and its drawbacks

Let us introduce an approach, which would perhaps naturally be considered at a first instance and focus for a moment on the gradient flow model (1). Given a certain gradient flow evolution $t \to x[a](t)$ depending on the unknown parameter function a, one might decide to design the recovery of a as an optimal control problem [6]: for instance, we may seek a parameter function $\widehat{a}$ which minimizes

$$\mathcal{E}^{[a]}(\widehat{a}) = \frac{1}{T}\int_0^T \left[\mathrm{dist}_{\mathcal{H}}(x^{[a]}(s) - x^{[\widehat{a}]}(s))^2 + \mathcal{R}(\widehat{a})\right] ds, \tag{5}$$

being $t \to x[\widehat{a}](t)$ the solution of gradient flow (1) for $\mathcal{J} = \mathcal{J}[\widehat{a}]$, i.e.,

$$\dot{x}^{[\widehat{a}]}(t) \in -\partial_x J(x[\widehat{a}](t), t), \tag{6}$$

and $\mathcal{R}(\cdot)$ is a suitable regularization function, which restricts the possible minimizers of (5) to a specific class. The first fundamental problem one immediately encounters with this formulation is the strongly nonlinear dependency of $t \to x^{[\widehat{a}]}(t)$ on $\widehat{a}$, which results in a strong non-convexity of the functional (5). This also implies that a direct

minimization of (5) would risk to lead to suboptimal solutions, and even the computation of a first order optimality condition in terms of Pontryagin's minimum principle would not characterize uniquely the minimal solutions. Besides this fundamental hurdles, the numerical implementation of either strategy (direct optimization or solution of the first order optimality conditions) is expected to be computationally unfeasible to reasonable degree of accuracy as soon as the underlying discretization dimension grows.

1.5 A variational approach towards learning parameter functions in nonlocal energies

Let us consider the framework of the example in Section 1.2. We restrict our attention to interaction kernels a belonging to the following *set of admissible kernels*

$$X = \big\{ b : \mathbb{R}_+ \to \mathbb{R} \,|\, b \in L_\infty(\mathbb{R}_+) \cap W^1_{\infty,\mathrm{loc}}(\mathbb{R}_+) \big\}.$$

In particular every $a \in X$ is weakly differentiable, and its local Lipschitz constant $\mathrm{Lip}_K(a)$ is finite for every compact set $K \subset \mathbb{R}_+$. Our goal is to learn the unknown influence function $a \in X$ from the observation of the dynamics of the empirical measure μ_N, defined by $\mu_N(t) = \frac{1}{N}\sum_{i=1}^N \delta_{x_i^{[a]}(t)}$, where $x_i^{[a]}(t)$ are driven by the interaction kernel a according to the equations of motion

$$\dot{x}_i^{[a]}(t) = \frac{1}{N}\sum_{j\neq i} a(|x_i^{[a]} - x_j^{[a]}|)(x_j^{[a]} - x_i^{[a]}), \quad i = 1, \dots, N. \tag{7}$$

Instead of the nonlinear optimal control problem above, we propose an alternative, direct approach which is both computationally very efficient and produces accurate approximations under reasonable assumptions. We consider a minimizer of the following *discrete error functional*

$$\mathcal{E}^{[a],N}(\widehat{a}) = \frac{1}{T}\int_0^T \frac{1}{N}\sum_{i=1}^N \left| \frac{1}{N}\sum_{j=1}^N \Big(\widehat{a}(|x_i^{[a]}(t) - x_j^{[a]}(t)|)(x_i^{[a]}(t) - x_j^{[a]}(t)) - \dot{x}_i^{[a]}(t) \Big) \right|^2 dt, \tag{8}$$

among all functions $\widehat{a} \in X$. The minimization of $\mathcal{E}^{[a],N}(\widehat{a})$ has a close connection to the optimal control problem (5):

Proposition 1.1. *If $a, \widehat{a} \in X$ then there exist a constant $C > 0$ depending on $T, \widehat{a}$ and μ_0^N and a compact set $K \subset \mathbb{R}_+$ such that*

$$\|x^{[a]}(t) - x^{[\widehat{a}]}(t)\|^2 \leq C\mathcal{E}^{[a],N}(\widehat{a}), \tag{9}$$

for all $t \in [0,T]$, and $x^{[a]}, x^{[\widehat{a}]}$ are the solutions to (22) for the interaction kernels a and $\widehat{a}$ respectively. (Here $\|x\| = \frac{1}{N}\sum_{i=1}^N |x_i|$ $\|x\|^2 = \frac{1}{N}\sum_{i=1}^N |x_i|^2$?, for $x \in \mathbb{R}^{d\times N}$.)

Therefore if $\widehat{a}$ makes $\mathcal{E}^{[a],N}(\widehat{a})$ small, the trajectories $t \to x^{[\widehat{a}]}(t)$ of the system (7) with interaction kernel $\widehat{a}$ instead of a are an approximation of the trajectories $t \to x^{[a]}(t)$ at finite time. Moreover, contrary to the optimal control approach, the functional $\mathcal{E}^{[a],N}$ can be easily computed from witnessed trajectories $x_i^{[a]}(t)$ and $\dot{x}_i^{[a]}(t)$. We may consider discrete-time approximations the time derivative $\dot{x}_i^{[a]}$ (e.g. by finite differences) and we shall assume that the data of the problem is the full set of observations $x_i^{[a]}(t)$ for $t \in [0,T]$, for a prescribed finite time horizon $T > 0$. Furthermore, being a simple quadratic functional, its minimizers can be efficiently numerically approximated on a finite element space: given a finite dimensional space $V \subset X$, we let

$$\widehat{a}_{N,V} = \arg\min_{\widehat{a} \in V} \mathcal{E}^{[a],N}(\widehat{a}). \tag{10}$$

The fundamental question to be addressed in this paper is

(Q) For which choice of the approximating spaces $V \in \Lambda$ (we assume here that Λ is a countable family of invading subspaces of X) does $\widehat{a}_{N,V} \to a$ for $N \to \infty$ and $V \to X$ and in which topology should convergence hold?

We show now how we address this issue in detail by a variational approach, seeking a limit functional for which techniques of Γ-convergence [11], whose general aim is establishing the convergence of minimizers for a sequence of equi-coercive functionals to minimizers of a target functional, may provide a limit for the sequence of minimizers $(\widehat{a}_{N,V})_{N \in \mathbb{N}, V \in \Lambda}$. Letting $F^{[a]}(z) = -a(|z|)z$, for $z \in \mathbb{R}^d$, we rewrite the functional (8) as follows:

$$\begin{aligned} \mathcal{E}^{[a],N}(\widehat{a}) &= \frac{1}{T}\int_0^T \frac{1}{N}\sum_{i=1}^N \left| \frac{1}{N}\sum_{j=1}^N \left(F^{[\widehat{a}]} - F^{[a]}\right)(x_i^{[a]} - x_j^{[a]}) \right|^2 dt \\ &= \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \left| \left(F^{[\widehat{a}]} - F^{[a]}\right) * \mu_N(t) \right|^2 d\mu_N(t)(x) dt, \end{aligned} \tag{11}$$

the notation here is getting a bit out of hand, with $d\mu_N(t)(x)$... The candidate for a Γ-limit functional is then

$$\mathcal{E}^{[a]}(\widehat{a}) = \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \left| \left(F^{[\widehat{a}]} - F^{[a]}\right) * \mu(t) \right|^2 d\mu(t)(x) dt, \tag{12}$$

where μ is a weak solution to the mean-field equation (2), as soon as the initial conditions $x_i^{[a]}(0)$ are identically and independently distributed according to a compactly supported probability measure $\mu(0) = \mu^0$.

Several issues need to be addressed at this point. The first one is to establish the space where a result of Γ-convergence may hold. We expect that such a space may *not* be independent of the initial probability measure μ^0. In fact, by Jensen inequality we have

$$\mathcal{E}^{[a]}(\widehat{a}) \qquad \leq \qquad \frac{1}{T}\int_0^T \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |\widehat{a}(|x-y|) - a(|x-y|)|^2 |x-y|^2 d\mu(t)(x) d\mu(t)(y) dt \tag{13}$$

$$\leq_{\text{is this an =?}} \quad \frac{1}{T}\int_0^T \int_{\mathbb{R}_+} \left|\widehat{a}(s) - a(s)\right|^2 s^2 d\varrho(t)(s)dt \tag{14}$$

where $\varrho(t)$ is the pushforward of $\mu(t)\otimes\mu(t)$ by the Euclidean distance map $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ defined by $(x, y) \mapsto d(x, y) = |x - y|$. In other words, $\varrho : [0, T] \to \mathcal{P}_1(\mathbb{R}_+)$ is defined for every Borel set $A \subset \mathbb{R}_+$ as $\varrho(t)(A) = (\mu(t) \otimes \mu(t))\big(d^{-1}(A)\big)$. The mapping $t \in [0, T] \mapsto \varrho(t)(A)$ is lower semi-continuous for every open set $A \subseteq \mathbb{R}_+$, and it is upper semi-continuous (see Lemma 3.1) for any compact set A. We may therefore define a probability measure $\overline{\rho}$ on the Borel σ-algebra on $\mathbb{R}_+$: for any open set $A \subseteq \mathbb{R}_+$ we define

$$\overline{\rho}(A) := \frac{1}{T}\int_0^T \varrho(t)(A)dt, \tag{15}$$

and extend this set function to a probability measure on all Borel sets. Finally we define

$$\rho(A) := \int_A s^2 d\overline{\rho}(s), \tag{16}$$

for all $A \subseteq \mathbb{R}_+$ Borel sets. Then one can reformulate (14) as follows

$$\mathcal{E}^{[a]}(\widehat{a}) \leq \int_{\mathbb{R}_+} \left|\widehat{a}(s) - a(s)\right|^2 d\rho(s) = \|\widehat{a} - a\|^2_{L_2(\mathbb{R}_+,\rho)}. \tag{17}$$

Notice that ρ is defined through $\mu(t)$ which depends on the initial probability measure μ^0.

To establish coercivity of the learning problem it is natural to assume that there exists $c_T > 0$ such that the following bound holds

$$c_T\|\widehat{a} - a\|^2_{L_2(\mathbb{R}_+,\rho)} \leq \mathcal{E}^{[a]}(\widehat{a}), \tag{18}$$

for all relevant $\widehat{a} \in X \cap L_2(\mathbb{R}_+, \rho)$. This crucial assumption eventually determines also the natural space $X \cap L_2(\mathbb{R}_+, \rho)$ for the solutions, which therefore depends on the choice of the initial conditions μ^0. In particular the constant $c_T \geq 0$ might not be nondegenerate for all the choices of μ^0 and one has to pick the initial distribution so that (18) can hold for $c_T > 0$. In Section 3.2 we show that for some specific choices of a and rather general choices of $\widehat{a} \in X$ one can construct probability measure valued trajectories $t \to \mu(t)$ which allow to validate (18).

We now introduce the key property that a family of approximation spaces V_N must possess in order to ensure that the minimizers of the functionals $\mathcal{E}_N$ converge to minimizers of $\mathcal{E}$ by Γ-convergence.

Definition 1.2. Let $M > 0$ and $K = [0, 2R]$ interval in $\mathbb{R}_+$ be given . We say that a family of closed subsets $V_N \subset X_{M,K}$, $N \in \mathbb{N}$ has the *uniform approximation property* in $L_\infty(K)$ if for all $b \in X_{M,K}$ there exists a sequence $(b_N)_{N\in\mathbb{N}}$ converging uniformly to b on K and such that $b_N \in V_N$ for every $N \in \mathbb{N}$.

We are ready to state the main result of the paper:

Theorem 1.3. *Assume $a \in X$, fix $\mu^0 \in \mathcal{P}_c(\mathbb{R}^d)$ and let $K = [0, 2R]$ be an interval in $\mathbb{R}_+$ with $R > 0$ as in Proposition 2.2. Set*

$$M \geq \|a\|_{L_\infty(K)} + \|a'\|_{L_\infty(K)}.$$

For every $N \in \mathbb{N}$, let $x_{0,1}^N, \dots, x_{0,N}^N$ be i.i. μ^0-distributed and define $\mathcal{E}_N$ as in (30) for the solution μ_N of system (21) with initial datum

$$\mu_N^0 = \frac{1}{N} \sum_{i=1}^N \delta_{x_{0,i}^N}, \quad x \in \mathbb{R}^d.$$

For $N \in \mathbb{N}$, let $V_N \subset X_{M,K}$ be a sequence of subsets with the uniform approximation property as in Definition 1.2 and consider

$$\widehat{a}_N \in \arg\min_{\widehat{a} \in V_N} \mathcal{E}^{[a],N}(\widehat{a}).$$

Then the sequence $(\widehat{a}_N)_{N \in \mathbb{N}}$ converges uniformly on K to some continuous function $\widehat{a} \in X_{M,K}$ such that $\mathcal{E}(\widehat{a}) = 0$. MM: I have to check: isn't this to subsequences???

If we additionally assume the coercivity condition (18), then it holds $\widehat{a} = a$ in $L_2(\mathbb{R}_+, \rho)$.

1.6 Numerical implementation of the variational approach

move some of this above, and have more about numerics here

The strength of the result from the variational approach followed in Section 1.5 is the total arbitrariness of the sequence V_N except for the assumed *uniform approximation property* and that the result holds - deterministically - with respect to uniform convergence, which is quite strong. The condition that the spaces V_N are to be picked as subsets of $X_{M,K}$ requires knowledge of $M \geq \|a\|_{L_\infty(K)} + \|a'\|_{L_\infty(K)}$. While this may be reasonable in some applications, when this information is not available the finite dimensional optimization (10) is not anymore a simple *unconstrained* least squares (as claimed in (10)), but a problem constrained by a uniform bound on both the solution and its gradient. A possible way to circumvent this problem is to choose M very large for N moderately small and then tune it down to the "right" level adaptively, for N growing. Or perhaps one may implement efficiently such a numerical optimization (10) with L_∞ constraints. We address these aspects in Section 5 with several numerical experiments. We shall see that a careful choice of the function spaces V_N enables us to implement the minimization problem (10) as a constrained L_2 problem, for which a plethora of very fast and efficient numerical schemes exist. We show that, as expected, if we let N grow, the minimizers $\widehat{a}_N$ approximates better and better the unknown potential a. Moreover, we present a very effective strategy for tuning properly the parameter M.

We conclude this introduction by mentioning that the variational approach is based on a compactness argument and, as a consequence, it does not provide any rate of convergence. This is another significant drawback of this technique. In our follow-up

paper [5] we are following the approach developed by DeVore et al. in [4, 3] towards universal algorithms for learning regression functions from independent samples drawn according to an unknown probability distribution. Then with high probability, and for suitable choices of approximating spaces V_N, we obtain that for every $\beta > 0$

$$\mathcal{P}_{\mu^0}\Big(\|a - \widehat{a}_{N,V_N}\|_{L_2(\mathbb{R}_+,\rho)} > (c_3\|a\|_\infty + |a|_{\mathcal{A}^s_\mu})\big(\tfrac{\log N}{N^3}\big)^{\frac{s}{2s+1}}\Big) \le c_4 N^{-\beta}\,, \tag{19}$$

if c_3 is chosen sufficiently large (depending on β), where $\mathcal{A}^s_\mu$ is a suitable functional class indicating how efficiently a can be approximated by piecewise polynomial functions in $L_2(\mathbb{R}_+, \rho)$.

2 Preliminaries

The space $\mathcal{P}(\mathbb{R}^n)$ is the set of probability measures on $\mathbb{R}^n$, while the space $\mathcal{P}_p(\mathbb{R}^n)$ is the subset of $\mathcal{P}(\mathbb{R}^n)$ whose elements have finite p-th moment, i.e., $\int_{\mathbb{R}^n} |x|^p d\mu(x) < +\infty$. We denote by $\mathcal{P}_c(\mathbb{R}^n)$ the subset of $\mathcal{P}_1(\mathbb{R}^n)$ which consists of all probability measures with compact support. For any $\mu \in \mathcal{P}(\mathbb{R}^{n_1})$ and Borel function $f : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$, we denote by $f_\#\mu \in \mathcal{P}(\mathbb{R}^{n_2})$ the *push-forward of μ through f*, defined by

$$f_\#\mu(B) := \mu(f^{-1}(B)) \quad \text{for every Borel set } B \text{ of } \mathbb{R}^{n_2}.$$

In particular, if one considers the projection operators π_1 and π_2 defined on the product space $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, for every $\rho \in \mathcal{P}(\mathbb{R}^{n_1} \times \mathbb{R}^{n_2})$ we call *first* (resp., *second*) *marginal* of ρ the probability measure $\pi_{1\#}\rho$ (resp., $\pi_{2\#}\rho$). Given $\mu \in \mathcal{P}(\mathbb{R}^{n_1})$ and $\nu \in \mathcal{P}(\mathbb{R}^{n_2})$, we denote with $\Gamma(\mu, \nu)$ the family of couplings between μ and ν, i.e. the subset of all probability measures in $\mathcal{P}(\mathbb{R}^{n_1} \times \mathbb{R}^{n_2})$ with first marginal μ and second marginal ν.

On the set $\mathcal{P}_p(\mathbb{R}^n)$ we shall consider the following distance, called the Wasserstein or Monge-Kantorovich-Rubinstein distance,

$$\mathcal{W}_p^p(\mu, \nu) = \inf_{\rho \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^{2n}} |x - y|^p d\rho(x, y)\,. \tag{20}$$

If $p = 1$, we have the following equivalent expression for the Wasserstein distance:

$$\mathcal{W}_1(\mu, \nu) = \sup\left\{\int_{\mathbb{R}^n} \varphi(x) d(\mu - \nu)(x) : \varphi \in \mathrm{Lip}(\mathbb{R}^n),\ \mathrm{Lip}_{\mathbb{R}^n}(\varphi) \le 1\right\},$$

where $\mathrm{Lip}_{\mathbb{R}^n}(\varphi)$ stands for the Lipschitz constant of φ on $\mathbb{R}^n$. We denote by $\Gamma_o(\mu, \nu)$ the set of optimal couplings for which the minimum is attained, i.e.,

$$\rho \in \Gamma_o(\mu, \nu) \iff \rho \in \Gamma(\mu, \nu) \text{ and } \int_{\mathbb{R}^{2n}} |x - y|^p d\rho(x, y) = \mathcal{W}_p^p(\mu, \nu).$$

It is well-known that $\Gamma_o(\mu, \nu)$ is non-empty for every $(\mu, \nu) \in \mathcal{P}_p(\mathbb{R}^n) \times \mathcal{P}_p(\mathbb{R}^n)$, hence the infimum in (20) is actually a minimum. For more details, see e.g. [2, 21].

For any $\mu \in \mathcal{P}_1(\mathbb{R}^d)$ and $f : \mathbb{R}^d \to \mathbb{R}^d$, the notation $f * \mu$ stands for the convolution of f and μ:

$$(f * \mu)(x) = \int_{\mathbb{R}^d} f(x-y) d\mu(y)\,.$$

This function is continuous and finite-valued whenever f is continuous and *sublinear*, i.e., there exists a constant $C > 0$ such that $|f(\xi)| \leq C(1 + |\xi|)$ for all $\xi \in \mathbb{R}^d$.

2.1 The mean-field limit equation and existence of solutions

As already stated in the introduction, our learning approach is based on the following underlying *finite time horizon initial value problem*: given $T > 0$ and $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$, consider a probability measure valued trajectory $\mu : [0, T] \to \mathcal{P}_1(\mathbb{R}^d)$ satisfying

$$\begin{cases} \dfrac{\partial \mu}{\partial t}(t) = -\nabla \cdot ((F^{[a]} * \mu(t))\mu(t)) \quad \text{for } t \in (0, T], \\ \mu(0) = \mu_0. \end{cases} \tag{21}$$

We consequently give our notion of solution for (21).

Definition 2.1. We say that a map $\mu : [0, T] \to \mathcal{P}_1(\mathbb{R}^d)$ is a solution of (21) with initial datum μ_0 if the following hold:

1. μ has uniformly compact support, i.e., there exists $R > 0$ such that $\mathrm{supp}(\mu(t)) \subset B(0, R)$ for every $t \in [0, T]$;
2. μ is continuous with respect to the Wasserstein distance $\mathcal{W}_1$;
3. μ satisfies (21) in the weak sense, i.e., for every $\phi \in \mathcal{C}_c^\infty(\mathbb{R}^d; \mathbb{R})$ it holds

$$\frac{d}{dt} \int_{\mathbb{R}^d} \phi(x) d\mu(t)(x) = \int_{\mathbb{R}^d} \nabla \phi(x) \cdot (F^{[a]} * \mu(t))(x) d\mu(t)(x).$$

The system (21) is closely related to the family of ODE's, indexed by $N \in \mathbb{N}$,

$$\begin{cases} \dot{x}_i^N(t) = \dfrac{1}{N} \displaystyle\sum_{j=1}^N F^{[a]}(x_i^N(t) - x_j^N(t)) \quad \text{for } t \in (0, T], \\ x_i^N(0) = x_{0,i}^N, \end{cases} \quad i = 1, \ldots, N, \tag{22}$$

which may be rewritten as

$$\begin{cases} \dot{x}_i^N(t) = (F^{[a]} * \mu^N(t))(x_i^N(t)) \\ x_i^N(0) = x_{0,i}^N, \end{cases} \quad i = 1, \ldots, N, \tag{23}$$

for $t \in (0,T]$, by means of the *empirical measure* $\mu^N : [0,T] \to \mathcal{P}_c(\mathbb{R}^d)$ defined as

$$\mu^N(t) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i^N(t)}. \tag{24}$$

As already explained in the introduction, we shall restrict our attention to interaction kernels belonging to the following *set of admissible kernels*

$$X = \big\{ b : \mathbb{R}_+ \to \mathbb{R} \,|\, b \in L_\infty(\mathbb{R}_+) \cap W^1_{\infty,\mathrm{loc}}(\mathbb{R}_+) \big\}.$$

The well-posedness of (23) is rather standard under the assumption $a \in X$. The well-posedness of system (21) and several crucial properties enjoyed by its solutions may also be proved as soon as $a \in X$. We refer the reader to [2] for results on existence and uniqueness of solutions for (21), and to [8] for generalizations in case of interaction kernels not belonging to the class X. In the following we report the main results, whose proofs are collected in the Appendices in order to keep this work self-contained and to allow explicit reference to constants.

Proposition 2.2. *Let $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$ be given. Let $(\mu_0^N)_{N\in\mathbb{N}} \subset \mathcal{P}_c(\mathbb{R}^d)$ be a sequence of empirical measures of the form*

$$\mu_0^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_{0,i}^N}, \qquad \text{for some } x_{0,i}^N \in \mathrm{supp}(\mu_0) + \overline{B(0,1)}$$

satisfying $\lim_{N\to\infty} \mathcal{W}_1(\mu_0, \mu_0^N) = 0$. For every $N \in \mathbb{N}$, denote with $\mu^N : [0,T] \to \mathcal{P}_1(\mathbb{R}^d)$ the curve given by (24) where $(x_1^N, \dots, x_N^N)$ is the unique solution of system (22).

Then, there exists $R > 0$ depending only on $T, a,$ and $\mathrm{supp}(\mu_0)$ such that the sequence $(\mu^N)_{N\in\mathbb{N}}$ converges in $\mathcal{P}_1(B(0,R))$ equipped with the Wasserstein metric $\mathcal{W}_1$, up to subsequences, to a solution μ of (21) with initial datum μ_0 satisfying

$$\mathrm{supp}(\mu^N(t)) \cup \mathrm{supp}(\mu(t)) \subseteq B(0,R), \qquad \text{for every } N \in \mathbb{N} \text{ and } t \in [0,T].$$

A proof of this standard result is reported in Appendix 6.2 together with the necessary technical lemmas in Appendix 6.1.

2.2 The transport map and uniqueness of mean-field solutions

Another way for building a solution of equation (21) is by means of the so-called *transport map*, i.e., the function describing the evolution in time of the initial measure μ_0. The transport map can be constructed by considering the following single-agent version of system (23),

$$\begin{cases} \dot\xi(t) = (F^{[a]} * \mu(t))(\xi(t)) & \text{for } t \in (0,T], \\ \xi(0) = \xi_0, \end{cases} \tag{25}$$

where ξ is a mapping from $[0,T]$ to $\mathbb{R}^d$ and $a \in X$. Here $\mu : [0,T] \to \mathcal{P}_1(\mathbb{R}^d)$ is a continuous map with respect to the Wasserstein distance $\mathcal{W}_1$ satisfying $\mu(0) = \mu_0$ and $\mathrm{supp}(\mu(t)) \subseteq B(0,R)$, where R is given by (47) from the choice of T, a and μ_0.

By Theorem 6.7 and Lemma 6.8, we can consider the family of flow maps $\mathcal{T}_t^\mu : \mathbb{R}^d \to \mathbb{R}^d$, indexed by $t \in [0,T]$ and the choice of the mapping μ, defined by

$$\mathcal{T}_t^\mu(\xi_0) = \xi(t),$$

where $\xi : [0,T] \to \mathbb{R}^d$ is the unique solution of (25) with initial datum ξ_0. The by now well-known result [7, Theorem 3.10] shows that the solution of (21) with initial value μ_0 is the unique fixed-point of the *push-foward map*

$$\Gamma[\mu](t) := (\mathcal{T}_t^\mu)_\# \mu_0. \tag{26}$$

A relevant, basic property of the transport map is proved in the following

Proposition 2.3. *$\mathcal{T}_t^\mu$ is a locally bi-Lipschitz map, i.e. it is a locally Lipschitz map, with locally Lipschitz inverse.*

Proof. The choice $r = R$ in Lemma 6.8 and the inequality (54) trivially implies the following stability estimate

$$\left|\mathcal{T}_t^\mu(x_0) - \mathcal{T}_t^\mu(x_1)\right| \le e^{T\,\mathrm{Lip}_{B(0,R)}(F^{[a]})}|x_0 - x_1|, \quad \text{for } |x_i| \le R\,, \quad i = 0,1\,. \tag{27}$$

i.e., $\mathcal{T}_t^\mu$ is locally Lipschitz.

In view of the uniqueness of the solutions to the ODE (25), it is also clear that, for any $t_0 \in [0,T]$, the inverse of $\mathcal{T}_{t_0}^\mu$ is given by the transport map associated to the backward-in-time ODE

$$\begin{cases} \dot\xi(t) = (F^{[a]} * \mu(t))(\xi(t)) & \text{for } t \in [0,t_0), \\ \xi(t_0) = \xi_0. \end{cases}$$

However, this problem in turn can be cast into the form of an usual IVP simply by considering the reverse trajectory $\nu_t = \mu_{t_0 - t}$. Then $y(t) = \xi(t_0 - t)$ solves

$$\begin{cases} \dot y(t) = -\big(F^{[a]} * \nu(t)\big)(y(t)) & \text{for } t \in (0,t_0], \\ y(0) = \xi(t_0). \end{cases}$$

The corresponding stability estimate for this problem then yields that the inverse of $\mathcal{T}_t^\mu$ exists and is locally Lipschitz (with the same local Lipschitz constant as $\mathcal{T}_t^\mu$). □

It is also known [7] that we have uniqueness and continuous dependence on the initial data for (21) (we report a proof of it in Appendix 6.4 for completeness):

Theorem 2.4. *Fix $T > 0$ and let $\mu : [0,T] \to \mathcal{P}_1(\mathbb{R}^d)$ and $\nu : [0,T] \to \mathcal{P}_1(\mathbb{R}^d)$ be two equi-compactly supported solutions of (21), for $\mu(0) = \mu_0$ and $\nu(0) = \nu_0$ respectively. Consider $R > 0$ such that*

$$\operatorname{supp}(\mu(t)) \cup \operatorname{supp}(\nu(t)) \subseteq B(0,R) \tag{28}$$

for every $t \in [0,T]$. Then, there exist a positive constant $\overline{C}$ depending only on T, a, and R such that

$$\mathcal{W}_1(\mu(t),\nu(t)) \leq \overline{C}\, \mathcal{W}_1(\mu_0,\nu_0) \tag{29}$$

for every $t \in [0,T]$. In particular, equi-compactly supported solutions of (21) are uniquely determined by the initial datum.

3 The learning problem for the kernel function

As already explained in the introduction, our goal is to learn $a \in X$ from observation of the dynamics of μ^N corresponding to system (22) with a as interaction kernel, μ_0^N as initial datum and T as finite time horizon.

We pick $\widehat{a}$ among those functions in X which would give rise to a dynamics close to μ^N: we choose $\widehat{a} \in X$ as a minimizer of the following *discrete error functional*

$$\mathcal{E}^{[a],N}(\widehat{a}) = \frac{1}{T}\int_0^T \frac{1}{N}\sum_{i=1}^N \left| \frac{1}{N}\sum_{j=1}^N \left(\widehat{a}(|x_i^N(t) - x_j^N(t)|)(x_i^N(t) - x_j^N(t)) - \dot{x}_i^N(t)\right)\right|^2 dt. \tag{30}$$

Proof of Proposition 1.1. Let us denote $x = x^{[a]}$ and $\widehat{x} = x^{[\widehat{a}]}$ and we estimate by Jensen or Hölder inequalities

$$\begin{aligned}
\|x(t) - \widehat{x}(t)\|^2 &= \left\| \int_0^t (\dot{x}(s) - \dot{\widehat{x}}(s))ds \right\|^2 \leq t\int_0^t \|\dot{x}(s) - \dot{\widehat{x}}(s)\|^2 ds \\
&= t\int_0^t \frac{1}{N}\sum_{i=1}^N \left| (F^{[a]} * \mu^N(x_i) - F^{[\widehat{a}]} * \widehat{\mu}^N(\widehat{x}_i))\right|^2 ds \\
&\leq 2t\int_0^t \left[\frac{1}{N}\sum_{i=1}^N \left| (F^{[a]} - F^{[\widehat{a}]}) * \mu^N(x_i))\right|^2 \right. \\
&\quad + \left| \frac{1}{N}\sum_{j=1}^N \widehat{a}(|x_i - x_j|)((\widehat{x}_j - x_j) + (x_i - \widehat{x}_i)) \right. \\
&\quad \left. \left. + (\widehat{a}(|\widehat{x}_i - \widehat{x}_j|) - \widehat{a}(|x_i - x_j|))\,(\widehat{x}_j - \widehat{x}_i)\right|^2 \right] ds \\
&\leq 2T^2 \mathcal{E}^{[a],N}(\widehat{a}) + \int_0^t 8T(\|\widehat{a}\|_{L_\infty(K)}^2 + (R\,\mathrm{Lip}_K(\widehat{a}))^2)\|x(s) - \widehat{x}(s)\|^2 ds,
\end{aligned}$$

for $K = [0, 2R]$ and $R > 0$ is as in Proposition 2.2 for a substituted by $\widehat{a}$. An application of Gronwall's inequality yields the estimate

$$\|x(t) - \widehat{x}(t)\|^2 \leq 2T^2 e^{8T^2(\|\widehat{a}\|^2_{L_\infty(K)} + (R\,\mathrm{Lip}_K(\widehat{a}))^2)} \mathcal{E}^{[a],N}(\widehat{a}),$$

which is the desired bound. □

As already mentioned in the introduction, for $N \to \infty$ a natural mean-field approximation to the learning problem is the functional

$$\mathcal{E}^{[a]}(\widehat{a}) = \frac{1}{T} \int_0^T \int_{\mathbb{R}^d} \left| \big((F^{[\widehat{a}]} - F^{[a]}) * \mu(t) \big)(x) \right|^2 d\mu(t)(x) dt,$$

where $\mu(t)$ is a weak solution to (21).

3.1 The measure $\overline{\rho}$

In order to rigorously introduce the coercivity condition (18), we need to explore finer properties of the family of measures $(\varrho(t))_{t\in[0,T]}$, where we recall that $\varrho(t)(A) = (\mu(t) \otimes \mu(t))\big(d^{-1}(A)\big)$ for A a Borel set of $\mathbb{R}_+$.

Lemma 3.1. *For every open set $A \subseteq \mathbb{R}_+$ the mapping $t \in [0,T] \mapsto \varrho(t)(A)$ is lower semi-continuous, whereas for any compact set A it is upper semi-continuous.*

Proof. As a first step we show that for every given sequence $(t_n)_{n\in\mathbb{N}}$ converging to $t \in [0,T]$ we have the weak convergence $\varrho(t_n) \rightharpoonup \varrho(t)$ for $n \to +\infty$. We first note that $\mu(t_n) \otimes \mu(t_n) \rightharpoonup \mu(t) \otimes \mu(t)$, since $\mu(t_n) \rightharpoonup \mu(t)$ because of the continuity of $\mu(t)$ in the Wasserstein metric $\mathcal{W}_1$. This implies the claimed weak convergence $\varrho(t_n) \rightharpoonup \varrho(t)$, since for any function $f \in \mathcal{C}(\mathbb{R}_+)$, $f \circ d \in \mathcal{C}(\mathbb{R}^d \times \mathbb{R}^d)$, and hence

$$\begin{aligned}\int_{\mathbb{R}_+} f\, d\varrho(t_n) &= \int_{\mathbb{R}^{2d}} (f \circ d)(x,y) d(\mu(t_n) \otimes \mu(t_n))(x,y) \\ &\overset{n\to\infty}{\longrightarrow} \int_{\mathbb{R}^{2d}} (f \circ d)(x,y) d(\mu(t) \otimes \mu(t))(x,y) = \int_{\mathbb{R}_+} f\, d\varrho(t).\end{aligned}$$

The claim now follows from general results for weakly* convergent sequences of Radon measures, see e.g. [1, Proposition 1.62]. □

Lemma 3.1 justifies the following

Definition 3.2. The probability measure $\overline{\rho}$ on the Borel σ-algebra on $\mathbb{R}_+$ is defined for any Borel set $A \subseteq \mathbb{R}_+$ as follows

$$\overline{\rho}(A) := \frac{1}{T} \int_0^T \varrho(t)(A) dt. \tag{31}$$

Notice that Lemma 3.1 shows that (31) is well-defined only for sets A that are open or compact in $\mathbb{R}_+$. This directly implies that $\overline{\rho}$ can be extended to any Borel set A, since both families of sets provide a basis for the Borel σ-algebra on $\mathbb{R}_+$. Moreover $\overline{\rho}$ is a regular measure on $\mathbb{R}_+$, since Lemma 3.1 also implies that for any Borel set A

$$\overline{\rho}(A) = \sup\{\overline{\rho}(F) : F \subseteq A,\ F \text{ compact}\} = \inf\{\overline{\rho}(G) : A \subseteq G,\ G \text{ open}\}\,.$$

The measure $\overline{\rho}$ measures which - and how much - regions of $\mathbb{R}_+$ (the set of inter-point distances) are explored during the dynamics of the system. Highly explored regions are where our learning process ought to be successful, since these are the areas where we do have enough samples from the dynamics to reconstruct the function a.

We now show the absolute continuity of $\overline{\rho}$ w.r.t. the Lebesgue measure on $\mathbb{R}_+$. First of all we observe the following:

Lemma 3.3. *Let μ_0 be absolutely continuous w.r.t. the d-dimensional Lebesgue measure $\mathcal{L}_d$. Then $\mu(t)$ is absolutely continuous w.r.t. $\mathcal{L}_d$ for every $t \in [0,T]$.*

Proof. Both μ_0 and $\mu(t)$ are supported in $B(0,R)$, with R as in (47). $\mu(t)$ is the push-forward of μ_0 under the locally bi-Lipschitz map $\mathcal{T}_t^\mu$. Since $\mathcal{T}_t^\mu$ has Lipschitz inverse on $B(0,R)$, this inverse maps $\mathcal{L}_d$-null sets to $\mathcal{L}_d$-null sets, so μ_0-null sets are not only $\mathcal{L}_d$-null sets by assumption, but are also $\mu(t)$-null sets. □

Lemma 3.4. *Let μ_0 be absolutely continuous w.r.t. $\mathcal{L}_d$. Then, for all $t \in [0,T]$, the measures $\varrho(t)$ and $\overline{\rho}$ are absolutely continuous w.r.t. $\mathcal{L}_{1\llcorner\mathbb{R}_+}$ (Lebesgue measure in $\mathbb{R}$ restricted to $\mathbb{R}_+$).*

Proof. Fix $t \in [0,T]$. By Lemma 3.3 we already know that $\mu(t)$ is absolutely continuous w.r.t. $\mathcal{L}_d$, and so $\mu(t) \otimes \mu(t)$ is absolutely continuous w.r.t. $\mathcal{L}_{2d}$. It hence remains to show that $\mathcal{L}_{2d}$ is absolutely continuous w.r.t. $\mathcal{L}_{1\llcorner\mathbb{R}_+}$, where d is the distance function, but this follows easily by observing that $d^{-1}(A) = 0$ for every $\mathcal{L}_{1\llcorner\mathbb{R}_+}$-null set A, and an application of Fubini's theorem. The absolute continuity of $\overline{\rho}$ now follows immediately from the one of $\varrho(t)$ for every t and its definition as an integral average (31). □

As an easy consequence of the fact that the dynamics of our system has support uniformly bounded in time, we get the following crucial properties of the measure $\overline{\rho}$.

Lemma 3.5. *The measure $\overline{\rho}$ is finite and has compact support.*

Proof. We have

$$\overline{\rho}(\mathbb{R}_+) = \frac{1}{T}\int_0^T \varrho(t)(\mathbb{R}_+)dt = \frac{1}{T}\int_0^T \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y| d\mu(t)(x) d\mu(t)(y) dt < +\infty,$$

since the distance function is continuous and the support of μ is uniformly bounded in time. This shows that $\overline{\rho}$ is bounded. Since the supports of the measures $\varrho(t)$ are the subsets of $K = \{|x-y| : x,y \in B(0,R)\} = [0,2R]$, where R is given by (47), by construction we also have $\operatorname{supp}\overline{\rho} \subseteq K$. □

Remark 1. While absolute continuity of μ_0 implies the same for $\overline{\rho}$, the situation is different for purely atomic measures μ_0^N: then $\mu^N(t)$ is also purely atomic for every t, and so is $\varrho^N(t) = d_\#(\mu^N(t) \otimes \mu^N(t))$. However $\overline{\rho}$ is in general not purely atomic, due to the averaging in time in its definition (31).

An easy consequence of Lemma 3.5 is that if $f \in X$, then

$$\|f\|^2_{L_2(\mathbb{R}_+,\rho)} = \int_{\mathbb{R}_+} \left|f(s)\right|^2 d\rho(s) \leq \mathrm{diam}(\mathrm{supp}(\rho))\|f\|^2_{L_\infty(\mathrm{supp}(\rho))}\,, \tag{32}$$

and therefore $X \subseteq L_2(\mathbb{R}_+, \rho)$.

3.2 On the coercivity assumption

With the measure $\overline{\rho}$ at disposal we define, as in (16), $\rho(A) = \int_A s^2 d\overline{\rho}(s)$ for all Borel sets $A \subset \mathbb{R}_+$. By means of ρ, we recall the estimate

$$\mathcal{E}^{[a]}(\widehat{a}) \leq \frac{1}{T}\int_0^T \int_{\mathbb{R}_+} \left|\widehat{a}(s) - a(s)\right|^2 s^2 d\varrho(t)(s)dt = \|\widehat{a} - a\|^2_{L_2(\mathbb{R}_+,\rho)}. \tag{33}$$

This suggested the coercivity condition (18): $\mathcal{E}^{[a]}(\widehat{a}) \geq c_T\|\widehat{a} - a\|^2_{L_2(\mathbb{R}_+,\rho)}$. The main reason this condition is of interest to us is:

Proposition 3.6. *Assume $a \in X$ and that the coercivity condition* (18) *holds. Then any minimizer of $\mathcal{E}$ in X coincides ρ-a.e. with a.*

Proof. Notice that $\mathcal{E}(a) = 0$, and since $\mathcal{E}^{[a]}(\widehat{a}) \geq 0$ for all $\widehat{a} \in X$ this implies that a is a minimizer of $\mathcal{E}$. Now suppose that $\mathcal{E}^{[a]}(\widehat{a}) = 0$ for some $\widehat{a} \in X$. By (18) we obtain that $\widehat{a} = a$ in $L_2(\mathbb{R}_+, \rho)$, and therefore they coincide ρ-almost everywhere. □

3.2.1 Coercivity is "generically" satisfied

We make the case that while "degeneracies" would cause our coercivity condition to fail, but in a "generic" case the coercivity inequality holds.

Letting $K(r) = (\widehat{a}(r) - a(r))r$, for $r \in \mathbb{R}_+$, the coercivity condition becomes

$$\begin{aligned}\frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \left|\int_{\mathbb{R}^d} K(|x-y|)\frac{x-y}{|x-y|}d\mu(t)(x)\right|^2 d\mu(t)(y) \\ \geq \frac{c_T}{T}\int_0^T \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |K(|x-y|)|^2\, d\mu(t)(x)d\mu(t)(y)\,.\end{aligned}$$

This condition is satisfied if it is satisfied for all t in any nontrivial time interval in $[0, T]$. We will therefore consider the inequality without the time average for a fixed t:

$$\int_{\mathbb{R}^d} \left|\int_{\mathbb{R}^d} K(|x-y|)\frac{x-y}{|x-y|}d\mu(t)(x)\right|^2 d\mu(t)(y) \geq c_T \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |K(|x-y|)|^2\, d\mu(t)(x)d\mu(t)(y)\,,$$

and this will imply the inequality with the time average if this inequality holds for t in a nontrivial interval within $[0,T]$. Furthermore, we restrict our attention to the case when $\mu(t)$ is a discrete measure $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ (we drop t since t is fixed), so that the inequality reads

$$\frac{1}{N}\sum_{i=1}^N \left| \frac{1}{n}\sum_{j=1}^N K(|x_i - x_j|)\frac{x_i - x_j}{|x_i - x_j|} \right|^2 \geq \frac{c_T'}{N^2}\sum_{i=1}^N\sum_{j=1}^N |K(|x_i - x_j|)|^2 \,. \tag{34}$$

Assume that $\mathbf{K} := (K(|x_i - x_j|))_{i,j=1,\dots,N}$ is a random matrix. This model is not unreasonable: after all $\mathbf{K}$ involves the difference of our estimator $\widehat{a}$ and the target influence function a: for large classes of estimators this difference is random with the samples used to construct the estimator, with fluctuations that are, say, Gaussian. More specifically let us assume that $\mathbf{K}$ has Gaussian rows, each with variance $\sigma^2 I_N$. Since the bounds we wish to obtain, and our estimates below, are scale invariant, we may, and will, assume $\sigma = 1$. We now show that the coercivity assumption is satisfied. First of all observe that the discrete coercivity inequality (34) above may be re-written the matrices above as:

$$\frac{1}{N}\sum_{i=1}^N \left| \frac{1}{N}\mathbf{K}\mathbf{X}_i \right|^2 \geq \frac{c_t'}{N^2}||\mathbf{K}||_{\mathbb{F}}^2 \,. \tag{35}$$

Then we estimate (we use the notation $B(:,j)$ the j-th column of a matrix B)

$$\mathbb{E}\left[|\mathbf{K}\mathbf{X}_i|^2\right] = \sum_{j=1}^N \mathbb{E}\left[|\langle \mathbf{K}(j,:), \mathbf{X}_i(:,j)\rangle|^2\right] \geq C\sum_{j=1}^N N|\mathbf{X}_i(:,j)|^2 = CN||\mathbf{X}_i||_{\mathbb{F}}^2 = CN^2 \,,$$

where in the second step we used standard properties of random projections against Gaussian vectors [19], and in the last step we used the fact that every row of $\mathbf{X}_i$ is a unit vector. These bounds in fact hold with high probability, and does the bound

$$\frac{1}{N}\sum_{i=1}^N \left| \frac{1}{N}\mathbf{K}\mathbf{X}_i \right|^2 \geq \frac{1}{N}\sum_{i=1}^N \frac{1}{N^2}||\mathbf{X}_i||_{\mathbb{F}}^2 \geq C \,.$$

One the other hand, since $\mathbb{E}[||\mathbf{K}||_{\mathbb{F}}^2] \leq CN^2$ by standard random matrix theory results (e.g. [19]), and in fact not just in expectation but in high probability, the left hand side of (35) is upper bounded by $c_t' C$. Choosing c_t' small enough, we obtain (35).

The argument may be generalized to other models of random matrices, for example with sub-Gaussian rows (for $\mathbf{K}$) and uniformly lower-bounded smallest singular values. One may also consider $\mathbf{X}_i$ random, and obtain similar results. Also, the continuous case is not substantially different from the discrete case, as it may be derived by smoothing discrete approximations. We do not pursue these generalizations, as our purpose here is to show that the coercivity assumption is satisfied "generically".

3.3 Existence of minimizers of $\mathcal{E}_N$

The following proposition, which is a straightforward consequence of Ascoli-Arzelá Theorem, indicates the right ambient space where to state an existence result for the minimizers of $\mathcal{E}_N$.

Proposition 3.7. *Fix $M > 0$ and $K = [0, 2R] \subset \mathbb{R}_+$ for any $R > 0$. Define the set*

$$X_{M,K} = \left\{ b \in W^1_\infty(K) : \|b\|_{L_\infty(K)} + \|b'\|_{L_\infty(K)} \leq M \right\}.$$

The space $X_{M,K}$ is relatively compact with respect to the uniform convergence on K.

Proof. Consider $(\widehat{a}_n)_{n\in\mathbb{N}} \subset X_{M,K}$. The Fundamental Theorem of Calculus (which is applicable for functions in W^1_∞, see [1, Theorem 2.8]) implies that the functions $\widehat{a}_n$ are all Lipschitz continuous with Lipschitz constant uniformly bounded by M, which in turn implies equi-continuity. They are moreover pointwise uniformly equibounded. Ascoli-Arzelá Theorem then implies the existence of a subsequence converging uniformly on K to some $\widehat{a} \in X_{M,K}$. □

Proposition 3.8. *Assume $a \in X$. Fix $M > 0$ and $K = [0, 2R] \subset \mathbb{R}_+$ for $R > 0$ as in Proposition 2.2. Let V be a closed subset of $X_{M,K}$ w.r.t. the uniform convergence. Then, the optimization problem*

$$\textit{minimize } \mathcal{E}^{[a],N}(\widehat{a}) \textit{ among all } \widehat{a} \in V$$

admits a solution.

Proof. Since $\inf \mathcal{E}_N \geq 0$, we can consider a minimizing sequence $(\widehat{a}_n)_{n\in\mathbb{N}} \subset V$, i.e., such that $\lim_{n\to\infty} \mathcal{E}_N(\widehat{a}_n) = \inf_V \mathcal{E}_N$. By Proposition 3.7 there exists a subsequence of $(\widehat{a}_n)_{n\in\mathbb{N}}$ (labelled again $(\widehat{a}_n)_{n\in\mathbb{N}}$) converging uniformly on K to a function $\widehat{a} \in V$ (since V is closed). We now show that $\lim_{n\to\infty} \mathcal{E}_N(\widehat{a}_n) = \mathcal{E}_N(\widehat{a})$, from which it follows that $\mathcal{E}_N$ attains its minimum in V.

As a first step, notice that the uniform convergence of $(\widehat{a}_n)_{n\in\mathbb{N}}$ to $\widehat{a}$ on K and the compactness of K imply that the functionals $F^{[\widehat{a}_n]}(x-y)$ converge uniformly to $F^{[\widehat{a}]}(x-y)$ on $B(0,R) \times B(0,R)$ (where R is as in (47)). Moreover, we have the uniform bound

$$\begin{aligned} \sup_{x,y\in B(0,R)} |F^{[\widehat{a}_n]}(x-y) - F^{[a]}(x-y)| &= \sup_{x,y\in B(0,R)} |\widehat{a}_n(|x-y|) - a(|x-y|)||x-y| \\ &\leq 2R \sup_{r\in K} |\widehat{a}_n(r) - a(r)| \\ &\leq 2R(M + \|a\|_{L_\infty(K)}). \end{aligned} \tag{36}$$

As the measures $\mu^N(t)$ are compactly supported in $B(0,R)$ uniformly in time, the boundedness (36) allows us to apply three times the dominated convergence theorem to yield

$$\lim_{n\to\infty} \mathcal{E}_N(\widehat{a}_n) = \lim_{n\to\infty} \frac{1}{T} \int_0^T \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \left(F^{[\widehat{a}_n]}(x-y) - F^{[a]}(x-y) \right) d\mu^N(t)(y) \right|^2 d\mu^N(t)(x) dt$$

$$
\begin{aligned}
&= \frac{1}{T}\int_0^T \lim_{n\to\infty} \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \left(F^{[\widehat{a}_n]}(x-y) - F^{[a]}(x-y) \right) d\mu^N(t)(y) \right|^2 d\mu^N(t)(x)dt \\
&= \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \left| \lim_{n\to\infty} \int_{\mathbb{R}^d} \left(F^{[\widehat{a}_n]}(x-y) - F^{[a]}(x-y) \right) d\mu^N(t)(y) \right|^2 d\mu^N(t)(x)dt \\
&= \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \left(F^{[\widehat{a}]}(x-y) - F^{[a]}(x-y) \right) d\mu^N(t)(y) \right|^2 d\mu^N(t)(x)dt \\
&= \mathcal{E}^{[a],N}(\widehat{a}),
\end{aligned}
$$

which proves the statement.

□

4 Γ-convergence of $\mathcal{E}_N$ to $\mathcal{E}$

This section is devoted to a proof of Theorem 1.3. We start with a technical lemma.

Lemma 4.1. *Under the assumptions of Theorem 1.3, let $(b_N)_{N\in\mathbb{N}} \subset X_{M,K}$ be a sequence of continuous functions uniformly converging to a function $b \in X_{M,K}$ on $K = [0, 2R]$ with $R > 0$ as in (47). Then it holds*

$$
\lim_{N\to\infty} \mathcal{E}_N(b_N) = \mathcal{E}(b).
$$

Proof. From [13, Lemma 3.3] it follows $\mathcal{W}_1(\mu^0, \mu_N^0) \to 0$ for $N \to \infty$. Hence, from (29) we have that $W_1(\mu(t), \mu_N(t)) \to 0$ for $N \to \infty$, uniformly for $t \in [0, T]$. Moreover, for all $x, y, y' \in B(0, R)$, by triangle inequality we have

$$
\begin{aligned}
|(F^{[a]} - F^{[b]})(x-y') - (F^{[a]} - F^{[b]})(x-y)| \\
\leq \left[2R(\mathrm{Lip}_K(a) + \mathrm{Lip}_K(b)) + \|a\|_{L_\infty(K)} + \|b\|_{L_\infty(K)}\right] |y - y'|,
\end{aligned}
$$

which implies the uniform Lipschitz continuity of $(F^{[a]} - F^{[b]})(x - \cdot)$ in $B(0, R)$, for fixed $x \in B(0, R)$ was:with respect to $x \in B(0, R)$. For every $\varepsilon > 0$ one can find $N_0(\varepsilon)$ such that, for all $N \geq N_0(\varepsilon)$ we have

$$
\sup_{x,y\in B(0,R)} |F^{[b_N]}(x-y) - F^{[b]}(x-y)| \leq 2R\|b_N - b\|_{L_\infty(K)} \leq \varepsilon/2,
$$

as well as

$$
\left| \int_{\mathbb{R}^d} \left(F^{[b]} - F^{[a]}\right)(x-y) d\mu^N(t)(y) - \int_{\mathbb{R}^d} \left(F^{[b]} - F^{[a]}\right)(x-y) d\mu(t)(y) \right| \leq \varepsilon/2,
$$

uniformly with respect to $t \in [0, T]$ and $x \in B(0, R)$. The first estimate follows from the uniform convergence of the b_N, while the second one follows from the uniform Lipschitz continuity of $(F^{[a]} - F^{[b]})(x - \cdot)$ for fixed $x \in B(0, R)$ was:with respect to $x \in B(0, R)$

and the uniform Wasserstein convergence of $\mu^N(t)$ to $\mu(t)$ with respect to $t \in [0,T]$. Hence for $N \geq N_0(\varepsilon)$ we obtain

$$\begin{aligned}
&\left|\left|\int_{\mathbb{R}^d} \big(F^{[b_N]} - F^{[a]}\big)(x-y)d\mu^N(t)(y)\right| - \left|\int_{\mathbb{R}^d} \big(F^{[b]} - F^{[a]}\big)(x-y)d\mu(t)(y)\right|\right| \\
&\leq \left|\int_{\mathbb{R}^d} \big(F^{[b_N]} - F^{[a]}\big)(x-y)d\mu^N(t)(y) - \int_{\mathbb{R}^d} \big(F^{[b]} - F^{[a]}\big)(x-y)d\mu(t)(y)\right| \\
&\leq \left|\int_{\mathbb{R}^d} \big(F^{[b_N]} - F^{[b]}\big)(x-y)d\mu^N(t)(y)\right| \\
&\qquad + \left|\int_{\mathbb{R}^d} \big(F^{[b]} - F^{[a]}\big)(x-y)d\mu^N(t)(y) - \int_{\mathbb{R}^d} \big(F^{[b]} - F^{[a]}\big)(x-y)d\mu(t)(y)\right| \\
&\leq 2R\|b_N - \widehat{a}\|_{L_\infty(K)} \int_{\mathbb{R}^d} d\mu^N(t)(y) + \frac{\varepsilon}{2} = \varepsilon\,.
\end{aligned}$$

Therefore, for every $t \in [0,T]$ and $x \in B(0,R)$,

$$\lim_{N\to\infty} \left|\int_{\mathbb{R}^d} \big(F^{[b_N]} - F^{[a]}\big)(x-y)d\mu^N(t)(y)\right|^2 = \left|\int_{\mathbb{R}^d} \big(F^{[b]} - F^{[a]}\big)(x-y)d\mu(t)(y)\right|^2. \quad (37)$$

Denote

$$\begin{aligned}
H_N(t,x) &= \left|\int_{\mathbb{R}^d} \big(F^{[b_N]} - F^{[a]}\big)(x-y)d\mu^N(t)(y)\right|^2, \qquad G_N(t) = \int_{\mathbb{R}^d} H_N(t,x)d\mu_N(t)(x)\,, \\
H(t,x) &= \left|\int_{\mathbb{R}^d} \big(F^{[b]} - F^{[a]}\big)(x-y)d\mu(t)(y)\right|^2, \qquad G(t) = \int_{\mathbb{R}^d} H(t,x)d\mu(t)(x),
\end{aligned}$$

and we estimate

$$\begin{aligned}
|G_N(t) - G(t)| &\leq \left|\int_{\mathbb{R}^d} H(t,x)d\mu_N(t)(x) - \int_{\mathbb{R}^d} H(t,x)d\mu(t)(x)\right| \\
&\qquad + \int_{\mathbb{R}^d} |H_N(t,x) - H(t,x)|\, d\mu_N(t)(x). \quad (38)
\end{aligned}$$

We now prove that the function H is Lipschitz continuous with respect to $x \in B(0,R)$ uniformly in $t \in [0,T]$. To do so, we write H as

$$H(t,x) = g\left(\int_{\mathbb{R}^d} f(x,y)d\mu(t)(y)\right),$$

where g is a differentiable function in $\mathbb{R}^d$ and $f(\cdot,y)$ is a Lipschitz continuous function uniformly in y with values in $\mathbb{R}^d$ (in this case $f(\cdot,y) = (F^{[a]} - F^{[b]})(\cdot - y)$). Since the measure $\mu(t)$ has support contained in $B(0,R)$, it follows

$$\left|\int_{\mathbb{R}^d} f(x,y)d\mu(t)(y)\right| \leq \int_{\mathbb{R}^d} |f(x,y)|d\mu(t)(y) \leq \sup_{x,y\in B(0,R)} |f(x,y)| =: S < +\infty.$$

Therefore, for every $x, x' \in B(0,R)$

$$\begin{aligned}\big|H(t,x) - H(t,x')\big| &\leq \mathrm{Lip}_{B(0,S)}(g)\left|\int_{\mathbb{R}^d} f(x,y)d\mu(t)(y) - \int_{\mathbb{R}^d} f(x',y)d\mu(t)(y)\right| \\ &\leq \mathrm{Lip}_{B(0,S)}(g)\int_{\mathbb{R}^d}\big|f(x,y)-f(x',y)\big|\,d\mu(t)(y) \\ &\leq \mathrm{Lip}_{B(0,S)}(g)\ \mathrm{Lip}_{B(0,R)}(f)\ |x-x'|,\end{aligned}$$

from which follows the Lipschitz continuity of $H(t,x)$ with respect to $x \in B(0,R)$ uniformly in $t \in [0,T]$.

From this uniform Lipschitz continuity and the uniform Wasserstein convergence of $\mu_N(t)$ to $\mu(t)$ with respect to $t \in [0,T]$, it follows that for every $\varepsilon > 0$ we can find $N_0(\varepsilon)$ such that for all $N \geq N_0(\varepsilon)$ it holds

$$\left|\int_{\mathbb{R}^d} H(t,x)d\mu_N(t)(x) - \int_{\mathbb{R}^d} H(t,x)d\mu(t)(x)\right| \leq \frac{\varepsilon}{2}, \tag{39}$$

uniformly with respect to $t \in [0,T]$. From (37) it follows also that for all $N \geq N_0(\varepsilon)$ we have

$$|H_N(t,x) - H(t,x)| \leq \frac{\varepsilon}{2}, \tag{40}$$

uniformly with respect to $t \in [0,T]$ and $x \in B(0,R)$. A combination of (38) with (39) and (40) yields $|G_N(t) - G(t)| \leq \varepsilon$ uniformly in $t \in [0,T]$. Hence

$$\lim_{N\to\infty}\int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d}\big(F^{[b_N]}-F^{[a]}\big)(x-y)d\mu^N(t)(y)\right|^2 d\mu_N(t)(x) = \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d}\big(F^{[b]}-F^{[a]}\big)(x-y)d\mu(t)(y)\right|^2 d\mu(t)(x),$$

holds uniformly in $t \in [0,T]$ does this just mean "holds for every $t \in [0,T]$"?.

To eventually show that $\lim_{N\to\infty}\mathcal{E}_N(b_N) = \mathcal{E}(b)$, we simply note that uniform convergence of G_N to G implies

$$\lim_{N\to\infty}\mathcal{E}_N(b_N) = \lim_{N\to\infty}\frac{1}{T}\int_0^T G_N(t)dt = \frac{1}{T}\int_0^T G(t)dt = \mathcal{E}(b).$$

□

Proof of Theorem 1.3. The sequence of minimizers $(\widehat{a}_N)_{N\in\mathbb{N}}$ is by definition a subset of $X_{M,K}$, hence by Proposition 3.7 it admits a subsequence $(\widehat{a}_{N_k})_{k\in\mathbb{N}}$ uniformly converging to a function $\widehat{a} \in X_{M,K}$.

To show the optimality of $\widehat{a}$ in $X_{M,K}$, let $b \in X_{M,K}$ be given. By Definition 1.2, we can find a sequence $(b_N)_{N\in\mathbb{N}}$ converging uniformly to b on K such that $b_N \in V_N$ for every $N \in \mathbb{N}$. Hence, by Lemma 4.1, we have

$$\mathcal{E}(b) = \lim_{N\to\infty}\mathcal{E}_N(b_N).$$

Now, by the optimality of $\widehat{a}_{N_k}$ in V_N and again by Lemma 4.1, it follows that

$$\mathcal{E}(b) = \lim_{N\to\infty} \mathcal{E}_N(b_N) = \lim_{k\to\infty} \mathcal{E}_{N_k}(b_{N_k}) \geq \lim_{k\to\infty} \mathcal{E}_{N_k}(\widehat{a}_{N_k}) = \mathcal{E}(\widehat{a})\,.$$

We can therefore conclude that for every $b \in X_{M,K}$

$$\mathcal{E}(b) \geq \mathcal{E}(\widehat{a})\,. \tag{41}$$

In particular, (41) applies to $b = a \in X_{M,K}$ (by the particular choice of M), which finally implies

$$0 = \mathcal{E}(a) \geq \mathcal{E}(\widehat{a}) \geq 0 \Longrightarrow \mathcal{E}(\widehat{a}) = 0,$$

showing that $\widehat{a}$ is also a minimizer of $\mathcal{E}$. When the coercivity condition (18) holds, by Proposition 3.6 it follows $\widehat{a} = a$ in $L_2(\mathbb{R}_+, \rho)$. □

5 Numerical experiments

In this section we report several numerical experiments to document the validity and feasibility of Theorem 1.3. We will first show how the reconstruction of a kernel gets better as the number of agents N increases, in accordance with the Γ-convergence result reported in the last section. This feature holds true also for interaction kernel not lying in the function space X, as shown in Figure 2. We will then investigate the relationship between the functional $\mathcal{E}_N$ and the $L_2(\mathbb{R}_+, \rho)$ error w.r.t. the potential a, and the behavior of $\mathcal{E}_N$ when we let the constraint M vary. Finally, we will see how we can get a very satisfactory reconstruction of the unknown interaction kernel by keeping N fixed and averaging the minimizers of the functional $\mathcal{E}_N$ obtained from several samples of the initial data distribution μ^0.

5.1 Numerical framework

All experiments will rely on a common numerical set-up, which we clarify in this section. All the initial data μ_0^N are drawn from a common probability distribution μ_0 which is the uniform distribution on the d-dimensional cube $[-L, L]^d$. For every μ_0^N, we simulate the evolution of the system starting from μ_0^N until time T, and we shall denote with R the maximal distance between particles reached during the time frame $[0, T]$. Notice that we have at our disposal only a finite sequence of snapshots of the dynamics: if we denote with $0 = t_0 < t_1 < \ldots < t_m = T$ the time instants at which these snapshots are taken, we can consider the *discrete-time error functional*

$$\overline{\mathcal{E}}_N(\widehat{a}) = \frac{1}{m}\sum_{k=1}^{m}\frac{1}{N}\sum_{j=1}^{N}\left|\frac{1}{N}\sum_{i=1}^{N}\widehat{a}(|x_j(t_k) - x_i(t_k)|)(x_j(t_k) - x_i(t_k)) - \dot{x}_i(t_k)\right|^2,$$

which is the time-discrete counterpart of the continuous-time error functional $\mathcal{E}_N(\widehat{a})$. As already mentioned in the introduction, velocities $\dot{x}_i(t_k)$ appearing in $\overline{\mathcal{E}}_N(\widehat{a})$ are computed as finite differences, i.e.,

$$\dot{x}_i(t_k) = \frac{x_i(t_k) - x_i(t_{k-1})}{t_k - t_{k-1}}, \text{ for every } k \geq 1.$$

About the reconstruction procedure, we fix the constraint M and consider the sequence of invading subspaces V_N generated by a linear B-spline basis with $D(N)$ elements supported on $[0, 2R]$, i.e., for every element $\widehat{a} \in V_N$ it holds

$$\widehat{a}(r) = \sum_{\lambda=1}^{D(N)} a_\lambda \varphi_\lambda(r), \quad \text{for every } r \in [0, R].$$

Since V_N has to be invading, $D(N)$ is a strictly increasing function of N. To avoid the inconvenience of having a reconstructed kernel with value at 0 and at R equal to 0, we consider the knots of the B-spline basis to be

$$\left[0, 0, 0, \frac{R}{(D(N)-2)}, \ldots, \frac{(D(N)-3)R}{(D(N)-2)}, R, R, R\right], \tag{42}$$

i.e., the spline reconstruction has in 0 and R two discontinuity points.

Whenever $\widehat{a} \in V_N$, we can rewrite the functional $\mathcal{E}_N(\widehat{a})$ as

$$\begin{aligned}
\overline{\mathcal{E}}_N(\widehat{a}) &= \frac{1}{m}\sum_{k=1}^{m}\frac{1}{N}\sum_{j=1}^{N}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{\lambda=1}^{D(N)} a_\lambda \varphi_\lambda(|x_j(t_k) - x_i(t_k)|)(x_j(t_k) - x_i(t_k)) - \dot{x}_i(t_k)\right|^2 \\
&= \frac{1}{m}\sum_{k=1}^{m}\frac{1}{N}\sum_{j=1}^{N}\left|\sum_{\lambda=1}^{D(N)} a_\lambda \frac{1}{N}\sum_{i=1}^{N} \varphi_\lambda(|x_j(t_k) - x_i(t_k)|)(x_j(t_k) - x_i(t_k)) - \dot{x}_i(t_k)\right|^2 \\
&= \frac{1}{mN}\left\|C\vec{a} - v\right\|_2^2,
\end{aligned}$$

where $\vec{a} = (a_1, \ldots, a_{D(N)})$, $v = (\dot{x}_1(t_1), \ldots, \dot{x}_N(t_1), \ldots, \dot{x}_1(t_m), \ldots, \dot{x}_N(t_m))$ and $C \in \mathbb{R}^{d \times Nm \times D(N)}$ satisfies for every $j = 1, \ldots, N$, $k = 1, \ldots, m$, $\lambda = 1, \ldots, D(N)$

$$C(jk, \lambda) = \frac{1}{N}\sum_{i=1}^{N} \varphi_\lambda(|x_j(t_k) - x_i(t_k)|)(x_j(t_k) - x_i(t_k)) \in \mathbb{R}^d.$$

We shall numerically implement the constrained minimization with the software CVX, which allows constraints and objectives to be specified using standard MATLAB expression syntax. In order to use it, we need to rewrite the constraint of our minimization problem, which reads

$$\|a\|_{L_\infty([0,R])} + \|a'\|_{L_\infty([0,R])} \leq M,$$

using only the minimization variable of the problem, which is the vector of coefficients of the B-spline basis $\vec{a}$. Notice that the property of being a linear B-spline basis implies that, for every $\lambda = 1, \dots, D(N) - 1$, the property $\mathrm{supp}(\varphi_\lambda) \cap \mathrm{supp}(\varphi_{\lambda+j}) \neq \emptyset$ holds if and only if $j = 1$. Hence, for every $a \in V_N$ we have

$$\|a\|_{L_\infty([0,R])} = \max_{r\in[0,R]} \left| \sum_{\lambda=1}^{D(N)} a_\lambda \varphi_\lambda(r) \right| \leq \max_{\lambda=1,\dots,D(N)-1} (|a_\lambda| + |a_{\lambda+1}|) \leq 2\|\vec{a}\|_\infty,$$

$$\|a'\|_{L_\infty([0,R])} = \max_{r\in[0,R]} \left| \sum_{\lambda=1}^{D(N)} a_\lambda \varphi'_\lambda(r) \right| \leq \max_{\lambda=1,\dots,D(N)-1} |a_{\lambda+1} - a_\lambda| = \|D\vec{a}\|_\infty,$$

where, in the last line, D is the matrix

$$D = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ & & & & & \\ 0 & 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

We numerically implement the constrained minimization problem

$$\min_{\widehat{a}\in V_N} \mathcal{E}_N(\widehat{a}) \quad \text{subject to} \quad \|\widehat{a}\|_{L_\infty([0,R])} + \|\widehat{a}'\|_{L_\infty([0,R])} \leq M,$$

in the following way

$$\min_{\vec{a}\in\mathbb{R}^{D(N)}} \frac{1}{mN} \|C\vec{a} - v\|_2^2 \quad \text{subject to} \quad 2\|\vec{a}\|_\infty + \|D\vec{a}\|_\infty \leq M. \tag{43}$$

The byproduct of the time discretization and the reformulation of the constraint is that minimizers of problem (43) may not be minimizers of the original one. This is the price to pay for the numerical implementation of the theoretical result. However, we will see how in practice the modifications we made are so mild that we obtain very satisfactory reconstructions of the unknown potential, and moreover we still observe the approximation properties proved in the previous sections.

5.2 Varying N

In Figure 1 we show the reconstruction of a truncated Lennard-Jones type interaction kernel obtained with different values of N. The following table reports the values of the different problem's parameters.

It is clearly visible how the the piecewise linear approximant (displayed in blue) gets closer and closer to the potential to be recovered (in red), as predicted by the theoretical

d	L	T	M	N	$D(N)$
2	3	0.5	100	$[10, 20, 40, 80]$	$2N$

Table 1: Parameter values for Figure 1 and Figure 2

results of the previous sections. What is however surprising is that the same behavior is witnessed in Figure 2, where the algorithm is applied to an interaction kernel a not belonging to the function space X (due to its singularity at the origin) with the same specifications reported in Table 1. The algorithm performs an excellent approximation despite the highly oscillatory nature of the function a.

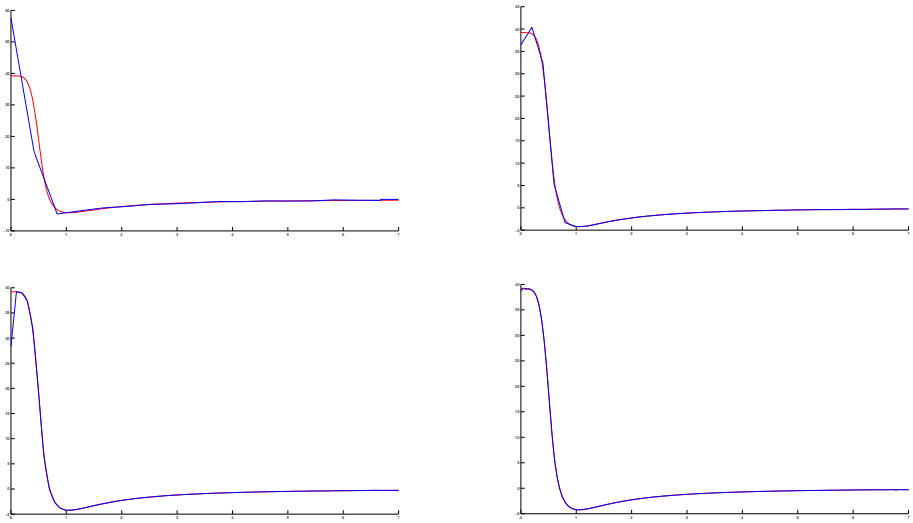

Figure 1: Iterative reconstruction of a potential with different values of N. In red: the unknown kernel. In blue: its reconstruction by minimization of $\mathcal{E}_N$. From left-top to right-bottom: reconstruction with $N = 10$, $N = 20$, $N = 40$, $N = 80$ agents.

5.3 The coercivity constant

We now turn our attention to the coercivity constant c_T appearing in (18) and thoroughly discussed in Section 3.2. In Figure 3 we see a comparison between the evolution of the value of the error functional $\overline{\mathcal{E}}_N(\widehat{a}_N)$ and of the $L_2(\mathbb{R}_+, \rho)$-error $\|a - \widehat{a}_N\|^2_{L_2(\mathbb{R}_+,\rho)}$ for different values of N.

In this experiment, the potential a to be retrieved is the truncated Lennard-Jones type interaction kernel of Figure 1 and the parameters used in the algorithm are reported in Table 2.

For every value of N, we have obtained the minimizer $\widehat{a}_N$ of problem (43) and we have computed the errors $\overline{\mathcal{E}}_N(\widehat{a}_N)$ and $\|a-\widehat{a}_N\|^2_{L_2(\mathbb{R}_+,\rho)}$. The $L_2(\mathbb{R}_+, \rho)$-error multiplied

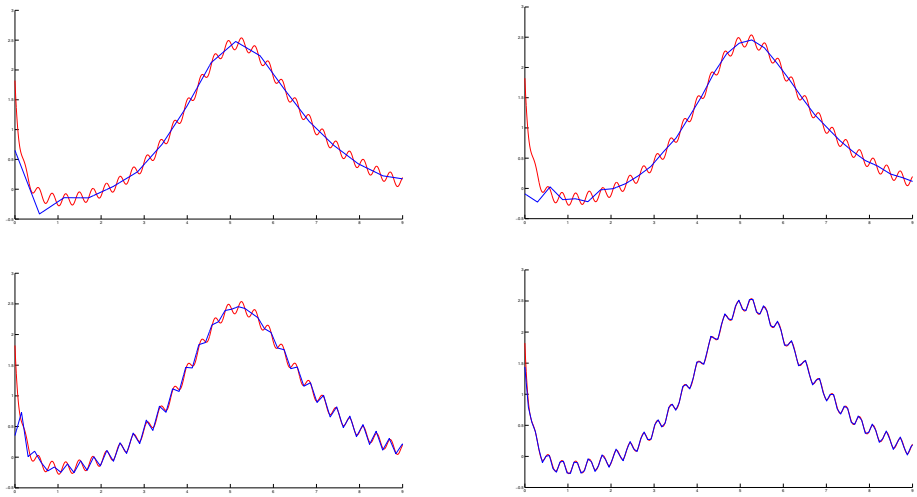

Figure 2: Iterative reconstruction of a potential with a singularity at the origin and highly oscillatory behavior. In red: the unknown kernel. In blue: its reconstruction by minimization of $\mathcal{E}_N$. From left-top to right-bottom: reconstruction with $N = 10$, $N = 20$, $N = 40$, $N = 80$ agents.

d	L	T	M	N	$D(N)$
2	5	0.5	100	$[3, 4, \ldots, 12]$	$3N - 5$

Table 2: Parameter values for Figure 3

by a factor $\frac{1}{10}$ lies entirely below the curve of $\overline{\mathcal{E}}_N(\widehat{a}_N)$, which let us empirically estimate the value of c_T around that value.

The major issue with this experiment is that the minimization procedure quickly reaches values near machine precision. This causes the inability of the approximation to comply with the increasing number of distances explored by the increasing number of agents, and leads to a resurgence of the $L_2(\mathbb{R}_+, \rho)$-error, which is defined as

$$\|\widehat{a} - a\|^2_{L_2(\mathbb{R}_+,\rho)} = \frac{1}{T}\int_0^T \int_{\mathbb{R}_+} \left|\widehat{a}(s) - a(s)\right|^2 s^2 d\varrho(t)(s)dt. \tag{44}$$

Hence, rather than seeing a smooth descent to zero for higher values of N, due to finite precision a saw-like pattern emerges.

5.4 Tuning the constraint M

Figure 5 shows what happens when we modify the value of M in problem (43). More specifically, we generate μ_0^N as explained in Section 5.1 once, and we simulate the system starting from μ_0^N until time T. With the data of this single evolution, we solve problem

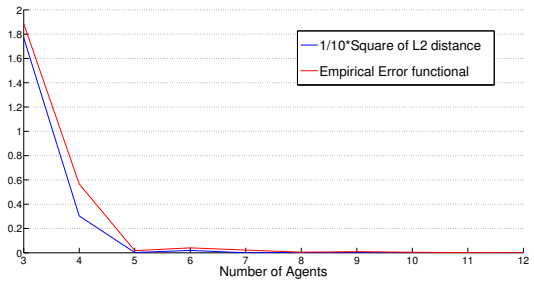

Figure 3: Plot of $\overline{\mathcal{E}}_N(\widehat{a}_N)$ and $\frac{1}{10}\|a - \widehat{a}_N\|^2_{L_2(\mathbb{R}_+,\rho)}$ for different values of N. In this experiment, we can estimate the constant c_T with the value $\frac{1}{10}$. I would make this a log plot, i.e. with teh y-axis in $\log_{10}$ (or $\log_2$) scale

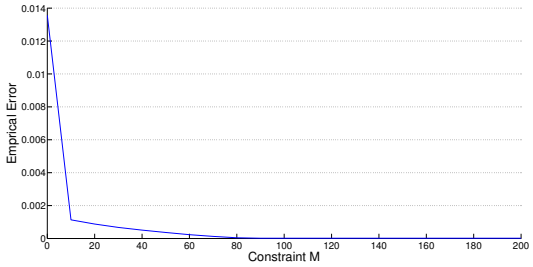

Figure 4: Values of $\overline{\mathcal{E}}_N$ for fixed $N = 50$ for different values of $M \in [0, 200]$.I would make this a log plot, i.e. with teh y-axis in $\log_{10}$ (or $\log_2$) scale

(43) for several values of M, denoting with $\widehat{a}_M$ the minimizer obtained with a specific value of M. On the left column of Figure 5 we show how the reconstruction $\widehat{a}_M$ gets closer and closer to the true potential a (in white) as M increases, while on the right column we see that the original trajectories (again, in white) used for the inverse problem are approximated better and better by those generated with $\widehat{a}_M$, if we let M grow. Table 3 reports the values of the parameters of these experiments.

	d	L	T	M	N	$D(N)$
First row	2	3	1	$2.7 \times [10, 15, \dots, 40]$	20	60
Second row	2	3	1	$1.25 \times [10, 15, \dots, 40]$	20	150

Table 3: Parameter values for Figure 5

Until now, we have no criteria to sieve those values of M which enable a successful reconstruction of a potential $a \in X$. To carry out such an analysis, we fix all the parameters of our problem as in Table 4 and again take a to be the truncated Lennard-Jones type potential of Figure 1. From what we said in the previous section regarding the continuous-time error functional $\mathcal{E}_N$, we can expect two things from the discrete-time one $\overline{\mathcal{E}}_N$ if we keep N fixed:

- that $\overline{\mathcal{E}}_N(\widehat{a}_M)$ will be decreasing as M increases. This follows from the fact that our constraint in problem (43) is an inequality, therefore the bigger M, the larger the class of potential minimizers;
- that there is a critical value $\overline{M} > 0$ such that $\overline{\mathcal{E}}_N(\widehat{a}_M)$ stabilizes as $M \geq \overline{M}$. Indeed, from the assumption $a \in X$ follows that after you hit $\overline{M} := \|a\|_{L_\infty(K)} + \|a'\|_{L_\infty(K)}$, the class of potential minimizers should not grow much more.

Figure 4 document precisely this expected behavior. This actually suggests that an

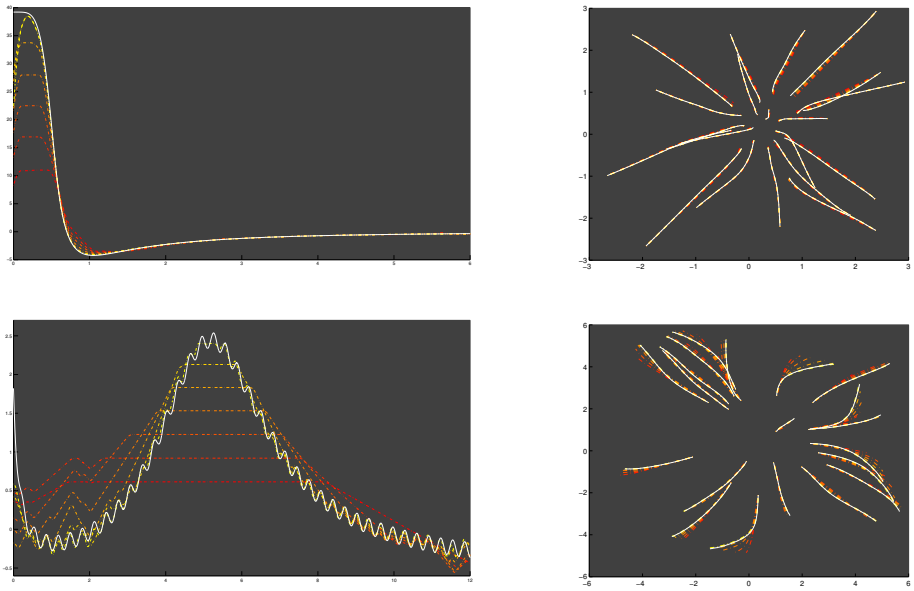

Figure 5: Iterative reconstruction of a potential with different values of M. On the left column: the true kernel in white and its reconstructions for different M; the brighter the curve, the larger the M. On the right column: the true trajectories of the agents in white, the trajectories associated to the reconstructed potentials with the same color.This pictures appear on gray/dark background when I compile, instead of white...

d	L	T	M	N	$D(N)$
2	3	0.5	$[0, 10, 20, \dots, 200]$	50	100

Table 4: Parameter values for Figure 4

effective strategy to find a sufficiently good constraint M *a posteriori* is to perform an experiment like the one in Figure 4 and to choose an M after which the curve of $\overline{\mathcal{E}}_N(\widehat{a}_M)$ flattens.

5.5 Reconstruction with N fixed

We now explore the possibility of a reconstruction strategy which does not rely on increasing the value of N. Indeed, problem (43) can swiftly become unfeasible when N is moderately large, also because the dimension of the invading subspaces V_N increase with N too. A strategy to overcome this limitation could be to consider, for a fixed N, several discrete initial data $(\mu_{0,i}^N)_{i=1}^{\Theta}$ all independently drawn from the same distribution μ_0 (in our case, the d-dimensional cube $[-L, L]^d$). For every $i = 1, \dots, \Theta$, we simulate the system until time T and, with the trajectories we obtain, we solve problem (43).

At the end of this procedure, we have a family of reconstructed potentials $(\widehat{a}_{N,i})_{i=1}^{\Theta}$, all approximating the same true kernel a. Averaging these potentials, we obtain a functional

$$\widehat{a}_N(r) = \frac{1}{\Theta}\sum_{i=1}^{\Theta}\widehat{a}_{N,i}(r), \qquad \text{for every } r \in [0,R],$$

which we claim to be a good approximation to the true kernel a. To motivate this claim, we report the outcome of an experiment whose data can be found in Table 5.

d	L	T	M	N	$D(N)$	Θ
2	2	0.5	1000	50	150	5

Table 5: Parameter values for Figure 6

Figure 6 shows the true potential a and the averaged reconstruction $\widehat{a}_N$. The black-dotted lines corresponds to the 95% confidence interval obtained from the averaged data.

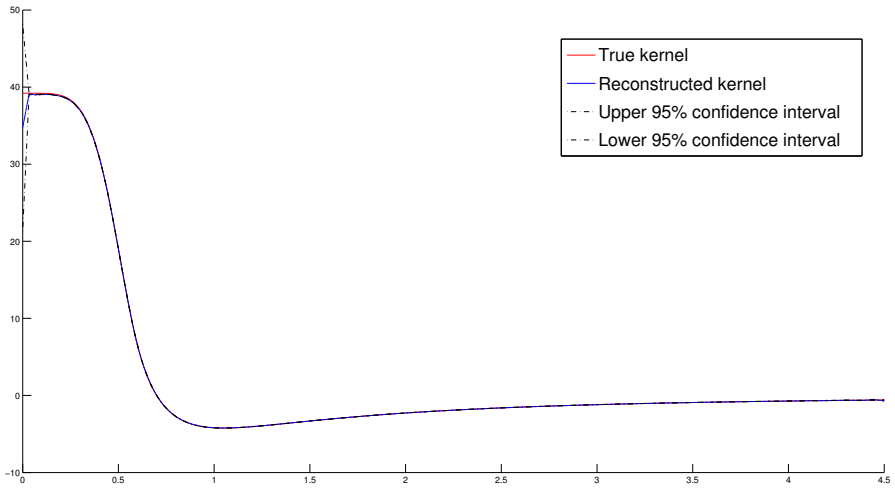

Figure 6: Reconstruction of a obtained by averaging 5 solutions of the minimization of $\mathcal{E}_N$ for $N = 50$. In red: the unknown kernel. In blue: the average of reconstructions. In black: 95% confidence interval.

6 Appendix

6.1 Technical lemmas for the mean-field limit

The following preliminary result tells us that solutions to system (22) are also solutions to systems (21), whenever conveniently rewritten.

Proposition 6.1. *Let $N \in \mathbb{N}$ be given. Let $(x_1^N, \dots, x_N^N) : [0,T] \to \mathbb{R}^{dN}$ be the solution of (22) with initial datum $x_0^N \in \mathbb{R}^{dN}$. Then the empirical measure $\mu_N : [0,T] \to \mathcal{P}_1(\mathbb{R}^d)$ defined as in (24) is a solution of (21) with initial datum $\mu_0 = \mu_N(0) \in \mathcal{P}_c(\mathbb{R}^d)$.*

Proof. It can be easily proved by arguing exactly as in [14, Lemma 4.3]. □

We are able to state several basic estimates that shall be useful towards an existence and uniqueness result for the solutions of system (22).

Lemma 6.2. *Let $a \in X$ and $\mu \in \mathcal{P}_1(\mathbb{R}^d)$. Then for all $y \in \mathbb{R}^d$ the following hold:*

$$|(F^{[a]} * \mu)(y)| \leq \|a\|_{L_\infty(\mathbb{R}_+)} \left(|y| + \int_{\mathbb{R}^d} |x| d\mu(x) \right).$$

Proof. Trivially follows from $a \in L_\infty(\mathbb{R}_+)$. □

Lemma 6.3. *If $a \in X$ then $F^{[a]} \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}^d)$.*

Proof. For any compact set $K \subset \mathbb{R}^d$ and for every $x, y \in K$ it holds

$$\begin{aligned} |F^{[a]}(x) - F^{[a]}(y)| &= |a(|x|)x - a(|y|)y| \\ &\leq |a(|x|)||x-y| + |a(|x|) - a(|y|)||y| \\ &\leq (|a(|x|)| + \mathrm{Lip}_K(a)|y|)|x-y|, \end{aligned}$$

and since $a \in L_\infty(\mathbb{R}_+)$ and $y \in K$, it follows that $F^{[a]}$ is locally Lipschitz with Lipschitz constant depending only on a and K. □

Lemma 6.4. *If $a \in X$ and $\mu \in \mathcal{P}_c(\mathbb{R}^d)$ then $F^{[a]} * \mu \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}^d)$.*

Proof. For any compact set $K \subset \mathbb{R}^d$ and for every $x, y \in K$ it holds

$$\begin{aligned} |(F^{[a]} * \mu)(x) - (F^{[a]} * \mu)(y)| &= \left| \int_{\mathbb{R}^d} a(|x-z|)(x-z) d\mu(z) - \int_{\mathbb{R}^d} a(|y-z|)(y-z) d\mu(z) \right| \\ &\leq \int_{\mathbb{R}^d} |a(|x-z|) - a(|y-z|)|x-z| d\mu(z) \\ &\quad + \int_{\mathbb{R}^d} |a(|y-z|)||x-y| d\mu(z) \\ &\leq \mathrm{Lip}_{\widehat{K}}(a)|x-y| \int_{\mathbb{R}^d} |x-z| d\mu(z) + \|a\|_{L_\infty(\mathbb{R}_+)}|x-y| \\ &\leq \left(\mathrm{Lip}_{\widehat{K}}(a)(|x|+1) + \|a\|_{L_\infty(\mathbb{R}_+)} \right) |x-y| \\ &\leq \left(C\mathrm{Lip}_{\widehat{K}}(a) + \|a\|_{L_\infty(\mathbb{R}_+)} \right) |x-y|, \end{aligned}$$

where C is a constant depending on K, and $\widehat{K}$ is a compact set containing both K and $\mathrm{supp}(\mu)$. □

Proposition 6.5. *If $a \in X$ then system (22) admits a unique global solution in $[0,T]$ for every initial datum $x_0^N \in \mathbb{R}^{dN}$.*

Proof. Rewriting system (22) in the form of (23), follows easily that the function $G : \mathbb{R}^{dN} \to \mathbb{R}^{dN}$ defined for every $(x_1, \dots, x_N) \in \mathbb{R}^{dN}$ as

$$G(x_1, \dots, x_N) = ((F^{[a]} * \mu_N)(x_1), \dots, (F^{[a]} * \mu_N)(x_N)),$$

where μ_N is the empirical measure given by (24), satisfies $G \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}^{dN})$. Indeed, for any $x_1, \dots, x_N, y_1, \dots, y_N \in K$ compact subset of $\mathbb{R}^d$, denoting with ν^N the empirical measure given by $y_1, \dots, y_N$, it simply suffices to write

$$\begin{aligned} |G(x_1, \dots, x_N) - G(y_1, \dots, y_N)| &\leq \sum_{i=1}^{N} |(F^{[a]} * \mu_N)(x_i) - (F^{[a]} * \nu^N)(y_i)| \\ &\leq \sum_{i=1}^{N} \Bigg(|(F^{[a]} * \mu_N)(x_i) - (F^{[a]} * \mu_N)(y_i)| \\ &\qquad + |(F^{[a]} * \mu_N)(y_i) - (F^{[a]} * \nu^N)(y_i)| \Bigg). \end{aligned}$$

Applying Lemma 6.4 to the first term and performing similar calculations to the ones in the proof of Lemma 6.3 on the second one, gives the desired result. The Cauchy-Lipschitz Theorem for ODE systems then yields the statement. □

Variants of the following result are [14, Lemma 6.7] and [7, Lemma 4.7]

Lemma 6.6. *Let $a \in X$ and let $\mu : [0,T] \to \mathcal{P}_c(\mathbb{R}^d)$ and $\nu : [0,T] \to \mathcal{P}_c(\mathbb{R}^d)$ be two continuous maps with respect to $\mathcal{W}_1$ satisfying*

$$\mathrm{supp}(\mu(t)) \cup \mathrm{supp}(\nu(t)) \subseteq B(0,R), \tag{45}$$

for every $t \in [0,T]$, for some $R > 0$. Then for every $r > 0$ there exists a constant $L_{a,r,R}$ such that

$$\|F^{[a]} * \mu(t) - F^{[a]} * \nu(t)\|_{L_\infty(B(0,r))} \leq L_{a,r,R} \mathcal{W}_1(\mu(t), \nu(t)) \tag{46}$$

for every $t \in [0,T]$.

Proof. Fix $t \in [0,T]$ and take $\pi \in \Gamma_o(\mu(t), \nu(t))$. Since the marginals of π are by definition $\mu(t)$ and $\nu(t)$, it follows

$$\begin{aligned} F^{[a]} * \mu(t)(x) - F^{[a]} * \nu(t)(x) &= \int_{B(0,R)} F^{[a]}(x-y) d\mu(t)(y) - \int_{B(0,R)} F^{[a]}(x-z) d\nu(t)(z) \\ &= \int_{B(0,R)^2} \Big(F^{[a]}(x-y) - F^{[a]}(x-z) \Big) \, d\pi(y,z) \end{aligned}$$

By using Lemma 6.3 and the hypothesis (45), we have

$$\|F^{[a]} * \mu(t) - F^{[a]} * \nu(t)\|_{L_\infty(B(0,r))} \leq \operatorname*{ess\,sup}_{x \in B(0,r)} \int_{B(0,R)^2} \Big| F^{[a]}(x-y) - F^{[a]}(x-z) \Big| \, d\pi(y,z)$$

$$\leq \mathrm{Lip}_{B(0,R+r)}(F^{[a]}) \int_{B(0,R)^2} |y-z| d\pi(y,z)$$
$$= \mathrm{Lip}_{B(0,R+r)}(F^{[a]}) \mathcal{W}_1(\mu(t),\nu(t)),$$

hence (46) holds with $L_{a,r,R} = \mathrm{Lip}_{B(0,R+r)}(F^{[a]})$. □

6.2 Proof of Proposition 2.2

Notice that for every $N \in \mathbb{N}$, by Proposition 6.1, μ_N is the unique solution of (21) with initial datum μ_N^0. We start by fixing $N \in \mathbb{N}$ and estimating the growth of $|x_i^N(t)|^2$ for $i = 1, \dots, N$. By using Lemma 6.2, we have

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}|x_i^N(t)|^2 &\leq \dot{x}_i^N(t) \cdot x_i^N(t) \\
&\leq \left| (F^{[a]} * \mu_N(t))(x_i(t)) \right| |x_i^N(t)| \\
&\leq \|a\|_{L_\infty(\mathbb{R}_+)} \left(|x_i^N(t)| + \frac{1}{N}\sum_{j=1}^N |x_j^N(t)| \right) |x_i^N(t)| \\
&\leq 2\|a\|_{L_\infty(\mathbb{R}_+)} \max_{j=1,\dots,N} |x_j^N(t)||x_i^N(t)| \\
&\leq 2\|a\|_{L_\infty(\mathbb{R}_+)} \max_{j=1,\dots,N} |x_j^N(t)|^2.
\end{aligned}$$

If we denote by $q(t) := \max_{j=1,\dots,N} |x_j^N(t)|^2$, then the Lipschitz continuity of q implies that q is a.e. differentiable. Stampacchia's Lemma [17, Chapter 2, Lemma A.4] ensures that for a.e. $t \in [0,T]$ there exists $k = 1, \dots, N$ such that

$$\dot{q}(t) = \frac{d}{dt}|x_k^N(t)|^2 \leq 4\|a\|_{L_\infty(\mathbb{R}_+)} q(t).$$

Hence, Gronwall's Lemma and the hypothesis $x_{0,i}^N \in \mathrm{supp}(\mu^0) + \overline{B(0,1)}$ for every $N \in \mathbb{N}$ and $i = 1, \dots, N$, imply that

$$q(t) \leq q(0) e^{4\|a\|_{L_\infty(\mathbb{R}_+)} t} \leq C_0 e^{4\|a\|_{L_\infty(\mathbb{R}_+)} t} \text{ for a.e. } t \in [0,T],$$

for some uniform constant C_0 depending only on μ^0. Therefore, the trajectory $\mu_N(\cdot)$ is bounded uniformly in N in a ball $B(0,R) \subset \mathbb{R}^d$, where

$$R = \sqrt{C_0} e^{2\|a\|_{L_\infty(\mathbb{R}_+)} T}. \tag{47}$$

This, in turn, implies that $\mu_N(\cdot)$ is Lipschitz continuous with Lipschitz constant uniform in N, since by the fact that $|x_i^N(t)| \leq R$ for a.e. $t \in [0,T]$, for all $N \in N$ and $i = 1, \dots, N$, and Lemma 6.2 follows

$$|\dot{x}_i^N(t)| = |(F^{[a]} * \mu_N(t))(x_i^N(t))|$$

$$\leq \|a\|_{L_\infty(\mathbb{R}_+)} \left(|x_i^N(t)| + \frac{1}{N} \sum_{j=1}^N |x_j^N(t)| \right)$$
$$\leq 2R\|a\|_{L_\infty(\mathbb{R}_+)}.$$

We have thus found a sequence $(\mu_N)_{N\in\mathbb{N}} \subset \mathcal{C}^0([0,T], \mathcal{P}_1(B(0,R)))$ for which the following holds:

- $(\mu_N)_{N\in\mathbb{N}}$ is equicontinuous and is contained in a closed subset of $\mathcal{P}_1(B(0,R))$ equipped with the $\mathcal{W}_1$ metric, because of the uniform Lipschitz constant $2R\|a\|_{L_\infty(\mathbb{R}_+)}$;
- for every $t \in [0,T]$, the sequence $(\mu_N(t))_{N\in\mathbb{N}}$ is relatively compact in $\mathcal{P}_1(B(0,R))$ equipped with the $\mathcal{W}_1$ metric. This holds because $(\mu_N(t))_{N\in\mathbb{N}}$ is a tight sequence, since $B(0,R)$ is compact, and hence relatively compact w.r.t. weak convergence due to Prokhorov's Theorem. By [2, Proposition 7.1.5] and the uniform integrability of the first moments of the family $(\mu_N(t))_{N\in\mathbb{N}}$ follows relative compactness also in the metric space $(\mathcal{P}_1(B(0,R)), \mathcal{W}_1)$.

Therefore, we can apply the Ascoli-Arzelá Theorem for functions with values in a metric space (see for instance, [16, Chapter 7, Theorem 18]) to infer the existence of a subsequence $(\mu^{N_k})_{k\in\mathbb{N}}$ of $(\mu_N)_{N\in\mathbb{N}}$ such that

$$\lim_{k\to\infty} \mathcal{W}_1(\mu^{N_k}(t), \mu(t)) = 0 \quad \text{uniformly for a.e. } t \in [0,T], \tag{48}$$

for some $\mu \in \mathcal{C}^0([0,T], \mathcal{P}_1(B(0,R)))$ with Lipschitz constant bounded by $2R\|a\|_{L_\infty(\mathbb{R}_+)}$. The hypothesis $\lim_{N\to\infty} \mathcal{W}_1(\mu_N^0, \mu^0) = 0$ now obviously implies $\mu(0) = \mu^0$.

We are now left with verifying that this curve μ is a solution of (21). For all $t \in [0,T]$ and for all $\varphi \in \mathcal{C}_c^1(\mathbb{R}^d; \mathbb{R})$, since it holds

$$\frac{d}{dt}\langle \varphi, \mu_N(t)\rangle = \frac{1}{N}\frac{d}{dt}\sum_{i=1}^N \varphi(x_i^N(t)) = \frac{1}{N}\sum_{i=1}^N \nabla\varphi(x_i^N(t)) \cdot \dot{x}_i^N(t),$$

by directly applying the substitution $\dot{x}_i^N(t) = (F^{[a]} * \mu_N(t))(x_i^N(t))$, we have

$$\langle \varphi, \mu_N(t) - \mu_N(0)\rangle = \int_0^t \left[\int_{\mathbb{R}^d} \nabla\varphi(x) \cdot (F^{[a]} * \mu_N(s))(x) d\mu_N(s)(x) \right] ds.$$

By Lemma 6.6, the inequality (48), and the compact support of $\varphi \in \mathcal{C}_c^1(\mathbb{R}^d; \mathbb{R})$, follows

$$\lim_{N\to\infty} \|\nabla\varphi \cdot (F^{[a]} * \mu_N(t) - F^{[a]} * \mu(t))\|_{L_\infty(\mathbb{R}^d)} = 0 \quad \text{uniformly for a.e. } t \in [0,T].$$

If we denote with $\mathcal{L}_{1\llcorner[0,t]}$ the Lebesgue measure on the time interval $[0,t]$, since the product measures $\frac{1}{t}\mu^N(s) \times \mathcal{L}_{1\llcorner[0,t]}$ converge in $\mathcal{P}_1([0,t]\times\mathbb{R}^d)$ to $\frac{1}{t}\mu(s) \times \mathcal{L}_{1\llcorner[0,t]}$, we finally get from the dominated convergence theorem that

$$\lim_{N\to\infty} \int_0^t \int_{\mathbb{R}^d} \nabla\phi(x) \cdot (F^{[a]} * \mu_N(s))(x) d\mu_N(s)(x) ds$$

$$= \int_0^t \int_{\mathbb{R}^d} \nabla\phi(x) \cdot (F^{[a]} * \mu(s))(x) d\mu(s)(x) ds,$$

which proves that μ is a solution of (21) with initial datum μ^0.

6.3 Existence and uniqueness of solutions for (25)

For the reader's convenience we start by briefly recalling some general, well-known results about solutions to Carathéodory differential equations. We fix a domain $\Omega \subset \mathbb{R}^d$, a Carathéodory function $g\colon [0,T] \times \Omega \to \mathbb{R}^d$, and $0 < \tau \leq T$. A function $y\colon [0,\tau] \to \Omega$ is called a solution of the Carathéodory differential equation

$$\dot{y}(t) = g(t, y(t)) \tag{49}$$

on $[0,\tau]$ if and only if y is absolutely continuous and (49) is satisfied a.e. in $[0,\tau]$. The following existence result holds.

Theorem 6.7. *Fix $T > 0$ and $y_0 \in \mathbb{R}^d$. Suppose that there exists a compact subset Ω of $\mathbb{R}^d$ such that $y_0 \in \mathrm{int}(\Omega)$ and there exists $m_\Omega \in L_1([0,T])$ for which it holds*

$$|g(t,y)| \leq m_\Omega(t), \tag{50}$$

for a.e. $t \in [0,T]$ and for all $y \in \Omega$. Then there exists a $\tau > 0$ and a solution $y(t)$ of (49) defined on the interval $[0,\tau]$ which satisfies $y(0) = y_0$. If there exists $C > 0$ such that the function g also satisfies the condition

$$|g(t,y)| \leq C(1 + |y|), \tag{51}$$

for a.e. $t \in [0,T]$ and every $y \in \Omega$, and it holds $B(0,R) \subseteq \Omega$, for $R > |y_0| + CTe^{CT}$, then the local solution $y(t)$ of (49) which satisfies $y(0) = y_0$ can be extended to the whole interval $[0,T]$. Moreover, for every $t \in [0,T]$, any solution satisfies

$$|y(t)| \leq \Big(|y_0| + Ct\Big) e^{Ct}. \tag{52}$$

Proof. Since $y_0 \in \mathrm{int}(\Omega)$, we can consider a ball $B(y_0, r) \subset \Omega$. The classical result [12, Chapter 1, Theorem 1] and (50) yield the existence of a local solution defined on an interval $[0,\tau]$ and taking values in $B(y_0, r)$.

If (51) holds, any solution of (49) with initial datum y_0 satisfies

$$|y(t)| \leq |y_0| + Ct + C \int_0^t |y(s)|\, ds$$

for every $t \in [0,\tau]$, therefore (52) follows from Gronwall's inequality. In particular the graph of a solution $y(t)$ cannot reach the boundary of $[0,T] \times B(0, |y_0| + CTe^{CT})$ unless $\tau = T$, therefore the continuation of the local solution to a global one on $[0,T]$ follows, for instance, from [12, Chapter 1, Theorem 4]. □

Gronwall's inequality easily gives us the following results on continuous dependence on the initial data.

Lemma 6.8. *Let g_1 and $g_2\colon [0,T]\times\mathbb{R}^n \to \mathbb{R}^n$ be Carathéodory functions both satisfying* (51) *for the same constant $C>0$. Let $r>0$ and define*

$$\rho_{r,C,T} := \Big(r + CT\Big)\, e^{CT}\,.$$

Assume in addition that there exists a constant $L>0$ satisfying

$$|g_1(t,y_1) - g_1(t,y_2)| \le L|y_1 - y_2|$$

for every $t\in[0,T]$ and every y_1, y_2 such that $|y_i|\le \rho_{r,C,T}$, $i=1,2$. Then, if $\dot y_1(t) = g_1(t,y_1(t))$, $\dot y_2(t) = g_2(t,y_2(t))$, $|y_1(0)|\le r$ and $|y_2(0)|\le r$, one has

$$|y_1(t)-y_2(t)| \le e^{Lt}\left(|y_1(0)-y_2(0)| + \int_0^t \|g_1(s,\cdot)-g_2(s,\cdot)\|_{L_\infty(B(0,\rho_{r,C,T}))}\,ds\right) \tag{53}$$

for every $t\in[0,T]$.

Proof. We can bound $|y_1(t)-y_2(t)|$ from above as follows:

$$\begin{aligned}
|y_1(t)-y_2(t)| &\le |y_1(0)-y_2(0)| + \int_0^t |\dot y_1(s) - \dot y_2(s)|ds\\
&= |y_1(0)-y_2(0)|\\
&\quad + \int_0^t |g_1(s,y_1(s)) - g_1(s,y_2(s)) + g_1(s,y_2(s)) - g_2(s,y_2(s))|ds\\
&\le |y_1(0)-y_2(0)| + \int_0^t \|g_1(s,\cdot)-g_2(s,\cdot)\|_{L_\infty(B(0,\rho_{r,C,T}))}\,ds\\
&\quad + L\int_0^t |y_1(s)-y_2(s)|ds.
\end{aligned}$$

Since the function $\alpha(t) = |y_1(0)-y_2(0)| + \int_0^t \|g_1(s,\cdot)-g_2(s,\cdot)\|_{L_\infty(B(0,\rho_{r,C,T}))}\,ds$ is increasing, an application of Gronwall's inequality gives (53), as desired. □

Proposition 6.9. *Fix $T>0$, $a\in X$, $\mu^0\in\mathcal{P}_c(\mathbb{R}^d)$, $\xi_0\in\mathbb{R}^d$ and let $R>0$ be given by Proposition 2.2 from the choice of T, a and μ^0. For every map $\mu:[0,T]\to\mathcal{P}_1(\mathbb{R}^d)$ which is continuous with respect to $\mathcal{W}_1$ such that*

$$\operatorname{supp}(\mu(t)) \subseteq B(0,R) \quad \text{for every } t\in[0,T],$$

there exists a unique solution of system (25) *with initial value μ^0 defined on the whole interval $[0,T]$.*

Proof. By Lemma 6.2 follows that, for any compact set $K \subset \mathbb{R}^d$ containing ξ_0, there exists a function $m_K \in L_1([0,T])$ for which the function $g(t,y) = (F^{[a]} * \mu(t))(y)$ satisfies (50). Moreover, for fixed t this function is locally Lipschitz continuous, as follows from Lemma 6.4, thus $g(t,y) = (F^{[a]} * \mu(t))(y)$ is a Carathéodory function.

From the hypothesis that the support of μ is contained in $B(0,R)$ and Lemma 6.2, follows the existence of a constant C depending on T, a and μ^0 such that

$$|(F^{[a]} * \mu(t))(y)| \leq C(1 + |y|)$$

holds for every $y \in \mathbb{R}^d$ and for every $t \in [0,T]$. Hence $F^{[a]} * \mu(t)$ is sublinear and (51) holds. By considering a sufficiently large compact set K containing ξ_0, Theorem 6.7 guarantees the existence of a solution of system (25) defined on $[0,T]$.

To establish uniqueness notice that, from Lemma 6.3, for every compact subset $K \in \mathbb{R}^d$ and any $x, y \in K$, it holds

$$\begin{aligned} |(F^{[a]} * \mu(t))(x) - (F^{[a]} * \mu(t))(y)| &\leq \left| \int_{\mathbb{R}^d} F^{[a]}(x-z) d\mu(t)(z) - \int_{\mathbb{R}^d} F^{[a]}(y-z) d\mu(t)(z) \right| \\ &\leq \int_{\mathbb{R}^d} \left| F^{[a]}(x-z) - F^{[a]}(y-z) \right| d\mu(t)(z) \\ &\leq \mathrm{Lip}_{\widehat{K}}(F^{[a]}) |x-y|, \end{aligned} \tag{54}$$

where $\widehat{K}$ is a compact set containing both K and $B(0,R)$. Hence, uniqueness follows from (54) and Lemma 6.8 by taking $g_1 = g_2$, $y_1(0) = y_2(0)$ and $r = |y_1(0)|$. □

6.4 Continuous dependence on the initial data

The following Lemma and (53) are the main ingredients of the proof of Theorem 2.4 on continuous dependance on initial data.

I'm confused by r and R in the Lemma below and its proof. Do we need r? Should $L_{a,R,R}$ be $L_{a,r,R}$?

Lemma 6.10. *Let $\mathcal{T}_1$ and $\mathcal{T}_2 \colon \mathbb{R}^n \to \mathbb{R}^n$ be two bounded Borel measurable functions. Then, for every $\mu \in \mathcal{P}_1(\mathbb{R}^n)$ one has*

$$\mathcal{W}_1((\mathcal{T}_1)_\# \mu, (\mathcal{T}_2)_\# \mu) \leq \|\mathcal{T}_1 - \mathcal{T}_2\|_{L_\infty(\mathrm{supp}\,\mu)}.$$

If in addition $\mathcal{T}_1$ is locally Lipschitz continuous, and μ, $\nu \in \mathcal{P}_1(\mathbb{R}^n)$ are both compactly supported on a ball $B(0,r)$ of $\mathbb{R}^n$, then

$$\mathcal{W}_1((\mathcal{T}_1)_\# \mu, (\mathcal{T}_1)_\# \nu) \leq \mathrm{Lip}_{B(0,r)}(E_1) \mathcal{W}_1(\mu, \nu).$$

Proof. See [7, Lemma 3.11] and [7, Lemma 3.13]. □

We can now prove Theorem 2.4.

Proof of Theorem 2.4. Let $\mathcal{T}_t^\mu$ and $\mathcal{T}_t^\nu$ be the flow maps associated to system (25) with measure μ and ν, respectively. By (26), the triangle inequality, Lemma 6.6, Lemma 6.10 and (27) we have for every $t \in [0,T]$

$$\begin{aligned}\mathcal{W}_1(\mu(t),\nu(t)) &= \mathcal{W}_1((\mathcal{T}_t^\mu)_\#\mu^0, (\mathcal{T}_t^\nu)_\#\nu_0)\\ &\le \mathcal{W}_1((\mathcal{T}_t^\mu)_\#\mu^0, (\mathcal{T}_t^\mu)_\#\nu_0) + \mathcal{W}_1((\mathcal{T}_t^\mu)_\#\nu_0, (\mathcal{T}_t^\nu)_\#\nu_0) \qquad (55)\\ &\le e^{T\,\mathrm{Lip}_{B(0,R)}(F^{[a]})}\mathcal{W}_1(\mu^0,\nu_0) + \|\mathcal{T}_t^\mu - \mathcal{T}_t^\nu\|_{L_\infty(B(0,R))}.\end{aligned}$$

Using (53) with $y_1(0) = y_2(0)$ we get

$$\|\mathcal{T}_t^\mu - \mathcal{T}_t^\nu\|_{L_\infty(B(0,r))} \le e^{t\,\mathrm{Lip}_{B(0,R)}(F^{[a]})} \int_0^t \|F^{[a]} * \mu(s) - F^{[a]} * \nu(s)\|_{L_\infty(B(0,R))}\, ds. \quad (56)$$

Combining (55) and (56) with Lemma 6.6, we have

$$\mathcal{W}_1(\mu(t),\nu(t)) \le e^{T\,\mathrm{Lip}_{B(0,R)}(F^{[a]})}\left(\mathcal{W}_1(\mu^0,\nu_0) + L_{a,R,R}\int_0^t \mathcal{W}_1(\mu(s),\nu(s))\, ds\right)$$

for every $t \in [0,T]$, where $L_{a,R,R}$ is the constant from Lemma 6.6. Gronwall's inequality now gives

$$\mathcal{W}_1(\mu(t),\nu(t)) \le e^{T\,\mathrm{Lip}_{B(0,R)}(F^{[a]})+L_{a,R,R}}\mathcal{W}_1(\mu^0,\nu_0),$$

which is exactly (29) with $\overline{C} = e^{T\,\mathrm{Lip}_{B(0,R)}(F^{[a]})+L_{a,R,R}}$.

Consider now two solutions of (21) with the same initial datum μ^0. Since, from Proposition 2.2 they both satisfy (28) for the given *a priori known* R given by (47), then (29) guarantees they both describe the same curve in $\mathcal{P}_1(\mathbb{R}^d)$. This concludes the proof. □

References

[1] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems.*, volume 254. Clarendon Press Oxford, 2000.

[2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures.* Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[3] P. Binev, A. Cohen, W. Dahmen, and R. DeVore. Universal algorithms for learning theory. II. Piecewise polynomial functions. *Constr. Approx.*, 26(2):127–152, 2007.

[4] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov. Universal algorithms for learning theory. I. Piecewise constant functions. *J. Mach. Learn. Res.*, 6:1297–1321, 2005.

[5] M. Bongini, M. Fornasier, M. Hansen, and M. Maggioni. Inferring interaction rules from observations of evolutive systems II: The universal learning approach. *in preparation*, 2016.

[6] A. Bressan and B. Piccoli. *Introduction to the mathematical theory of control*, volume 2 of *AIMS Series on Applied Mathematics*. American Institute of Mathematical Sciences (AIMS), Springfield, MO, 2007.

[7] J. Cañizo, J. Carrillo, and J. Rosado. A well-posedness theory in measures for some kinetic models of collective motion. *Math. Models Methods Appl. Sci.*, 21(3):515–539, 2011.

[8] J. A. Carrillo, Y.-P. Choi, and M. Hauray. The derivation of swarming models: Mean-field limit and Wasserstein distances. In *Collective Dynamics from Bacteria to Crowds: An Excursion Through Modeling, Analysis and Simulation Series*, volume 553, pages 1–46. CISM International Centre for Mechanical Sciences, 2014.

[9] J. A. Carrillo, M. Fornasier, G. Toscani, and F. Vecil. Particle, kinetic, and hydrodynamic models of swarming. In *Mathematical modeling of collective behavior in socio-economic and life sciences*, pages 297–336. Springer, 2010.

[10] F. Cucker and S. Smale. Emergent behavior in flocks. *IEEE Trans. Automat. Control*, 52(5):852–862, 2007.

[11] G. Dal Maso. *An introduction to* Γ*-convergence*. Progress in Nonlinear Differential Equations and their Applications, 8. Birkhäuser Boston, Inc., Boston, MA, 1993.

[12] A. Filippov. *Differential equations with discontinuous right-hand sides*. Kluwer Academic Publishers, 1988.

[13] M. Fornasier and J.-C. Hütter. Consistency of probability measure quantization by means of power repulsion-attraction potentials. Submitted, 2015.

[14] M. Fornasier and F. Solombrino. Mean-field optimal control. *ESAIM Control Optim. Calc. Var.*, 20(4):1123–1152, 2014.

[15] J. E. Herbert-Reada, A. Pernab, R. P. Mannb, T. M. Schaerfa, D. J. T. Sumpterb, and A. J. W. Warda. Inferring the rules of interaction of shoaling fish. *PNAS*, 108(46):18726?18731, 2011.

[16] J. L. Kelley. *General topology*. Springer-Verlag, 1955.

[17] D. Kinderlehrer and G. Stampacchia. *An Introduction to Variational Inequalities and their Applications*. Academic Press, New York, NY, 1980.

[18] R. Mann. Bayesian inference for identifying interaction rules in moving animal groups. *PLoS ONE*, 6(8):e22827. doi:10.1371/journal.pone.0022827, 2011.

[19] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027, 2010.

[20] A. T. Vicsek, E. Czirók, O. Ben-Jacob, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.*, 75(6):1226–1229, 1995.

[21] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.