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 their 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 that 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 means of several numerical experiments.

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

Contents

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, like tissues and 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. [28] and

Cucker-Smale [13], 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 relatively little effort in the applied mathematics community towards matching data from real-life cases.

In this paper we aim at bridging in the specific setting of first order models, the well-developed theory of dynamical systems and mean-field equations with classical approaches of approximation theory, nonlinear time series analysis, and machine learning. We provide a mathematical framework for the reliable identification of the governing forces 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 governing laws of evolution for specific models in social dynamics governed by nonlocal interactions. The models considered in the scope of this paper are deterministic, however we intend in follow up work to extend our results towards stochastic dynamical systems.

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 nonlocally interacting particles

Let us introduce an example of the general framework described above. It is actually the main focus of this paper. Assume that $x = (x_1, \dots, x_N) \in \mathcal{H} \equiv \mathbb{R}^{d\times N}$ and that

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

where $A : \mathbb{R}_+ \to \mathbb{R}$ is a suitable nonlinear interaction kernel function, which, for simplicity we assume to be smooth (see below more precise conditions), 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

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

Under suitable assumptions of local Lipschitz continuity and boundedness of the interaction function

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

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, see [2] and Proposition 2.2 below, that for $N \to \infty$ a mean-field approximation holds: if the initial conditions $x_i(0)$ 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 $\mu^N(t) = \frac{1}{N}\sum_{i=1}^N \delta_{x_i(t)}$ weakly converges for $N \to \infty$ to the probability measure-valued trajectory $t \mapsto \mu(t)$ satisfying the equation

$$\partial_t \mu(t) = -\nabla \cdot ((F^{[a]} * \mu(t))\mu(t)), \quad \mu(0) = \mu_0, \tag{4}$$

in weak sense, where $F^{[a]}(z) = -a(|z|)z = -A'(|z|)/|z|$, 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 [10, 11] for recent overviews on this subject. Despite the tremendous theoretical success of such research direction in terms of mathematical results on well-posedness and asymptotic behavior of solutions, 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 return to consider again an abstract energy $\mathcal{J}^{[a]}$ and let us assume that it is dependent on a parameter function a, as indicated in the superscript. 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 abstract energy, but also the corresponding evolutions $t \mapsto x^{[a]}(t)$ driven according to (1), for fixed initial conditions $x^{[a]}(0) = x_0$. (Here we assume that the class of a is such that the evolutions exist and they are essentially well-posed; we explicitly stress again the dependency on a with a superscript $[a]$, which below we may remove as soon as such dependency is clear from the context.) The fundamental question to be here 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 relatively little is done in the specific directions we mentioned in the example above. We refer, for instance, to [3, 4, 12, 24, 26, 23] and references therein, for groundbreaking 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 perhaps would be naturally considered at a first instance, and focus for a moment on the gradient flow model (1). Given a certain gradient flow evolution $t \mapsto x^{[a]}(t)$ depending on the unknown parameter function a, one might decide to design the recovery of a as an optimal control problem [8]: 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 $t \mapsto x^{[\widehat{a}]}(t)$ the solution of gradient flow (1) for $\mathcal{J} = \mathcal{J}^{[\widehat{a}]}$, i.e.,

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

and $\mathcal{R}(\cdot)$ is a suitable regularization functional, 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 $t \mapsto x^{[\widehat{a}]}(t)$ 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 these 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 number of particles N is significantly large (the well-known term *curse of dimensionality* conied by Richard E. Bellman for optimal control problems).

1.5 A variational approach towards learning parameter functions in nonlocal energies

Let us now consider again the more specific 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 interaction 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

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

Instead of the nonconvex optimal control problem above, we propose an alternative, direct approach which is both computationally very efficient and guarantees accurate approximations under reasonable assumptions. In particular, we consider as an estimator of the kernel a 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 competitor functions $\widehat{a} \in X$. Actually, the minimization of $\mathcal{E}^{[a],N}$ has a close connection to the optimal control problem, as it also promotes the minimization of the discrepancy $\mathrm{dist}_{\mathcal{H}}(x^{[a]}(s) - x^{[\widehat{a}]}(s))^2$ in (5) (here we remind that in this setting $\mathcal{H}$ is the Euclidean space $\mathbb{R}^{d\times N}$):

Proposition 1.1. *If $a, \widehat{a} \in X$ then there exist a constant $C > 0$ depending on T, a, and $x^{[a]}(0)$ such that*

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

for all $t \in [0,T]$, where $x^{[a]}$ and $x^{[\widehat{a}]}$ are the solutions of (7) for the interaction kernels a and $\widehat{a}$ respectively. (Here $\|x\|^2 = \frac{1}{N}\sum_{i=1}^N |x_i|^2$, 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 system (7) with interaction kernel $\widehat{a}$ instead of a are as well a good approximation of the trajectories $t \mapsto x^{[a]}(t)$ at finite time. The proof of this statement follows by Jensen's inequality and an application of Gronwall's lemma, as reported in detail in Section 3.

For simplicity of notations, we may choose to ignore below the dependence on a of the trajectories, and write $x \equiv x^{[a]}$ when such a dependence is clear from the context. Additionally, whenever we consider the limit $N \to \infty$, we may denote the dependency of the trajectory on the number of particles $N \in \mathbb{N}$ by setting $x^N \equiv x \equiv x^{[a]}$.

Contrary to the optimal control approach, the functional $\mathcal{E}^{[a],N}$ is convex and can be easily computed from witnessed trajectories $x_i(t)$ and $\dot{x}_i(t)$. We may even consider discrete-time approximations of the time derivatives $\dot{x}_i$ (e.g., by finite differences) and we shall assume that the data of the problem is the full set of observations $x_i(t)$ 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 mathematical question 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 [14], 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 clear characterization of the limits for the sequence of minimizers $(\widehat{a}_{N,V})_{N\in\mathbb{N},V\in\Lambda}$. Recalling again that $F^{[a]}(z) = -a(|z|)z$, for $z \in \mathbb{R}^d$, we rewrite the functional (8) as follows:

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

for $\mu^N(t) = \frac{1}{N}\sum_{i=1}^N \delta_{x_i(t)}$. This formulation of the functional makes it easy to recognize that the candidate for a Γ-limit 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 $x_i(0)$ are identically and independently distributed according to a compactly supported probability measure $\mu(0) = \mu_0$.

Although all of this is very natural, 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 and the identification of a can take place. As the trajectories $t \mapsto x(t)$ do not explore the whole space in finite time, we expect that such a space may *not* be independent of the initial probability measure μ_0, as we clarify immediately. By Jensen inequality we have

$$\begin{aligned} \mathcal{E}^{[a]}(\widehat{a}) &\leq \frac{1}{T}\int_0^T \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |\widehat{a}(|x-y|) - a(|x-y|)|^2 |x-y|^2 d\mu(t)(x)d\mu(t)(y)dt \quad (13) \\ &= \frac{1}{T}\int_0^T \int_{\mathbb{R}_+} \left|\widehat{a}(s) - a(s)\right|^2 s^2 d\varrho(t)(s)dt \quad (14) \end{aligned}$$

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 for any compact set A (see Lemma 3.1). We may therefore define a time-averaged probability measure $\overline{\rho}$ on the Borel σ-algebra of $\mathbb{R}_+$ by averaging $\varrho(t)$ over $t \in [0,T]$: 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 any Borel set $A \subseteq \mathbb{R}_+$, to take into account the polynomial weight s^2 as appearing in (14). Then one can reformulate (14) in a very compact form 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 essential to assume that there exists $c_T > 0$ such that also the following additional 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 non-degenerate 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 $t \mapsto \mu(t)$ which allow to validate (18).
In order to ensure compactness of the sequence of minimizers of $\mathcal{E}^{[a],N}$, we shall need to restrict the sets of possible solutions to classes of the type

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

where $M > 0$ is some predetermined constant and $K \subset \mathbb{R}_+$ is a suitable compact set.

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 $\mathcal{E}^{[a],N}$ over V_N converge to minimizers of $\mathcal{E}^{[a]}$.

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 $x^N_{0,1}, \dots, x^N_{0,N}$ be i.i. μ_0-distributed and define $\mathcal{E}^{[a],N}$ as in (11) *for the solution μ^N of the equation* (4) *with initial datum*

$$\mu^N_0 = \frac{1}{N} \sum_{i=1}^N \delta_{x^N_{0,i}}.$$

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 $(\widehat{a}_N)_{N\in\mathbb{N}}$ has a subsequence converging uniformly on K to some continuous function $\widehat{a} \in X_{M,K}$ such that $\mathcal{E}^{[a]}(\widehat{a}) = 0$.
If we additionally assume the coercivity condition (18)*, then $\widehat{a} = a$ in $L_2(\mathbb{R}_+, \rho)$. Moreover, in this latter case, if there exist rates $\alpha, \beta > 0$, constants $C_1, C_2 > 0$, and a sequence $(a_N)_{N\in\mathbb{N}}$ of elements $a_N \in V_N$ such that*

$$\|a - a_N\|_{L_\infty(K)} \leq C_1 N^{-\alpha}, \tag{19}$$

and

$$\mathcal{W}_1(\mu_0^N, \mu_0) \leq C_2 N^{-\beta}, \tag{20}$$

then there exists a constant $C_3 > 0$ *such that*

$$\|a - \widehat{a}_N\|^2_{L_2(\mathbb{R}_+,\rho)} \leq C_3 N^{-\min\{\alpha,\beta\}}, \tag{21}$$

for all $N \in \mathbb{N}$. *In particular, in this case, it is the entire sequence* $(\widehat{a}_N)_{N\in\mathbb{N}}$ *(and not only subsequences) to converge to* a *in* $L_2(\mathbb{R}_+, \rho)$.

We remark that the $L_2(\mathbb{R}_+, \rho)$ used in our results is useful when ρ has positive density on large intervals of $\mathbb{R}_+$. Notice that the main result, under the validity of the coercivity condition, not only ensures the identification of a on the support of ρ, but it also provides a prescribed rate of convergence. For functions a in $X_{M,K}$ and for finite element spaces V_N of continuous piecewise linear functions constructed on regular meshes of size N^{-1} a simple sequence $(a_N)_{N\in\mathbb{N}}$ realizing (19) with $\alpha = 1$ and $C_1 = M$ is the piecewise linear approximation to a which interpolates a on the mesh nodes. For the approximation estimate (20) there are plenty of results concerning such rates and we refer to [16] and references therein. Roughly speaking, for μ_0^N the empirical measure obtained by sampling N times independently from μ_0, the bound (20) holds with high probability for a certain N for β of order $1/d$ (more precisely see [16, Theorem 1]), which is a manifestation of the aforementioned curse of dimensionality. While it is in general relatively easy to increase α as the smoothness a increases, and doing so independently of d, since a is a function of one variable only, obtaining $\beta > 1/d$ is in general not possible unless μ_0 has very special properties, see [18, Section 4.4 and Section 4.5].

1.6 Numerical implementation of the variational approach

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 the uniform convergence, which is quite strong. However, the condition that the spaces V_N are to be picked as subsets of $X_{M,K}$ requires the prior knowledge of $M \geq \|a\|_{L_\infty(K)} + \|a'\|_{L_\infty(K)}$. Hence, the finite dimensional optimization (10) is not anymore a simple *unconstrained* least squares (as claimed in the paragraph before (10)), but a problem constrained by a uniform bound on both the solution and its gradient. Nevertheless, as we clarify in Section 5, for $M > 0$ fixed and choosing V_N made of piecewise linear continuous functions, imposing the uniform L_∞ bounds in the least square problem does not constitute a severe difficulty. Also the tuning of the parameter $M > 0$ turns out to be rather simple. In fact, for N fixed the minimizers $\widehat{a}_N \equiv \widehat{a}_{N,M}$ have the property that the map

$$M \mapsto \mathcal{E}^{[a],N}(\widehat{a}_{N,M})$$

is monotonically decreasing as a function of the constraint parameter M and it becomes constant for $M \geq M^*$, for $M^* > 0$ empirically not depending on N. We claim that this special value M^* is indeed the "right" parameter for the L_∞ bound. For such a

choice, we show also numerically that, as expected, if we let N grow, the minimizers $\widehat{a}_N$ approximates better and better the unknown potential a.

Despite the fact that both the tuning of $M > 0$ and the constrained minimization over $X_{M,K}$ requiring L_∞ bounds are not severe issues, it would be way more efficient to perform a unconstrained least squares over X. In our follow-up paper [7] we extend the approach developed by DeVore et al. in [6, 5] towards universal algorithms for learning regression functions from independent samples drawn according to an unknown probability distribution. This extension presents several challenges including the lack of independence of the samples collected in our framework and the nonlocality of the scalar products of the corresponding least squares. This has a price to pay, i.e., that the spaces V_N need to be carefully chosen and the result of convergence holds only with high probability. For the development of the latter results, we need to address several variational and measure theoretical properties of the model which are considered in details in this first paper as reported below.

2 Preliminaries

2.1 Optimal transport and Wasserstein distances

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 μ 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 $\mathcal{P}_p(\mathbb{R}^n)$ which consists of all probability measures with compact support. For any $\mu \in \mathcal{P}(\mathbb{R}^{n_1})$ and a 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 p_1 and p_2 defined on the product space $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, for every $\pi \in \mathcal{P}(\mathbb{R}^{n_1} \times \mathbb{R}^{n_2})$ we call *first* (resp., *second*) *marginal* of π the probability measure $p_{1\#}\pi$ (respectively, $p_{2\#}\pi$). 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,

$$\mathcal{W}_p^p(\mu, \nu) = \inf_{\pi \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^{2n}} |x - y|^p d\pi(x, y)\,. \tag{22}$$

If $p = 1$, we have the following equivalent characterization of the 1-Wasserstein distance:

$$\mathcal{W}_1(\mu, \nu) = \sup \left\{ \int_{\mathbb{R}^n} \varphi(x) d(\mu - \nu)(x) : \varphi \in \mathrm{Lip}(\mathbb{R}^n),\ \mathrm{Lip}_{\mathbb{R}^n}(\varphi) \le 1 \right\}, \tag{23}$$

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.,

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

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 (22) is actually a minimum. For more details, see e.g. [2, 29].

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 $|f(\xi)| \le C(1+|\xi|)$ for all $\xi \in \mathbb{R}^d$.

2.2 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

$$\begin{cases} \dfrac{\partial \mu}{\partial t}(t) = -\nabla \cdot ((F^{[a]} * \mu(t))\mu(t)) \quad \text{for } t \in (0,T], \\ \mu(0) = \mu_0. \end{cases} \tag{24}$$

We consequently give our notion of solution for (24).

Definition 2.1. We say that a map $\mu : [0,T] \to \mathcal{P}_1(\mathbb{R}^d)$ is a solution of (24) 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 (24) 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 equation (24) is closely related to the family of ODEs, indexed by $N \in \mathbb{N}$,

$$\begin{cases} \dot{x}_i^N(t) = \dfrac{1}{N}\displaystyle\sum_{j=1}^N F^{[a]}(x_i^N(t) - x_j^N(t)) \quad \text{for } t \in (0,T], \\ x_i^N(0) = x_{0,i}^N, \end{cases} \quad i = 1,\dots,N, \tag{25}$$

which may be rewritten as

$$\begin{cases} \dot{x}_i^N(t) = (F^{[a]} * \mu^N(t))(x_i^N(t)) \\ x_i^N(0) = x_{0,i}^N, \end{cases} \qquad i = 1, \dots, N, \tag{26}$$

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

$$\mu^N(t) = \frac{1}{N} \sum_{j=1}^{N} \delta_{x_j^N(t)}. \tag{27}$$

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

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

The well-posedness of (26) is rather standard under the assumption $a \in X$. The well-posedness of system (24) 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 (24), and to [10] for generalizations in case of interaction kernels not necessarily belonging to the class X. Nevertheless, in the following we recall 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*

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

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 (27) where $(x_1^N, \dots, x_N^N)$ is the unique solution of system (25).

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, up to extraction of subsequences, in $\mathcal{P}_1(B(0, R))$ equipped with the Wasserstein metric $\mathcal{W}_1$ to a solution μ of (24) 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 6.3 together with the necessary technical lemmas in Appendix 6.2.

2.3 The transport map and uniqueness of mean-field solutions

Another way for building a solution of equation (24) 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 (26),

$$\begin{cases} \dot{\xi}(t) = (F^{[a]} * \mu(t))(\xi(t)) \quad \text{for } t \in (0,T], \\ \xi(0) = \xi_0, \end{cases} \tag{28}$$

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 $\mathrm{supp}(\mu(t)) \subseteq B(0,R)$, for a given $R > 0$.

According to Proposition 6.10, if μ is any solution of (24), 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 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 (28) with initial datum ξ_0. The by now well-known result [9, Theorem 3.10] shows that the solution of (24) with initial value μ_0 is also the unique fixed-point of the *push-foward map*

$$\Gamma[\mu](t) := (\mathcal{T}_t^\mu)_\# \mu_0. \tag{29}$$

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. Let $R > 0$ be sufficiently large such that $\mathrm{supp}(\mu_0) \subseteq B(0,R)$. The choice $r = R$ in Proposition 6.3, Lemma 6.5, and Lemma 6.6 imply the following stability estimate

$$\left|\mathcal{T}_t^\mu(x_0) - \mathcal{T}_t^\mu(x_1)\right| \le e^{T\,\mathrm{Lip}_{B(0,R)}(F^{[a]})}|x_0 - x_1|, \quad \text{for } |x_i| \le R\,, \quad i = 0,1\,. \tag{30}$$

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

In view of the uniqueness of the solutions to the ODE (28), 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

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

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

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

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, see, e.g., [9], that one has also uniqueness and continuous dependence on the initial data for (24) (we report a proof of these properties in Appendix 6.3 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 (24), for $\mu(0) = \mu_0$ and $\nu(0) = \nu_0$ respectively. Let $R > 0$ be such that for every $t \in [0,T]$*

$$\mathrm{supp}(\mu(t)) \cup \mathrm{supp}(\nu(t)) \subseteq B(0,R)\,. \tag{31}$$

Then, there exist a positive constant $\overline{C}$ depending only on T, a, and R such that

$$\mathcal{W}_1(\mu(t),\nu(t)) \le \overline{C}\,\mathcal{W}_1(\mu_0,\nu_0) \tag{32}$$

for every $t \in [0,T]$. In particular, equi-compactly supported solutions of (24) 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 (25) 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: roughly speaking we choose $\widehat{a}_N \in X$ as a minimizer of the following *discrete error functional*

$$\mathcal{E}^{[a],N}(\widehat{a}) = \frac{1}{T}\int_0^T \frac{1}{N}\sum_{i=1}^N \left| \frac{1}{N}\sum_{j=1}^N \Big(\widehat{a}(|x_i^{[a]}(t) - x_j^{[a]}(t)|)(x_i^{[a]}(t) - x_j^{[a]}(t)) - \dot{x}_i^{[a]}(t)\Big)\right|^2 dt. \tag{33}$$

Let us remind that, by Proposition 1.1, this optimization guarantees also that any minimizer $\widehat{a}_N$ produces very good trajectory approximations $x^{[\widehat{a}_N]}(t)$ to the "true" ones $x^{[a]}(t)$ at least at finite time $t \in [0,T]$.

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

$$\begin{aligned}
\|x(t) - \widehat{x}(t)\|^2 &= \left\| \int_0^t (\dot{x}(s) - \dot{\widehat{x}}(s))ds \right\|^2 \le t\int_0^t \|\dot{x}(s) - \dot{\widehat{x}}(s)\|^2 ds \\
&= t\int_0^t \frac{1}{N}\sum_{i=1}^N \left| (F^{[a]} * \mu^N(x_i) - F^{[\widehat{a}]} * \widehat{\mu}^N(\widehat{x}_i))\right|^2 ds \\
&\le 2t\int_0^t \left[\frac{1}{N}\sum_{i=1}^N \left| (F^{[a]} - F^{[\widehat{a}]}) * \mu^N(x_i))\right|^2 \right.
\end{aligned}$$

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

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

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

which is the desired bound. □

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 $f\in\mathcal{C}(\mathbb{R}_+)$, it holds $f\circ d\in\mathcal{C}(\mathbb{R}^d\times\mathbb{R}^d)$, and hence

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

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

$$\overline{\rho}(A):=\frac{1}{T}\int_0^T\varrho(t)(A)dt. \tag{34}$$

Notice that Lemma 3.1 shows that (34) 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 (60). The measure $\mu(t)$ is the pushforward of μ_0 under the locally bi-Lipschitz map $\mathcal{T}_t^\mu$, see Proposition 2.3. 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 (34). □

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. *Let $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$. Then 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 (60), 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 it 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 (34). For example, one obtains

$$\frac{1}{T}\int_0^T \delta(t)dt = \frac{1}{T}\mathcal{L}_{1\llcorner[0,T]}\,,$$

as becomes immediately clear when integrating a continuous function against those kind of measures.

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 $A \subset \mathbb{R}_+$. An easy consequence of Lemma 3.5 is that if $a \in X$, then

$$\|a\|^2_{L_2(\mathbb{R}_+,\rho)} = \int_{\mathbb{R}_+} \big|a(s)\big|^2 d\rho(s) \leq \mathrm{diam}(\mathrm{supp}(\rho))\|a\|^2_{L_\infty(\mathrm{supp}(\rho))}\,, \tag{35}$$

and therefore $X \subseteq L_2(\mathbb{R}_+,\rho)$. As already mentioned in the introduction, for $N \to \infty$ a natural mean-field approximation to the learning functional is given by

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

where $\mu(t)$ is a weak solution to (24). By means of ρ, we recall the estimate from the Introduction

$$\mathcal{E}^{[a]}(\widehat{a}) \leq \frac{1}{T}\int_0^T \int_{\mathbb{R}_+} \big|\widehat{a}(s) - a(s)\big|^2 s^2 d\varrho(t)(s)dt = \|\widehat{a} - a\|^2_{L_2(\mathbb{R}_+,\rho)}. \tag{36}$$

This inequality suggested in turn 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 $\mathcal{E}^{[a]}$ in X coincides ρ-a.e. with a.*

Proof. Notice that $\mathcal{E}^{[a]}(a) = 0$, and since $\mathcal{E}^{[a]}(\widehat{a}) \geq 0$ for all $\widehat{a} \in X$ this implies that a is a minimizer of $\mathcal{E}^{[a]}$. 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, i.e., $c_T = 0$, in a "generic" case the coercivity inequality holds. On the one hand, we show that if we could model the misfit $\mathcal{K}(r) = (a(r) - \widehat{a}(r))r$ to behave randomly, in a sufficiently independent manner, over a finite set of trajectory distances, then the coercivity condition holds with high probability. While the needed independence assumptions will typicallybe too strong to be applicable in practice, the arguments we provide are by far not the most general possible, and we view them as one possible notion of a "generic" case. On the other hand, in the next section we also present a more rigorous *deterministic* argument to verify the coercivity condition for very particular choices of a.

With the notation of the misfit just introduced, the coercivity condition reads

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

If the inequality holds without the time average for a fixed t_0,

$$\int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \mathcal{K}(|x-y|)\frac{x-y}{|x-y|} d\mu(t_0)(x) \right|^2 d\mu(t_0)(y) \geq c'_{t_0} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |\mathcal{K}(|x-y|)|^2 \, d\mu(t_0)(x) d\mu(t_0)(y)\,,$$

then by a continuity argument it can be extended to a nontrivial time interval. We will therefore freeze time and investigate the inequality at this fixed time. Additionally, for the moment we restrict our attention to the case where $\mu(t_0)$ is a discrete measure $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ (we drop t_0 since it is now fixed), so that the inequality reads

$$\frac{1}{N}\sum_{i=1}^N \left| \frac{1}{N}\sum_{j=1}^N \mathcal{K}(|x_i - x_j|)\frac{x_i - x_j}{|x_i - x_j|} \right|^2 \geq \frac{c'_{t_0}}{N^2}\sum_{i=1}^N\sum_{j=1}^N |\mathcal{K}(|x_i - x_j|)|^2 \,. \tag{37}$$

We argue now that this ("instantaneous") inequality holds with high probability as soon as the matrix $\mathbf{K} := (\mathcal{K}(|x_i - x_j|))_{i,j=1,\dots,N}$ is modeled as a random matrix. Although it is not completely plausible to argue statistical independence of the entries of such a matrix because it comes from evaluating a smooth function over distances of non-random points, this model is not completely unreasonable: after all $\mathbf{K}$ involves the difference of our estimator $\widehat{a}$ and the target influence function a. For least squares estimators this difference is random with the samples used to construct the estimator, often with nearly independent, perhaps even Gaussian, fluctuations. We assume that $\mathbf{K}$ has independent 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, with a constant $c'_{t_0} = O(1/N)$. Let $\mathbf{X}_i \in \mathbb{R}^{N\times d}$ be

the matrix whose j-th row is the (fixed) vector $\frac{x_i - x_j}{|x_i - x_j|} \in \mathbb{R}^d$, and let $\mathbf{K}(i,:) \in \mathbb{R}^N$ be the i-th row of $\mathbf{K}$. The coercivity inequality (37) may be re-written as:

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{N}\mathbf{K}(i,:)\mathbf{X}_i\right|^2 \geq \frac{c_t'}{N^2}||\mathbf{K}||_{\mathbb{F}}^2\,. \tag{38}$$

Then we estimate

$$\begin{aligned}
\mathbb{E}\left[|\mathbf{K}(i,:)\mathbf{X}_i|^2\right] &= \sum_{l=1}^{d}\sum_{j,j'=1}^{N}\mathbb{E}\left[\mathcal{K}(|x_i - x_j|)\mathcal{K}(|x_i - x_{j'}|)|\right]\left(\frac{x_i - x_j}{|x_i - x_j|}\right)_l\left(\frac{x_i - x_{j'}}{|x_i - x_{j'}|}\right)_l \\
&= \sum_{l=1}^{d}\sum_{j=1}^{N}\mathbb{E}\left[\mathcal{K}(|x_i - x_j|)^2\right]\left(\frac{x_i - x_j}{|x_i - x_j|}\right)_l^2 \\
&= \sum_{j=1}^{N}|\mathbf{X}_i(j,:)|^2 = ||\mathbf{X}_i||_{\mathbb{F}}^2 = N\,,
\end{aligned}$$

where we used independence, and in the last step we used the fact that every row of $\mathbf{X}_i$ is a unit vector. By concentration one readily obtains that with high probability

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{N}\mathbf{K}(i,:)\mathbf{X}_i\right|^2 \geq \frac{C}{N}\,.$$

One the other hand, since $\mathbb{E}[||\mathbf{K}||_{\mathbb{F}}^2] \leq CN^2$ by standard random matrix theory results (e.g. [27]), and in fact not just in expectation but also with high probability, the right hand side of (38) is bounded by $c_{t_0}'C$ from above. Choosing c_{t_0}' small enough (and at least as small as $O(1/N)$, as a function of N), we obtain (38) with high-probability.

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, sufficiently uncorrelated with $\mathbf{K}$, 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 "generically" satisfied. A model where the behavior of the coercivity constant c_t' would be quite different as N grows, is the following: we assume that $\mathcal{K}(|x_i - x_j|)$ is distributed as $\frac{\eta_{ij}}{|x_i - x_j|^\alpha}$, where η_{ij} are i.i.d. standard normal distributions, and furthermore we assume that as N grows the quantity $\frac{1}{N}\sum_{j=1}^{N}|x_i - x_j|^{-\alpha}$ grows as $N^{\gamma-1}$, for some $\gamma \geq 1$, and for every $i = 1, \ldots, N$ fixed. Repeating the calculation above we obtain that the coercivity condition holds with constant that scales as $O(N^{\gamma-1})$, in particular is $O(1)$ independently of N for $\gamma = 1$. The first assumption may be motivated that estimators of the influence function may have performance proportional to the gradient of the influence function itself, and such gradient may decay with distance; the second assumption is about the scaling of the "bulk" of the system as N grows: for $\gamma = 0$ such size is independent of N,

for $\gamma > 0$ it grows with N. Note that the case $\gamma = 1$ is indeed very natural: the quantity $\frac{1}{N}\sum_{j=1}^{N} |x_i - x_j|^{-\alpha}$ is expected to approach the corresponding integral in the mean-field limit, which is independent of N. Under this natural scaling, the coercivity constant is independent of N, suggesting it holds in the limit as well.

3.2.2 The deterministic case

We construct now deterministic examples of trajectories $t \to \mu(t)$ for which the coercivity condition (18) holds. We start with the simple case of two particles, i.e., $N = 2$, for which no specific assumptions on $a, \widehat{a}$ are required to verify (18) other than their boundedness in 0. Again it is convenient to write $\mathcal{K}(r) = (a(r) - \widehat{a}(r))r$, so that the coercivity condition in this case can be reformulated as

$$\frac{1}{T}\int_0^T \frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{N}\sum_{j=1}^{N}\mathcal{K}(|x_i - x_j|)\frac{x_i - x_j}{|x_i - x_j|}\right|^2 dt \geq \frac{c_T}{N^2 T}\int_0^T \sum_{i=1}^{N}\sum_{j=1}^{N}|\mathcal{K}(|x_i - x_j|)|^2 dt. \tag{39}$$

Now, let us observe more closely the integrand on the left-hand-side, and for $\widehat{i} \neq i$, $i, \widehat{i} \in \{1, 2\}$ and $N = 2$, and we obtain

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^{2}\left|\frac{1}{2}\sum_{j=1}^{2}\mathcal{K}(|x_i - x_j|)\frac{x_i - x_j}{|x_i - x_j|}\right|^2 &= \frac{1}{2}\sum_{i=1}^{2}\left|\frac{1}{2}\sum_{j\neq i}^{2}\mathcal{K}(|x_i - x_j|)\frac{x_i - x_j}{|x_i - x_j|}\right|^2 \\
&= \frac{1}{4}\sum_{i=1}^{2}\left|\mathcal{K}(|x_i - x_{\widehat{i}}|)\frac{x_i - x_{\widehat{i}}}{|x_i - x_{\widehat{i}}|}\right|^2 \\
&= \frac{1}{4}\sum_{i=1}^{2}\left|\mathcal{K}(|x_i - x_{\widehat{i}}|)\right|^2 = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}|\mathcal{K}(|x_i - x_j|)|^2.
\end{aligned}$$

Integrating over time the latter equality yields (39) for $N = 2$ with an actual equal sign and $c_T = 1$. Notice that here we have not made any specific assumptions on the trajectories $t \mapsto x_i(t)$. Let us then consider the case of $N = 3$ particles. Already in this simple case the angles between particles may be rather arbitrary and analyzing the many possible configurations becomes an involved exercise. (Notice that we circumvented this problem in the random model in section 3.2.1 thanks to the assumed independence of the entries of the rows of $\mathbf{K}$.) To simplify the problem we assume that $d = 2$ and that at a certain time t the particles are disposed precisely at the vertexes of a equilateral triangle of edge length r. This makes the computation of the angles very simple. We also assume that $\mathcal{K}$ gets its maximal absolute value precisely at r, hence

$$\frac{1}{9}\sum_{i=1}^{3}\sum_{j=1}^{3}|\mathcal{K}(|x_i - x_j|)|^2 \leq \|\mathcal{K}\|_\infty^2 = \mathcal{K}(r)^2.$$

Notice that, independently of the behavior of the particles at any other time $t \in [0,T]$, it holds also

$$\frac{1}{9T}\int_0^T \sum_{i=1}^3\sum_{j=1}^3 |\mathcal{K}(|x_i - x_j|)|^2 dt \leq \|\mathcal{K}\|_\infty^2 = \mathcal{K}(r)^2. \tag{40}$$

A direct computation in this case of particles disposed at the vertexes of a equilateral triangle shows that

$$\frac{1}{3}\sum_{i=1}^3 \left| \frac{1}{3}\sum_{j=1}^3 \mathcal{K}(|x_i - x_j|)\frac{x_i - x_j}{|x_i - x_j|}\right|^2 = \frac{1}{3}\mathcal{K}(r)^2,$$

and therefore

$$\frac{1}{3}\sum_{i=1}^3 \left| \frac{1}{3}\sum_{j=1}^3 \mathcal{K}(|x_i - x_j|)\frac{x_i - x_j}{|x_i - x_j|}\right|^2 \geq \frac{1}{18}\sum_{i=1}^3\sum_{j=1}^3 |\mathcal{K}(|x_i - x_j|)|^2.$$

Unfortunately the assumption that $\mathcal{K}$ achieves its maximum in absolute value at r does not allow us yet to conclude by a simple integration over time the coercivity condition as we did for the case of two particles. In order to extend the validity of the inequality to arbitrary functions taking maxima at other points, we need to integrate over time by assuming now that the particles are vertexes of equilateral triangles with time dependent edge length, say from $r = 0$ growing in time up to $r = 2R > 0$. This will allow the trajectories to explore any possible distance within a given interval and to capture the maximal absolute value of any kernel. More precisely, let us now assume that $\mathcal{K}$ is an arbitrary bounded continuous function, achieving its maximal absolute value over $[0, 2R]$, say at $r_0 \in (0, 2R)$ and we can assume that this is obtained corresponding to the time t_0 when the particles form precisely the equilateral triangle of side length r_0. Now we need to make a stronger assumption on $\widehat{a}$, i.e., we require $\widehat{a}$ to belong to a class of equi-continuous functions, for instance functions which are Lipschitz continuous with uniform Lipschitz constant (such as the functions in $X_{M,K}$). Under this equi-continuity assumption, there exist $\varepsilon > 0$ and a constant $c_{T,\varepsilon} > 0$ independent of $\mathcal{K}$ (but perhaps depending only on its modulus of continuity) such that

$$\begin{aligned}
&\frac{1}{T}\int_0^T \frac{1}{3}\sum_{i=1}^3 \left| \frac{1}{3}\sum_{j=1}^3 \mathcal{K}(|x_i - x_j|)\frac{x_i - x_j}{|x_i - x_j|}\right|^2 dt \\
\geq\ & \frac{1}{T}\int_{t_0-\varepsilon}^{t_0+\varepsilon} \frac{1}{3}\sum_{i=1}^3 \left| \frac{1}{3}\sum_{j=1}^3 \mathcal{K}(|x_i - x_j|)\frac{x_i - x_j}{|x_i - x_j|}\right|^2 dt \\
\geq\ & \frac{c_{T,\varepsilon}}{3}\mathcal{K}(r_0) \geq \frac{c_{T,\varepsilon}}{18T}\int_0^T \sum_{i=1}^3\sum_{j=1}^3 |\mathcal{K}(|x_i - x_j|)|^2 dt.
\end{aligned}$$

In the latter inequality we used (40). Hence, also in this case, one can construct examples for which the coercivity assumption is verifiable. Actually this construction can be extended to any group of N particles disposed on the vertexes of regular polygons. As an example of how one should proceed, let us consider the case of $N = 4$ particles disposed instantanously at the vertexes of a square of side length $\sqrt{2}r > 0$. In this case one directly verfies that

$$\frac{1}{4}\sum_{i=1}^{4}\left|\frac{1}{4}\sum_{j=1}^{4}\mathcal{K}(|x_i - x_j|)\frac{x_i - x_j}{|x_i - x_j|}\right|^2 = \frac{1}{16}(\mathcal{K}(2r) + \sqrt{2}\mathcal{K}(\sqrt{2}r))^2. \tag{41}$$

Let us assume that the maximal absolute value of $\mathcal{K}$ in absolute value is attained precisely at $\sqrt{2}r$. Then the minimum of the expression on the right-hand side of (41) is attained for the case where $\mathcal{K}(2r) = -\mathcal{K}(\sqrt{2}r)$ yielding the following estimate from below

$$\frac{1}{4}\sum_{i=1}^{4}\left|\frac{1}{4}\sum_{j=1}^{4}\mathcal{K}(|x_i - x_j|)\frac{x_i - x_j}{|x_i - x_j|}\right|^2 \geq \frac{3 - 2\sqrt{2}}{16}\mathcal{K}(\sqrt{2}r)^2.$$

Hence, also in this case, we can apply the continuity argument above to eventually show the coercivity condition. Similar procedures can be followed for any $N \geq 5$. However, as $N \to \infty$ one can show numerically that the lower bound vanishes quite rapidly, making it impossible, perhaps not surprisingly, to conclude the coercivity condition for the uniform distribution over the circle.

All the examples presented so far are based on discrete measures $\mu^N = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$ supported on particles lying on the vertexes of polytopes. However, one can consider an approximated convolution identity g_ε for which $g_\varepsilon \to \delta_0$ for $\varepsilon \to 0$, where δ_0 is a Dirac delta in 0, and the regularized probability measure

$$\mu_\varepsilon(x) = g_\varepsilon * \mu^N(x) = \frac{1}{N}\sum_{i=1}^{N} g_\varepsilon(x - x_i).$$

This diffuse measure approximates μ^N in the sense that $\mathcal{W}_1(\mu_\varepsilon, \mu^N) \to 0$ for $\varepsilon \to 0$, hence, in particular, integrals against Lipschitz functions can be well-approximated, i.e.,

$$\left|\int_{\mathbb{R}^d}\varphi(x)d\mu^N(x) - \int_{\mathbb{R}^d}\varphi(x)d\mu_\varepsilon(x)\right| \leq \mathrm{Lip}(\varphi)\mathcal{W}_1(\mu_\varepsilon, \mu^N).$$

Under the additional assumption that $\mathrm{Lip}_K(\widehat{a}) \sim \|\widehat{a}\|_{L_\infty(K)}$ (and this is true whenever $\widehat{a}$ is a piecewise polynomial function over a finite partition of $\mathbb{R}_+$, with the constant of the equivalence depending on the particular partition) one can extend the validity of the coercivity condition for μ^N (39) to μ_ε as follows

$$\frac{1}{T}\int_0^T\int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d}\mathcal{K}(|x - y|)\frac{y - x}{|y - x|}d\mu_\varepsilon(x)\right|^2 d\mu_\varepsilon(y)dt$$

$$\geq \frac{c_{T,\varepsilon}}{T} \int_0^T \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\mathcal{K}(|x-y|)|^2 d\mu_\varepsilon(x) d\mu_\varepsilon(y) dt,$$

for a constant $c_{T,\varepsilon} > 0$ for $\varepsilon > 0$ small enough.

In these latter sections we showed that the coercivity condition holds for "generic" cases as well as for highly structured deterministic ones. In practice we can numerically verify that it holds in many situations, see Section 5.3, and from now on we assume it without further concerns.

3.3 Existence of minimizers of $\mathcal{E}^{[a],N}$

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 $\mathcal{E}^{[a],N}$.

Proposition 3.7. *Fix $M > 0$ and $K = [0, 2R] \subset \mathbb{R}_+$ for any $R > 0$. Recall 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 (applicable for functions in W^1_∞, see [1, Theorem 2.8]) implies that the functions $\widehat{a}_n$ are all Lipschitz continuous with uniformly bounded Lipschitz constant, and are therefore equicontinuous. Since they are also pointwise uniformly equi-bounded, by Ascoli-Arzelá Theorem there exists 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*

$$\min_{\widehat{a}\in V} \mathcal{E}^{[a],N}(\widehat{a})$$

admits a solution.

Proof. For proving the statement we apply the direct method of calculus of variations. Since $\inf \mathcal{E}^{[a],N} \geq 0$, we can consider a minimizing sequence $(\widehat{a}_n)_{n\in\mathbb{N}} \subset V$, i.e., such that $\lim_{n\to\infty} \mathcal{E}^{[a],N}(\widehat{a}_n) = \inf_V \mathcal{E}^{[a],N}$. 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 $\lim_{n\to\infty} \mathcal{E}^{[a],N}(\widehat{a}_n) = \mathcal{E}^{[a],N}(\widehat{a})$, from which it follows that $\mathcal{E}^{[a],N}$ 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 (60)). Moreover, we have the uniform bound

$$\begin{aligned}\sup_{x,y\in B(0,R)}|F^{[\widehat{a}_n]}(x-y)-F^{[a]}(x-y)| &= \sup_{x,y\in B(0,R)}|\widehat{a}_n(|x-y|)-a(|x-y|)||x-y| \\ &\le 2R\sup_{r\in K}|\widehat{a}_n(r)-a(r)| \\ &\le 2R(M+\|a\|_{L_\infty(K)}).\end{aligned} \tag{42}$$

As the measures $\mu^N(t)$ are compactly supported in $B(0,R)$ uniformly in time, the boundedness (42) allows us to apply three times the dominated convergence theorem to yield

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

which proves the statement.

□

4 Γ-convergence of $\mathcal{E}^{[a],N}$ to $\mathcal{E}^{[a]}$

This section is devoted to a proof of Theorem 1.3.

4.1 Uniform convergence estimates

We start with a technical result.

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 and $b\in X_{M,K}$, for $K=[0,2R]$ with $R>0$ as in (60). Then we have the estimate*

$$\left|\mathcal{E}^{[a],N}(b_N)-\mathcal{E}^{[a]}(b)\right|\le c_1\mathcal{W}_1(\mu_0^N,\mu_0)+c_2\|b_N-b\|_{L_\infty(K)}, \tag{43}$$

where the constants are explicitly given by $c_1=32\overline{C}R(2R+1)M^2$ and $c_2=16R^2M$.

Proof. By (32), $\mathcal{W}_1(\mu(t),\mu^N(t))\le\overline{C}\mathcal{W}_1(\mu_0,\mu_0^N)$ uniformly in $t\in[0,T]$. For all $x,y,y'\in B(0,R)$, by the triangle inequality we have

$$|(F^{[a]}-F^{[b]})(x-y')-(F^{[a]}-F^{[b]})(x-y)|$$

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

which implies the Lipschitz continuity of $(F^{[a]}-F^{[b]})(x-\cdot)$ in $B(0,R)$ for fixed $x\in B(0,R)$. Since $a,b\in X_{M,K}$, this implies

$$\mathrm{Lip}_{B(0,R)}|(F^{[a]}-F^{[b]})(x-\cdot)|\leq 2(2R+1)M, \tag{44}$$

uniformly with respect to $x\in B(0,R)$. Consequently, we have

$$\begin{aligned}&\left|\int_{\mathbb{R}^d}\big(F^{[b]}-F^{[a]}\big)(x-y)d\mu^N(t)(y)-\int_{\mathbb{R}^d}\big(F^{[b]}-F^{[a]}\big)(x-y)d\mu(t)(y)\right|\\ &\qquad\leq \mathrm{Lip}_{B(0,R)}|(F^{[a]}-F^{[b]})(x-\cdot)|\,\mathcal{W}_1(\mu^N(t),\mu(t))\leq 2\overline{C}(2R+1)M\,\mathcal{W}_1(\mu_0^N,\mu_0),\end{aligned}$$

uniformly with respect to $t\in[0,T]$ and $x\in B(0,R)$. Furthermore, we also have

$$\sup_{x,y\in B(0,R)}|F^{[b_N]}(x-y)-F^{[b]}(x-y)| \;\leq\; 2R\|b_N-b\|_{L_\infty(K)}, \tag{45}$$

$$\sup_{x,y\in B(0,R)}|F^{[a]}(x-y)-F^{[b]}(x-y)| \;\leq\; 2R\|a-b\|_{L_\infty(K)}. \tag{46}$$

Hence we further obtain

$$\begin{aligned}&\left|\left|\int_{\mathbb{R}^d}\big(F^{[b_N]}-F^{[a]}\big)(x-y)d\mu^N(t)(y)\right|-\left|\int_{\mathbb{R}^d}\big(F^{[b]}-F^{[a]}\big)(x-y)d\mu(t)(y)\right|\right| \qquad (47)\\ &\leq\left|\int_{\mathbb{R}^d}\big(F^{[b_N]}-F^{[a]}\big)(x-y)d\mu^N(t)(y)-\int_{\mathbb{R}^d}\big(F^{[b]}-F^{[a]}\big)(x-y)d\mu(t)(y)\right|\\ &\leq\left|\int_{\mathbb{R}^d}\big(F^{[b_N]}-F^{[b]}\big)(x-y)d\mu^N(t)(y)\right|\\ &\qquad+\left|\int_{\mathbb{R}^d}\big(F^{[b]}-F^{[a]}\big)(x-y)d\mu^N(t)(y)-\int_{\mathbb{R}^d}\big(F^{[b]}-F^{[a]}\big)(x-y)d\mu(t)(y)\right|\\ &\leq 2R\|b_N-b\|_{L_\infty(K)}\int_{\mathbb{R}^d}d\mu^N(t)(y)+2(2R+1)M\,\mathcal{W}_1(\mu^N(t),\mu(t))\\ &\leq 2R\|b_N-b\|_{L_\infty(K)}+2\overline{C}(2R+1)M\,\mathcal{W}_1(\mu_0^N,\mu_0).\end{aligned}$$

Let

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

$$H(t,x)=\left|\int_{\mathbb{R}^d}\big(F^{[b]}-F^{[a]}\big)(x-y)d\mu(t)(y)\right|^2,\qquad G(t)=\int_{\mathbb{R}^d}H(t,x)d\mu(t)(x).$$

Then immediately it follows

$$|G_N(t)-G(t)|\leq\left|\int_{\mathbb{R}^d}H(t,x)d\mu^N(t)(x)-\int_{\mathbb{R}^d}H(t,x)d\mu(t)(x)\right|$$

$$+ \int_{\mathbb{R}^d} |H_N(t,x) - H(t,x)| \, d\mu^N(t)(x). \tag{48}$$

From (46) and (44) we obtain

$$\begin{aligned} \mathrm{Lip}_{B(0,R)} H(t,\cdot) &\leq 2 \left(\sup_{x,y \in B(0,R)} |F^{[a]}(x-y) - F^{[b]}(x-y)| \right) \cdot \mathrm{Lip}_{B(0,R)} (F^{[a]} - F^{[b]})(\cdot - y) \\ &\leq 4R \|a-b\|_{L_\infty(K)} \cdot 2(2R+1)M \,, \end{aligned}$$

and therefore

$$\begin{aligned} \left| \int_{\mathbb{R}^d} H(t,x) d\mu^N(t)(x) - \int_{\mathbb{R}^d} H(t,x) d\mu(t)(x) \right| &\leq \mathrm{Lip}_{B(0,R)} H(t,\cdot) \mathcal{W}_1(\mu^N(t), \mu(t)) \\ &\leq 8R(2R+1)\overline{C}M \|a-b\|_{L_\infty(K)} \mathcal{W}_1(\mu_0^N, \mu_0) \\ &\leq 16R(2R+1)\overline{C}M^2 \mathcal{W}_1(\mu_0^N, \mu_0) \end{aligned} \tag{49}$$

uniformly in $t \in [0,T]$. Similarly, (45), (44), and (47) imply

$$\begin{aligned} \big|H_N(t,x) - H(t,x)\big| &\leq \Big(2R \|b_N - b\|_{L_\infty(K)} + 2\overline{C}(2R+1)M \, \mathcal{W}_1(\mu_0^N, \mu_0) \Big) \\ &\qquad \times 2R \Big(\|b_N - a\|_{L_\infty(K)} + \|b-a\|_{L_\infty(K)} \Big) \\ &\leq 8RM \Big(2R \|b_N - b\|_{L_\infty(K)} + 2\overline{C}(2R+1)M \, \mathcal{W}_1(\mu_0^N, \mu_0) \Big) \end{aligned} \tag{50}$$

uniformly in $t \in [0,T]$ and $x \in B(0,R)$. A combination of (48) with (49) and (50) yields

$$|G_N(t) - G(t)| \leq 32\overline{C}R(2R+1)M^2 \mathcal{W}_1(\mu_0^N, \mu_0) + 16R^2 M \|b_N - b\|_{L_\infty(K)}$$

uniformly in $t \in [0,T]$. Thus we finally arrive at

$$\begin{aligned} \big|\mathcal{E}^{[a],N}(b_N) - \mathcal{E}^{[a]}(b)\big| &= \left| \frac{1}{T} \int_0^T \big(G_N(t) - G(t) \big) dt \right| \\ &\leq 32\overline{C}R(2R+1)M^2 \mathcal{W}_1(\mu_0^N, \mu_0) + 16R^2 M \|b_N - b\|_{L_\infty(K)}. \end{aligned}$$

This proves the claim.

□

As a corollary, we now immediately obtain the following convergence result.

Lemma 4.2. *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* (60)*. Then it holds*

$$\lim_{N\to\infty} \mathcal{E}^{[a],N}(b_N) = \mathcal{E}^{[a]}(b).$$

Proof. This follows immediately from the estimate (43), upon noticing $\mathcal{W}_1(\mu_0, \mu_0^N) \to 0$ for $N \to \infty$ as a consequence of the Glivenko-Cantelli theorem, see for instance [19, Lemma 3.3]. □

4.2 Proof of the main result

We are now ready to present the proof of our main result Theorem 1.3.

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 $N \in \mathbb{N}$. Lemma 4.2 implies

$$\lim_{N\to\infty} \mathcal{E}^{[a],N}(b_N) = \mathcal{E}^{[a]}(b),$$

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

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

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

$$\mathcal{E}^{[a]}(b) \geq \mathcal{E}^{[a]}(\widehat{a})\,. \tag{51}$$

In particular, (51) applies to $b = a \in X_{M,K}$ (by the particular choice of M), which finally implies

$$0 = \mathcal{E}^{[a]}(a) \geq \mathcal{E}^{[a]}(\widehat{a}) \geq 0 \text{ or } \mathcal{E}^{[a]}(\widehat{a}) = 0,$$

showing that $\widehat{a}$ is also a minimizer of $\mathcal{E}^{[a]}$. When the coercivity condition (18) holds, it follows that $\widehat{a} = a$ in $L_2(\mathbb{R}_+, \rho)$. Assume now that (19) and (20) hold together with (18). Then, by these latter conditions, two applications of (43), the minimality of $\widehat{a}_N$, and the optimality of a in the sense that $\mathcal{E}^{[a]}(a) = 0$, we obtain the following chain of estimates

$$\begin{aligned}
\|\widehat{a}_N - a\|^2_{L_2(\mathbb{R}_+,\rho)} &\leq \frac{1}{c_T}\mathcal{E}^{[a]}(\widehat{a}_N) \\
&\leq \frac{1}{c_T}\Big(\mathcal{E}^{[a],N}(\widehat{a}_N) + (\mathcal{E}^{[a]}(\widehat{a}_N) - \mathcal{E}^{[a],N}(\widehat{a}_N))\Big) \\
&\leq \frac{1}{c_T}\Big(\mathcal{E}^{[a],N}(\widehat{a}_N) + c_1\mathcal{W}_1(\mu_0^N, \mu_0)\Big) \\
&\leq \frac{1}{c_T}\Big(\mathcal{E}^{[a],N}(a_N) + c_1\mathcal{W}_1(\mu_0^N, \mu_0)\Big) \\
&\leq \frac{1}{c_T}\big(2c_1\mathcal{W}_1(\mu_0^N, \mu_0) + c_2\|a - a_N\|_{L_\infty(K)}\big) \\
&\leq C_3 N^{-\min\{\alpha,\beta\}}.
\end{aligned}$$

This concludes the proof. □

5 Numerical experiments

In this section we report several numerical experiments to document the validity and applicability of Theorem 1.3. We will first show how the reconstruction of the unknown kernel a 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 at least some interaction kernels not lying in the function space X, as shown in Figure 2. We will then investigate empirically the validity of the coercivity condition (18) comparing the functional $\mathcal{E}^{[a],N}(\widehat{a}_N)$ with $\|a-\widehat{a}_N\|^2_{L_2(\mathbb{R}_+,\rho^N)}$ where ρ^N is constructed as ρ but referring to the empirical measures μ^N. Then we address the behavior of $\mathcal{E}^{[a],N}(\widehat{a}_{N,M})$ for N fixed, while we let the constraint constant M vary (here $\widehat{a}_{N,M} \equiv \widehat{a}_N$). Finally, we show how we can get a very satisfactory reconstruction of the unknown interaction kernel by keeping N fixed and averaging the minimizers of the functional $\mathcal{E}^{[a],N}$ obtained from several samples of the initial data distribution μ_0.

5.1 Numerical framework

All experiments 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*

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

which is the time-discrete counterpart of the continuous-time error functional $\mathcal{E}^{[a],N}$. As already mentioned in the introduction, derivatives $\dot{x}_i(t_k)$ appearing in $\mathcal{E}^{[a],N}_{\Delta}$ are actually approximated as well by finite differences: in our experiments we will use the simplest approximation

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

Regarding the reconstruction procedure, we fix the constraint level $M>0$ and consider the sequence of invading subspaces V_N of $X_{M,K}$ ($K=[0,2R]$ here) generated by a B-spline basis with $D(N)$ elements supported on $[0,2R]$: for every element $\widehat{a}\in V_N$ it holds

$$\widehat{a}(r) = \sum_{\lambda=1}^{D(N)} a_\lambda \varphi_\lambda(r), \qquad r\in[0,2R].$$

In order for V_N to increase in N and invade $X_{M,K}$, we let $D(N)$ be a strictly increasing function of N. For the sake of simplicity, we shall employ a *linear uniform* B-spline basis supported on the interval $[0, 2R]$ with 0-smoothness conditions at the boundary, see [15].

Whenever $\widehat{a} \in V_N$, we can rewrite the functional $\mathcal{E}_\Delta^{[a],N}$ as

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

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

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

We shall numerically implement the constrained minimization with the software CVX [22, 21], 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

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

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 $\operatorname{supp}(\varphi_\lambda) \cap \operatorname{supp}(\varphi_{\lambda+j}) \neq \emptyset$ holds if and only if $j = 1$. Hence, for every $a \in V_N$ we have

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

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

where, in the last line, $\mathbf{D}$ is the standard finite difference matrix

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

We therefore replace the constrained minimization problem

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

by

$$\min_{\vec{a}\in\mathbb{R}^{D(N)}} \frac{1}{mN} \|\mathbf{C}\vec{a} - \vec{v}\|_2^2 \quad \text{subject to} \quad 2\|\vec{a}\|_\infty + \|\mathbf{D}\vec{a}\|_\infty \leq M\,, \tag{52}$$

which has weaker constraints, but is amenable to numerical solution. The byproduct of the time discretization and the reformulation of the constraint is that minimizers of problem (52) may not be precisely the minimizers of the original one. This is the price to pay for this simple numerical implementation of the L_∞-constraints. Despite such a crude discrete model, we still observe all the approximation properties proved in the previous sections and the implementation results both simple and effective.

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. Table 1 reports the values of the different parameters.

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.

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 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. In particular, the algorithm performs an excellent approximation despite the highly oscillatory nature of the function a and produce a natural numerical homogeneization when the discretization is not fine enough.

5.3 Numerical validation of the coercivity condition

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 $\mathcal{E}_\Delta^{[a],N}(\widehat{a}_N)$ and of the $L_2(\mathbb{R}_+, \rho^N)$-error $\|a - \widehat{a}_N\|^2_{L_2(\mathbb{R}_+,\rho^N)}$ 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.

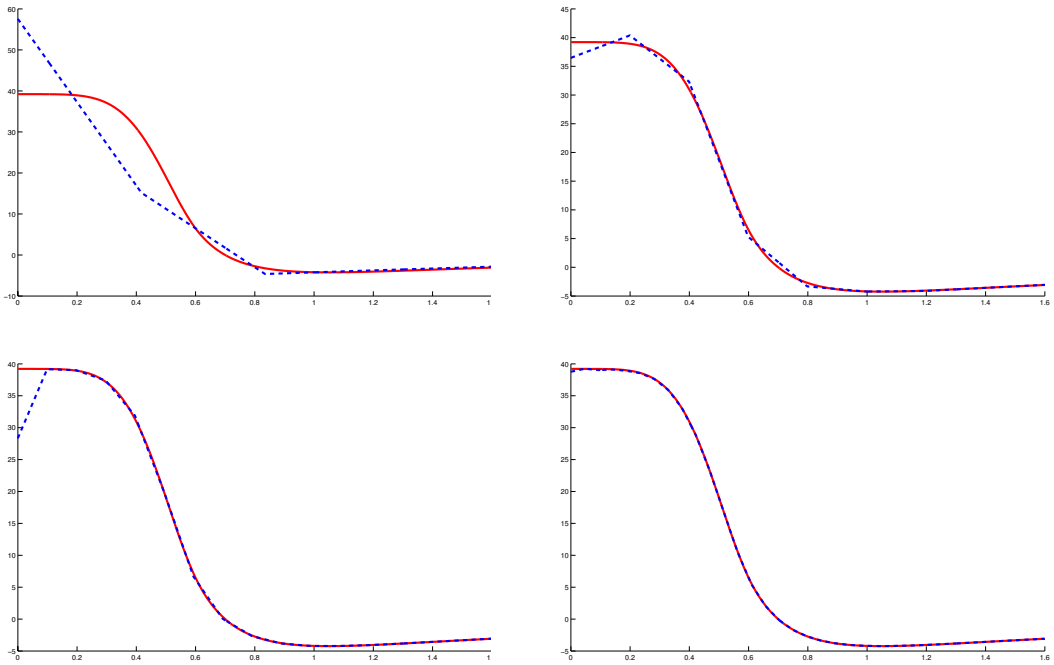

Figure 1: Iterative reconstruction of a potential with different values of N. In red: the unknown kernel. In blue: its reconstruction by minimization of $\mathcal{E}^{[a],N}$. From left-top to right-bottom: reconstruction with $N = 10, 20, 40, 80$ agents. We notice that the uniform convergence at 0 is slower in view of the quadratic polynomial weight s^2 as in (16) and because less information is actually gathered around 0.

For every value of N, we have obtained the minimizer $\widehat{a}_N$ of problem (52) and we have computed the errors $\mathcal{E}^{[a],N}(\widehat{a}_N)$ and $\|a-\widehat{a}_N\|^2_{L_2(\mathbb{R}_+,\rho^N)}$. The $L_2(\mathbb{R}_+,\rho^N)$-error multiplied by a factor $\frac{1}{10}$ lies entirely below the curve of $\mathcal{E}^{[a],N}(\widehat{a}_N)$, which let us empirically estimate the value of c_T around that value (see Figures 5.3 and 5.3).

5.4 Tuning the constraint M

Figure 5 shows what happens when we modify the value of M in problem (52). 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 (52) for several values of M and we denote with $\widehat{a}_M \equiv \widehat{a}_{N,M} \equiv \widehat{a}_N$ the minimizer obtained with a specific value of M. On the left side 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 side we illustrate how the original trajectories (again, in white) used for the inverse problem are approximated better and better by those generated with the computed approximation $\widehat{a}_M$, if we let M grow. Table 3 reports the values of the parameters of

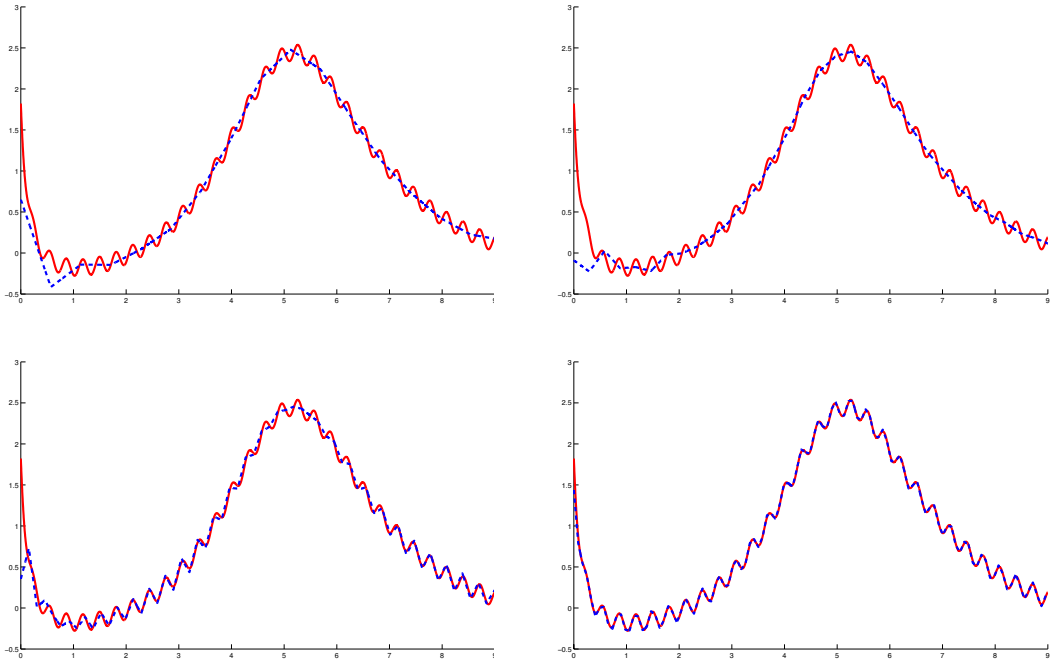

Figure 2: Iterative reconstruction of a potential with a singularity at the origin and highly oscillatory behavior. In red: the unknown kernel. In blue: its reconstruction by minimization of $\mathcal{E}^{[a],N}$. From left-top to right-bottom: reconstruction with $N = 10, 20, 40, 80$ agents.

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

Table 2: Parameter values for Figure 3.

these experiments.

So far we have no *a priori* criteria to sieve those values of M, which enable a successful reconstruction of a potential $a \in X$. However, the tuning *a posteriori* of the parameter $M > 0$ turns out to be rather easy. In fact, for N fixed the minimizers $\widehat{a}_{N,M}$ have the property that the map

$$M \mapsto \mathcal{E}^{[a],N}(\widehat{a}_{N,M})$$

is monotonically decreasing as a function of the constraint parameter M and it becomes constant for $M \geq M^*$, for $M^* > 0$ which, as shown empirically, does not depend on N. This special value M^* is indeed the "right" parameter for the L_∞ bound. For such a choice, we show that, if we let N grow, the minimizers $\widehat{a}_N$ approximates better and better the unknown potential a. Figure 6 documents precisely this expected behavior.

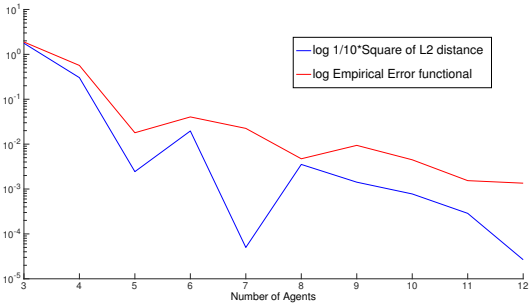

Figure 3: Plot in logarithmic scale of $\mathcal{E}^{[a],N}(\widehat{a}_N)$ and $\frac{1}{10}\|a-\widehat{a}_N\|^2_{L_2(\mathbb{R}_+,\rho^N)}$ for different values of N. In this experiment, we can estimate the constant c_T with the value $\frac{1}{10}$.

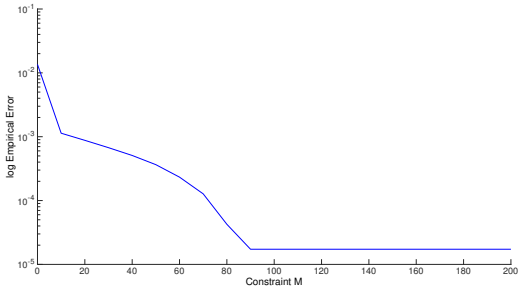

Figure 4: Values in logarithmic scale of $\mathcal{E}^{[a],N}_{\Delta}(\widehat{a}_{N,M})$ for fixed $N=50$ for different values of $M\in[0,200]$.

	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.

5.5 Montecarlo-like reconstructions for N fixed

We mimic now the mean-field reconstruction strategy, by multiple randomized draw of N particles as initial conditions i.i. distributed according to μ_0 for N fixed and relatively small. Indeed, problem (52) can swiftly become computationally unfeasible when N is moderately large, also because the dimension of the approximating subspaces V_N needs to increase with N too. We consider, for a fixed N, several discrete initial data $(\mu^N_{0,\theta})_{\theta=1}^{\Theta}$ all independently drawn from the same distribution μ_0 (in our case, the d-dimensional cube $[-L,L]^d$). For every $\theta=1,\dots,\Theta$, we simulate the system until time T and, with the trajectories we obtained, we solve problem (52). At the end of this procedure, we have a family of reconstructed potentials $(\widehat{a}_{N,\theta})_{\theta=1}^{\Theta}$, all approximating the same true kernel a. Empirically averaging these potentials, we obtain an approximation

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

which we claim to be a better approximation to the true kernel a than any single snapshots. To support this claim, we report in Figure 7 the outcome of an experiment whose data can be found in Table 4.

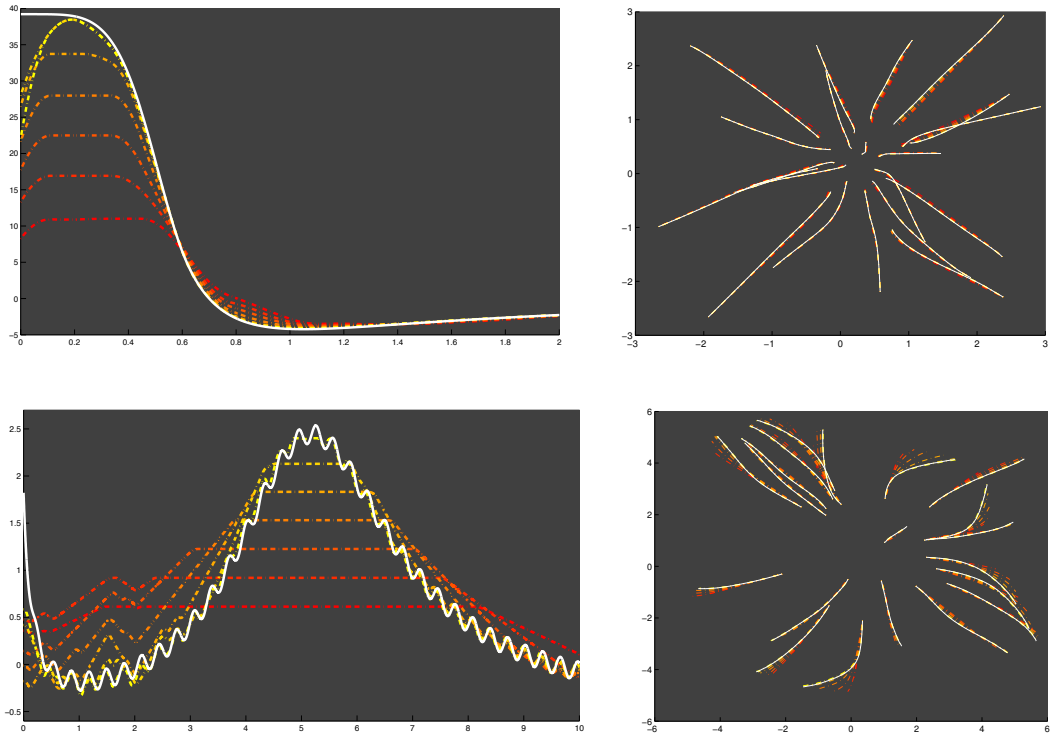

Figure 5: Different reconstructions of a potential for different values of M. On the left column: the true kernel in white and its reconstructions for different M; the brighter the curve, the larger the M. On the right column: the true trajectories of the agents in white, the trajectories associated to the reconstructed potentials with the same color.

Acknowledgement

Mattia Bongini, Massimo Fornasier, and Markus Hansen acknowledge the financial support of the ERC-Starting Grant (European Research Council, 306274) High-Dimensional Sparse Optimal Control (HDSPCONTR). Mauro Maggioni acknowledges the support of ONR-N00014-12-1-0601 and NSF-ATD/DMS-12-22567.

6 Appendix

Although similar results on the limit relationship between ODE systems of the type (25) and their mean-field equations (24) appear in different forms in other papers, see, e.g., [2, 9, 10, 20], in this Appendix we collect them for our specific setting in a nutshell for the sake of being self-contained and for the convenience of those readers less familiar with these properties of evolutive systems.

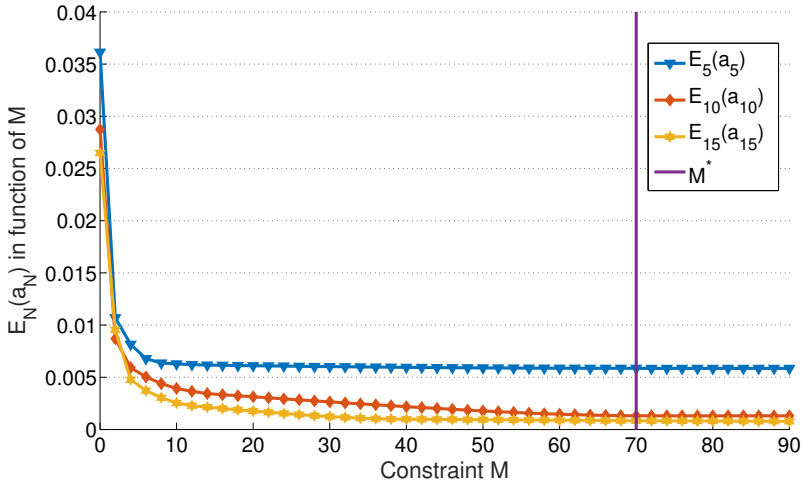

Figure 6: Behavior of the error $\mathcal{E}^{[a],N}(\widehat{a}_{N,M})$ as a function of the constraint M for different values of N.

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

Table 4: Parameter values for the experiment of Figure 7.

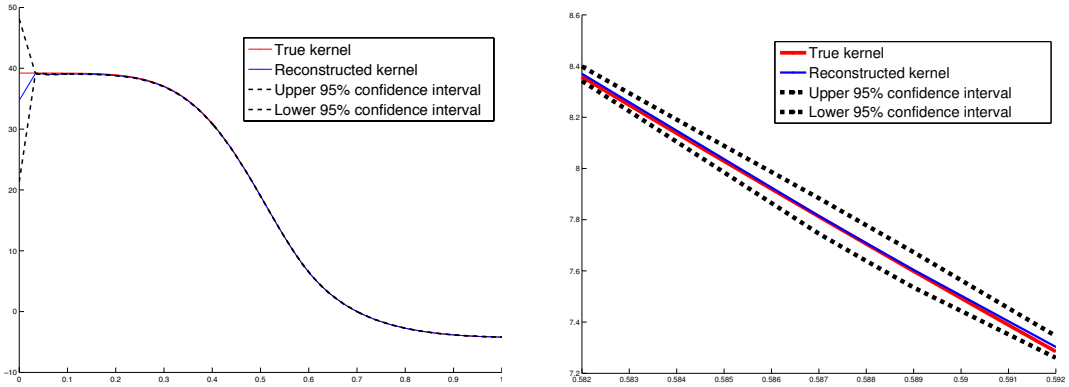

Figure 7: Reconstruction of a obtained by averaging 5 solutions of the minimization of $\mathcal{E}_{\Delta}^{[a],N}$ for $N = 50$. In red: the unknown kernel. In blue: the average of reconstructions. In black: 95% confidence interval for the parameter estimates returned by the Matlab function `normfit`. The figure on the right shows a zoom of the left figure.

6.1 Standard results on existence and uniqueness for ODE

For the reader's convenience and for the sake of a self-contained presentation, 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 $g\colon [0,T]\times\Omega \to \mathbb{R}^d$, i.e. the function g is continuous in y and measurable in t, and $0 < \tau \leq T$. A function $y\colon [0,\tau] \to \Omega$ is called a solution of the Carathéodory differential equation

$$\dot{y}(t) = g(t, y(t)) \tag{53}$$

on $[0,\tau]$ if and only if y is absolutely continuous and (53) is satisfied a.e. in $[0,\tau]$. The following well-known local existence result holds, see [17, Chapter 1, Theorem 1] .

Theorem 6.1. *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*

$$|g(t,y)| \leq m_\Omega(t), \tag{54}$$

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 (53) defined on the interval $[0,\tau]$ which satisfies $y(0) = y_0$.

The result can be extended to a global existence as follows.

Theorem 6.2. *Consider an interval $[0,T]$ on the real line and a Carathéodory function $g\colon [0,T]\times\mathbb{R}^d \to \mathbb{R}^d$. Assume that there exists a constant $C > 0$ such that the function g satisfies the condition*

$$|g(t,y)| \leq C(1+|y|), \tag{55}$$

for a.e. $t \in [0,T]$ and every $y \in \mathbb{R}^d$. Then there exists a solution $y(t)$ of (53) defined on the whole interval $[0,T]$, which satisfies $y(0) = y_0$. Moreover, for every $t \in [0,T]$, any solution satisfies

$$|y(t)| \leq \Big(|y_0| + Ct\Big)\, e^{Ct}. \tag{56}$$

Proof. Set $\rho := (|y_0| + CT)\, e^{CT}$. Consider now a ball $\Omega \subset \mathbb{R}^n$ centered at 0 with radius strictly greater than ρ. Existence of a local solution defined on an interval $[0,\tau]$ and taking values in Ω follows now easily from (55) and Theorem 6.1. If (55) holds, any solution of (53) 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 (56) 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 [17, Chapter 1, Theorem 4]. □

A further application of Gronwall's inequality yields the following results on continuous dependence on the initial data.

Proposition 6.3. *Let g_1 and $g_2\colon [0,T]\times\mathbb{R}^n\to\mathbb{R}^n$ be Carathéodory functions both satisfying (55) 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

$$|y_1(t)-y_2(t)|\le e^{Lt}\left(|y_1(0)-y_2(0)|+\int_0^t\|g_1(s,\cdot)-g_2(s,\cdot)\|_{L_\infty(B(0,\rho_{r,C,T}))}\,ds\right)\tag{57}$$

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 (57), as desired. □

6.2 Technical results for the mean-field limit

Let us start this section with some lemmas concerning the growth and the Lipschitz continuity of the right-hand side of (25) .

Lemma 6.4. *Let $a\in X$ and $\mu\in\mathcal{P}_1(\mathbb{R}^d)$. Then for all $y\in\mathbb{R}^d$ the following hold:*

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

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

Lemma 6.5. *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 6.6. *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

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

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

Proposition 6.7. *Let us fix $N \in \mathbb{N}$ and $a \in X$. Then the system (25) admits a unique global solution in $[0,T]$ for every initial datum $x_0^N \in \mathbb{R}^{d\times N}$.*

Proof. Let us define the function $g : \mathbb{R}^{d\times N} \to \mathbb{R}^{d\times N}$ defined for every $x = (x_1, \dots, x_N) \in \mathbb{R}^{d\times N}$ 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 (27). The system (25) in the form (26) can be rewritten compactly as

$$\dot{x}(t) = g(x(t)).$$

The function g is clearly a Carathéodory function and, by Lemma 6.4, it clearly satisfies a sublinear growth condition of the type (55). Moreover it is also locally Lipschitz continuous: 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

$$|g(x_1, \dots, x_N) - g(y_1, \dots, y_N)| \leq \sum_{i=1}^N |(F^{[a]} * \mu^N)(x_i) - (F^{[a]} * \nu^N)(y_i)|$$

$$\leq \sum_{i=1}^{N} \Bigg(|(F^{[a]} * \mu^N)(x_i) - (F^{[a]} * \mu^N)(y_i)| + |(F^{[a]} * \mu^N)(y_i) - (F^{[a]} * \nu^N)(y_i)| \Bigg).$$

Applying Lemma 6.6 to the first term and performing similar calculations to the ones in the proof of Lemma 6.5 on the second one, gives the desired result. We conclude the existence of a unique global solution by an application of Theorem 6.2 and its uniqueness follows from Lemma 6.3. □

The following preliminary result tells us that solutions to system (25) are also solutions to the equation (24), whenever conveniently rewritten.

Proposition 6.8. *Let $N \in \mathbb{N}$ be given and $a \in X$. Let $(x_1^N, \dots, x_N^N) : [0,T] \to \mathbb{R}^{d\times N}$ be the solution of (25) with initial datum $x_0^N \in \mathbb{R}^{d\times N}$. Then the empirical measure $\mu^N : [0,T] \to \mathcal{P}_1(\mathbb{R}^d)$ defined as in (27) is a solution of (24) with initial datum $\mu_0 = \mu^N(0) \in \mathcal{P}_c(\mathbb{R}^d)$.*

Proof. It can be proved by testing the equation (24) against a continuously differentiable function, arguing exactly as in [20, Lemma 4.3]. □

6.3 Existence and uniqueness of solutions for (24)

Variants of the following result are [20, Lemma 6.7] and [9, Lemma 4.7]

Lemma 6.9. *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*

$$\operatorname{supp}(\mu(t)) \cup \operatorname{supp}(\nu(t)) \subseteq B(0,R), \tag{58}$$

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

$$\|F^{[a]} * \mu(t) - F^{[a]} * \nu(t)\|_{L_\infty(B(0,r))} \leq L_{a,r,R} \mathcal{W}_1(\mu(t), \nu(t)) \tag{59}$$

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 6.5 and the hypothesis (58), we have

$$\begin{aligned}\|F^{[a]} * \mu(t) - F^{[a]} * \nu(t)\|_{L_\infty(B(0,r))} &\leq \operatorname*{ess\,sup}_{x\in B(0,r)} \int_{B(0,R)^2} \left|F^{[a]}(x-y) - F^{[a]}(x-z)\right| d\pi(y,z) \\ &\leq \mathrm{Lip}_{B(0,R+r)}(F^{[a]}) \int_{B(0,R)^2} |y-z| d\pi(y,z) \\ &= \mathrm{Lip}_{B(0,R+r)}(F^{[a]}) \mathcal{W}_1(\mu(t),\nu(t)),\end{aligned}$$

hence (59) holds with $L_{a,r,R} = \mathrm{Lip}_{B(0,R+r)}(F^{[a]})$. □

We show now the proof of Proposition 2.2 which states the existence of solutions for (24) .

Proof of Proposition 2.2. Let us define the quantity $\mathcal{X}_N(t) := \max_{i=1,\dots,N} |x_i^N(t)|$. By integration of (26) we obtain

$$\begin{aligned}|x_i^N(t)| &\leq |x_{0,i}^N| + \int_0^t (F^{[a]} * \mu^N(s))(x_i^N)| ds \\ &\leq |x_{0,i}^N| + \int_0^t \frac{1}{N} \sum_{j=1}^N |a(|x_i - x_j|)||x_j - x_i| ds \\ &\leq |x_{0,i}^N| + \|a\|_{L_\infty(\mathbb{R}_+)} \int_0^t \frac{1}{N} \sum_{j=1}^N (|x_j| + |x_i|) ds,\end{aligned}$$

implying

$$\mathcal{X}_N(t) \leq \mathcal{X}_N(0) + 2\|a\|_{L_\infty(\mathbb{R}_+)} \int_0^t \mathcal{X}_N(s) ds.$$

Hence, Gronwall's Lemma and the hypothesis $x_{0,i}^N \in \mathrm{supp}(\mu_0)$ for every $N \in \mathbb{N}$ and $i = 1, \dots, N$, imply that

$$\mathcal{X}_N(t) \leq \mathcal{X}_N(0) e^{2\|a\|_{L_\infty(\mathbb{R}_+)} t} \leq C_0 e^{2\|a\|_{L_\infty(\mathbb{R}_+)} t} \text{ for a.e. } t \in [0,T],$$

for some uniform constant C_0 depending only on μ_0. Therefore, the support of the empirical measure $\mu^N(\cdot)$ is bounded uniformly in N in a ball $B(0,R) \subset \mathbb{R}^d$, where

$$R = C_0 e^{2\|a\|_{L_\infty(\mathbb{R}_+)} T}. \tag{60}$$

Now, notice that from (23) it follows

$$\mathcal{W}_1(\mu^N(t), \mu^N(s)) \leq \frac{1}{N} \sum_{i=1}^N |x_i^N(t) - x_i^N(s)|,$$

and the local Liptschitz contiunuity of $\mu^N(t)$ follows from the one of $x_i^N(t)$: indeed $|x_i^N(t)| \leq R$ for a.e. $t \in [0,T]$, for all $N \in N$ and $i = 1, \dots, N$, and Lemma 6.4 yields

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

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

Hence, the sequence $(\mu^N)_{N\in\mathbb{N}} \subset \mathcal{C}^0([0,T], \mathcal{P}_1(B(0,R)))$ is equicontinuous, because equi-Lipschitz continuous, and equibounded in the complete 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, [25, 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

$$\lim_{k\to\infty} \mathcal{W}_1(\mu^{N_k}(t), \mu(t)) = 0 \quad \text{uniformly for a.e. } t \in [0,T], \tag{61}$$

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 $\lim_{N\to\infty} \mathcal{W}_1(\mu_0^N, \mu_0) = 0$ now obviously implies $\mu(0) = \mu_0$. In particular it holds

$$\lim_{k\to\infty} \langle \varphi, \mu^N(t) - \mu^N(0) \rangle = \langle \varphi, \mu(t) - \mu_0 \rangle \tag{62}$$

for all $\varphi \in \mathcal{C}_c^1(\mathbb{R}^d; \mathbb{R})$.

We are now left with verifying that this curve μ is a solution of (24). For all $t \in [0,T]$ and for all $\varphi \in \mathcal{C}_c^1(\mathbb{R}^d; \mathbb{R})$, 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 6.9, the inequality (59), and the compact support of $\varphi \in \mathcal{C}_c^1(\mathbb{R}^d; \mathbb{R})$, it 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

$$\lim_{N\to\infty} \int_0^t \int_{\mathbb{R}^d} \nabla\phi(x) \cdot (F^{[a]} * \mu^N(s))(x) d\mu^N(s)(x) ds$$
$$= \int_0^t \int_{\mathbb{R}^d} \nabla\phi(x) \cdot (F^{[a]} * \mu(s))(x) d\mu(s)(x) ds, . \tag{63}$$

The statement now follows from combination of (62) and (63). □

Proposition 6.10. *Fix $T > 0$, $a \in X$, $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$, $\xi_0 \in \mathbb{R}^d$ and $R > 0$. 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 (28) with initial value ξ_0 defined on the whole interval $[0,T]$.

Proof. The statement follows again by a proper combination of Lemma 6.4 and Lemma 6.6 with Theorem 6.2 for the existence, and the uniqueness similarly follows from Proposition 6.3.

□

The following Lemma and (57) are the main ingredients of the proof of Theorem 2.4 on continuous dependance on initial data and uniqueness of solutions for (24).

Lemma 6.11. *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*

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

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 $\mathbb{R}^n$ for $r > 0$, then

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

Proof. See [9, Lemma 3.11] and [9, 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 (28) with measure μ and ν, respectively. By (29), the triangle inequality, Lemma 6.9, Lemma 6.11 and (30) we have for every $t \in [0,T]$

$$\begin{aligned} \mathcal{W}_1(\mu(t), \nu(t)) &= \mathcal{W}_1((\mathcal{T}_t^\mu)_\# \mu_0, (\mathcal{T}_t^\nu)_\# \nu_0) \\ &\leq \mathcal{W}_1((\mathcal{T}_t^\mu)_\# \mu_0, (\mathcal{T}_t^\mu)_\# \nu_0) + \mathcal{W}_1((\mathcal{T}_t^\mu)_\# \nu_0, (\mathcal{T}_t^\nu)_\# \nu_0) \\ &\leq e^{T \operatorname{Lip}_{B(0,R)}(F^{[a]})} \mathcal{W}_1(\mu_0, \nu_0) + \|\mathcal{T}_t^\mu - \mathcal{T}_t^\nu\|_{L_\infty(B(0,R))}. \end{aligned} \tag{64}$$

Using (57) with $y_1(0) = y_2(0)$ we get

$$\|\mathcal{T}_t^\mu - \mathcal{T}_t^\nu\|_{L_\infty(B(0,r))} \leq e^{t \operatorname{Lip}_{B(0,R)}(F^{[a]})} \int_0^t \|F^{[a]} * \mu(s) - F^{[a]} * \nu(s)\|_{L_\infty(B(0,R))} \, ds. \tag{65}$$

Combining (64) and (65) with Lemma 6.9, we have

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

for every $t \in [0,T]$, where $L_{a,R,R}$ is the constant from Lemma 6.9. Gronwall's inequality now gives

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

which is exactly (32) with $\overline{C} = e^{T \operatorname{Lip}_{B(0,R)}(F^{[a]})+L_{a,R,R}}$.

Consider now two solutions of (24) with the same initial datum μ_0. By definition they both satisfy (31) for some $R > 0$ and (32) guarantees they both describe the same trajectory 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] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the National Academy of Sciences*, 105(4):1232–1237, 2008.

[4] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic. Empirical investigation of starling flocks: a benchmark study in collective animal behaviour. *Animal Behaviour*, 76(1):201–215, 2008.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] A. Cavagna, A. Cimarelli, I. Giardina, A. Orlandi, G. Parisi, A. Procaccini, R. Santagati, and F. Stefanini. New statistical tools for analyzing the structure of animal groups. *Mathematical Biosciences*, 214(1-2):32–37, 2008.

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

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

[15] De Boor, Carl. B(asic)-Spline Basics. Technical report, Wisconsin University–Madison Mathematics Research Center, August 1986.

[16] S. Dereich, M. Scheutzow, and R. Schottstedt. Constructive quantization: approximation by empirical measures. *Ann. Inst. Henri Poincaré (B)*, 49(4):1183–1203, 2013.

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

[18] M. Fornasier, J. Hakovec, and J. Vybral. Particle systems and kinetic equations modeling interacting agents in high dimension. *Multiscale Modeling & Simulation*, 9(4):1727–1764, 2011.

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

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

[21] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. `http://stanford.edu/~boyd/graph_dcp.html`.

[22] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, Mar. 2014.

[23] 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.

[24] H. Hildenbrandt, C. Carere, and C. Hemelrijk. Self-organized aerial displays of thousands of starlings: a model. *Behavioral Ecology*, 21(6):1349–1359, 2010.

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

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

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

[28] 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.

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