Interaction Kernel Learning in Multi-Agent Dynamical System

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

1 New model

The scope of this note is to support a model which differs from the one proposed at Duke in two fundamental aspects:

1.) In the dynamical systems we are considering, of the type

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

we cannot distinguish between different particles. Hence assuming an asymmetric distribution of initial conditions $(x_1(0), \dots, x_N(0)) \sim \pi_0$, where $\pi_0 \neq \bigotimes_{i=1}^N \mu_0$ is perhaps a "conceptual crime." Hence we are forced to consider distributions of the type $\pi_0 = \bigotimes_{i=1}^N \mu_0$ leading to the well-established theory of the so-called BBGKY-hierarchy and mean-field limits. This implies that the only relevant underlying dynamics is the transport of the indicated probability μ_0 along the characteristics induced by (1) (more explanations below).

2.) While at Duke there has been an attempt to parallel the work of Binev, Cohen, Dahmen, DeVore and Temlyakov on piecewise polynomial approximations induced by least squares. It is quite clear to us (as explained below) that the reference infinite dimensional variational problem approximated by the finite dimensional sampling counterpart is **not** a least squares in the strict sense that the theory by BCDDT has to be significantly (?) modified and adapted. This means no straightforward application of it!

1.1 Mean-field limit

Let us reformulate (1) as a discrete instance of a mean-field PDE. For that we introduce the empirical measure $\mu_t^N = \frac{1}{N}\sum_{j=1}^N \delta_{x_i(t)}$ supported on the trajectories of (1). We can then rewrite (1) in the following form

$$\begin{aligned}\dot{x}_i &= \frac{1}{N}\sum_{j=1}^{N} a(|x_i - x_j|)\frac{x_j - x_i}{|x_j - x_i|} \\ &= \int_{\mathbb{R}^d} a(|x_i - y|)\frac{y - x_i}{|y - x_i|} d\mu_t^N(y) = \big(F[a] * \mu_t^N\big)(x_i)\,, \end{aligned} \tag{2}$$

where we introduced the notation

$$F[a](\xi) = \frac{\xi}{|\xi|} a(|\xi|)\,, \quad \xi \in \mathbb{R}^d \setminus \{0\}\,.$$

Based on this formulation and given a probability measure-valued mapping $\mu : [0, T] \to P_1(\mathbb{R}^d)$ (with values on the probability with bounded first moments) and one point $x_0 \in \mathbb{R}^d$ we define the corresponding trajectory in $\mathbb{R}^d$ as the solution of the IVP

$$\begin{cases} \dot{x}(t) = \big(F[a] * \mu\big)(x)\,, \\ x(0) = x_0 \end{cases} \tag{3}$$

We denote the associated flow map by $\mathcal{T}_t^{\mu}(x_0) := x(t)$; accordingly, for $\mu_0 \in P_c(\mathbb{R}^d)$ (i.e. μ_0 is a compactly supported probability measure on $\mathbb{R}^d$), we define further the fixed point solution $\mu : [0, T] \to P_c(\mathbb{R}^d)$ of

$$\mu(t) \equiv \mu_t = \mathcal{T}_t^{\mu} \# \mu_0 \tag{4}$$

via the push-forward of μ_0 by means of the transport/flow map $\mathcal{T}_t^{\mu}(x_0)$ defined according to (3). It's possible to prove (see Section 8 of Ambrosio, Gigli, Sararé 2008) that μ is also the solution to the equation

$$\int_{\mathbb{R}^d} \varphi(x) d\mu_t(x) - \int_{\mathbb{R}^d} \varphi(x) d\mu_0(x) = \int_0^t \int_{\mathbb{R}^d} \nabla\varphi(x) \cdot \big(F[a] * \mu_s\big)(x) d\mu_s(x) ds \tag{5}$$

$\forall \varphi \in C^1(\mathbb{R}^d)$ or formally by saying that μ_t is the weak solution of the PDE

$$\frac{\partial \mu_t}{\partial t} = -\nabla \cdot \Big[\big(F[a] * \mu_t\big)\mu_t\Big]\,. \tag{6}$$

The equation (6) is called the mean-field equation associated to (1) for $N \to \infty$ and it can be derived also via BBGKY-hierarchy (which clarifies how the indistinguishability of particles results in the evolution of a single probability μ_t).

1.2 Stability of solution of (5) and (6)

Let us now introduce the space X of admissible potentials a as follows:

$$X = \big\{a : \mathbb{R}_+ \to \mathbb{R} \mid F[a] \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}^d)\big\}\,. \tag{7}$$

We also introduce the 1-Wasserstein distance on $P_1(\mathbb{R}^d)$,

$$W_1(\mu,\nu) = \sup_{\mathrm{Lip}\varphi\leq 1} \left| \int_{\mathbb{R}^d} \varphi d\mu - \int_{\mathbb{R}^d} \varphi d\nu \right| , \quad \mu,\nu \in P_1(\mathbb{R}^d) . \tag{8}$$

Given $\mu_0, \nu_0 \in P_c(\mathbb{R}^d)$ (compact support particularly entails finite first moment), and denoting by μ_t and ν_t, respectively, the corresponding solutions of (4)–(6) then it is possible to show that

$$W_1(\mu_t, \nu_t) \leq C(T) W_1(\mu_0, \nu_0) \tag{9}$$

for all $t \in [0,T]$. Notice that, as $\mu_t^N = \frac{1}{N}\sum_{j=1}^N \delta_{x_i(t)}$ is a particular solution of (4)–(6), for $x_1, x_2, \ldots, x_N, \ldots$ i.i. μ_0-distributed points, we conclude

$$W_1(\mu_t, \mu_t^N) \leq C(T) W_1(\mu_0, \mu_0^N) \overset{N\to\infty}{\longrightarrow} 0 \tag{10}$$

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

For future reference we note that the family of measures $\mu = (\mu_t)_t$ is equi-compactly supported, i.e. there exists a compact set K_μ such that

$$\operatorname{supp} \mu_t \subset K_\mu \quad \text{for all} \quad t \in [0,T] . \tag{11}$$

Moreover, their first moments are bounded.

1.3 Learning a

We are considering the following "error function"

$$\begin{aligned} E_N(\widehat{a}) &= \frac{1}{T}\int_0^T \frac{1}{N}\sum_{i=1}^N \left\| \frac{1}{N}\sum_{j=1}^N \big(F[\widehat{a}] - F[a]\big)(x_i - x_j) \right\|_{\mathbb{R}^d}^2 dt \\ &\equiv \frac{1}{T}\int_0^T \frac{1}{N}\sum_{i=1}^N \left\| \frac{1}{N}\sum_{j=1}^N \big(\widehat{a}(|x_i - x_j|) - a(|x_i - x_j|)\big) \frac{x_i - x_j}{|x_i - x_j|} \right\|_{\mathbb{R}^d}^2 dt , \end{aligned} \tag{12}$$

and we define a minimizer $\widehat{a}_V^N$ of it as

$$\widehat{a}_V^N = \arg\min\nolimits_{\widehat{a}\in V} E_N(\widehat{a})$$

for arbitrary subspaces $V \subset X$.

In order to understand the approximation properties of $\widehat{a}_V^N$ towards a for $N \to \infty$ and $V \uparrow X$ we need to rewrite (12) in a proper form:

$$E_N(\widehat{a}) = \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \left\| \big(F[a] - F[\widehat{a}]\big) * \mu_t^N(x) \right\|_{\mathbb{R}^d}^2 d\mu_t^N(x) . \tag{13}$$

This suggests a very specific counterpart of E_N for $N \to \infty$ (in the spirit of the mean-field limits) given by

$$E(\widehat{a}) = \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \left\| \big(F[\widehat{a}] - F[a]\big) * \mu_t(x) \right\|_{\mathbb{R}^d}^2 d\mu_t(x) , \tag{14}$$

where μ_t is the solution to the mean-field equations (4)–(6) associated to (1).

Let us look at E more carefully and consider afterwards certain additional coercivity assumptions on it: Using the obvious estimate $\big\|F[a](\xi)\big\|_{\mathbb{R}^d} \leq \big|a(|\xi|)\big|$ we obtain

$$E(\widehat{a}) \leq \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \bigg(\int_{\mathbb{R}^d} \big|\widehat{a}(|x-y|) - a(|x-y|)\big| d\mu_t(y)\bigg)^2 d\mu_t(x)dt\,.$$

Now observe that for any $\nu \in P(\mathbb{R}^d)$ the following estimate holds

$$\int_{\mathbb{R}^d} |f(x)|d\nu(x) \leq \bigg(\int_{\mathbb{R}^d} |f(x)|^2 d\nu(x)\bigg)^{1/2}\,.$$

Hence we further estimate

$$E(\widehat{a}) \leq \frac{1}{T}\int_0^T \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \big|\widehat{a}(|x-y|) - a(|x-y|)\big|^2 d\mu_t(y)d\mu_t(x)dt\,.$$

Using distance the map

$$d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+\,, \qquad (x,y) \mapsto d(x,y) = |x-y|\,,$$

we define by retract the probability measure-valued mapping $\varrho : [0,T] \to P(\mathbb{R}_+)$

$$\varrho_t = d\#(\mu_t \otimes \mu_t)\,, \tag{15}$$

which is explicitly defined via

$$\varrho_t(A) = (\mu_t \otimes \mu_t)\big(d^{-1}(A)\big) \tag{16}$$

for Borel-sets $A \subset \mathbb{R}_+$.

Lemma 1.1. *For every open set $A \subset \mathbb{R}_+$ the mapping $t \mapsto \varrho_t(A)$ is lower semi-continuous, wheres for compact A it is upper semi-continuous.*

Proof. **Step 1:** As a first step we show that for every given sequence $(t_n)_n$ converging to $t > 0$ we have weak convergence $\varrho_{t_n} \rightharpoonup \varrho_t$. For this in turn we first prove weak convergence $\mu_{t_n} \otimes \mu_{t_n} \rightharpoonup \mu_t \otimes \mu_t$.

It is a basic property of the space $C(\mathbb{R}^d \times \mathbb{R}^d)$ that it coincides with the inductive tensor product $C(\mathbb{R}^d) \otimes_\varepsilon C(\mathbb{R}^d)$. In particular, functions of the form $h = \sum_{j=1}^J f_j \otimes g_j$ with $f_j, g_j \in C(\mathbb{R}^d)$, $j = 1, \ldots, J$, $J \in \mathbb{N}$, are a dense subspace of $C(\mathbb{R}^{2d})$. Thus to prove weak convergence of measures on $\mathbb{R}^{2d}$, we can restrict the considerations to such functions. Due to linearity of the integrals this can be further reduced to simple tensor products $h = f \otimes g$.

However, for such simple tensor products we can simply apply Fubini's Theorem and the weak convergence $\mu_{t_n} \rightharpoonup \mu_t$ (which in turn is a consequence of the continuity of μ w.r.t. the Wasserstein metric W_1), and find

$$\int_{\mathbb{R}^{2d}} f \otimes g\, d(\mu_{t_n} \otimes \mu_{t_n}) = \int_{\mathbb{R}^d} f d\mu_{t_n} \cdot \int_{\mathbb{R}^d} g d\mu_{t_n} \overset{n\to\infty}{\longrightarrow} \int_{\mathbb{R}^d} f d\mu_t \cdot \int_{\mathbb{R}^d} g d\mu_t\,.$$

This already implies the claimed weak convergence $\varrho_{t_n} \rightharpoonup \varrho_t$. In detail: Given a function $f \in C(\mathbb{R}_+)$. Then we obtain

$$\int_{\mathbb{R}_+} f \, d\varrho_{t_n} = \int_{\mathbb{R}^{2d}} (f \circ d)(x,y) d(\mu_{t_n} \otimes \mu_{t_n})(x,y)$$
$$\overset{n\to\infty}{\longrightarrow} \int_{\mathbb{R}^{2d}} (f \circ d)(x,y) d(\mu_t \otimes \mu_t)(x,y) = \int_{\mathbb{R}_+} f \, d\varrho_t \,,$$

simply observe that the continuity of d implies continuity of $f \circ d$.

bf Step 2: The claim now follows from general results for weakly* convergent sequences of Radon measures, see e.g. [**?**, Proposition 1.62] □

This lemma justifies defining another probability measure ρ,

$$\rho = \frac{1}{T}\int_0^T \varrho_t dt \,, \qquad \rho(A) = \frac{1}{T}\int_0^T \varrho_t(A) dt \,. \tag{17}$$

More precisely, it is defined in this way on compact and open sets, and thereafter extended to all Borel-measurable sets. Then E can be estimated from above as follows,

$$E(\widehat{a}) \leq \frac{1}{T}\int_0^T \int_{\mathbb{R}_+} \left|\widehat{a}(s) - a(s)\right|^2 d\varrho_t(s) dt \equiv \int_{\mathbb{R}_+} \left|\widehat{a}(s) - a(s)\right|^2 d\rho \equiv \|\widehat{a} - a\|^2_{L_2(\mathbb{R}_+,\rho)} \,. \tag{18}$$

Equation (18) suggests an additional relevant condition to ensure that a is actually the unique minimizer of E on $X \cap L_2(\mathbb{R}_+, \rho)$. We assume that there exists a constant $c > 0$ such that

$$E(\widehat{a}) \geq c\|\widehat{a} - a\|^2_{L_2(\mathbb{R}_+,\rho)} \,. \tag{19}$$

Since $E(a) = 0$ and $E(\widehat{a}) \geq 0$ for all $\widehat{a} \in X$ then a is a minimizer of E. Clearly, if $a \in X \cap L_2(\mathbb{R}_+, \rho)$ and $E(\widehat{a}) = 0$ for some $\widehat{a} \in L_2(\mathbb{R}_+, \rho)$, by (19) we obtain that $\widehat{a} = a$ in $L_2(\mathbb{R}_+, \rho)$.

1.4 Properties of ρ

Before we proceed to our main result, we shall have a closer look at the measures ϱ_t and ρ.

Lemma 1.2. *Let μ_0 be absolutely continuous w.r.t. the d-dimensional Lebesgue measure $\mathcal{L}^d$. Then for every $t \in [0,T]$ also the measures μ_t are absolutely continuous w.r.t. $\mathcal{L}^d$.*

Proof. **Step 1:** As a first step, we note that the transport map $\mathcal{T}_t^\mu$ is locally Bi-Lipschitz, i.e. it is a bijective locally Lipschitz map, and its inverse is locally Lipschitz as well. Bijectivity is a consequence of the uniqueness of the solution to the corresponding ODE.

Note that with a being bounded on $\mathbb{R}_+$ also $F[a]$ is bounded on $\mathbb{R}^d$, which in turn yields boundedness of $F[a] * \mu_t$ (uniformly in t; see [**?**, Lemma 6.4]). Moreover, for fixed t this function is locally Lipschitz continuous, thus $g(t,x) = (F[a] * \mu_t)(x)$ is a Carathéodory function. In particular, we have

$$|g(t,x_1) - g(t,x_2)| \leq C_{a,\mu}|x_1 - x_2|$$

for almost every t and x_1, x_2 with $|x_i| \leq c_{r,a,T} = (r + T\|a\|_\infty)\exp(T\|a\|_\infty)$. This ultimately implies the stability estimate

$$\left|\mathcal{T}_t^\mu x_0 - \mathcal{T}_t^\mu x_1\right| \leq \exp\bigl(TC_{a,\mu}\bigr)|x_0 - x_1|\,, \qquad |x_i| \leq r\,, \quad i = 0,1\,,$$

shown e.g. in [**?**, Lemma 6.3], i.e. $\mathcal{T}_t^\mu$ is locally Lipschitz.

In view of the uniqueness of the solutions to the ODE, it is furthermore clear that the inverse of $\mathcal{T}_{t_0}^\mu$ is given by the transport map associated to the backward ODE

$$\dot{x}(t) = \bigl(F[a] * \mu\bigr)(x)\,, \quad x(t_0) = x_0\,.$$

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

$$\dot{y}(t) = -\bigl(F[a] * \nu\bigr)(x)\,, \quad y(0) = x(t_0)\,.$$

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

Step 2: Now let a Lebesgue null-set $A \subset \mathbb{R}^d$ be given. Put $B = (\mathcal{T}_0^\mu)^{-1}(A) = \mathcal{T}_0^\mu(A)$, the (pre-)image of A under the transport map $\mathcal{T}_t^\mu$. The claim now follows from showing $\mathcal{L}^d(B) = 0$, as then by assumption also $\mu_0(B) = 0$, which by definition yields

$$0 = \mu_0(B) = \mu_0\bigl((\mathcal{T}_0^\mu)^{-1}(A)\bigr) \equiv \mu_t(A)\,.$$

Moreover, we can reduce this further to consider only $B \cap K_\mu$ with K_μ from (11), since $\mu_t(B \setminus K_\mu) = 0$ for all t. Hence we no longer need to distinguish between local and global Lipschitz maps.

It thus remains to show that the image of a Lebesgue null-set under a Lipschitz map is again a null-set. To see this, recall that a measurable set A has Lebesgue meaure zero if, and only if for every $\varepsilon > 0$ there exists a family of balls $B_1, B_2, \ldots$ (or, equivalently, with cubes) such that

$$A \subset \bigcup_n B_n \qquad \text{and} \qquad \sum_n \mathcal{L}^d(B_n) < \varepsilon\,.$$

Let L be the Lipschitz constant of $\mathcal{T}_t^\mu$ on K, and $d(B_n)$ the diameter. Then clearly the image of B_n under $\mathcal{T}_t^\mu$ is contained in a ball of diameter at most $Ld(B_n)$. Denote those balls by $\widetilde{B}_n$. Then it immediately follows

$$\mathcal{T}_t^\mu(A) \subset \bigcup_n \widetilde{B}_n \qquad \text{as well as} \qquad \sum_n \mathcal{L}^d(\widetilde{B}_n) = L^d \sum_n \mathcal{L}^d(B_n) < L^d\varepsilon\,.$$

Thus we have found a cover for $\mathcal{T}_t^\mu(A)$ with the required property for $L^d\varepsilon$, which finally yields $\mathcal{L}^d(\mathcal{T}_t^\mu(A)) = 0$. □

Lemma 1.3. *Let μ_0 be absolutely continuous w.r.t. $\mathcal{L}^d$. Then for all $t \in [0,T]$, the measure ϱ_t is absolutely continuous w.r.t. $\mathcal{L}^1|_{\mathbb{R}_+}$, and this remains true for the measure ρ.*

Proof. Fix $t \in [0,T]$. By Lemma 1.2 we already know that μ_t is absolutely continuous w.r.t. $\mathcal{L}^d$. This immediately implies that $\mu_t \otimes \mu_t$ is absolutely continuous w.r.t. $\mathcal{L}^{2d}$. It hence remains to show that $d\#\mathcal{L}^{2d}$ is absolutely continuous w.r.t. $\mathcal{L}^1|_{\mathbb{R}_+}$.

Let $A \subset \mathbb{R}_+$ be a Lebesgue null-set, and put $B = d^{-1}(A) \subset \mathbb{R}^{2d}$. Moreover, we denote $B_x = \{y \in \mathbb{R}^d : |x-y| \in A\}$. Then clearly $B_{x+z} = z + B_x$. Moreover, using Fubin's Theorem we obtain

$$\mathcal{L}^{2d}(B) = \int_{\mathbb{R}^d} \mathcal{L}^d(B_x) d\mathcal{L}^d(x)\,.$$

It thus remains $\mathcal{L}^d(B_x) = 0$ for one (and thus for all, due to translation invariance of $\mathcal{L}^d$) $x \in \mathbb{R}^d$. However, to calculate $\mathcal{L}^d(B_0)$, we can transform to polar coordinates, and once more using Fubini's Theorem we obtain

$$\mathcal{L}^d(B_x) = \int_{\mathbb{R}^d} \chi_{B_0}(y) d\mathcal{L}^d(y) = \int_{S^d}\int_{\mathbb{R}_+} \chi_A(r) dr d\omega = \Omega_d \mathcal{L}^1(A) = 0\,,$$

where Ω_d is the surface measure of the unit sphere S_d. This proves the absolute continuity of ϱ_t, since

$$\mathcal{L}^1(A) = 0 \Longrightarrow \mathcal{L}^{2d}(d^{-1}(A)) \Longrightarrow (\mu_t \otimes \mu_t)(d^{-1}(A)) = 0 \iff \varrho_t(A) = 0\,.$$

The absolute continuity of ρ now follows immediately from the one of ϱ_t for every t and its definition as an integral average. □

Lemma 1.4. *The measure ρ has compact support.*

Proof. The supports of the measures ϱ_t are subsets of $B = d(K_\mu, K_\mu) = \{|x-y| : x, y \in K_\mu\}$, where K_μ is the set introduced in (11). Due to continuity of d this set B is again a compact subset of $\mathbb{R}_+$. We then immediately obtain $\operatorname{supp}\rho \subset B$. □

Remark 1. While absolute continuity of μ_0 implies the same for ρ, the situation is different for purely atomic measures μ_0. On the one hand, also μ_t then is purely atomic for every t, and this remains true for ϱ_t. However, due to the averaging involved in the definition of ρ it generally cannot be atomic. For example, we obtain

$$\frac{1}{T}\int_0^T \delta_t dt = \frac{1}{T}\mathcal{L}^1|_{[0,T]}\,,$$

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

2 The main result

Theorem 2.1. *Assume $a \in X \cap L_2(\mathbb{R}_+,\rho)$ as well as*

$$a \in W^{1,p}_{loc}(\mathbb{R}_+) \tag{20}$$

for some $1 \leq p \leq \infty$. Further assume that E satisfies (19). Let $x_1, x_2, \dots, x_N, \dots$ i.i. μ_0-distributed for some $\mu_0 \in P_c(\mathbb{R}^d)$, and define a sequence of finite-dimensional subspaces $V_M \subset L_2(\mathbb{R}_+,\rho)$ for $M = 2, 3, \dots$ such that for all $b \in X \cap L_2(\mathbb{R}_+,\rho)$ with $\|b\|_{1,p} \leq \|a\|_{1,p}$

$$\exists b_M \in V_M, \|b_M\|_{1,p} \leq \|a\|_{1,p} \quad s.t. \quad b_M \to b\,,$$

with local convergence in $W^{1,p}$.

We define

$$E_N(\widehat{a})_p = \begin{cases} \frac{1}{T}\int_0^T \frac{1}{N}\sum_{i=1}^N \Big\| \frac{1}{N}\sum_{j=1}^N \big(\widehat{a}(|x_i - x_j|) - a(|x_i - x_j|)\big)\frac{x_i - x_j}{|x_i - x_j|}\Big\|^2 dt\,, \\ \qquad \text{if } \widehat{a} \in V_N, \|\widehat{a}\|_{1,p} \leq \|a\|_{1,p}\,, \\ +\infty\,, \\ \qquad \text{if } \widehat{a} \in L_2(\mathbb{R}_+,\rho), \text{but } \widehat{a} \text{ does not satisfy the conditions above.} \end{cases} \tag{21}$$

Accordingly, we define

$$\widehat{a}_N = \arg\min_{\widehat{a} \in L_2(\mathbb{R}_+,\rho)} E_N(\widehat{a})_p \tag{22}$$

(notice that $\widehat{a}_N \in V_N$ and $\|\widehat{a}_N\|_{1,p} \leq \|a\|_{1,p}$ by definition).

Then the sequence $(\widehat{a}_N)_N$ converges uniformly to some continuous function $\overline{a}$ with $E(\overline{a}) = 0$. If we additionally assume the coercivity condition (19), then we have $\overline{a} = a$. Furthermore, the sequence $(\widehat{a}'_N)_N$ has a subsequence weakly converging in $L_p(\mathbb{R}_+,\rho)$ to $\overline{a}'$ for every $1 < p < \infty$.

We start with a technical lemma.

Lemma 2.2. *Let $(a_N)_N \subset L_2(\mathbb{R}_+,\rho)$ be a sequence of continuous, weakly differentiable functions such that for all N*

$$\|a_N\|_\infty \leq C_0 \qquad \text{and} \qquad \|a'_N\|_{L_\infty(\mathbb{R}_+)} \leq C_1\,.$$

Then there exists a uniformly convergent subsequence, denoted by $(a_{N_k})_{k\in\mathbb{N}}$, with limit $\overline{a}$, and it holds

$$\liminf_{k\to\infty} E_{N_k}(a_{N_k}) \geq E(\overline{a})\,. \tag{23}$$

Proof. **Step 1:** By assumption, the functions a_N are uniformly bounded. We shall prove that they are also equi-continuous, then by the Arzelà-Ascoli Theorem there exists a subsequence uniformly converging to some continuous function $\overline{a}$.

To see the equi-continuity we apply the Fundamental Theorem of Calculus (which is applicable for functions from $W^{1,p}$, see [**?**, Theorem 2.8]) to obtain

$$a_N(x) - a_N(y) = \int_{[x,y]} a'_N(t)dt\,.$$

This implies

$$\big|a_N(x) - a_N(y)\big| \leq \int_{[x,y]} |a_N'(t)|dt \leq |x-y|^{1/p'} \|a_N'\|_{L_p(\mathbb{R}_+)} \,.$$

In particular, the functions a_N all are Lipschitz continuous with Lipschitz constant uniformly bounded by C_1, which in turn implies equi-continuity.

Step 2: We notice that $W_1(\mu_t, \mu_t^N) \to 0$ for $N \to \infty$ (see (10)) particularly implies weak convergence of the sequence of measures $(\mu_t^N)_N$ towards μ_t for every $t \in [0,T]$. Thus applying the result from *arXiv-article*, we obtain

$$\begin{aligned}\liminf_k \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} &\Big\| \big(F[a_{N_k}] - F[a]\big) * \mu_t^{N_k}(x) \Big\|^2 d\mu_t^{N_k}(x)dt \\ &\geq \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \liminf_{\substack{k\to\infty,\\ x'\to x}} \Big\| \big(F[a_{N_k}] - F[a]\big) * \mu_t^{N_k}(x') \Big\|^2 d\mu_t(x)dt \\ &= \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \liminf_{k\to\infty} \Big\| \big(F[a_{N_k}] - F[a]\big) * \mu_t^{N_k}(x) \Big\|^2 d\mu_t(x)dt \,,\end{aligned}$$

where the last line is due to continuity of

$$x' \mapsto \Big\| \big(F[a_{N_k}] - F[a]\big) * \mu_t^{N_k}(x') \Big\| \,.$$

It remains to deal with

$$\liminf_{k\to\infty} \bigg\| \int_{\mathbb{R}^d} \big(F[a_{N_k}] - F[a]\big)(x-y) d\mu_t^{N_k}(y) \bigg\|^2 \,.$$

Step 3: Combining the uniform convergence of a_{N_k} and the weak convergence of $\mu_t^{N_k}$ we see that the limit $k \to \infty$ exists: For $k \geq k_0(\delta)$ we have

$$\|F[a_{N_k}] - F[\overline{a}]\|_{L_\infty(\mathbb{R}^d)} = \|a_{N_k} - \overline{a}\|_{L_\infty(\mathbb{R}_+)} \leq \delta/2 \,,$$

as well as

$$\bigg\| \int_{\mathbb{R}^d} \big(F[\overline{a}] - F[a]\big)(x-y) d\mu_t^{N_k}(y) - \int_{\mathbb{R}^d} \big(F[\overline{a}] - F[a]\big)(x-y) d\mu_t(y) \bigg\| \leq \delta/2 \,,$$

note that continuity of a and $\overline{a}$ implies continuity of $F[a]$ and $F[\overline{a}]$. Hence for $k \geq k_0(\delta)$

we obtain

$$\begin{aligned}
&\left|\left\|\int_{\mathbb{R}^d}\big(F[a_{N_k}]-F[a]\big)(x-y)d\mu_t^{N_k}(y)\right\|-\left\|\int_{\mathbb{R}^d}\big(F[\overline{a}]-F[a]\big)(x-y)d\mu_t(y)\right\|\right|\\
&\leq\left\|\int_{\mathbb{R}^d}\big(F[a_{N_k}]-F[a]\big)(x-y)d\mu_t^{N_k}(y)-\int_{\mathbb{R}^d}\big(F[\overline{a}]-F[a]\big)(x-y)d\mu_t(y)\right\|\\
&\leq\left\|\int_{\mathbb{R}^d}\big(F[a_{N_k}]-F[\overline{a}]\big)(x-y)d\mu_t^{N_k}(y)\right\|\\
&\qquad+\left\|\int_{\mathbb{R}^d}\big(F[\overline{a}]-F[a]\big)(x-y)d\mu_t^{N_k}(y)-\int_{\mathbb{R}^d}\big(F[\overline{a}]-F[a]\big)(x-y)d\mu_t(y)\right\|\\
&\leq\int_{\mathbb{R}^d}\frac{\delta}{2}d\mu_t^{N_k}(y)+\frac{\delta}{2}=\delta\,,
\end{aligned}$$

which implies

$$\lim_{k\to\infty}\left\|\int_{\mathbb{R}^d}\big(F[a_{N_k}]-F[a]\big)(x-y)d\mu_t^{N_k}(y)\right\|=\left\|\int_{K_\varepsilon^t}\big(F[\overline{a}]-F[a]\big)(x-y)d\mu_t(y)\right\|.$$

As this holds for every (fixed) t, this implies the claim. □

Lemma 2.3. *Let $(a_N)_N\subset L_2(\mathbb{R}_+,\rho)$ be a sequence of continuous functions which converges pointwise $\mathcal{L}^1$-almost everywhere to some function $\overline{a}$. Then for every $\varepsilon>0$ there exist sets $K_\varepsilon^t\subset\mathbb{R}^d$, $t\in[0,T]$, such that $\mu_t(\mathbb{R}^d\setminus K_\varepsilon^t)\leq\varepsilon$ and*

$$\begin{aligned}
&\liminf_{N\to\infty}E_N(a_N)\\
&\qquad\geq\frac{1}{T}\int_0^T\int_{\mathbb{R}^d}\left(\left\|\int_{K_\varepsilon^t}\big(F[\overline{a}]-F[a]\big)(x-y)d\mu_t(y)\right\|-4\varepsilon\|a\|_\infty\right)^2d\mu_t(x)\,.
\end{aligned}\tag{24}$$

Proof. **Step 1:** Step 2 of the proof of Lemma 2.2 can be re-used here.

Since $\operatorname{supp}\mu_t^N\subset A$ (compact) uniformly w.r.t. N and t *(reference)* by Egorov's Theorem for all $\varepsilon>0$ and all $t>0$ there exist measurable sets $K_\varepsilon^t\subset A$ with $\mu_t(A\setminus K_\varepsilon^t)\leq\varepsilon$ such that $F[\widehat{a}_N]\rightrightarrows F[\overline{a}]$ uniformly on K_ε^t. Splitting the integration domain accordingly and using triangle inequality we find

$$\begin{aligned}
&\left\|\int_{\mathbb{R}^d}\big(F[a_N]-F[a]\big)(x-y)d\mu_t^N(y)\right\|\\
&=\left\|\int_{K_\varepsilon^t}\big(F[a_N]-F[a]\big)(x-y)d\mu_t^N(y)+\int_{A\setminus K_\varepsilon^t}\big(F[a_N]-F[a]\big)(x-y)d\mu_t^N(y)\right\|\\
&\geq\underbrace{\left\|\int_{K_\varepsilon^t}\big(F[a_N]-F[a]\big)(x-y)d\mu_t^N(y)\right\|}_{I_N}-\underbrace{\left\|\int_{A\setminus K_\varepsilon^t}\big(F[a_N]-F[a]\big)(x-y)d\mu_t^N(y)\right\|}_{II_N}.
\end{aligned}$$

For the term I_N we further find by combining the uniform convergence of a_N on K^t_ε and the weak convergence of μ^N_t that the limit $N \to \infty$ exists: For $N \geq N_0(\delta)$ we have

$$\|F[a_N] - F[\overline{a}]\| = \|a_N - \overline{a}\|_{L_\infty(K^t_\varepsilon)} \leq \delta/2\,,$$

as well as

$$\left\| \int_{K^t_\varepsilon} \big(F[\overline{a}] - F[a]\big)(x-y)d\mu^N_t(y) - \int_{K^t_\varepsilon} \big(F[\overline{a}] - F[a]\big)(x-y)d\mu_t(y) \right\| \leq \delta/2\,,$$

note that continuity of a and $\overline{a}$ (the latter holds only on K^t_ε!) implies continuity of $F[a]$ and $F[\overline{a}]$. Hence for $N \geq N_0(\delta)$ we obtain

$$\begin{aligned}
&\left| I - \left\| \int_{K^t_\varepsilon} \big(F[\overline{a}] - F[a]\big)(x-y)d\mu_t(y) \right\| \right| \\
&\leq \left\| \int_{K^t_\varepsilon} \big(F[a_N] - F[a]\big)(x-y)d\mu^N_t(y) - \int_{K^t_\varepsilon} \big(F[\overline{a}] - F[a]\big)(x-y)d\mu_t(y) \right\| \\
&\leq \left\| \int_{K^t_\varepsilon} \big(F[a_N] - F[\overline{a}]\big)(x-y)d\mu^N_t(y) \right\| \\
&\qquad + \left\| \int_{K^t_\varepsilon} \big(F[\overline{a}] - F[a]\big)(x-y)d\mu^N_t(y) - \int_{K^t_\varepsilon} \big(F[\overline{a}] - F[a]\big)(x-y)d\mu_t(y) \right\| \\
&\leq \int_{K^t_\varepsilon} \frac{\delta}{2} d\mu^N_t(y) + \frac{\delta}{2} \leq \delta\,,
\end{aligned}$$

which implies

$$\lim_{N\to\infty} I_N = \left\| \int_{K^t_\varepsilon} \big(F[\overline{a}] - F[a]\big)(x-y)d\mu_t(y) \right\|.$$

For the term II_N we first obtain

$$II_N \leq 2\|a\|_\infty \int_{A\setminus K^t_\varepsilon} d\mu^N_t(y) \leq 2\|a\|_\infty \int_{\mathbb{R}^d} \psi_\varepsilon(y) d\mu^N_t(y)\,,$$

where ψ_ε is a bounded continuous (bump) function approximating $\chi_{A\setminus K^t_\varepsilon}$ in $L_1(\mu_t)$ from above, i.e. $\psi_\varepsilon \geq \chi_{A\setminus K^t_\varepsilon}$ and

$$\|\psi_\varepsilon - \chi_{A\setminus K^t_\varepsilon}\|_{L_1(\mu_t)} \leq \varepsilon\,.$$

But then the weak convergence of μ^N_t implies

$$\int_{\mathbb{R}^d} \psi_\varepsilon(y) d\mu^N_t(y) \longrightarrow \int_{\mathbb{R}^d} \psi_\varepsilon(y) d\mu_t(y) \leq 2\varepsilon$$

by choice of ψ_ε and K^t_ε. Put together, we thus have

$$\limsup_{N\to\infty} II_N \leq 4\|a\|_\infty \varepsilon\,.$$

In turn, this yields

$$\liminf_N \left\| \int_{\mathbb{R}^d} \big(F[a_N]-F[a]\big)(x-y)d\mu_t^N(y) \right\| \geq \left\| \int_{K_\varepsilon^t} \big(F[\overline{a}]-F[a]\big)(x-y)d\mu_t(y) \right\| - 4\|a\|_\infty \varepsilon \,,$$

so far for every choice of $\varepsilon > 0$ and $t > 0$. Integrating over t then yields (24). □

Lemma 2.4. *Let $(a_N)_N \subset L_2(\mathbb{R}_+, \rho)$ be a sequence of continuous, weakly differentiable functions such that for all N*

$$\|a_N\|_{W_1^1(\mathbb{R}_+,\rho)} \leq C_0 \,,$$

and let $(a_N)_N$ be uniformly bounded at a point.

Then there exists a subsequence $(a_{N_k})_k$ which converges $\mathcal{L}^1$-a.e.

Proof. If we denote by ϕ_t the Radon-Nicodym derivative of ϱ_t w.r.t. $\mathcal{L}^1$, for arbitrary fixed $t \in [0, T]$ we can apply Helly's Selection Theorem to the functions $(a_N \phi_t)_N$. Then there exists a subsequence $(a_{N_k}\phi_t)_k$ which converges pointwise $\mathcal{L}^1$-a.e. to some function $\overline{a}\phi_t$ of bounded variation. □

Proof of Theorem 2.1. By the respective construction of the functionals E_N, the sequence of minimizers $(\widehat{a}_N)_N$ satisfies the assumptions of one of the above lemmas. Hence it has a pointwise ρ-a.e. or uniformly, respectively, convergent subsequence with limit function $\overline{a}$. We wish to show that $\overline{a}$ is a minimizer of E, then by (19) we have $\overline{a} = a$. Since in this way for every subsequence we can extract a subsubsequence converging to the same limit function a, we can infer that the entire sequence converges to a. For simplicity we will only treat the case $p = \infty$; the case $1 < p < \infty$ can be reduced to the proof for $p = 1$ in view of $W_p^1 \hookrightarrow W_1^1$; finally, the case $p = 1$ uses the same arguments as $p = \infty$, with Lemma 2.2 replaced by Lemmas 2.3 and 2.4.

Now let $b \in X \cap L_2(\mathbb{R}_+, \rho)$ with $\|b\|_{1,\infty} \leq \|a\|_{1,\infty}$ be given. By assumption there exists $b_N \in V_N$ with $\|b_N\|_{1,\infty} \leq \|b\|_{1,\infty} \leq \|a\|_\infty$ and $b_N \to b$. Hence, by the same arguments as in the proof of Lemma 2.2, combining the uniform convergence of a subsequence b_{N_k} and the weak convergence of μ_t^N, we obtain

$$E(b) = \lim_{k\to\infty} E_{N_k}(b_{N_k}).$$

Now, we can argue

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

The first inequality is due to optimality of $\widehat{a}_{N_k}$ and the second is due to Lemma 2.2. We therefore can conclude the fundamental estimate

$$E(b) \geq E(\overline{a}) \,,$$

which particularly applies to $b = a \in X \cap L_2(\mathbb{R}_+, \rho)$. This finally implies

$$0 = E(a) \geq E(\overline{a}) \geq 0 \Longrightarrow E(\overline{a}) = 0.$$

In case (19) holds, yields $\overline{a} = a$, that condition immediately further implies $\overline{a} = a$. □

3 DeVore et.al.

The setting: $X \subset \mathbb{R}^d$ bounded domain, $Y = \mathbb{R}$, and $Z = X \times Y$. Moreover, ρ is an unknown probability measure on Z, and ρ_X the corresponding marginal measure on X, i.e. $\rho_X(A) = \rho(A \times Y)$. For $f \in L_2(X, \rho_X)$, $f : X \to Y$, define the functional

$$\mathcal{E}(f) = \int_Z (y - f(x))^2 d\rho \,.$$

Objective: Find the minimizer of $\mathcal{E}$, or equivalently, minimize $\|f - f_\rho\|_{L_2(\rho)}$.

Data: Independent samples (x_j, y_j) with distribution ρ.

Approach: For a given subspace V define f_V as the minimizer of the approximate (empirical) functional

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{j=1}^{m} (y_j - f(x_j))^2 \,,$$

which clearly can be rewritten as integration against the empirical measure $\rho_m = \frac{1}{m} \sum_{j=1}^{m} \delta_{x_j}$, with the identification $y_j = f_\rho(x_j)$.

i) In our setting, together with the coercivity assumption, we know that the only minimizer of our functional E is the function a. In that case, minimizing E is equivalent to minimizing $\| \cdot - a\|_{L_2(\rho)}$, in the sense that they have the same unique global minimizer (at least ρ-a.e.) (which coincides with the exact solution of our learning problem).

ii) In the DeVore-Article, while the measure ρ is not known, it is assumed that one can obtain independent samples. For us, we may sample the initial values according to μ_0, which in turn leads to independent samples of μ_t, but do we get also samples for our measure ρ that way?

4 Analysis

We start with an estimate for the sampling error, i.e. the error contribution incurred on a fixed partition Λ due to inexact projections.

Theorem 4.1 (Sampling error)**.** *Let Λ be a fixed partition. Then*

$$P\big(\|P_\Lambda a - a_M^N\|_{L_2(\rho)} > \eta\big) \le 4M \exp\Big(-\frac{3N^2\eta^2}{256MB_a^2}\Big)$$

where $a_M^N = \arg\min_{b\in V_M} E_N(b)$, $M = \#\Lambda$ and $B_a = \|a\|_\infty$.

Proof. We follow the proof of DeVore et. al.

Step 1: By definition we can write

$$\|P_\Lambda a - a_M^N\|_{L_2(\rho)}^2 = \sum_{I\in\Lambda} |c_I - c_I^N|^2 \rho(I)\,.$$

With the function a, also the coefficients c_I and $c_{I,N}$ are always bounded (**add details**). Therefore, if we split the partition into

$$\Lambda^- = \Big\{I \in \Lambda : \rho(I) \le \frac{\eta^2}{8NM^2}\Big\}$$

and $\Lambda^+ = \Lambda \setminus \Lambda^-$. Then we first obtain

$$\sum_{I\in\Lambda^-} |c_I - c_I^N|^2 \le \frac{\eta^2}{2}$$

(i.e. with probability 1). To prove the theorem it is hence sufficient to show

$$P\Big(|c_I - c_I^N|^2 \ge \frac{\eta^2}{2M\rho(I)}\Big) \le 4\exp\Big(-\frac{3N^2\eta^2}{256MB_a^2}\Big)\,,$$

because of the union bound and $\#\Lambda^+ \le \#\Lambda = M$. To see this estimate, we write $\rho^N(I) = (1+\beta_I)\rho(I)$, so that in case $|\beta_I| \le \frac{1}{2}$ we find

$$\begin{aligned}|c_I - c_I^N| &= \Big|\frac{\alpha_I}{\rho_I} - \frac{\alpha_I^N}{\rho_I^N}\Big| = \frac{1}{\rho(I)(1+\beta_I)}|\alpha_I^N - \alpha_I - \beta_I\alpha_I| \\ &\le \frac{2}{\rho(I)}\big(|\alpha_I^N - \alpha_I| + |\alpha_I\beta_I|\big)\,.\end{aligned}$$

If we further require

$$|\alpha_I - \alpha_I^N| \le \frac{\eta\sqrt{\rho(I)}}{4\sqrt{2M}}$$

as well as

$$|\rho^N(I) - \rho(I)| \le \min\Big(\frac{1}{2}\rho(I), \frac{\eta\rho(I)^{3/2}}{4\sqrt{2M}|\alpha_I|}\Big)$$

we conclude (note $\alpha_I\beta_I = \frac{\alpha_I}{\rho(I)}(\rho^N(I) - \rho(I)))$

$$\begin{aligned}|c_I - c_I^N| &\leq \frac{2}{\rho(I)}\big(|\alpha_I^N - \alpha_I| + |\alpha_I\beta_I|\big)\\ &\leq \frac{2}{\rho(I)}\frac{\eta\sqrt{\rho(I)}}{4\sqrt{2M}} + \frac{2}{\rho(I)}\frac{|\alpha_I|}{\rho(I)}\rho^N(I) - \rho(I)|\\ &\leq \frac{\eta}{2\sqrt{2M\rho(I)}} + \min\Big(\frac{|\alpha_I|}{\rho(I)}, \frac{\eta}{2\sqrt{2M\rho(I)}}\Big) \leq \frac{\eta}{\sqrt{2M\rho(I)}}\,.\end{aligned}$$

We thus obtain

$$\begin{aligned}&P\Big(|c_I^N - c_I|^2 \geq \frac{\eta^2}{2M\rho(I)}\Big)\\ &\quad\leq P\Big(|\alpha_I^N - \alpha_I| \geq \frac{\eta\sqrt{\rho(I)}}{4\sqrt{2N}}\Big) + P\Big(|\rho^N(I) - \rho(I)| \geq \min\big(\frac{1}{2}\rho(I), \frac{\eta\rho(I)^{3/2}}{4\sqrt{2M}|\alpha_I|}\big)\Big)\,.\end{aligned} \tag{25}$$

Step 2: These probabilities we now estimate with the help of Bernstein's inequality. We recall, if ζ_i are m independent realizations of a bounded random variable ζ with $|\zeta(\omega) - \mathbb{E}(\zeta)| \leq B$ and $\mathrm{Var}(\zeta) = \sigma^2$, then for every $\varepsilon > 0$ it follows

$$P\Big(\Big|\frac{1}{m}\sum_{i=1}^m \zeta_i - \mathbb{E}(\zeta)\Big| \geq \varepsilon\Big) \leq 2e^{-\frac{m\varepsilon^2}{2(\sigma^2 + B\varepsilon/3)}}\,.$$

needs reference

Since we have no direct access to ρ-distributed random variables, we have to rely on random variables distributed according to the initial distribution μ_0. We therefore write

$$\begin{aligned}&\rho(I) - \rho^N(I)\\ &\quad= \frac{1}{T}\int_0^T \int_{\mathbb{R}^{2d}} \chi_{d^{-1}(I)} d(\mu_t \otimes \mu_t)dt - \frac{1}{N^2}\sum_{i,j}\frac{1}{T}\int_0^T \chi_I(|\mathcal{T}_t^\mu x_i - \mathcal{T}_t^\mu x_j|)dt\\ &\quad= \frac{1}{T}\int_0^T \int_{\mathbb{R}^{2d}} \chi_{\Phi_t^{-1}(I)}(x,y) d(\mu_0 \otimes \mu_0)(x,y)\,dt - \frac{1}{N^2}\sum_{i,j}\frac{1}{T}\int_0^T \chi_{\Phi_t^{-1}(I)}(x_i, x_j)dt\\ &\quad= \int_{\mathbb{R}^{2d}} \frac{1}{T}\int_0^T \chi_{\Phi_t^{-1}(I)}(x,y)dt\, d(\mu_0 \otimes \mu_0) - \frac{1}{N^2}\sum_{i,j}\frac{1}{T}\int_0^T \chi_{\Phi_t^{-1}(I)}(x_i, x_j)dt\,,\end{aligned}$$

where $\Phi = (\mathcal{T}_t^\mu \otimes \mathcal{T}_t^\mu) \circ d$, i.e. $\Phi_t(x,y) = |\mathcal{T}_t^\mu x - \mathcal{T}_t^\mu y|$. Therein, in the last line the inner integral has to be interpreted as a Bochner-integral. Hence defining

$$\zeta = \frac{1}{T}\int_0^T \chi_{\Phi_t^{-1}(I)} dt\,,$$

again in the sense of a Bochner integral in $L_1(\mathbb{R}^{2d}, \mu_0 \otimes \mu_0)$, we can write

$$\rho(I) - \rho^N(I) = \int_{\mathbb{R}^{2d}} \zeta(x,y) d(\mu_0 \otimes \mu_0)(x,y) - \frac{1}{N^2} \sum_{i,j} \zeta(x_i, x_j)\,.$$

Let x_i, $i = 1, \dots, N$, be independent samples of μ_0. Then the pairs (x_i, x_j), $i, j = 1, \dots, N$, are independent w.r.t. $\mu_0 \otimes \mu_0$. Thus we can apply Bernstein's inquality to ζ, which clearly is bounded by 1 (thus $B = 2$ here), with $\mathbb{E}(\zeta) = \rho(I)$ and $\mathrm{Var}(\zeta) = \rho(I)(1 - \rho(I)) \le \rho(I)$. We then obtain in the case $\frac{1}{2}\rho(I) \ge \frac{\eta\rho(I)^{3/2}}{4\sqrt{2M}|\alpha_I|}$

$$\begin{aligned} P\Big(|\rho(I) - \rho^N(I)| \ge \frac{\eta\rho(I)^{3/2}}{4\sqrt{2M}|\alpha_I|}\Big) &\le 2\exp\Big(-\frac{N^2\eta^2\rho(I)^3}{64M|\alpha_I|^2(\rho(I) + \frac{2}{3}\frac{\eta\rho(I)^{3/2}}{4\sqrt{2M}|\alpha_I|})}\Big) \\ &\le 2\exp\Big(-\frac{N^2\eta^2\rho(I)^3}{256MB_a^2\rho(I)^3/3}\Big) = 2\exp\Big(-\frac{3N^2\eta^2}{256MB_a^2}\Big) \end{aligned}$$

where we used the simple estimate $|\alpha_I| \le B_a\rho(I)$. In the case $\frac{1}{2}\rho(I) \le \frac{\eta\rho(I)^{3/2}}{4\sqrt{2M}|\alpha_I|}$ we find

$$\begin{aligned} P\big(|\rho(I) - \rho^N(I)| \ge \tfrac{1}{2}\rho(I)\big) &\le 2\exp\Big(-\frac{N^2\rho(I)^2}{8(\rho(I) + \rho(I)/3}\Big) \\ &= 2\exp\Big(-\frac{3}{32}N^2\rho(I)\Big) \le 2\exp\Big(-\frac{3N^2\eta^2}{256MB_a^2}\Big)\,, \end{aligned}$$

where in the last step we used $I \in \Lambda^+$. This takes care of the second probability in (25). To deal with the first one, we now apply Bernstein's inequality with

$$\zeta_a(x,y) = \frac{1}{T}\int_0^T a\big(\Phi_t\big)\chi_{\Phi_t^{-1}(I)} dt\,,$$

again in the sense of Bochner integrals. With this definition we can write

$$\alpha_I - \alpha_I^N = \int_{\mathbb{R}^{2d}} \zeta_a(x,y) d(\mu_0 \otimes \mu_0)(x,y) - \frac{1}{N^2}\sum_{i,j} \zeta_a(x_i, x_j)\,.$$

Moreover, we have $\mathbb{E}(\zeta_a) = \alpha_I$, $|\zeta(x,y) - \mathbb{E}(\zeta)| \le 2B_a$ and $\mathrm{Var}(\zeta) \le B_a^2\rho(I)$. Thus Bernstein's inequality yields

$$\begin{aligned} P\Big(|\rho(I) - \rho^N(I)| \ge \frac{\eta\sqrt{\rho(I)}}{4\sqrt{2M}}\Big) &\le 2\exp\Big(-\frac{N^2\eta^2\rho(I)}{64M(B_a^2\rho(I) + \frac{2B_a\eta\sqrt{\rho(I)}}{12\sqrt{2M}})}\Big) \\ &\le 2\exp\Big(-\frac{N^2\eta^2\rho(I)}{64M(B_a^2\rho(I) + \frac{4B_a^2\rho(I)}{12})}\Big) = 2\exp\Big(-\frac{3N^2\eta^2}{256MB_a^2}\Big)\,. \end{aligned}$$

Here we once more used $I \in \Lambda^+$. This completes the proof. □

TODO: Include Lemma which shows that ζ and ζ_a are well-defined functions from L_1 (particularly: measurable); equivalently that the mentioned Bochner integral exist.