RKHS regularization of singular local stochastic volatility McKean-Vlasov models

Christian Bayer*[1], Denis Belomestny†[2], Oleg Butkovsky‡[1], and John Schoenmakers§[1]

[1]Weierstrass Institute, Mohrenstrasse 39, 10117 Berlin, Germany.
[2]Duisburg-Essen University, Essen

March 3, 2022

Abstract

Motivated by the challenges related to the calibration of financial models, we consider the problem of solving numerically a singular McKean-Vlasov equation

$$dS_t = \sigma(t, S_t) S_t \frac{\sqrt{v}_t}{\sqrt{\mathsf{E}[v_t|S_t]}} dW_t,$$

where W is a Brownian motion and v is an adapted diffusion process. This equation can be considered as a singular local stochastic volatility model. Whilst such models are quite popular among practitioners, unfortunately, its well-posedness has not been fully understood yet and, in general, is possibly not guaranteed at all. We develop a novel regularization approach based on the reproducing kernel Hilbert space (RKHS) technique and show that the regularized model is well-posed. Furthermore, we prove propagation of chaos. We demonstrate numerically that a thus regularized model is able to perfectly replicate option prices due to typical local volatility models. Our results are also applicable to more general McKean–Vlasov equations.

1. Introduction

The present article is motivated by [GHL12], wherein Guyon and Henry-Labordère proposed a particle method for the calibration of local stochastic

*christian.bayer@wias-berlin.de
†denis.belomestny@uni-due.de
‡oleg.butkovskiy@gmail.com
§john.schoenmakers@wias-berlin.de

volatility models (e.g. stock price models). Let us recall that *local volatility models*

$$dS_t = \sigma(t, S_t) S_t dW_t, \tag{1.1}$$

where W denotes a one-dimensional Brownian motion under a risk-neutral measure and S the forward price of a stock, can replicate any sufficiently regular implied volatility surface, provided that we choose the local volatility according to *Dupire's formula*, symbolically, $\sigma \equiv \sigma_{\mathrm{Dup}}$ [Dup94]. Unfortunately, it is well understood that Dupire's model exhibits unrealistic random price behavior despite perfect fits to market prices of options. On the other hand, *stochastic volatility models*

$$dS_t = \sqrt{v_t} S_t dW_t \tag{1.2}$$

for a suitably chosen stochastic variance process v_t, may lead to realistic (in particular, time-homogeneous) dynamics, but are typically difficult or impossible to fit to observed implied volatility surfaces. We refer to [Gat11] for an overview of stochastic and local volatility models.

Local stochastic volatility models can combine the advantages of both local and stochastic volatility models. Indeed, if the forward price is given by

$$dS_t = \sqrt{v_t} \sigma(t, S_t) S_t dW_t, \tag{1.3}$$

then it exactly fits the observed market option prices provided that

$$\sigma_{\mathrm{Dup}}(t, x)^2 = \sigma(t, x)^2 \mathsf{E}\left[v_t \mid S_t = x \right]. \tag{1.4}$$

This is a simple consequence of the celebrated Gyöngy's Markovian projection theorem [Gyo86, Theorem 4.6], see also [BS13, Corollary 3.7]. With this choice of σ we have

$$dS_t = \sigma_{\mathrm{Dup}}(t, S_t) S_t \frac{\sqrt{v_t}}{\sqrt{\mathsf{E}\left[v_t \mid S_t \right]}} dW_t, \tag{1.5}$$

Note that v in (1.5) can be any positive adapted stochastic process. In a sense, (1.5) may be considered as an inversion of the Markovian projection due to [Gyo86], applied to Dupire's local volatility (asset) model, i.e. (1.1) with $\sigma \equiv \sigma_{\mathrm{Dup}}$.

Thus, the stochastic local volatility model of McKean–Vlasov type (1.5) solves the smile calibration problem. However, equation (1.5) is singular in a sense explained below and very hard to analyze and to solve. Even the problem of proving existence or uniqueness for (1.5) (under various assumptions on v) turned out to be notoriously difficult and only a few results are available; we refer to [LSZ20] for an extensive discussion and literature review.

Let us recall that the theory of standard McKean–Vlasov equations of the form

$$dZ_t = \widetilde{H}\left(t, Z_t, \mu_t\right) dt + \widetilde{F}\left(t, Z_t, \mu_t\right) dW_t \tag{1.6}$$

with $\mu_t = \mathrm{Law}(Z_t)$, is well understood under appropriate regularity conditions, in particular, Lipschitz continuity of $\widetilde{H}$ and $\widetilde{F}$ w.r.t. the standard Euclidean distances in the first two arguments and w.r.t. the Wasserstein distance in μ_t, see [Fun84, CD16a, MV16]. Denoting $Z_t := (X_t, Y_t)$, it is not difficult to see that the conditional expectation $(x, \mu_t) \mapsto \mathsf{E}\left[A(Y_t) \mid X_t = x\right]$ is, unfortunately, not Lipschitz continuous in the above sense. Therefore, the standard theory does not apply to (1.5).

There are a number of results available in the literature where the Lipschitz condition on drift and diffusion is not imposed. Bossy and Jabir [BJ17] considered singular MV systems of the form:

$$dX_t = \mathsf{E}[\ell(X_t)|Y_t]dt + \mathsf{E}[\gamma(X_t)|Y_t]dW_t, \tag{1.7a}$$
$$dY_t = b(X_t, Y_t)dt + \sigma(Y_t)dB_t, \tag{1.7b}$$

or, alternatively, the seemingly even less regular equation

$$dX_t = \sigma(p(t, X_t))dW_t, \tag{1.8}$$

where $p(t, \cdot)$ denotes the density of X_t. [BJ17] establishes well-posedness of (1.7) and (1.8) under suitable regularity conditions (in particular, ellipticity) based on energy estimates of the corresponding non-linear PDEs. Interestingly, these techniques break down when the roles of X and Y are reversed in (1.7), i.e., when $\mathsf{E}[\gamma(X_t)|Y_t]$ is replaced by $\mathsf{E}[\gamma(Y_t)|X_t]$ in (1.7a) – and similarly for the drift term. Hence, the results of [BJ17] do not imply well-posedness of (1.5).

In [LSZ20], the authors studied the following two-dimensional SDE,

$$dX_t = b_1(X_t)\frac{h(Y_t)}{\mathsf{E}[h(Y_t)|X_t]}\,dt + \sigma_1(X_t)\frac{f(Y_t)}{\sqrt{\mathsf{E}[f^2(Y_t)|X_t]}}\,dW_t, \tag{1.9a}$$
$$dY_t = b_2(Y_t)\,dt + \sigma_2(Y_t)\,dB_t, \tag{1.9b}$$

where W and B are two independent one-dimensional Brownian motions. Clearly, this can be seen as (1.5) with a non-zero drift and with the process v chosen in a special way. The authors proved strong existence and uniqueness of solutions to (1.9) in the *stationary* case. In particular, this imposes strong conditions on b_1 and b_2, but also requires the initial value (X_0, Y_0) to be random and have the stationary distribution. Existence and uniqueness of (1.9) in the general case (without the stationarity assumptions) remains open.

Finally, let us mention [JZ20, Theorem 2.2], which established weak existence of the solutions to (1.5) for the case when v is a jump process taking finitely many values.

Another question apart from well-posedness of these singular McKean–Vlasov equations is how to solve them numerically (in a certain sense). Let us recall that even for standard SDEs with singular or irregular drift, where

existence/uniqueness is known for quite some time, the convergence of the corresponding Euler scheme with non-vanishing rate has been established only very recently [BDG19, JM21]. The situation with the singular McKean–Vlasov equations presented above is much more complicated and very few results are available in the literature. In particular, the results of [LSZ20] do not provide a way to construct a numerical algorithm for solving (1.5) even in the stationary case considered there.

We study the problem of numerically solving singular McKean-Vlasov (MV) equations of a more general form than (1.5):

$$dX_t = H\left(t, X_t, Y_t, \mathsf{E}[A_1(Y_t)|X_t]\right) dt + F\left(t, X_t, Y_t, \mathsf{E}[A_2(Y_t)|X_t]\right) dW_t, \quad (1.10)$$

where H, F, A_1, A_2 are sufficiently regular functions, W is a d-dimensional Brownian motion, and Y is a given stochastic process, for example, a diffusion process. Note that if one considers the Euler scheme, then a key issue is how to approximate the conditional expectation $\mathsf{E}[A_i(Y_t)|X_t = x]$, $i = 1, 2$, $x \in \mathbb{R}^d$.

One approach to tackle this problem was suggested by Guyon and Henry-Labordère [GHL12] (see also [AKH02]). They used the "identity"

$$\mathsf{E}[A(Y_t) \mid X_t = x] \text{“=”} \frac{\mathsf{E}A(Y_t)\delta_x(X_t)}{\mathsf{E}\delta_x(X_t)},$$

where δ_x is the Dirac delta function concentrated at x. This suggests the following approximation:

$$\mathsf{E}[A(Y_t) \mid X_t = x] \approx \frac{\sum_{i=1}^N A(Y_t^{i,N}) k_\varepsilon(X_t^{i,N} - x)}{\sum_{i=1}^N k_\varepsilon(X_t^{i,N} - x)}. \quad (1.11)$$

Here $\varepsilon > 0$ is a small parameter, $k_\varepsilon(\cdot) \approx \delta_0(\cdot)$ is a regularizing kernel, and $(X^{i,N}, Y^{i,N})_{i=1\dots N}$ is a particle system. While this method provides a way of constructing solutions to (1.10), it has an important disadvantage. One has to take $\varepsilon > 0$ small enough, but then (1.11) completely ignores the complicated structure of dependence of Y on X outside a tiny region $(x - C\varepsilon, x + C\varepsilon)$ for a certain $C > 0$ (indeed $k_\varepsilon(X_t^{i,N} - x) \approx 0$ outside that region).

As an alternative to [GHL12] we propose in this paper a novel approach based on ridge regression in the context of reproducing kernel Hilbert spaces (RKHS) which, in particular, does not have this disadvantage.

Let us recall that an RKHS $\mathcal{H}$ is a Hilbert space of real valued functions $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$, such that the evaluation map $\mathcal{H} \ni f \to f(x)$ is continuous for every $x \in \mathcal{X}$. This crucial property implies that there exists a positive symmetric kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that, for every $x \in \mathcal{X}$, $k_x := k(\cdot, x)$ $\in \mathcal{H}$ and one has that $\langle f, k_x \rangle_{\mathcal{H}} = f(x)$, for all $f \in \mathcal{H}$. As a main feature, any positive definite kernel k uniquely determines a RKHS $\mathcal{H}$ and the other

way around. For a detailed introduction and further properties of RKHS we refer to the literature, for example [SC08, Chapter 4]. We recall that the RKHS framework is popular in machine learning, where it is widely used for computing conditional expectations.

Consider a pair of random variables (X, Y) taking values in $\mathcal{X} \times \mathcal{X}$ with finite second moments and denote $\nu := \mathrm{Law}(X, Y)$. Suppose that $A\colon \mathcal{X} \to \mathbb{R}$ is sufficiently regular and $\mathcal{H}$ is large enough so that we have $\mathsf{E}\left[A(Y) \mid X = \cdot\right] \in \mathcal{H}$. Then, formally,

$$\begin{aligned} c_A^\nu(\cdot) := \int_{\mathcal{X}\times\mathcal{X}} k(\cdot, x) A(y) \nu(dx, dy) &= \int_{\mathcal{X}} k(\cdot, x) \nu(dx, \mathcal{X}) \int_{\mathcal{X}} A(y) \nu(dy|x) \\ &= \int_{\mathcal{X}} k(\cdot, x) \mathsf{E}\left[A(Y)|X = x\right] \nu(dx, \mathcal{X}) \\ &=: \mathcal{C}^\nu \mathsf{E}\left[A(Y)|X = \cdot\right], \end{aligned}$$

where

$$\mathcal{C}^\nu f := \int_{\mathcal{X}} k(\cdot, x) f(x) \nu(dx, \mathcal{X}), \qquad \text{for} \qquad f \in \mathcal{H}.$$

Unfortunately, in general, the operator $\mathcal{C}^\nu$ is not invertible. As $\mathcal{C}^\nu$ is positive definite, it is, however, possible to *regularize* the inversion by replacing $\mathcal{C}^\nu$ by $\mathcal{C}^\nu + \lambda I_{\mathcal{H}}$ for some $\lambda > 0$. Indeed, it turns out that

$$m_A^\lambda(\cdot; \nu) := (\mathcal{C}^\nu + \lambda I_{\mathcal{H}})^{-1} c_A^\nu, \tag{1.12}$$

is the solution to the minimization problem

$$m_A^\lambda(\cdot; \nu) := \operatorname*{arg\,min}_{f \in \mathcal{H}} \big(\mathsf{E}(A(Y) - f(X))^2 + \lambda \|f\|_{\mathcal{H}}^2 \big), \tag{1.13}$$

see Proposition 3.1. On the other hand one also has

$$\mathsf{E}[A(Y)|X = \cdot] = \operatorname*{arg\,min}_{f \in L_2(\mathbb{R}^d, \mathrm{Law}(X))} \mathsf{E}(A(Y) - f(X))^2,$$

and therefore it is natural to expect that if $\lambda > 0$ is small enough and $\mathcal{H}$ is large enough, then $m_A^\lambda(\cdot; \nu) \approx \mathsf{E}[A(Y)|X = \cdot]$, i.e. $m_A^\lambda(\cdot; \nu)$ is close to the true conditional expectation.

The main result of the article is that the regularized MV system obtained by replacing the conditional expectations with their regularized versions (1.12) in (1.10) is well-posed and propagation of chaos holds for the corresponding particle system, see Theorem 2.1 and Theorem 2.2.

To establish these theorems, we study the joint regularity of $m_A^\lambda(x; \nu)$ in the space variable x, and the measure ν for fixed $\lambda > 0$. These type of results are almost absent in the literature on RKHS and we here fill this gap. In particular, we prove that under suitable conditions, $m_A^\lambda(x; \nu)$ is Lipschitz in both arguments, i.e. w.r.t. the standard Euclidean norm in x and the

Wasserstein-1-norm in ν, and, can be calculated numerically in an efficient way, see Section 2. Additionally, in Section 3 we study the convergence of $m_A^\lambda(\cdot;\nu)$ in (1.12) to the true conditional expectation for fixed ν as $\lambda \downarrow 0$.

Let us note that, as a further nice feature of the RKHS approach compared to the δ-like) kernel method of [GHL12], one may incorporate, at least in principle, possible global prior information concerning properties of $\mathsf{E}[A(Y)|X = \cdot]$ into the choice of the RKHS generating kernel k (e.g. smoothness, tail or growth behavior). This degree of freedom is similar to, for example, how one can choose the basis functions in the usual regression methods for American options. We also note that the Lipschitz constants for $m_A^\lambda(\cdot;\nu)$ with respect to both arguments are expressed in bounds related to A and the kernel k, only, see Theorem 2.3. In contrast, if we would have dealt with standard ridge regression, i.e. ridge regression based on a system of basis functions, we would have to control bounds for possibly infinitely many basis functions, which is considered to be a far more delicate task.

Thus, the contribution of the current work is fourfold. First, we propose a RKHS-based approach to regularize (1.10) and prove the well-posedeness of the regularized equation. Second, we show convergence of the approximation (1.13) to the true conditional expectation as $\lambda \downarrow 0$. Third, we suggest a particle based approximation of the regularized equation and analyze its convergence. Finally, we apply our algorithm to the problem of smile calibration in finance and illustrate its performance on simulated data. In particular, we validate our results by solving numerically a regularized version of (1.5) (with m_A^λ in place of the conditional expectation). We show that our system is indeed an approximate solution to (1.5) in the sense that we get very close fits of the implied volatility surface — the final goal of the smile calibration problem.

The rest of the paper is organized as follows. Our main theoretical results are given in Section 2. Convergence properties of the regularized conditional expectation m_A^λ are established in Section 3. A numerical algorithm for solving (1.10) and an efficient implementable approximation of m_A^λ are discussed in Section 4. Section 5 contains numerical examples. The results of the paper are summarized in Section 6. Finally, all the proofs are placed in Section 7.

Convention on constants. Throughout the paper C denotes a positive constant whose value may change from line to line. The dependence of constants on parameters if needed will be indicated, e.g, $C(\lambda)$.

Acknowledgements. The authors are grateful to Peter Friz and Mykhaylo Shkolnikov for useful discussions. CB, OB, and JS are supported by the DFG Research Unit FOR 2402.

2. Main results

We begin by introducing the basic notation. For $a \in \mathbb{R}$ we denote $a_+ := \max(a, 0)$. Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. For $d \in \mathbb{N}$, let $\mathcal{X} \subset \mathbb{R}^d$ be an open subset, and $\mathcal{P}_2(\mathcal{X})$ be the set of all probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with finite second moment. If $\mu, \nu \in \mathcal{P}_2(\mathcal{X})$, $p \in [1, 2]$, then we denote the *Wasserstein-p* (Kantorovich) distance between them by

$$\mathbb{W}_p(\mu, \nu) := \inf(\mathsf{E}|X - Y|^p)^{1/p},$$

where the infimum is taken over all random variables X, Y with $\mathrm{Law}(X) = \mu$, $\mathrm{Law}(Y) = \nu$.

Let $\mathcal{C}^1(\mathcal{X}, \mathbb{R})$ be the space of all functions $f\colon \mathcal{X} \to \mathbb{R}$ such that

$$\|f\|_{\mathcal{C}^1} := \sup_{x \in \mathcal{X}} |f(x)| + \sup_{\substack{x \in \mathcal{X} \\ i=1,\dots,d}} |\partial_{x_i} f(x)| < \infty.$$

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric, positive definite kernel, and $\mathcal{H}$ be a reproducing kernel Hilbert space of functions $f\colon \mathcal{X} \to \mathbb{R}$ associated with the kernel k. That is, for any $x \in \mathcal{X}$, $f \in \mathcal{H}$ one has

$$f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}}.$$

In particular, $\langle k(x, \cdot), k(y, \cdot)\rangle_{\mathcal{H}} = k(x, y)$, for any $x, y \in \mathcal{X}$. We refer to [SC08, Chapter 4] for further properties of RKHS.

Let $A\colon \mathcal{X} \to \mathbb{R}$ be a measurable function such that $|A(x)| \le C(1 + |x|)$ for some universal constant $C > 0$ and all $x \in \mathcal{X}$. For $\nu \in \mathcal{P}_2(\mathcal{X} \times \mathcal{X})$, $\lambda \ge 0$ consider the following optimization problem (*ridge regression*)

$$m_A^\lambda(\cdot; \nu) := \arg\min_{f \in \mathcal{H}} \left\{ \int_{\mathcal{X} \times \mathcal{X}} |A(y) - f(x)|^2 \, \nu(dx, dy) + \lambda \|f\|_{\mathcal{H}}^2 \right\}. \tag{2.1}$$

We fix $T > 0$, $d \in \mathbb{N}$ and consider the system

$$dX_t = H(t, X_t, Y_t, \mathsf{E}[A_1(Y_t)|X_t])dt + F(t, X_t, Y_t, \mathsf{E}[A_2(Y_t)|X_t])dW_t^X \tag{2.2a}$$

$$dY_t = b(t, Y_t)dt + \sigma(t, Y_t)dW_t^Y, \tag{2.2b}$$

where $H\colon [0, T] \times \mathcal{X} \times \mathcal{X} \times \mathbb{R} \to \mathbb{R}^d$, $F\colon [0, T] \times \mathcal{X} \times \mathcal{X} \times \mathbb{R} \to \mathbb{R}^d \times \mathbb{R}^d$, $A_i\colon \mathcal{X} \to \mathbb{R}$, $b\colon [0, T] \times \mathcal{X} \to \mathbb{R}^d$, $\sigma\colon [0, T] \times \mathcal{X} \to \mathbb{R}^d \times \mathbb{R}^d$ are measurable functions, W^X, W^Y are two (possibly correlated) d-dimensional Brownian motions on $(\Omega, \mathcal{F}, \mathsf{P})$, and $t \in [0, T]$. It is also implicitly assumed that the space $\mathcal{X} \times \mathcal{X}$ is invariant for the process (X, Y) (which is trivially the case when $\mathcal{X} = \mathbb{R}^d$, however for our purposes we will be mostly interested in the case when $\mathcal{X} = \mathbb{R}^d_+$).

As explained above, denoting $\mu_t := \mathrm{Law}(X_t, Y_t)$, we see that the functional $(x, \mu_t) \mapsto \mathsf{E}\left[A_i(Y_t) \mid X_t = x\right]$ is not Lipschitz continuous even if A_i is

smooth. Therefore the classical results on well-posedness of McKean–Vlasov equations are not applicable to (2.2).

The main idea of our approach is to replace the conditional expectation by an approximation which has "nice" properties (in particular, it is Lipschitz). This would imply strong existence and uniqueness of the new system. Furthermore, we will demonstrate numerically that the solution to the new system is still "close" to the solution of (2.2) in a certain sense. Thus, we consider the following system:

$$d\widehat{X}_t = H(t, \widehat{X}_t, Y_t, m^{\lambda}_{A_1}(\widehat{X}_t; \widehat{\mu}_t))dt + F(t, \widehat{X}_t, Y_t, m^{\lambda}_{A_2}(\widehat{X}_t; \widehat{\mu}_t))dW^X_t, \quad (2.3a)$$

$$dY_t = b(t, Y_t)dt + \sigma(t, Y_t)dW^Y_t \quad (2.3b)$$

$$\widehat{\mu}_t = \mathrm{Law}(\widehat{X}_t, Y_t). \quad (2.3c)$$

where $t \in [0, T]$. We would need the following assumption on the kernel k.

Assumption K. The kernel k is twice continuously differentiable in both variables, $k(x, x) > 0$ for all $x \in \mathcal{X}$, and

$$D^2_k := \sup_{\substack{(x,y)\in\mathcal{X}\times\mathcal{X} \\ 1\le i,j\le d}} \max\big\{|\partial_{x_i}\partial_{y_j}k^2(x, y)|, |\partial_{x_i}\partial_{y_j}k(x, y)|, |\partial_{x_i}k(x, y)|, |\partial_{y_j}k(x, y)|, |k(x, y)|\big\} < \infty$$

Now we are ready to state our main results. Their proofs are given in Section 7.

Theorem 2.1. *Suppose that Assumption K is satisfied for the kernel k, the space $\mathcal{X} \times \mathcal{X}$ is invariant for the process $(\widehat{X}, Y)$ and*

(1) *$A_i \in \mathcal{C}^1(\mathcal{X}, \mathbb{R})$, $i = 1, 2$;*

(2) *there exists a constant $C > 0$ such that for any $t \in [0, T]$, $x, y, x', y' \in \mathbb{R}^d$, $z, z' \in \mathbb{R}$,*

$$\begin{aligned} &|H(t, x, y, z) - H(t, x', y', z')| + |F(t, x, y, z) - F(t, x', y', z')| \\ &\quad + |b(t, y) - b(t, y')| + |\sigma(t, y) - \sigma(t, y')| \\ &\le C(|x - x'| + |y - y'| + |z - z'|); \end{aligned}$$

(3) *for any fixed $x, y, \in \mathbb{R}^d$, $z \in \mathbb{R}$ one has*

$$\int_0^T (|H(t, x, y, z)|^2 + |F(t, x, y, z)|^2 + |b(t, y)|^2 + |\sigma(t, y)|^2)\, dt < \infty;$$

(4) *$\mathsf{E}|\widehat{X}_0|^2 < \infty$, $\mathsf{E}|Y_0|^2 < \infty$.*

Then for any $\lambda > 0$ the system (2.3) with the initial condition $(\widehat{X}_0, Y_0)$ has a unique strong solution.

To analyze a numerical scheme solving (2.3), we consider a particle system

$$\begin{aligned} dX_t^{N,n} &= H\big(t, X_t^{N,n}, Y_t^{N,n}, m_{A_1}^{\lambda}(X_t^{N,n}; \mu_t^N)\big)dt \\ &\quad + F\big(t, X_t^{N,n}, Y_t^{N,n}, m_{A_2}^{\lambda}(X_t^{N,n}; \mu_t^N)\big)dW_t^{X,n} \qquad (2.4a) \\ dY_t^{N,n} &= b(t, Y_t^{N,n})\,dt + \sigma(t, Y_t^{N,n})\,dW_t^{Y,n} \qquad (2.4b) \\ \mu_t^N &= \frac{1}{N}\sum_{n=1}^{N} \delta_{(X_t^{N,n}, Y_t^{N,n})}, \qquad (2.4c) \end{aligned}$$

where $N \in \mathbb{N}$, $n = 1, \dots, N$, $t \in [0, T]$, and the pairs of $d \times d$-dimensional Brownian motions $(W^{X,n}, W^{Y,n})$, $n = 1, \dots, N$, are jointly independent and have the same law as (W^X, W^Y). The following propagation of chaos result holds; it establishes both weak and strong convergence of $X^{N,n}$.

Theorem 2.2. *Assume that all the conditions of Theorem 2.1 are satisfied. Suppose additionally that the functions H, F, b, σ are locally bounded and the initial conditions $(X_0^{N,n}, Y_0^{N,n})$ are jointly independent and have the same law as $(\widehat{X}_0, Y_0)$. Moreover, suppose that $\mathsf{E}|\widehat{X}_0|^q < \infty$, $\mathsf{E}|Y_0|^q < \infty$ for some $q > 4$. Then there exists a constant $C = C(\lambda, T, \mathsf{E}|\widehat{X}_0|^q, \mathsf{E}|Y_0|^q) > 0$ such that for any $N \in \mathbb{N}$*

$$\mathsf{E} \sup_{0 \le t \le T} |X_t^{N,1} - \overline{X}_t|^2 + \sup_{0 \le t \le T} \mathsf{E}[\mathbb{W}_2(\mu_t^N, \widehat{\mu}_t)^2] \le C N^{-1/2}, \qquad (2.5)$$

where the process $\overline{X}$ solves (2.3) with $W^{X,n}$, $W^{Y,n}$ in place of W^X, W^Y, respectively.

A crucial step which allowed us to obtain these results is the Lipschitz continuity of m^λ. The following holds.

Theorem 2.3. *Assume that the kernel k satisfies Assumption K. Let $A \in \mathcal{C}^1(\mathcal{X}, \mathbb{R})$. Then for any $x, y \in \mathcal{X}$, $\mu, \nu \in \mathcal{P}_2(\mathcal{X} \times \mathcal{X})$ on has*

$$|m_A^\lambda(x; \mu) - m_A^\lambda(y; \nu)| \le C_1 \mathbb{W}_1(\mu, \nu) + C_2 |x - y|,$$

where

$$C_1 := \left(\frac{D_k}{\lambda^2} + \frac{1}{\lambda}\right) d D_k^2 \|A\|_{\mathcal{C}^1} \quad \text{and} \quad C_2 := \frac{\sqrt{d}}{\lambda} D_k^2 \|A\|_{\mathcal{C}^1}.$$

This result is interesting for at least two reasons. First, it shows that m_A^λ is Lipschitz continuous in both arguments, provided that the kernel k is smooth enough. That is, the Lipschitz continuity property depends on $\mathcal{H}$ only through the smoothness of the kernel k. Second, this result gives an explicit dependence of the corresponding Lipschitz constant on λ and k.

Remark 2.4. Let us stress that Theorem 2.1 establishes the existence and uniqueness of (2.2) only for a fixed regularisation parameter $\lambda > 0$ and can not be used to study the limiting case $\lambda \to 0$. Indeed, it follows from Theorem 2.3, that as $\lambda \to 0$, the Lipschitz constants of m_A^λ blows up. Yet, we will demonstrate numerically in Section 5, that, actually, as $\lambda \to 0$ the solution to (2.2) does not blow up; on the contrary it weakly converges to a limit; this hints that (at least) weak existence of solutions to (1.10) should hold. Verifying this theoretically remains however an important open problem.

3. Approximation of conditional expectations

In this section we study the approximation m_A^λ introduced in (2.1) in more detail. Throughout this section we fix an open set $\mathcal{X} \subset \mathbb{R}^d$, a measure $\nu \in \mathcal{P}_2(\mathcal{X} \times \mathcal{X})$, and impose the following relatively weak assumptions on the function $A\colon \mathcal{X} \to \mathbb{R}$ and the positive kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Assumption A°. The function A is sublinear, i.e. there exists a constant $C > 0$ such that for all $x \in \mathcal{X}$ one has $|A(x)| \le C(1 + |x|)$.

Assumption K°. The kernel $k(\cdot, \cdot)$ is continuous on $\mathcal{X} \times \mathcal{X}$ and satisfies $0 < k(x, x) \le C(1 + |x|^2)$ for some $C > 0$.

It is easy to see that Assumption K° implies for any $x \in \mathcal{X}$

$$\|k(x, \cdot)\|_{\mathcal{H}}^2 = \langle k(x, \cdot), k(x, \cdot)\rangle_{\mathcal{H}} = k(x, x) \le C(1 + |x|^2). \tag{3.1}$$

Due to Assumption K° and [SC08, Lemma 4.33], $\mathcal{H}$ is a separable RKHS and one has for any $f \in \mathcal{H}$, $x \in \mathcal{X}$,

$$|f(x)| = |\langle k(x, \cdot), f\rangle_{\mathcal{H}}| \le \|k(x, \cdot)\|_{\mathcal{H}}\|f\|_{\mathcal{H}} \le C(1 + |x|)\|f\|_{\mathcal{H}}, \tag{3.2}$$

where we also used (3.1). Hence, every $f \in \mathcal{H}$ is sublinear and, as a consequence, for any fixed $\nu \in \mathcal{P}_2(\mathcal{X} \times \mathcal{X})$, the objective functional in (2.1) is finite.

It is also easy to see that (3.2) and (3.1) imply that for any $x, y \in \mathcal{X}$

$$k(x, y) \le C(1 + |x|)\|k(\cdot, y)\|_{\mathcal{H}} = C(1 + |x|)(1 + |y|). \tag{3.3}$$

Therefore, the Bochner integrals

$$c_A^\nu := \int_{\mathcal{X}\times\mathcal{X}} k(\cdot, x)A(y)\nu(dx, dy), \quad \text{and} \quad \mathcal{C}^\nu f := \int_{\mathcal{X}\times\mathcal{X}} k(\cdot, x)f(x)\nu(dx, dy). \tag{3.4}$$

are well defined functions in $\mathcal{H}$ for every $f \in \mathcal{H}$. Moreover, operator $\mathcal{C}^\nu\colon \mathcal{H} \to \mathcal{H}$ is symmetric and positive definite since

$$\langle g, \mathcal{C}^\nu f\rangle = \int_{\mathcal{X}} \langle g, k(\cdot, x)\rangle\, f(x)\nu(dx, \mathcal{X}) = \int_{\mathcal{X}} g(x)f(x)\nu(dx, \mathcal{X}).$$

Thus, by the Hellinger-Toeplitz theorem, $\mathcal{C}^\nu$ is a bounded self-adjoint linear operator on $\mathcal{H}$. As a consequence, for any $\lambda \geq 0$, the operator $\mathcal{C}^\nu + \lambda I_{\mathcal{H}}$ is a bounded self-adjoint operator on $\mathcal{H}$ with spectrum contained in the interval $[\lambda, \|\mathcal{C}^\nu\| + \lambda]$. Hence, if $\lambda > 0$, then $(\mathcal{C}^\nu + \lambda I_{\mathcal{H}})^{-1}$ exists and is a bounded self-adjoint operator on $\mathcal{H}$ with norm

$$\|(\mathcal{C}^\nu + \lambda I_{\mathcal{H}})^{-1}\|_{\mathcal{H}} \leq \lambda^{-1}. \tag{3.5}$$

We are now ready to state the following useful representation for the solution to (2.1).

Proposition 3.1. *Under Assumptions A°,K°, for any fixed $\nu \in \mathcal{P}_2(\mathcal{X} \times \mathcal{X})$ and $\lambda > 0$, the solution to (2.1) can be represented as*

$$m_A^\lambda(\cdot; \nu) = (\mathcal{C}^\nu + \lambda I_{\mathcal{H}})^{-1} c_A^\nu. \tag{3.6}$$

This representation may be seen as an infinite sample version of the usual solution representation for a ridge regression problem based on finite samples. We thus consider it as not essentially new, but, in order to keep our paper as self contained as possible we present a proof of it in Section 7. Proposition 3.1 allows us to prove Lipschitz continuity of m_A^λ, that is Theorem 2.3.

Let us now proceed with investigating when the function $m_A^\lambda = m_A^\lambda(\cdot; \nu)$ is a "good" approximation to the true conditional expectation

$$m_A = m_A(x, \nu) := \mathsf{E}_{(X,Y)\sim\nu}[A(Y)|X = x] \tag{3.7}$$

for small enough $\lambda > 0$.

Assume from now on that the measure $\nu \in \mathcal{P}_2(\mathcal{X} \times \mathcal{X})$ is nondegenerate with respect to the first variable. That is, for every open set $U \subset \mathcal{X}$ one has $\nu(U, \mathcal{X}) > 0$. Consider the Hilbert space $\mathcal{L}_2^\nu := L_2(\mathcal{X}, \nu(dx, \mathcal{X}))$. For $f \in \mathcal{L}_2^\nu$ put

$$T^\nu f := \int_{\mathcal{X}} k(\cdot, x) f(x) \nu(dx, \mathcal{X}). \tag{3.8}$$

Recalling (3.3), it is easy to see that T^ν is a linear operator $\mathcal{L}_2^\nu \to \mathcal{L}_2^\nu$. Note that that $\mathcal{H} \subset \mathcal{L}_2^\nu$ due to (3.2); thus, $\mathcal{C}^\nu$ is the restriction of T^ν to $\mathcal{H}$. Further, since $k(x, y) \leq \sqrt{k(x, x)}\sqrt{k(y, y)}$, the kernel k is Hilbert-Schmidt on $\mathcal{L}_2(\mathcal{X} \times \mathcal{X}, \nu(dx, \mathcal{X})\nu(dy, \mathcal{X}))$, i.e.

$$\int k^2(x, y)\nu(dx, \mathcal{X})\nu(dy, \mathcal{X}) < \infty,$$

due to Assumption K°. As a consequence of the standard results from functional analysis, one then has (see for example [Kre89]):

(i) the operator T^ν is self-adjoint and compact;

(ii) there exists an orthonormal system $(a_n)_{n\in\mathbb{N}}$ in $\mathcal{L}_2^\nu$ of continuous eigenfunctions corresponding to nonnegative eigenvalues σ_n of T^ν and $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots$;

(iii) If $J := \{n \in \mathbb{N} : \sigma_n > 0\}$, one has

$$T^\nu f = \sum_{n\in J} \sigma_n \langle f, a_n\rangle_{\mathcal{L}_2^\nu} a_n, \quad f \in \mathcal{L}_2^\nu$$

with $\lim_{n\to\infty} \sigma_n = 0$ if $J = \mathbb{N}$.

It is easy to see that a generalization of Mercer's theorem to unbounded domains [Sun05] implies the following statement.

Proposition 3.2. *Let k be a kernel satisfying Assumption K°. Then k has a series representation*

$$k(x,y) = \sum_{n\in J} \sigma_n a_n(x) a_n(y), \quad x, y \in \mathcal{X} \tag{3.9}$$

with uniform convergence on compact sets. Moreover, $(\widetilde{a}_n)_{n\in J}$ with $\widetilde{a}_n := \sqrt{\sigma_n} a_n$ is an orthonormal basis of $\mathcal{H}$ and the scalar product in $\mathcal{H}$ takes the form

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{n\in J} \frac{\langle f, a_n\rangle_{\mathcal{L}_2^\nu} \langle g, a_n\rangle_{\mathcal{L}_2^\nu}}{\sigma_n} \quad \text{for} \quad f, g \in \mathcal{H}. \tag{3.10}$$

Now we are ready to present the main result of this section next result, which quantifies the convergence properties of $m_A^\lambda(\cdot,\nu)$ as $\lambda \to 0$ for a fixed measure ν. Recall the notation (3.7). Let $P_{\bar{\mathcal{H}}}$ denote the orthogonal projection of $\mathcal{L}_2^\nu$ onto the closure $\overline{\mathcal{H}}$ of $\mathcal{H}$ in $\mathcal{L}_2^\nu$.

Theorem 3.3. *Assume that the kernel k satisfies Assumption K° and the function A is bounded measurable. Then $m_A \in \mathcal{L}_2^\nu$ and for any $\lambda > 0$*

$$\left\| P_{\bar{\mathcal{H}}} m_A - m_A^\lambda(\cdot;\nu) \right\|_{\mathcal{L}_2^\nu}^2 = \sum_{n\in J} \frac{\lambda^2}{(\sigma_n + \lambda)^2} \langle m_A, a_n\rangle_{\mathcal{L}_2^\nu}^2 . \tag{3.11}$$

In particular, $\left\| P_{\bar{\mathcal{H}}} m_A - m_A^\lambda(\cdot;\nu) \right\|_{\mathcal{L}_2^\nu} \to 0$ as $\lambda \downarrow 0$.

If, moreover, $P_{\bar{\mathcal{H}}} m_A \in \mathcal{H}$ one has

$$\left\| P_{\bar{\mathcal{H}}} m_A - m_A^\lambda(\cdot;\nu) \right\|_{\mathcal{H}}^2 = \sum_{n\in J} \frac{\lambda^2}{(\sigma_n + \lambda)^2 \sigma_n} \langle m_A, a_n\rangle_{\mathcal{L}_2^\nu}^2 , \tag{3.12}$$

and thus $\left\| P_{\bar{\mathcal{H}}} m_A - m_A^\lambda(\cdot;\nu) \right\|_{\mathcal{H}} \to 0$ for $\lambda \downarrow 0$.

Theorem 3.3 establishes convergence of $m_A^\lambda(\cdot;\nu)$ as $\lambda \to 0$ though without a rate. Its proof is placed in Section 7. Additional assumptions are needed to guarantee a certain convergence rate. This is done in the following corollary.

Corollary 3.4. *Suppose that the conditions of Theorem 3.3 are satisfied, and that moreover for some* $\theta \in (0, 1]$,

$$\sum_{n\in J} \sigma_n^{-\theta} \langle P_{\bar{\mathcal{H}}} m_A, a_n\rangle^2_{\mathcal{L}_2^\nu} < \infty. \tag{3.13}$$

Then

$$\left\| P_{\bar{\mathcal{H}}} m_A - m_A^\lambda(\cdot;\nu)\right\|^2_{\mathcal{L}_2^\nu} \le \left(1 - \frac{\theta}{2}\right)^2 \left(\frac{\lambda\theta}{2-\theta}\right)^\theta \sum_{n\in J} \sigma_n^{-\theta} \langle m_A, a_n\rangle^2_{\mathcal{L}_2^\nu}. \tag{3.14}$$

In particular, if $\theta = 1$, *that is* $P_{\bar{\mathcal{H}}} m_A \in \mathcal{H}$, *we get*

$$\left\| P_{\bar{\mathcal{H}}} m_A - m_A^\lambda(\cdot;\nu)\right\|_{\mathcal{L}_2^\nu} \le \frac{\sqrt{\lambda}}{2} \left\| P_{\bar{\mathcal{H}}} m_A\right\|_{\mathcal{H}}. \tag{3.15}$$

Proof. Inequality (3.14) follows from (3.11), (3.13), and the fact that the maximum of the function $x \mapsto \lambda^2 x^\theta/(x+\lambda)^2$, $x > 0$, is equal to

$$(1 - \theta/2)^2(\lambda\theta/(2-\theta))^\theta.$$

Inequality (3.15) follows from (3.14) and (3.10). □

Remark 3.5. If operator T^ν defined in (3.8) is injective, then $P_{\bar{\mathcal{H}}} = I_{\mathcal{L}_2^\nu}$. In this case $J = \mathbb{N}$ and Theorem 3.3 and Corollary 3.4 quantify the convergence to the true conditional expectation.

Thus, in this section we have shown that, under certain conditions, $m_A^\lambda(\cdot,\nu)$ may converge at least in $\mathcal{L}_2^\nu$-sense to the true conditional expectation $m_A(\cdot,\nu)$ as $\lambda \to 0$. This makes the heuristic discussion around (1.12) and (1.13) in Section 1 more rigorous.

Remark 3.6. Note that the measure $\widehat{\mu}_t$ in the solution of (2.3) depends on λ, so in fact $\widehat{\mu}_t = \widehat{\mu}_t^\lambda$. Therefore, even when $m_A^\lambda(\cdot,\nu) \to m_A(\cdot,\nu)$ for fixed ν and $\lambda \downarrow 0$, the question whether $m_{A_i}^\lambda(\cdot,\widehat{\mu}_t^\lambda)$ converges in some sense is still not answered. We believe that this question is intimately linked to the problem of existence of a solution to (2.2). As already explained, this is an unsolved open problem and therefore considered out of our scope. However, loosely speaking, assuming that the latter system has indeed a solution (in some sense) with solution measure μ_t say, it is natural to expect that for a suitable "rich enough" RKHS, $m_{A_i}^\lambda(\cdot,\mu_t) \to m_{A_i}(\cdot,\mu_t)$ (the true conditional expectation) as $\lambda \downarrow 0$.

Thus, as follows from the discussion above, the space $\mathcal{H}$ has to be large enough, otherwise there is no hope of convergence of $m_{A_i}^\lambda(\cdot,\widehat{\mu}_t^\lambda)$ to the true conditional expectation. Fortunately, there is a great flexibility for the choice of the kernel k and thus RKHS $\mathcal{H}$. For instance, starting with a simply structured infinite dimensional kernel, k_∞ (e.g., the Gaussian kernel), in general

it may happen that $m^{\lambda}_{A_i}$, $i = 1, 2$, have poor approximation properties in the RKHS $\mathcal{H}_\infty$ generated by k_∞. In such a case we may add another kernel to it, which incorporates possible prior information of $m^{\lambda}_{A_i}$ such as shape or growth behavior. For example, suppose one anticipates that $m^{\lambda}_{A_i}$ follows "roughly" some functions in the linear span of some suitably chosen set of basis functions, say, $\psi_1, ..., \psi_K$. One then may consider the RKHS $\mathcal{H} := \mathcal{H}_\infty \oplus \mathcal{H}_\psi$ generated by the kernel

$$k(x, y) := k_\infty(x, y) + k_\psi(x, y) := k_\infty(x, y) + \sum_{k=1}^{K} \psi_k(x)\psi_k(y), \tag{3.16}$$

where k_ψ generates $\mathcal{H}_\psi$, and, without loss of generality, $\mathcal{H}_\infty \cap \mathcal{H}_\psi = \{0\}$.

Of course, in this line of reasoning, we think of K being a "very low" number. The simplest extension one may think of is adding a constant, i.e. $K = 1$ and $\psi_1 \equiv c \neq 0$. Then, clearly, $1 \in \mathcal{H}$ and, for instance, if X and Y are independent, one then has that $\mathbb{E}[Y|X = \cdot] = \mathbb{E}[Y] \in \mathcal{H}$.

As another example, in the context of (1.3) one may think of a given stochastic volatility process v_t such that (1.3) with $\sigma = 1$ explains the market prices up to a large extend already. One then may expect that, in the solution of (1.5), $\mathsf{E}[v_t|S_t = \cdot]$ is roughly proportional to $\sigma^2_{\mathrm{Dup}}(t, \cdot)$. This suggest to chose a (time dependent) kernel of the form (3.16) with $\psi_1(t, \cdot) := \sigma^2_{\mathrm{Dup}}(t, \cdot)$. The advantage of a suitable kernel extension is best seen in the case where $\psi_1(t, \cdot) = \sigma^2_{\mathrm{Dup}}(t, \cdot)/\sigma^2(t, \cdot)$ due to some oracle. Then m^{λ}_{A} in the solution of the regularized version of (1.5) is expected to be found "almost" in the one dimensional space $\mathcal{H}_\psi$.

4. Numerical algorithm

Let us now describe in details our numerical algorithm to construct solutions to (1.10). We begin by discussing an efficient way of calculating m^{λ}_{A}.

4.1 Estimation of the conditional expectation

Let us recall that in order to solve the particle system (2.4) we need to compute

$$m^{\lambda}_{A}(\cdot; \mu^N_t) = \arg\min_{f \in \mathcal{H}} \Big\{ \frac{1}{N} \sum_{n=1}^{N} |A(Y^{N,n}_t) - f(X^{N,n}_t)|^2 + \lambda \|f\|^2_{\mathcal{H}} \Big\}. \tag{4.1}$$

for t belonging to a certain partition of $[0, T]$ and fixed large $N \in \mathbb{N}$; here $A = A_1$ or $A = A_2$. It follows from the representer theorem for RKHS [SHS01, Theorem 1] that m^{λ}_{A} has the following representation:

$$m^{\lambda}_{A}(\cdot; \nu^N_t) = \sum_{i=1}^{N} \alpha_i k(X^{N,i}_t, \cdot), \tag{4.2}$$

for some $\alpha = (\alpha_1, \dots, \alpha_N)^T \in \mathbb{R}^N$. Note that the optimal α can be calculated explicitly by plugging the representation (4.2) into the above minimization problem in place of f and minimizing over α. However, computing the optimal α directly takes $O(N^3)$ operations, which is prohibitively expensive keeping in mind that the number of particles N is going to be very large. Furthermore, even evaluating (4.2) at $X_t^{N,n}$, $n = 1, \dots, N$, for a given $\alpha \in \mathbb{R}^N$ is rather expensive, it requires $O(N^2)$ operations, and thus is impossible to implement.

To develop an implementable algorithm, let us note that many particles $X_t^{N,i}$ — and, as a consequence, the implied basis functions $k(X_t^{N,i}, \cdot)$ — will be close to each other. Therefore, we can considerably reduce the computational cost by only using $L \ll N$ rather than N basis functions as suggested in (4.2). More precisely, we choose $Z^1, \dots, Z^L$ among $X_t^{1,N}, \dots, X_t^{N,N}$ – e.g., by random choice or taking every $\frac{N}{L}$th point among the ordered sequence $X_t^{N,(1)}, \dots, X_t^{N,(N)}$ in case that X is one-dimensional – and approximate

$$\sum_{i=1}^{N} \alpha_i k(X_t^{N,i}, \cdot) \approx \sum_{j=1}^{L} \beta_j k(Z^j, \cdot), \tag{4.3}$$

where $\beta = (\beta_1, \dots, \beta_L)^T \in \mathbb{R}^L$. It is easy to see that

$$\begin{aligned} \Big\| \sum_{j=1}^{L} \beta_j k(Z^j, \cdot) \Big\|_{\mathcal{H}} &= \Big\langle \sum_{j=1}^{L} \beta_j k(Z^j, \cdot), \sum_{j=1}^{L} \beta_j k(Z^j, \cdot) \Big\rangle_{\mathcal{H}} \\ &= \sum_{j,k=1}^{L} \beta_j \beta_k \langle k(Z^j, \cdot), k(Z^k, \cdot) \rangle_{\mathcal{H}} \\ &= \sum_{j,k=1}^{L} \beta_j \beta_k k(Z^j, Z^k) = \beta^\top R \beta, \end{aligned}$$

where $R := (k(Z^j, Z^k))_{j,k=1,\dots,L}$ is an $L \times L$ matrix. Thus, recalling (4.1), we see that we have to solve

$$\arg\min_{\beta \in \mathbb{R}^L} [\frac{1}{N} (G - K\beta)^\top (G - K\beta) + \lambda \beta^\top R \beta],$$

where $G := (A(Y_t^{N,n}))_{n=1,\dots,N}$, $K := (k(Z^j, X_t^{N,n}))_{n=1,\dots,N, j=1,\dots,L}$ is an $N \times L$ matrix. Differentiating with respect to β, we get that the optimal value $\widehat{\beta} = \widehat{\beta}((X_t^N), (Y_t^N))$ satisfies

$$(K^\top K + N\lambda R)\widehat{\beta} = K^\top G, \tag{4.4}$$

and we approximate expectation as

$$m_A^\lambda(x; \mu_t^N) \approx \sum_{j=1}^{L} \widehat{\beta}_j k(Z^j, x) =: \widehat{m}_A^\lambda(x; \mu_t^N). \tag{4.5}$$

Remark 4.1. Let us see how many operations we need to calculate $\widehat{\beta}$, taking into account that $L \ll N$. We need $O(NL)$ to calculate K, $O(L^2)$ to calculate R, $O(NL^2)$ to calculate $K^\top K$ (this is the bottleneck); $O(L^3)$ to invert $K^\top K + N\lambda R$ and $O(NL)$ to calculate $K^\top G$ and solve (4.4). Thus, in total we would need $O(NL^2)$ operations.

4.2 Solving the McKean–Vlasov equation

With the function $\widehat{m}_A^\lambda$ in hand, we consider now the Euler scheme for the particle system (2.4). We fix time interval T, the number of time steps M, and, for simplicity, we consider a uniform time increment $\delta := T/M$. Let $\Delta W_i^{X,n}$ and $\Delta W_i^{Y,n}$ denote independent copies of $W^X_{(i+1)\delta} - W^X_{i\delta}$ and $W^Y_{(i+1)\delta} - W^Y_{i\delta}$, respectively, $n = 1, \dots, N$, $i = 1, \dots M$. Note that for stochastic volatility models, the Brownian motions driving the stock price and the variance process are usually correlated. We now define $\widetilde{X}_0^n = X_0^n$, $\widetilde{Y}_0^n = Y_0^n$, and for $i = 0, \dots, M-1$

$$\widetilde{X}^n_{i+1} = \widetilde{X}^n_i + H\left(i\delta, \widetilde{X}^n_i, \widetilde{Y}^n_i, \widehat{m}^\lambda_{A_1}(\widetilde{X}^n_i; \widetilde{\mu}^N_i)\right)\delta + F\left(i\delta, \widetilde{X}^n_i, \widetilde{Y}^n_i, \widehat{m}^\lambda_{A_2}(\widetilde{X}^n_i; \widetilde{\mu}^N_i)\right)\Delta W_i^{X,n} \tag{4.6a}$$

$$\widetilde{Y}^n_{i+1} = \widetilde{Y}^n_i + b(i\delta, \widetilde{Y}^n_i)\delta + \sigma(i\delta, \widetilde{Y}^n_i)\Delta W_i^{Y,n}, \tag{4.6b}$$

where $\widetilde{\mu}^N_i = \frac{1}{N}\sum_{n=1}^N \delta_{(\widetilde{X}^{N,n}_i, \widetilde{Y}^{N,n}_i)}$. Note that after each iteration we might need to update the values of $(\widetilde{X}^n_i, \widetilde{Y}^n_i)$ in order to ensure that they lie in $\mathcal{X} \times \mathcal{X}$ (this can be done, e.g., by replacing them by the closest to them point of $\mathcal{X} \times \mathcal{X}$).

We see that at each discretization time step of (4.6) we need to compute the approximations of the conditional expectations $\widehat{m}^\lambda_{A_r}(\widetilde{X}^n_i; \widetilde{\mu}^N_i)$, $r = 1, 2$. This is done using the algorithm discussed in Section 4.1, and takes $O(NL^2)$ operations, see Remark 4.1. Thus the total number of operations needed to implement (4.6) is $O(MNL^2)$.

5. Numerical examples and applications to local stochastic volatility models

As a main application of the regularisation approach presented above, we consider the problem of calibration of stochastic volatility models to market data. Fix $T > 0$. Let $C(t, K)$ be the price at time 0 of a European call option on a non-dividend paying stock X with strike K and maturity t, $K > 0$, $t \in [0, T]$. We assume that the market prices $(C(t, K))_{t>0, K>0}$ are given. To simplify the calculations, we suppose that the interest rate $r = 0$.

We study Local Stochastic Volatility (LSV) models. That is, we assume that the stock price X follows the dynamics

$$dX_t = \sqrt{Y_t}\sigma_{LV}(t, X_t)X_t dW_t^X, \quad t \in [0, T], \tag{5.1}$$

where W^X is a Brownian motion and (Y_t) is a strictly positive volatility process, both being adapted to some filtration $(\mathcal{F}_t)_{t\geq 0}$. As discussed in the introduction, if the function σ_{LV} is given by

$$\sigma^2_{\mathrm{LV}}(t, x) := \frac{\sigma^2_{\mathrm{Dup}}(t, x)}{\mathsf{E}\left[Y_t \middle| X_t = x\right]}, \quad t \in [0, T], x > 0,$$

where σ_{Dup} is the Dupire local volatility

$$\sigma^2_{\mathrm{Dup}}(t, x) := \frac{2\partial_t C(t, x)}{x^2 \partial_{xx} C(t, x)}, \quad t \in [0, T], x > 0, \tag{5.2}$$

then the model (5.1) is able to perfectly replicate the given call option prices (for any choice of the volatility process Y) [Dup94, Gyo86]. That is, one has the identity

$$C(t, K) = \mathsf{E}(X_t - K)_+, \quad t \in [0, T],\ K > 0. \tag{5.3}$$

In particular, the choice $Y \equiv 1$ recovers the local volatility model. In case where Y is a diffusion process

$$dY_t = b(t, Y_t)dt + \sigma(t, Y_t)dW_t^Y, \tag{5.4}$$

where W^Y is a Brownian motion possibly correlated with W^X, we see that the model (5.1)-(5.4) is a special case of the general McKean-Vlasov equation (2.2) with $\mathcal{X} = \mathbb{R}_+$.

To solve (5.1)-(5.4), we implement the algorithm described in Section 4: see (4.6) together with (4.5). To validate the results, we compare the call option prices obtained by the algorithm (that is $N^{-1}\sum_{n=1}^N (\widetilde{X}_M^n - K)_+$) with the given prices $C(T, K)$ for various $T > 0$ and $K > 0$. If the algorithm is correct and if $\widetilde{\mu}_M^N \approx \mathrm{Law}(X_T, Y_T)$, then, according to (5.3) one must have

$$C(t, K) \approx N^{-1}\sum_{n=1}^N (\widetilde{X}_M^n - K)_+ =: \widetilde{C}(t, K). \tag{5.5}$$

On the other hand, if the algorithm is not correct and the true law $\mathrm{Law}(X_T, Y_T)$ is very different from $\widetilde{\mu}_M^N$, then (5.5) will not hold.

We verify (5.5) in two different setups. First, we consider the Black–Scholes (BS) dynamics for the market; that is we assume $\sigma_{Dup} \equiv const$ for $const = 0.3$ and $S_0 = 1$. Second, we consider the local volatility dynamics

for the market, that is we set $C(t,K) := \mathsf{E}(S_t - K)_+$, where S_t follows the Heston model

$$dS_t = \sqrt{v_t} S_t \, dW_t, \tag{5.6a}$$
$$dv_t = \kappa(\theta - v_t)\, dt + \xi \sqrt{v_t}\, dB_t, \tag{5.6b}$$

with the following parameters: $\kappa = 1.5768$, $\theta = 0.0484$, $\xi = 0.5751$, and correlation $\rho = -0.7$ between the driving Brownian motions W and B, with initial values $S_0 = 1$, $v_0 = 0.1024$, cf. similar parameter choices in [FO09]. We solved (5.6) with the standard Euler method with 10^8 trajectories and 10^3 time steps. We calculate then σ_{Dup} from $C(t,K)$ using (5.2).

As our back-bone stochastic volatility model for Y, we choose a Heston-type model but with different parameters then the data-generating Heston model. That is, we set in (5.4) $b(t,x) = \lambda(\mu - x)$, $\sigma(t,x) = \eta\sqrt{x}$, $Y_0 = 0.0144$, $\lambda = 1$, $\mu = 0.0144$, $\eta = 0.5751$. We assume that W^X and W^Y are uncorrelated. Hence, the backbone model exhibits smaller initial as well as long-term variance, slower speed of mean-reversion, no correlation, but the same vol-of-vol as compared to the price-generating model. In particular, as the variance process has different parameters compared to the price-generating stochastic volatility model, a non-trivial local volatility function is required in order to match the implied volatility. Hence, even though the generating model is of the same *class*, the calibration problem is still non-trivial, and involves a singular MKV SDE.

We took $\mathcal{H}$ to be RKHS associated with the Gaussian kernel k with variance 5. We fix the number of time steps $M = 500$, $\lambda = 10^{-5}$, $L = 40$. At each time step of the Euler scheme we choose $(Z^j)_{j=1,\dots,L}$ by the following rule:

$$Z_j \text{ is } j \cdot 100/(L+1) \text{ percentile of the sequence } \{X_t^{N,n}\}_{n\in 1,\dots N}. \tag{5.7}$$

Figure 1 compares the theoretical and the calculated prices (in terms of implied volatilties) in the Black-Scholes (a) and Heston (b-d) settings for various strikes and maturities. That is, we first calculate $C(t,K)$ using the Black-Scholes model ("Black-Scholes setting") or (5.6) ("Heston setting"); then we calculate σ^2_{Dup} by (5.2); then we calculate $\widetilde{X}^n_M$, $n = 1,\dots,N$ using the algorithm (4.6) with $H \equiv 0$, $A_2(x) = x$, and

$$F(t,x,y,z) := x\sigma^2_{Dup}(t,x)\frac{\sqrt{y}}{\sqrt{z}};$$

then we calculate $\widetilde{C}(t,K)$ using (5.5); finally we transform the prices $C(T,K)$ and $\widetilde{C}(t,K)$ to the implied volatilities. As usual, for $K > S_0$ (out-of-the-money options) the implied volatilities are calculated from the call option prices $C(t,K)$ and $\widetilde{C}(t,K)$, and for $K < S_0$ (in-the-money options) the implied volatilities are calculated from the put option prices $P(t,K)$ and $\widetilde{P}(t,K)$ defined similarly.

We plot at Figure 1 implied volatilties for a wide range of strikes and maturities. More precisely, we consider all strikes K such that $\mathsf{P}(S_T < K) \in [0.02, 0.98]$ — this corresponds to all but very far in–the–money and out–of–the–money options. Pricing of the options with very far in or out of the money strikes is discussed later. One can from Figure 1 that already for $N = 10^4$ trajectories, identity (5.5) holds up to a small error for all the considered strikes and maturities. This error further diminishes as the number of trajectories increases. At $N = 10^6$ the true implied volatility curve and the one calculated from our approximation model become almost indistinguishable.

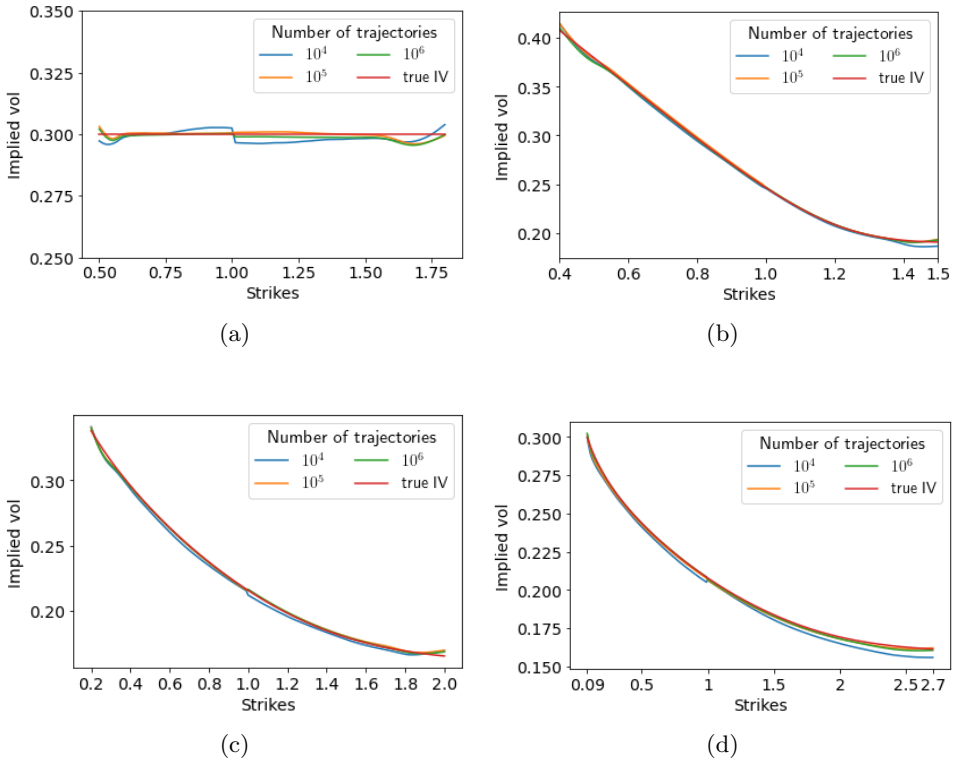
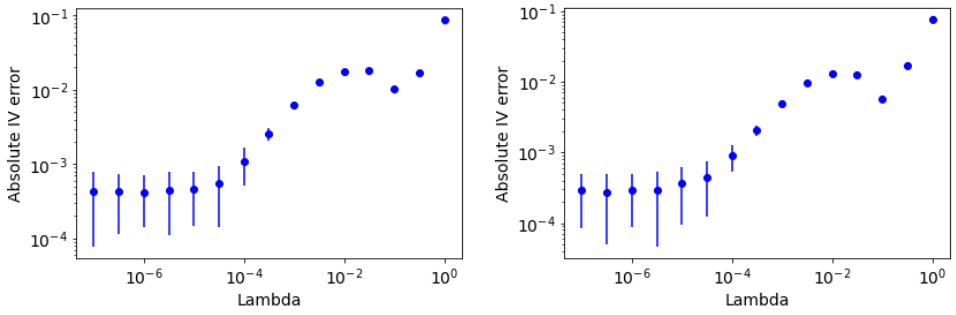


Figure 1: Fit of the smile vs number of trajectories. (a): Black-Scholes setting, $T = 1$ year. (b): Heston setting, $T = 1$ year. (c): Heston setting, $T = 4$ years. (d): Heston setting, $T = 10$ years.

Now let us discuss the stability of our model as the regularization parameter $\lambda \to 0$. We study the absolute error in the implied volatility of the 1 year ATM call option for various $\lambda \in [10^{-7}, 1]$ in the Black–Scholes and Heston settings described above. We took $N = 10^6$ trajectories and $L = 40$ Z_js at each step according to (5.7); we performed 100 repetitions at each considered value of λ. The results are presented at Figure 2. We see that in

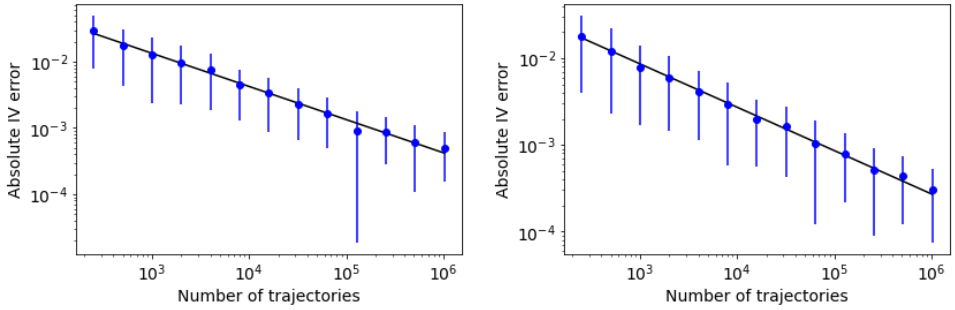
both settings, initially, the error drops as λ decreases, then it stabilizes once $\lambda \leq 10^{-5}$. Therefore for all our calculations we took $\lambda = 10^{-5}$. It is clear that the error does not blow up as λ becomes very small; the remaining error is due to other factors (numbers of trajectories and time steps being not big enough, etc). This indicates that (at least in our setting) the solution to the approximating equation (2.2) does converge weakly to the solution of (1.10) as $\lambda \to 0$.



(a) (b)

Figure 2: Mean absolute implied volatility error for different values of λ. (a): Black-Scholes setting. (b): Heston setting.

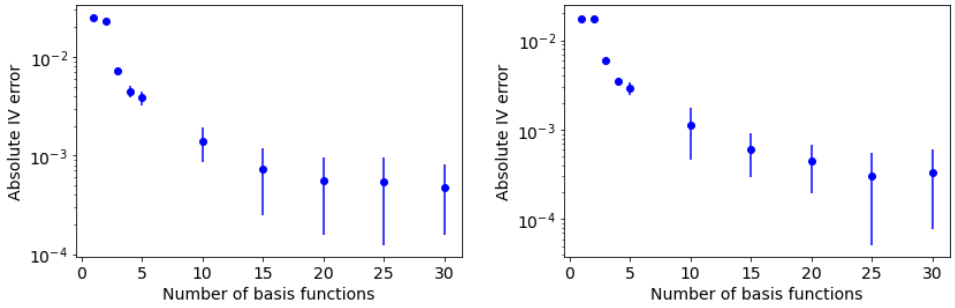
Let us see how the error in call option prices in (5.5) (and thus the distance between the laws of the true and approximated solutions) depends on the number of trajectories N. Recall, that it follows from Theorem 2.2 that this error should decrease as $N^{-1/4}$ (note the square in the left–hand side of (2.5). Figure 3 shows how the absolute error in the implied volatility of 1 year ATM call option decreases as the number of trajectories increases in (a) Black–Scholes setting and (b) Heston setting. We took $\lambda = 10^{-5}$, $L = 40$, $N \in [250, 2^{12} \cdot 250]$. We performed 100 repetitions at each value of N. We see the error decreases as $O(N^{-1/2})$ in both settings, which is even better than predicted by theory.



(a) (b)

Figure 3: Mean absolute implied volatility error vs number of trajectories. The black line is the approximation: error= $CN^{-1/2}$ (a): Black-Scholes setting; $C = 0.423$. (b): Heston setting; $C = 0.272$.

We also investigate the dependence of error in the implied volatility on the number of basis functions L in the representation (4.5). Recall that since the number of operations depends on L quadratically (it equals $O(MNL^2)$), it is extremely expensive to set L to be large. At Figure 4 we plotted the dependence of the absolute error in the implied volatility of 1 year ATM call option on L. We used $N = 10^6$ trajectories, $\lambda = 10^{-5}$, $L \in [1 \dots 30]$ and did 100 repetitions at each value of the number of basis functions. We see that as the number of basis functions increases, the error first drops significantly, but then stabilizes at $L \approx 20$.

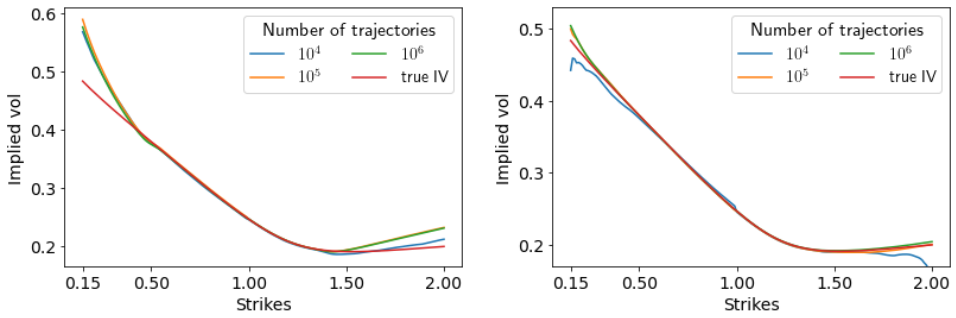


(a) (b)

Figure 4: Mean absolute implied volatility error vs number of basis functions. (a): Black-Scholes setting. (b): Heston setting.

As discussed above, Fig. 1 shows that our approximating model (2.3) agrees with the original model (1.10) and is able to callibrate the market correctly for a wide range of maturities T and strikes K with $\mathsf{P}(S_T < K) \in [0.02, 0.98]$. Now let us discuss the pricing of very far in-the-money or out-of-the-money options when $\mathsf{P}(S_T < K) \notin [0.02, 0.98]$. At Fig. 5(a),

we zoom out Fig. 1(b) to see the fit of the smile for all K such that $\mathsf{P}(S_T < K) \in [0.0001, 0.9999]$. We see that for very in/out-of-the money strikes with $\mathsf{P}(S_T < K) < 0.005$ or $\mathsf{P}(S_T > K) > 0.995$, the approximated model does not converge to the correct price as number of trajectories goes to infinity. This error is due to our way of choosing Z_js. Recall that we have selected them using (5.7). This means however that there will be no Z_j next to the very far in/out of the money strikes. Indeed, with $L = 40$ for $T = 1$ the left-most Z_j corresponds to 2.44th percentile of $\{X_T^{N,n}\}_{n\in 1,\dots N}$, which is approximately 0.451. One can see that this is quite far from 0.15 (the smallest strike considered in Fig. 5) and therefore it is not surprising that the approximation (4.3) fails. Similarly, the right-most Z_j corresponds to 97.56th percentile of $\{X_T^{N,n}\}_{n\in 1,\dots N}$, which is 1.440 and is far from 2, the largest strike considered in Fig. 5.



(a) (b)

Figure 5: Fit of the smile for very far in/out-of-the-money options. (a): Heston setting, Z_js are chosen according to (5.7). (b): Heston setting, Z_js are chosen according to (5.7) and (5.8).

A naive solution to this problem would be just to increase L, which is the number of Z_js that has to be taken in (4.3). Note however, that 0.15 corresponds to 0.01th percentile of $\{X_T^{N,n}\}_{n\in 1,\dots N}$, and to cover it one needs to take $L \approx 10^4$. Recalling that the number of operations of our algorithm depends quadratically on L, one can see that this new algorithm would require $6 \cdot 10^4$ more time than the current version with $L = 40$ and therefore is not feasible.

Therefore we suggest here another approach. We add to Z_js chosen according to (5.7), $2\widetilde{L}$ more elements: corresponding to $\widetilde{L}$ very small and $\widetilde{L}$ very large percentiles. More precisely, we took $\widetilde{L} = 5$ and added

$$Z_r \text{ which is } r \cdot 0.1 \text{ and } (100 - r \cdot 0.1) \text{ perecentiles of } \{X_T^{N,n}\}_{n\in 1,\dots N},\ r = 1, \dots, 5. \tag{5.8}$$

Thus, in total we have now 50 different Z_j in representation (4.3).

One can see from Fig. 5(b) that this choice have drastically increased the accuracy of the approximation of the smile for very far in/out-of-the money options, increasing the total time only by 60%. A small price to pay is that some of the near-the-money options has now a slightly larger error than it was in the initial way of choosing Z_js according to (5.7), see Table 1.

Strike K	0.15	0.25	0.75	1	1.25	1.75	2
$\mathsf{P}(S_T < K)$	0.0001	0.0017	0.1595	0.4574	0.8599	0.9988	0.9999
True IV	0.4832	0.4512	0.3103	0.2465	0.2025	0.1935	0.1995
IV error (A)	0.0786	0.0339	0.0010	0.0012	0.0006	0.0149	0.0268
IV error (B)	0.0233	0.0043	0.0007	0.0012	0.0008	0.0021	0.0067

Table 1: Comparison of two methods of choosing basis functions: only by (5.7) vs by (5.7) and (5.8). IV error denotes the average absolute error of implied volatilities, where Z_js is chosen by (5.7) (A); Z_js is chosen by (5.7) and (5.8) (B).

6. Conclusion and outlook

In this paper, we study the problem of calibrating local stochastic volatility models via the particle approach pioneered in [GHL12]. We suggest a novel RKHS based regularization method and prove that this regularization guarantees well-posedness of the underlying McKean-Vlasov SDE and the propagation of chaos property. Our numerical results suggest that the proposed approach is rather efficient for the calibration of various local stochastic volatility models and can outperform widely used local kernel methods. There are still some questions left open here. First, it remains unclear whether the regularised McKean-Vlasov SDE remains well-posed when the regularisation parameter λ tends to zero. This limiting case needs a separate study. Another important issue is the choice of RKHS and the number of basis functions which ideally should be adapted to the problem at hand. This problem of adaptation is left for future research.

7. Proofs

In this section we present the proofs of the results from Section 2 and Section 3.

Proof of Proposition 3.1. Let $I \subset \mathbb{N}$ and let $e := (e_i)_{i\in I}$ be a total orthonormal system in $\mathcal{H}$ (note that I is finite if $\mathcal{H}$ is finite dimensional). Define the

vector $\gamma^\nu \in \ell_2(I)$ by

$$\gamma_i^\nu := \langle e_i, c_A^\nu \rangle_{\mathcal{H}} = \int_{\mathcal{X}\times\mathcal{X}} \big\langle e_i, k(\cdot, x)\big\rangle_{\mathcal{H}} A(y)\nu(dx, dy)$$
$$= \int_{\mathcal{X}\times\mathcal{X}} e_i(x)A(y)\nu(dx, dy), \quad i \in I. \tag{7.1}$$

Since the operator $\mathcal{C}^\nu$ is bounded it may be described by the (possibly infinite) symmetric matrix

$$B^\nu := \big(\langle e_i, \mathcal{C}^\nu e_j\rangle_{\mathcal{H}}\big)_{(i,j)\in I\times I} = \Big(\int_{\mathcal{X}} e_i(x)e_j(x)\,\nu(dx, \mathcal{X})\Big)_{(i,j)\in I\times I}, \tag{7.2}$$

which acts as a bounded positive semi-definite operator on $\ell_2(I)$. Denote

$$\beta^\nu = (B + \lambda I)^{-1}\gamma^\nu. \tag{7.3}$$

For $f \in \mathcal{H}$ write $f = \sum_{i=1}^\infty \beta_i e_i$. Then, recalling (7.1) and (7.2), we derive

$$\begin{aligned}
&\arg\min_{f\in\mathcal{H}}\Big\{\int_{\mathcal{X}\times\mathcal{X}} |A(y) - f(x)|^2\,\nu(dx, dy) + \lambda\|f\|_{\mathcal{H}}^2\Big\}\\
&\quad= \arg\min_{\beta\in\ell_2(I)}\Big\{\int_{\mathcal{X}\times\mathcal{X}} |A(y) - \sum_{i=1}^\infty \beta_i e_i|^2\,\nu(dx, dy) + \lambda\|\beta\|_{\ell_2(I)}^2\Big\}\\
&\quad= \arg\min_{\beta\in\ell_2(I)}\Big\{-2\langle\beta, \gamma^\nu\rangle_{\ell_2(I)} + \langle\beta, (B + \lambda I)\beta\rangle_{\ell_2(I)}\Big\}\\
&\quad= \arg\min_{\beta\in\ell_2(I)}\Big\{-2\langle\beta - \beta^\nu, \gamma^\nu\rangle_{\ell_2(I)} + \langle\beta - \beta^\nu, (B + \lambda I)(\beta - \beta^\nu)\rangle_{\ell_2(I)}\\
&\qquad\qquad + 2\langle\beta - \beta^\nu, (B + \lambda I)\beta^\nu\rangle_{\ell_2(I)}\Big\}\\
&\quad= \arg\min_{\beta\in\ell_2(I)}\Big\{\langle\beta - \beta^\nu, (B + \lambda I)(\beta - \beta^\nu)\rangle_{\ell_2(I)}\Big\}\\
&\quad= \beta^\nu,
\end{aligned}$$

where we used the fact that $B + \lambda I$ is strictly positive definite and the definition (7.3). To complete the proof it remains to note that

$$\sum_{i=1}^\infty \beta_i^\nu e_i = (\mathcal{C}^\nu + \lambda I_{\mathcal{H}})^{-1} c_A^\nu,$$

which shows (3.6). □

Proof of Theorem 2.3. Let us write

$$|m_A^\lambda(x;\mu) - m_A^\lambda(y;\nu)| \le |m_A^\lambda(x;\mu) - m_A^\lambda(x;\nu)| + |m_A^\lambda(x;\nu) - m_A^\lambda(y;\nu)|$$
$$= I_1 + I_2. \tag{7.4}$$

Working with respect to the orthonormal basis introduced in the proof of Proposition 3.1, see (7.3), we derive for the first term in (7.4)

$$\begin{aligned} I_1 &= |\langle k(x,\cdot), m_A^\lambda(\cdot;\mu) - m_A^\lambda(\cdot;\nu)\rangle_{\mathcal{H}}| \\ &\le \|k(x,\cdot)\|_{\mathcal{H}} \|m_A^\lambda(\cdot;\mu) - m_A^\lambda(\cdot;\nu)\|_{\mathcal{H}} \\ &\le \sqrt{k(x,x)}\|\beta^\mu - \beta^\nu\|_{\ell_2(I)} \\ &\le D_k \|\beta^\mu - \beta^\nu\|_{\ell_2(I)} \end{aligned} \tag{7.5}$$

where we used (3.1) and Assumption K.

Denote $Q^\nu := B^\nu + \lambda I$ and $Q^\mu := B^\mu + \lambda I$. Recalling that they are bounded $\ell_2(I) \to \ell_2(I)$ operators with bounded inverses, it easy to see that

$$\|(Q^\mu)^{-1} - (Q^\nu)^{-1}\|_{\ell_2(I)} \le \|(Q^\mu)^{-1}\|_{\ell_2(I)} \|(Q^\nu)^{-1}\|_{\ell_2(I)} \|Q^\mu - Q^\nu\|_{\ell_2(I)}.$$

Therefore

$$\begin{aligned} \|\beta^\mu - \beta^\nu\|_{\ell_2(I)} &= \|(Q^\mu)^{-1}\gamma^\mu - (Q^\nu)^{-1}\gamma^\nu\|_{\ell_2(I)} \\ &\le \big\|\big((Q^\mu)^{-1} - (Q^\nu)^{-1}\big)\gamma^\mu\big\|_{\ell_2(I)} + \big\|(Q^\nu)^{-1}(\gamma^\mu - \gamma^\nu)\big\|_{\ell_2(I)} \\ &\le \big\|(Q^\mu)^{-1}\big\|_{\ell_2(I)} \big\|(Q^\nu)^{-1}\big\|_{\ell_2(I)} \big\|Q^\mu - Q^\nu\big\|_{\ell_2(I)} \big\|\gamma^\mu\big\|_{\ell_2(I)} \\ &\quad + \big\|(Q^\nu)^{-1}\big\|_{\ell_2(I)} \big\|\gamma^\mu - \gamma^\nu\big\|_{\ell_2(I)} \\ &\le \frac{1}{\lambda^2}\big\|B^\mu - B^\nu\big\|_{\ell_2(I)} \big\|\gamma^\mu\big\|_{\ell_2(I)} + \frac{1}{\lambda}\big\|\gamma^\mu - \gamma^\nu\big\|_{\ell_2(I)}. \end{aligned} \tag{7.6}$$

Now observe that for any $i, j \in I$

$$\begin{aligned} (B_{ij}^\mu - B_{ij}^\nu)^2 &= \Big(\int_{\mathcal{X}} e_i(x)e_j(x)\big(\mu(dx,\mathcal{X}) - \nu(dx,\mathcal{X})\big)\Big)^2 \\ &= \int_{\mathcal{X}}\int_{\mathcal{X}} e_i(x)e_j(x)e_i(y)e_j(y) \\ &\quad \times \big(\mu(dx,\mathcal{X}) - \nu(dx,\mathcal{X})\big)\big(\mu(dy,\mathcal{X}) - \nu(dy,\mathcal{X})\big). \end{aligned}$$

Hence, by using the identity

$$\sum_{i\in I} e_i(x)e_i(y) = \sum_{i\in I} \big\langle k(x,\cdot), e_i\big\rangle_{\mathcal{H}} \big\langle k(y,\cdot), e_i\big\rangle_{\mathcal{H}} = \big\langle k(x,\cdot), k(y,\cdot)\big\rangle_{\mathcal{H}} = k(x,y), \tag{7.7}$$

we get

$$\begin{aligned} \big\|B^\mu - B^\nu\big\|_{\ell_2(I)}^2 &\le \big\|B^\mu - B^\nu\big\|_{HS}^2 \\ &= \int_{\mathcal{X}} \big(\mu(dx,\mathcal{X}) - \nu(dx,\mathcal{X})\big) \int_{\mathcal{X}} k^2(x,y)\big(\mu(dy,\mathcal{X}) - \nu(dy,\mathcal{X})\big). \end{aligned} \tag{7.8}$$

By the duality principle, for every smooth $h : \mathcal{X} \to \mathbb{R}$ one has

$$\Big| \int_{\mathcal{X}} h(x)\big(\mu(dx,\mathcal{X}) - \nu(dx,\mathcal{X})\big)\Big| = \Big| \int_{\mathcal{X}\times\mathcal{X}} h(x)\big(\mu(dx,dy) - \nu(dx,dy)\big)\Big| \\ \leq \sup_{x\in\mathcal{X}} \big|\partial_x h(x)\big| \mathbb{W}_1(\mu,\nu).$$

So we continue (7.8) in the following way:

$$\|B^\mu - B^\nu\|^2_{\ell_2(I)} \leq \mathbb{W}_1(\mu,\nu) \sup_{x\in\mathcal{X}} \Big| \int_{\mathcal{X}} \partial_x k^2(x,y)\big(\mu(dy,\mathcal{X}) - \nu(dy,\mathcal{X})\big)\Big|, \quad (7.9)$$

and for each particular $x \in \mathcal{X}$ we have similarly

$$\Big| \int_{\mathcal{X}} \partial_x k^2(x,y)\big(\mu(dy,\mathcal{X}) - \nu(dy,\mathcal{X})\big)\Big| \leq \sum_{i=1}^d \Big| \int_{\mathcal{X}} \partial_{x_i} k^2(x,y)\big(\mu(dy,\mathcal{X}) - \nu(dy,\mathcal{X})\big)\Big| \\ \leq \sum_{i=1}^d \sup_{y\in\mathcal{X}} |\partial_y \partial_{x_i} k^2(x,y)| \mathbb{W}_1(\mu,\nu) \\ \leq d^2 D_k^2 \mathbb{W}_1(\mu,\nu),$$

where the last inequality follows from by Assumption K. Combining this with (7.9), we deduce

$$\|B^\mu - B^\nu\|_{\ell_2(I)} \leq D_k \mathbb{W}_1(\mu,\nu) d. \quad (7.10)$$

By a similar argument, using (7.7), we derive

$$\big\|\gamma^\mu - \gamma^\nu\big\|^2_{\ell_2(I)} \\ \leq \sum_{i\in I} \int_{\mathcal{X}\times\mathcal{X}} \int_{\mathcal{X}\times\mathcal{X}} e_i(x) e_i(x') A(y) A(y') (\mu-\nu)(dx,dy)(\mu-\nu)(dx',dy') \\ \leq \int_{\mathcal{X}\times\mathcal{X}} \int_{\mathcal{X}\times\mathcal{X}} k(x,x') A(y) A(y') (\mu-\nu)(dx,dy)(\mu-\nu)(dx',dy') \\ \leq d^2 \mathbb{W}_1^2(\mu,\nu) \|A\|^2_{\mathcal{C}^1} D_k^2, \quad (7.11)$$

where again Assumption K was used. Next note that

$$\|\gamma^\mu\|^2_{\ell_2(I)} = \int_{\mathcal{X}\times\mathcal{X}} \int_{\mathcal{X}\times\mathcal{X}} k(x,x') A(y) A(y') \mu(dx,dy)\mu(dx',dy') \\ \leq \int_{\mathcal{X}\times\mathcal{X}} \int_{\mathcal{X}\times\mathcal{X}} \big|A(y)\big| \sqrt{k(x,x)} \big|A(y')\big| \sqrt{k(x',x')} \mu(dx,dy)\mu(dx',dy') \\ = \Big(\int_{\mathcal{X}\times\mathcal{X}} \big|A(y)\big| \sqrt{k(x,x)} \mu(dx,dy)\Big)^2 \\ \leq \int_{\mathcal{X}\times\mathcal{X}} \big|A(y)\big|^2 \mu(dx,dy) \int_{\mathcal{X}\times\mathcal{X}} k(x,x)\mu(dx,dy) \\ \leq D_k^2 \|A\|^2_{\mathcal{C}^1} \quad (7.12)$$

due to Assumption K. Substituting now (7.10), (7.11), and (7.12) into (7.6) and then into (7.5), we finally get

$$I_1 \le (\lambda^{-1} D_k + 1)\lambda^{-1} D_k^2 \mathbb{W}_1(\mu,\nu) d \|A\|_{\mathcal{C}^1} \tag{7.13}$$

Now let us bound I_2 in (7.4). We clearly have

$$I_2 = |\langle k(x,\cdot) - k(y,\cdot), m_A^\lambda(\cdot;\nu)\rangle| \le \|k(x,\cdot) - k(y,\cdot)\|_{\mathcal{H}} \|m_A^\lambda(\cdot;\nu)\|_{\mathcal{H}} \tag{7.14}$$

Note that

$$\begin{aligned}
&\|k(x,\cdot) - k(y,\cdot)\|_{\mathcal{H}}^2 \\
&\quad = \langle k(x,\cdot) - k(y,\cdot), k(x,\cdot) - k(y,\cdot)\rangle_{\mathcal{H}} \\
&\quad = k(x,x) - k(x,y) - (k(y,x) - k(y,y)) \\
&\quad = \Big(\int_0^1 \partial_2 k(x, x+\xi(y-x))\, d\xi\Big)^\top (x-y) - \Big(\int_0^1 \partial_2 k(y, x+\xi(y-x))\, d\xi\Big)^\top (x-y) \\
&\quad = (x-y)^\top \Big(\int_0^1 \int_0^1 \partial_1\partial_2 k(x+\eta(y-x), x+\xi(y-x))\, d\xi d\eta\Big)^\top (x-y),
\end{aligned}$$

with ∂_1, ∂_2 denoting the vector of derivatives of k with respect to the first and second argument, respectively. Recalling Assumption K, we derive

$$\|k(x,\cdot) - k(y,\cdot)\|_{\mathcal{H}}^2 \le d D_k^2 |x-y|^2 . \tag{7.15}$$

Further, using (7.12), we see that

$$\|m_A^\lambda(\cdot;\nu)\|_{\mathcal{H}} = \|\beta^\nu\|_{\ell_2(I)} \le \|(B^\nu + \lambda I)^{-1}\|_{\ell_2(I)} \|\gamma^\nu\|_{\ell_2(I)} \le \lambda^{-1} D_k \|A\|_{\mathcal{C}^1}.$$

Combining this with (7.15) and substituting into (7.14), we get

$$I_2 \le \sqrt{d}\lambda^{-1} D_k^2 \|A\|_{\mathcal{C}^1} |x-y|.$$

This, together with (7.13) and (7.4), finally yields

$$|m_A^\lambda(x;\mu) - m_A^\lambda(y;\nu)| \le C_1 \mathbb{W}_1(\mu,\nu) + C_2 |x-y|,$$

where $C_1 = (\lambda^{-1} D_k + 1)\lambda^{-1} D_k^2 d \|A\|_{\mathcal{C}^1}$ and $C_2 = \sqrt{d}\lambda^{-1} D_k^2 \|A\|_{\mathcal{C}^1}$. This completes the proof of the theorem. □

Now we are ready to prove the main results of Section 2. They would follow from Theorem 2.3 obtained above.

Proof of Theorem 2.1. It follows from Theorem 2.3, and the assumptions of the theorem, and the fact that $\mathbb{W}_1$-metric can be bounded from above by the $\mathbb{W}_2$-metric, that the drift and diffusion of (2.3) are Lipschitz and satisfy the conditions of [CD16a, Theorem 4.21]. Hence it has a unique strong solution. □

Proof of Theorem 2.2. We see that Theorem 2.3 and the conditions of the theorem implies that all the assumptions of [CD16b, Theorem 2.12] hold. This implies (2.5). □

Proof of Theorem 3.3. Consider the operator $\mathcal{C}^\nu$ in the orthonormal basis $(\widetilde{a}_n)_{n\in J}$ of $\mathcal{H}$. Put

$$D^\nu := (\langle \widetilde{a}_i, \mathcal{C}^\nu \widetilde{a}_j\rangle_{\mathcal{H}})_{(i,j)\in J\times J} = (\langle \widetilde{a}_i, T^\nu \widetilde{a}_j\rangle_{\mathcal{H}})_{(i,j)\in J\times J} = (\sigma_j \delta_{ij})_{(i,j)\in J\times J},$$

since $\widetilde{a}_j$ is an eigenvector of T^ν with eigenvalue σ_j. Since $\mathcal{C}^\nu$ is diagonal in this basis, we see that for $\lambda > 0$ one has for $i \in J$

$$(\mathcal{C}^\nu + \lambda I_{\mathcal{H}})^{-1}\widetilde{a}_i = (\sigma_i + \lambda)^{-1}\widetilde{a}_i. \tag{7.16}$$

Consider also the function c_A^ν in this basis. We write for $i \in J$ similar to (7.1)

$$\eta_i^\nu := \langle c_A^\nu, \widetilde{a}_i\rangle_{\mathcal{H}} = \int_{\mathcal{X}\times\mathcal{X}} \widetilde{a}_i(x)A(y)\nu(dx,dy), \quad i \in I$$

and we clearly have $c_A^\nu = \sum_{i\in J}\eta_i^\nu \widetilde{a}_i$. Then, using Proposition 3.1 and (7.16) we derive for $\lambda > 0$

$$\begin{aligned} m_A^\lambda(\cdot;\nu) = (\mathcal{C}^\nu + \lambda I_{\mathcal{H}})^{-1} c_A^\nu &= \sum_{i\in J}\eta_i^\nu (\mathcal{C}^\nu + \lambda I_{\mathcal{H}})^{-1}\widetilde{a}_i \\ &= \sum_{i\in J}\eta_i^\nu (\sigma_i + \lambda)^{-1}\widetilde{a}_i. \end{aligned} \tag{7.17}$$

Next, since $m_A \in \mathcal{L}_2^\nu$, we have

$$P_{\overline{\mathcal{H}}} m_A = \sum_{i\in J} \left\langle \mathsf{E}_{(X,Y)\sim\nu}\left[A(Y)|X=\cdot\right], a_i\right\rangle_{\mathcal{L}_2^\nu} a_i. \tag{7.18}$$

Further, for $i \in J$ we deduce

$$\begin{aligned} \left\langle \mathsf{E}_{(X,Y)\sim\nu}[A(Y)|X=\cdot], a_i\right\rangle_{\mathcal{L}_2^\nu} &= \int_{\mathcal{X}} \mathsf{E}_{(X,Y)\sim\nu}\left[A(Y)|X=x\right] a_i(x)\nu(dx,\mathcal{X}) \\ &= \mathsf{E}_{(X,Y)\sim\nu}(a_i(X)\mathsf{E}[A(Y)|X]) \\ &= \mathsf{E}_{(X,Y)\sim\nu} a_i(X)A(Y) \\ &= \sigma_i^{-1/2}\eta_i^\nu, \end{aligned}$$

where we used that $\widetilde{a}_n = \sqrt{\sigma_n}a_n$. Substituting this into (7.18) and combining with (7.17), we get

$$P_{\overline{\mathcal{H}}} m_A - m_A^\lambda = \sum_{i\in J}(\eta_i^\nu \sigma_i^{-1} - \eta_i^\nu(\sigma_i+\lambda)^{-1})\widetilde{a}_i = \sum_{i\in J}\eta_i^\nu \frac{\lambda}{\sigma_i(\sigma_i+\lambda)}\widetilde{a}_i.$$

Thus

$$\left\|P_{\overline{\mathcal{H}}}m_A - m_A^\lambda\right\|^2_{\mathcal{L}_2^\nu} = \sum_{i\in J}(\eta_i^\nu)^2\frac{\lambda^2}{\sigma_i(\sigma_i+\lambda)^2} = \sum_{i\in J}\langle m_A, a_i\rangle^2_{\mathcal{L}_2^\nu}\frac{\lambda^2}{(\sigma_i+\lambda)^2},$$

which is (3.11). Similarly, recalling (3.10), we get

$$\left\|P_{\overline{\mathcal{H}}}m_A - m_A^\lambda\right\|^2_{\mathcal{H}} = \sum_{i\in J}(\eta_i^\nu)^2\frac{\lambda^2}{\sigma_i^2(\sigma_i+\lambda)^2} = \sum_{i\in J}\langle m_A, a_i\rangle^2_{\mathcal{L}_2^\nu}\frac{\lambda^2}{\sigma_i(\sigma_i+\lambda)^2},$$

which is finite whenever $P_{\overline{\mathcal{H}}}m_A \in \mathcal{H}$, that is, $\sum_{i\in J}\langle m_A, a_i\rangle^2_{\mathcal{L}_2^\nu}\sigma_i^{-1} < \infty$. This shows (3.12). It is easily seen by dominated convergence that the l.h.s. of (3.11) goes to zero, and, in the case $P_{\overline{\mathcal{H}}}m_A \in \mathcal{H}$ the l.h.s. of (3.12) goes to zero as well. □

References

[AKH02] Fabio Antonelli and Arturo Kohatsu-Higa. Rate of convergence of a particle method to the solution of the mckean–vlasov equation. *The Annals of Applied Probability*, 12(2):423–476, 2002.

[BDG19] Oleg Butkovsky, Konstantinos Dareiotis, and Máté Gerencsér. Approximation of sdes – a stochastic sewing approach. *arXiv preprint arXiv:1909.07961*, 2019.

[BJ17] Mireille Bossy and Jean-François Jabir. On the wellposedness of some McKean models with moderated or singular diffusion coefficient. In *International Symposium on BSDEs*, pages 43–87. Springer, 2017.

[BS13] Gerard Brunick and Steven Shreve. Mimicking an Itô process by a solution of a stochastic differential equation. *Ann. Appl. Probab.*, 23(4):1584–1628, 2013.

[CD16a] René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I.* Springer, Probability Theory and Stochastic Modelling 83, 2016.

[CD16b] René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications II.* Springer, Probability Theory and Stochastic Modelling 84, 2016.

[Dup94] Bruno Dupire. Pricing with a smile. *Risk*, 7:18–20, 1994.

[FO09] Fang Fang and Cornelis W Oosterlee. A novel pricing method for European options based on Fourier-cosine series expansions. *SIAM Journal on Scientific Computing*, 31(2):826–848, 2009.

[Fun84] Tadahisa Funaki. A certain class of diffusion processes associated with nonlinear parabolic equations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 67(3):331–348, 1984.

[Gat11] Jim Gatheral. *The volatility surface: a practitioner's guide*, volume 357. John Wiley & Sons, 2011.

[GHL12] Julien Guyon and Pierre Henry-Labordère. Being particular about calibration. *Risk Magazine*, 2012.

[Gyo86] István Gyongy. Mimicking the one-dimensional marginal distributions of processes having an Itô differential. *Probab. Theory Relat. Fields*, 71(4):501–516, 1986.

[JM21] Benjamin Jourdain and Stéphane Menozzi. Convergence rate of the Euler-Maruyama scheme applied to diffusion processes with L_Q –L_ρ drift coefficient and additive noise. *arXiv preprint arXiv:2105.04860*, 2021.

[JZ20] Benjamin Jourdain and Alexandre Zhou. Existence of a calibrated regime switching local volatility model. *Mathematical Finance*, 30(2):501–546, 2020.

[Kre89] Erwin Kreyszig. *Introductory functional analysis with applications.* New York etc.: John Wiley &— Sons, 1989.

[LSZ20] Daniel Lacker, Mykhaylo Shkolnikov, and Jiacheng Zhang. Inverting the Markovian projection, with an application to local stochastic volatility models. *Annals of Probability*, 48(5):2189–2211, 2020.

[MV16] Yuliya S. Mishura and Alexander Yu. Veretennikov. Existence and uniqueness theorems for solutions of McKean–Vlasov stochastic equations. *arXiv preprint arXiv:1603.02212*, 2016.

[SC08] Ingo Steinwart and Andreas Christmann. *Support vector machines.* Springer Science & Business Media, 2008.

[SHS01] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

[Sun05] Hongwei Sun. Mercer theorem for RKHS on noncompact sets. *Journal of Complexity*, 21:337 – 349, 2005.