Adv. Appl. Probab. 1–33 (2026)
*doi:*10.1017/apr.2026.10057

NUMERICAL APPROXIMATION OF MCKEAN–VLASOV SDES VIA STOCHASTIC GRADIENT DESCENT

ANKUSH AGARWAL,* *Western University*
ANDREA AMATO,** *** AND
STEFANO PAGLIARANI ,** **** *Università di Bologna*
GONÇALO DOS REIS ,***** *The University of Edinburgh*

Abstract

We propose a novel approach to numerically approximate McKean–Vlasov stochastic differential equations (MV-SDEs) using stochastic gradient descent (SGD) while avoiding the use of interacting particle systems (IPSs) and the associated simulation costs required to achieve the 'propagation of chaos' limit. The SGD technique is deployed to solve a Euclidean minimization problem, obtained by first representing the MV-SDE as a minimization problem over the set of continuous functions of time, and then approximating the domain with a finite-dimensional sub-space. Convergence is established by proving certain intermediate stability and moment estimates of the relevant stochastic processes, including the tangent processes. Numerical experiments illustrate the competitive performance of our SGD-based method compared with the IPS benchmarks. This work offers a theoretical foundation for using the SGD method in the context of numerical approximation of MV-SDEs, and provides analytical tools to study its stability and convergence.

Keywords: Nonlinear stochastic differential equations; convergence; polynomial approximation

2020 Mathematics Subject Classification: Primary 65C30
Secondary 65C05; 65C35

1. Introduction

In this work, we propose a novel technique for numerically approximating the solution of a class of McKean–Vlasov stochastic differential equations (MV-SDEs) using stochastic gradient descent (SGD). We cast the solution of the MV-SDEs as the fixed point of a certain functional for which we find the solution via minimizing a related objective function. As the possible domain is infinite-dimensional, we look for an approximation to the true solution in

Received 21 October 2023; accepted 3 January 2026.
* Postal address: Statistical and Actuarial Sciences, Western University, Canada. Email: ankush.agarwal@glasgow.ac.uk
** Postal address: Università di Bologna, Italy
*** Email address: andrea.amato6@studio.unibo.it
**** Email address: stefano.pagliarani9@unibo.it
***** Postal address: School of Mathematics, The University of Edinburgh, UK. Email: g.dosreis@ed.ac.uk

© The Author(s), 2026. Published by Cambridge University Press on behalf of Applied Probability Trust. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

a finite-dimensional domain via the method of SGD. We show that the approximation in the finite-dimensional domain approaches the true solution in the limit.

MV-SDEs play an important role in a variety of fields, including, but not limited to, kinetic theory of gases and plasmas [12, 38], chemotaxis modelling [45], population dynamics including crowd [39] and pedestrian [37] dynamics, machine learning (ML) [10, 11], finance and optimal control [9], quantum mechanics [27], and mean-field games [9]. Typically, the solution of an MV-SDE is approximated via an N_0-dimensional interacting particle system (IPS) and it is well-known, via the so-called 'propagation of chaos' results (e.g. [40, 46]), that the empirical law of the IPS converges to the law of the MV-SDE as $N_0 \to \infty$. Numerically, one can then use time-discretization and simulation to obtain accurate approximate solutions of the MV-SDE via the IPS ([2, 8, 48). The applicability of this approach has been recently extended across a variety of settings under varying regularity and growth assumptions on its coefficients; see, for example, [7, 17–20, 28, 42, 44, 50].

Even though the 'IPS' approach is very powerful and can be applied to many different classes of MV-SDEs, depending on the type of interaction kernel, the computational cost of simulating an IPS can be expensive. As such, there have been many other works in the literature to explore alternative techniques for numerically solving for the solution of an MV-SDE. In [47], an iterative multilevel Monte Carlo scheme of creating a particle system is combined with Picard iteration to reduce the computational cost. In [25], the technique of analytical expansions is used to derive accurate approximation of the density of the solution of an MV-SDE with elliptic constant diffusion coefficient and bounded drift. In another approach, [3, 13, 26] use the Fokker–Planck equation representation of the density of the solution to apply numerical techniques used to solve partial differential equations (PDEs). In [1], the authors consider a class of MV-SDEs for which the solution can be readily computed using the Picard iteration. They use a Fourier-transform approach to solve a fixed-point equation, which also allowed them to incorporate Lévy jumps in their analysis in a relatively straightforward manner. In [32], the authors used a multilevel Picard iteration to develop numerical estimates of the solution of a class of MV-SDEs in high dimensions. They are able to prove that their numerical approach is computationally efficient and avoids the curse of dimensionality unlike the IPS approach. In a way, many of the methods mentioned are of decoupling-type: in a first step the measure component (or a functional thereof) is approximated, said approximation is plugged back into the McKean–Vlasov equation's dynamics yielding an SDE that is then approximated in the usual fashion – such approaches have been used for importance sampling [21] or to deploy Koopman transfer operator approaches to the nonlinear semi-group operator of MV-SDEs [33]. In [15] the authors developed a cubature-method-based approach to solve decoupled forward–backward MV-SDEs. Noteworthy of mention is the recent work of [14] in the context of ergodic simulation of the invariant distribution of McKean–Vlasov equations – a setting we do not explore. There, one replaces the measure argument in the coefficients of the MV-SDE by the process' occupation (empirical) measure. This transforms the MV-SDE into a *self-interacting diffusion* that has full history dependence in its coefficients. The methodology remains then single-trajectory (no interacting particles) and thus avoids the classical particle-system propagation-of-chaos costs.

Recently, there have been tremendous developments in using machine-learning techniques to efficiently solve nonlinear PDEs in high dimensions owing to the seminal work [29]. Typically in this approach, an appropriate approximation of the forward process related to the nonlinear PDE and an approximation of the solution in terms of its gradient are provided as inputs to a feedforward neural network. The network is then trained to learn the mapping

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

between the forward process and the gradient of the solution using a SGD method, which minimizes a suitable loss function based on the terminal condition of the nonlinear PDE. It has been proven under different settings of nonlinear PDEs (for example, [30, 31) that the machine-learning-based method reliant on SGD avoids the curse of dimensionality. This approach of using machine learning to estimate the solution of a nonlinear PDE has also been successfully extended to solve McKean–Vlasov stochastic control problems, see, for example, [10, 11]. Our proposed approach to solve an MV-SDE using SGD is inspired by these aforementioned works (and the references therein). We do not rely on a neural network to learn the solution but instead use SGD to learn optimal weights in a polynomial basis function representation of the unknown functions in the considered MV-SDE. The particular class of MV-SDE we consider in this work (2) allows us to develop a more targeted approach, which is naturally more efficient than a general neural-network-based estimation approach.

For clarity of exposition, we work with a class of MV-SDEs in which the coefficients are separable with respect to their dependence on the measure and state of the MV-SDE, namely

$$\mathrm{d}X_t = \sum_{j=1}^{K} \Big(\mathrm{E}[\varphi_{1,j}(X_t)]\alpha_j(t, X_t)\,\mathrm{d}t + \mathrm{E}[\varphi_{2,j}(X_t)]\beta_j(t, X_t)\,\mathrm{d}W_t \Big), \quad X_0 = \xi. \tag{1}$$

Based on the recent work [5], a wide class of coefficients can be projected on a Fourier basis giving rise to the class of MV-SDEs in (1). We show that the unknown functions of time $\bar{\gamma}_{1,j}(t) := \mathrm{E}[\varphi_{1,j}(X_t)]$ and $\bar{\gamma}_{2,j}(t) := \mathrm{E}[\varphi_{2,j}(X_t)]$ satisfy a fixed-point equation, which we then pose as a minimization problem over the space of continuous functions. Next, we reformulate the infinite-dimensional minimization problem as a finite-dimensional one by looking for an approximate solution in a suitable finite-dimensional sub-domain, a typical choice being the space of nth-order polynomials. Once a basis for the latter is fixed, the objective function can be parametrized as a function on a Euclidean space. To minimize it we propose an algorithm based on SGD, where the stochastic gradient of the objective function depends on the solution to a Markovian SDE and on its gradient with respect to the parameters. To compute the gradient in SGD, we employ the Euler discretization of the relevant SDEs. To the best of our knowledge, our approach is the first to deploy an SGD-type algorithm to solve MV-SDEs. Moreover, we also provide a 'mini-batch' version of our SGD-based algorithm that greatly improves the efficiency of the algorithm.

In this work, we essentially build a theoretical foundation for the study of the SGD approach applied to the numerical approximation of MV-SDEs. We provide a convergence study and establish the related properties for such a purpose. Concretely, the convergence of our method is reliant on stability estimates of the related Markovian SDE and tangent processes, including a general interest result on the higher-order time-regularity property of the maps $t \mapsto \mathrm{E}[\varphi_{\cdot,\cdot}(X_t)]$. The moment estimates and Lipschitz property of the tangent process with respect to the unknown function also allows us to establish an explicit formula for the gradient of the loss function. Moreover, by performing a suitable penalization of the loss function, we are able to obtain Lipschitz continuity of its gradient. These intermediate results allow us to use the existing convergence results of descent methods by [6, 41] to show convergence of our algorithm to a stationary point. As our minimization problem is non-convex, classical results that ensure convergence of SGD algorithms to the global minimum do not apply to our setting. On the contrary, we rely on the developments in [6, 41] for the convergence of SGD algorithms to stationary points, in non-convex settings. We observe that the study of gradient-descent algorithms is an active research field, also with attempts of proving convergence to the global minimum in the non-convex setting. For instance, in [23], the authors prove global convergence

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

for a particular version of the SGD algorithm that makes use of mini-batches and a randomized initial point, under suitable geometric assumptions for the objective function. Checking these assumptions in our setting would be quite complex, thus, for the purpose of readability, we only verify the assumptions in [6, 41] to ensure convergence to a stationary point. Through extensive numerical studies, we demonstrate the excellent performance of our SGD algorithm in solving well-studied MV-SDEs. In particular, we measure the accuracy of our method by computing relative error with respect to the benchmarks obtained by the IPS approach, and show that we achieve very low error values with a small computational budget.

Overall, the results of this paper can be viewed as a proof of concept for the application of SGD algorithms as a valid tool to deliver numerical approximations of MV-SDEs without the simulation of IPSs. In order to reduce the technicalities and keep the proofs as transparent as possible, the convergence analysis is performed here in the case of bounded coefficients. However, it could be extended to the case of $\alpha_j(t,\cdot)$, $\beta_j(t,\cdot)$ with sub-linear growth and $\varphi_{1,j}$ and $\varphi_{2,j}$ with polynomial growth. We discuss this issue further below (see Remark 2). In particular, the numerical tests presented in Section 4.2 confirm that the approach is viable also when the coefficients $\varphi_{1,j}$ and $\varphi_{2,j}$ have quadratic growth.

An extension of this methodology to the case of Hölder continuous drift coefficients seems feasible, up to dealing with some technical issues arising from the fact that the gradient process becomes a diffusion with singular (distributional) drift coefficients. We also expect this approach to be suitable for extensions to the jump-diffusion framework. Finally, an interesting and challenging extension regards adapting this methodology to frameworks where the simulation of the associated IPS is less amenable, e.g. conditional MV-SDEs or MV-SDEs with moderate interaction among others.

The rest of the paper is organized as follows. At the start of Section 2, we provide the main setting of our considered MV-SDEs and for the convenience of the reader Tables 1 and 2 summarize the notation used throughout. We formulate the fixed-point problem for the unknown measure-dependent function in the MV-SDEs and then cast it as a minimization problem over the space of continuous functions. In Section 2.2, we introduce the finite-dimensional approximation of the original domain and provide an upper bound for the error we incur by solving the minimization problem on the reduced domain. We propose our SGD-based algorithm to solve the minimization problem at the beginning of Section 3, where we also provide a 'mini-batch' version of it. The convergence proof of the algorithm is presented in Section 3.3 along with the necessary intermediate results. The results from the extensive set of numerical experiments are presented in Section 4. The proofs of all the main theoretical results are relegated to the Appendix.

2. McKean–Vlasov SDEs with separable coefficients

Let $T > 0$ be a fixed time horizon. All the SDEs in this article are considered up to time T. We consider the class of $\mathbb{R}^d$-valued MV-SDEs in (1), where the interaction through the state is separated from the interaction through the law. Their dynamics can be compactly written as

$$\begin{cases} \mathrm{d}X_t = \bar{\gamma}(t)\big(\alpha(t, X_t)\,\mathrm{d}t + \beta(t, X_t)\,\mathrm{d}W_t\big), \quad X_0 = \xi, \\ \bar{\gamma}(t) = \mathrm{E}[\varphi(X_t)], \end{cases} \tag{2}$$

for measurable maps $\alpha : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{K\times d}$, $\beta : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{K\times d\times q}$, $\varphi : \mathbb{R}^d \to \mathbb{R}^K$, ξ a random variable, and with W being a q-dimensional Brownian motion (column vector). The full probabilistic framework can be found in the Appendix.

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

TABLE 1. Main notation for spaces and classic operators.

Notation	Description
$\mathbb{N}$	Set of natural numbers starting at 1
$\mathbb{N}_0 := \mathbb{N} \cup \{0\}$	Set of natural numbers including 0
$\mathbb{R}$	Set of real numbers
$\mathbb{R}_+ := [0, \infty)$	Set of non-negative real numbers
$\langle x, y \rangle$	Euclidean scalar product for $x, y \in \mathbb{R}^m$
$\lvert x \rvert$	Euclidean norm for $x \in \mathbb{R}^m$
$\mathrm{Tr}(A)$	Trace operator of a square matrix A with real entries
$L^p([0, T], \mathbb{R}^m)$	(or $L^p([0, T])$ in short) is the space of Borel-measurable functions $f : [0, T] \to \mathbb{R}^m$ with $\lVert f \rVert_{L^p([0,T])} = \left(\int_0^T \lvert f(t) \rvert^p \, \mathrm{d}t \right)^{\frac{1}{p}} < \infty$, for $p \geq 1$.
$L^\infty([0, T], \mathbb{R}^m)$	(or $L^\infty([0, T])$ in short) is the space of Borel-measurable functions $f : [0, T] \to \mathbb{R}^m$ with $\lVert f \rVert_{L^\infty([0,T])} = \sup_{t \in [0,T]} \lvert f(t) \rvert < \infty$
$C([0, T], \mathbb{R}^m)$	Space of continuous functions $f : [0, T] \to \mathbb{R}^m$ endowed with supremum norm $\lVert f \rVert_\infty = \sup_{t \in [0,T]} \lvert f(t) \rvert$
$C_b^n([0, T])$	Space of n-times continuously differentiable functions $f : [0, T] \to \mathbb{R}^m$ with bounded derivatives
$C_b^{n,q}([0, T] \times \mathbb{R}^d)$	Space of continuous functions $f = f(t, x) : [0, T] \times \mathbb{R}^d \to \mathbb{R}^m$, n-times differentiable in t, q-times in x, with bounded and continuous derivatives
$\partial_{x_i} f$	Partial derivative of $f = f(x) : \mathbb{R}^m \to \mathbb{R}^n$ with respect to component x_i
$\nabla_x f$	Jacobian $(n \times m)$-matrix. If $n = 1$, it is the gradient vector
$H_x f$	Hessian $(n \times m \times m)$-tensor of f. If $n = 1$, it is the Hessian matrix
$(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in [0,T]}, \mathrm{P})$	Filtered probability space, with $(\mathcal{F}_t)_{t \in [0,T]}$ satisfying the usual assumptions of completeness and continuity
$\mathrm{E}[\cdot]$	Expectation operator under measure P
L^p	Space of random variables ξ such that $\lVert \xi \rVert_{L^p} = \mathrm{E}[\lvert \xi \rvert^p]^{\frac{1}{p}} < \infty$, for $p \geq 1$
$\mathbb{L}^p_{[0,T]}$	Space of progressively measurable processes X such that $\lVert X \rVert_{\mathbb{L}^p_{[0,T]}} = \left(\int_0^T \mathrm{E}[\lvert X_t \rvert^p] \, \mathrm{d}t \right)^{\frac{1}{p}} < \infty$, for $p \geq 1$
$\mathcal{S}^p_{[0,T]}$	Space of progressively measurable processes X such that $\lVert X \rVert_{\mathcal{S}^p_{[0,T]}} = \mathrm{E}[\sup_{t \in [0,T]} \lvert X_t \rvert^p]^{\frac{1}{p}} < \infty$, for $p \geq 1$
$\mathcal{S}^\infty_{[0,T]}$	Space of progressively measurable processes X such that $\lVert X \rVert_{\mathcal{S}^\infty_{[0,T]}} = \operatorname{ess\,sup}_{\omega \in \Omega} \sup_{0 \leq t \leq T} \lvert X_t \rvert < \infty$

Remark 1. The framework (2) of $\bar{\gamma}$ maps can be extended to the type $\bar{\gamma}(\cdot) = h\big(\mathrm{E}[\varphi(X_\cdot)]\big)$ for some invertible and suitably regular function h. We present our results under the setting of (2) only for the sake of simplicity.

Our starting point could also have been a MV-SDE with more general coefficients, to which we could have applied a projection over a Fourier basis (see [5]), but such an

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

TABLE 2. Main notation for primary processes and auxiliary functions.

Notation	Description
X	Solution of the reference MV-SDE in (2)
Z^γ	Solution of (5) given an input function $\gamma \in L^\infty([0,T],\mathbb{R}^K)$
$\mathscr{L}$	Lifting map (13), $\mathbb{R}^{(n+1)\times K} \ni a \mapsto \mathscr{L}a := a_0 g_0 + \cdots + a_n g_n \in L^\infty([0,T])$, with $g_0, \ldots, g_n$ being linearly independent vectors of $C([0,T],\mathbb{R})$
F	Functional (7) to be minimized
$\mathbf{h}$, H	Auxiliary truncation and penalization maps (16) and (17)
G	Functional (19) (see also (18)) to be minimized (built from F)
Z^a	Solution of (20), i.e. Z^γ in (5) when $\gamma = \mathbf{h} \circ \mathscr{L}a$ for given $a \in \mathbb{R}^{(n+1)\times K}$
$Y^{a;k,j}$, $Y^{k,j}$	Tangent process $\partial_{a_{k,j}} Z^a$ in (21)
v, $v_{k,j}$	Gradient functional for the SGD algorithm in (24)

exercise would still have resulted in the class of MV-SDEs considered here. For example, in Section 4.3 we show, following [5, Section 2], that an MV-SDE with drift coefficient given by $\mathrm{E}\left[\exp\left(-(X_t - x)^2/2\right)\right]|_{x=X_t}$ can be approximated by a system in the form of (2), to which we apply our SGD algorithm.

To proceed further, we first present the well-posedness and regularity assumptions.

Assumption 1. *The random variable* $\xi \in L^2$.

Assumption 2. *The functions* $\alpha, \beta \in C([0,T]\times\mathbb{R}^d)$ *and there exists* $R > 0$ *such that*

$$|\alpha(t,x) - \alpha(t,y)| + |\beta(t,x) - \beta(t,y)| + |\varphi(x) - \varphi(y)| \le R|x-y|, \quad x, y \in \mathbb{R}^d,\ t \in [0,T], \tag{3}$$

and α, β, *and* φ *are bounded by* R.

The following well-posedness result is classical.

Proposition 1. *Let Assumptions 1 and 2 hold true. Then, there exists a unique solution* $X = (X_t)_{t\in[0,T]}$ *to MV-SDE (2) in* $\mathcal{S}^2_{[0,T]}$. *Furthermore,* $\bar\gamma \in C_b([0,T])$.

The above result follows from [9, Theorem 4.21 (p. 236)]). In particular, the fact that $\bar\gamma$ is bounded and continuous is straightforward as φ is bounded and continuous.

Remark 2. (*Growth assumptions*.) We expect the conclusion of Proposition 1 to still hold true when replacing the condition in Assumption 2 of boundedness on $\alpha(t,\cdot)$ and $\beta(t,\cdot)$ with sub-linear growth, the latter implied by the Lipschitz condition (3). However, we are not aware of this result in the literature. The proof would take advantage of the specific structure of the coefficients, which are separable with respect to the state and measure component. In addition, we expect the result to be true by assuming the function φ only locally Lipschitz continuous with polynomial growth, provided the initial datum has suitably many finite moments.

Remark 3. (*Regularity assumptions*.) Concerning the well-posedness of MV-SDEs and the moment estimates of its solutions, Lipschitz continuity of the coefficients is a rather standard assumption. The literature on MV-SDEs also contains results for less regular coefficients (see [16] and the references therein), but these typically require additional structural assumptions,

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

e.g. ellipticity conditions for the diffusion coefficient. In addition, we rely on the Lipschitz continuity of $\alpha(t,\cdot)$ and $\beta(t,\cdot)$ in the convergence analysis of the SGD algorithm (see Section 1). Thus, we prefer to introduce now this regularity requirement.

Example 1. *Assumption 2 requires the coefficient functions $\alpha(t,\cdot)$, $\beta(t,\cdot)$, and φ to be bounded and Lipschitz continuous. Among the elementary functions, this class includes several trigonometric functions (e.g. the model in Section 4.1), bounded rational functions, Gaussian-type functions, as well as sums and products of these. On the other hand, polynomial functions do not satisfy Assumption 2. However, affine functions $\alpha(t,\cdot)$, $\beta(t,\cdot)$, and polynomial-type functions φ are allowed in the extended framework discussed in Remark 2 (e.g. the model in Section 4.2).*

We now establish a time-regularity result for $\bar{\gamma}$, which can be achieved by requiring more regularity on the function φ and on the coefficients α, β. We will see in Section 2.2 that one can take advantage of this higher regularity to have a faster convergence of the solution to the finite-dimensional minimization problem to the true $\bar{\gamma}$ (see Proposition 4).

Assumption 3. *There exists $N \in \mathbb{N}$ and $R > 0$ such that $\alpha, \beta \in C_b^{N-1,2(N-1)}([0,T]\times\mathbb{R}^d)$ and $\varphi \in C_b^{2N}(\mathbb{R}^d)$, with α, β, and φ and all their derivatives bounded by R.*

Proposition 2. (Higher-order smoothness of $\bar{\gamma}$) *Let Assumptions 1–3 with some $N \in \mathbb{N}$ hold true. Then, the function $\bar{\gamma}$ defined in (2) is in $C_b^N([0,T])$ and, for any $m = 1,\dots,N$, we have*

$$\frac{\mathrm{d}^m}{\mathrm{d}t^m}\bar{\gamma}(t) = \mathrm{E}\big[\big(\mathscr{K}^m\varphi\big)(t,X_t)\big], \quad t\in[0,T], \tag{4}$$

where $\mathscr{K}$ is the Kolmogorov operator acting on a function $u \in C_b^{1,2}([0,T]\times\mathbb{R}^d)$ as defined as

$$\mathscr{K}u(t,x) = \partial_t u + \frac{1}{2}\mathrm{Tr}\Big(\bar{\gamma}(t)\beta(t,x)\beta^\top(t,x)\bar{\gamma}^\top(t)H_x u(t,x)\Big) + \big\langle\bar{\gamma}(t)\alpha(t,x),\,\nabla_x u(t,x)\big\rangle,$$

with $\mathscr{K}^m u$ being the m-times composition of $\mathscr{K}$ acting on u as

$$[0,T]\times\mathbb{R}^d \ni (t,x) \mapsto \big(\mathscr{K}^m u\big)(t,x) = \underbrace{\mathscr{K}\big(\cdots\mathscr{K}(\mathscr{K}u)\cdots\big)}_{m\text{ times}}(t,x).$$

The proof is found in Appendix B.

Remark 4. By inspecting the proof of Proposition 2 and by Remark 2, one can see that the same conclusion could be derived lifting the boundedness condition in Assumption 3, if the derivatives of $\alpha(t,\cdot)$, $\beta(t,\cdot)$, and φ have polynomial growth, provided that X_0 (thus, X_t) has suitably many finite moments.

Example 2. *All the elementary functions satisfying Assumption 2 discussed in Example 1 are smooth with bounded derivatives and thus fulfil Assumption 3. On the other hand, affine $\alpha(t,\cdot)$, $\beta(t,\cdot)$, and polynomial φ are allowed in the extended framework discussed in Remark 4.*

2.1. Fixed-point equation for $\bar{\gamma}$

We introduce a closely related SDE to (2), which takes function γ as input. Namely, given $\gamma \in L^\infty([0,T],\mathbb{R}^K)$ denote by Z^γ the solution to the Markovian SDE,

$$\mathrm{d}Z_t = \gamma(t)\big(\alpha(t,Z_t)\,\mathrm{d}t + \beta(t,Z_t)\,\mathrm{d}W_t\big), \quad Z_0 = \xi. \tag{5}$$

The following result is a direct application of Lemma 3.

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

Proposition 3. *Let Assumptions 1 and 2 hold true for some $R > 0$. Then (5) is strongly well-posed for any $\gamma \in L^\infty([0,T],\mathbb{R}^K)$. Furthermore, for any $\kappa > 0$, there exists $C > 0$ dependent on κ, T, R, and K, such that*

$$\|Z^\gamma - Z^{\widetilde{\gamma}}\|_{\mathcal{S}^2_{[0,T]}} \le C\|\gamma - \widetilde{\gamma}\|_\infty,$$

for any $\gamma, \widetilde{\gamma} \in L^\infty([0,T],\mathbb{R}^K)$ such that $\|\gamma\|_\infty, \|\widetilde{\gamma}\|_\infty \le \kappa$.

Now, define the operator $\Psi : C\big([0,T],\mathbb{R}^K\big) \to C\big([0,T],\mathbb{R}^K\big)$ given as

$$\Psi(\gamma)(t) := \mathrm{E}[\varphi(Z_t^\gamma)], \quad t \in [0,T],$$

which is well-defined under Assumptions 1 and 2.

Let us start by observing that solving the MV-SDE (2) is equivalent to finding a fixed point for Ψ. Indeed, if $\bar{\gamma}$ is a solution to

$$\gamma = \Psi(\gamma), \tag{6}$$

then the process $Z^{\bar{\gamma}}$ solves (2) and vice versa. Also note that, if $\bar{\gamma}$ solves (6), then $\bar{\gamma}$ solves the minimization problem

$$\min_{\gamma \in C([0,T])} F^2(\gamma), \tag{7}$$

where

$$F : C\big([0,T],\mathbb{R}^K\big) \to \mathbb{R}, \quad F(\gamma) := \|\Psi(\gamma) - \gamma\|_{L^2_{([0,T])}}. \tag{8}$$

Therefore, under Assumptions 1 and 2, Proposition 1 implies that there exists a unique $\bar{\gamma} \in C\big([0,T],\mathbb{R}^K\big)$ that solves (7), in particular,

$$\min_{\gamma \in C([0,T])} F^2(\gamma) = F^2(\bar{\gamma}) = 0. \tag{9}$$

Lemma 1. *Let Assumptions 1 and 2 hold true for some $R > 0$. Then, the functional F in (7) is locally Lipschitz continuous in sup-norm $\|\cdot\|_\infty$. Precisely, for any $\kappa > 0$ there exists $C > 0$ such that*

$$|F(\gamma) - F(\hat{\gamma})| \le C\|\gamma - \hat{\gamma}\|_\infty, \quad \gamma, \hat{\gamma} \in C([0,T],\mathbb{R}^K), \quad \text{with } \|\gamma\|_\infty, \|\hat{\gamma}\|_\infty \le \kappa, \tag{10}$$

where $C > 0$ depends on κ, T, R, and K.

Proof. Given $\kappa > 0$, take $\gamma, \hat{\gamma} \in C([0,T],\mathbb{R}^K)$ with $\|\gamma\|_\infty, \|\hat{\gamma}\|_\infty \le \kappa$. By the reverse triangle inequality we have

$$\Big| \|\Psi(\gamma) - \gamma\|_{L^2([0,T])} - \|\Psi(\hat{\gamma}) - \hat{\gamma}\|_{L^2([0,T])} \Big| \le \|\Psi(\gamma) - \Psi(\hat{\gamma})\|_{L^2([0,T])} + \|\gamma - \hat{\gamma}\|_{L^2([0,T])}. \tag{11}$$

The second term in the right-hand side of (11) is dominated by $\sqrt{T}\|\gamma - \hat{\gamma}\|_\infty$. For the first term, by the definition of Ψ and by the Lipschitz continuity of φ, we obtain

$$\begin{aligned} \|\Psi(\gamma) - \Psi(\hat{\gamma})\|^2_{L^2([0,T])} &= \int_0^T \Big| \mathrm{E}\big[\varphi(Z_t^\gamma) - \varphi(Z_t^{\hat{\gamma}})\big] \Big|^2 \,\mathrm{d}t \le C \int_0^T \mathrm{E}\Big[\big| Z_t^\gamma - Z_t^{\hat{\gamma}} \big|^2 \Big] \,\mathrm{d}t \\ &\le C\,T \|\gamma - \hat{\gamma}\|^2_\infty, \end{aligned}$$

where the last inequality follows from Proposition 3. Thus, by combining the two bounds for the terms in the right-hand side of (11), we get the result in (10). □

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

A quick inspection of the above proof shows that (10) also holds with $\|\cdot\|_\infty$ replaced by $\|\cdot\|_{L^2([0,T])}$ since the same is also true for Proposition 3.

Remark 5. Depending on the numerical accuracy requirements, instead of $F(\gamma)$ defined above, we could also use an objective function $\widehat{F}(\gamma)$ given as

$$\widehat{F}^2(\gamma) = \int_0^T w(s)|\Psi(\gamma)(s) - \gamma(s)|^2 \, \mathrm{d}s,$$

with a kernel (weight) function w over [0,T]. For example, we can take $w(t) = C_1 e^{C_2 t}$ for $C_1, C_2 > 0$. Such a choice of objective function $\widehat{F}(\gamma)$ in the minimization problem (7) would force the SGD algorithm to be more accurate towards the end of the interval [0,T].

2.2. Projection on a finite-dimensional domain

In order to solve (7) numerically using a gradient-descent method, it is imperative that we first narrow down the infinite-dimensional domain $C([0, T], \mathbb{R}^K)$ to a finite-dimensional one. This approach is a natural choice in our setting of solving (7) as the solution $\bar{\gamma}$ is a continuous map. Hereafter, we fix $n \in \mathbb{N}$ and a family $g_0, \ldots, g_n$ of linearly independent vectors of $C([0, T], \mathbb{R})$. Define the finite-dimensional sub-space of $C([0, T], \mathbb{R}^K)$ as

$$S_n := \{a_0 g_0 + \cdots + a_n g_n : a = (a_0, \ldots, a_n) \in \mathbb{R}^{(n+1)K}\}, \tag{12}$$

and the distance between $\gamma \in C([0, T], \mathbb{R}^K)$ and S_n as

$$d_n(\gamma) := \inf\{\|\gamma - p\|_\infty : p \in S_n\}.$$

Also introduce a linear lifting operator

$$\mathscr{L}: \mathbb{R}^{(n+1)\times K} \to S_n, \quad \mathscr{L}a = \mathscr{L}(a_0, \ldots, a_n) := a_0 g_0 + \cdots + a_n g_n. \tag{13}$$

Remark 6. The operator $\mathscr{L}$ is an isomorphism between finite-dimensional normed vector spaces. In particular, both $\mathscr{L}$ and its inverse $\mathscr{L}^{-1}$ are continuous and bounded operators. Furthermore, by a compactness argument, it can be easily seen that for any $\gamma \in C([0, T], \mathbb{R}^K)$, there exists $p_n(\gamma) \in S_n$ such that

$$\|\gamma - p_n(\gamma)\|_\infty = d_n(\gamma). \tag{14}$$

In other words, d_n is well-defined as a minimum and such a minimum is attained at $p = p_n(\gamma)$. It is important to observe that both $p_n(\gamma)$ and $d_n(\gamma)$ only depend on S_n and not on the particular choice of the basis g_i, $i = 0, \ldots, n$.

Example 3. (Polynomial of best approximation.) *Consider $n \in \mathbb{N}$ and $g_i(t) = t^i$ for $i = 0, \ldots, n$. Then the elements of S_n read as*

$$p(t) = \sum_{i=0}^{n} a_i t^i, \quad t \in [0, T].$$

This choice clearly leads to $S_n = \mathcal{P}_n(\mathcal{R}^K)$, the space of $\mathbb{R}^K$-valued polynomials of order n. We stress that $S_n = \mathcal{P}_n(\mathcal{R}^K)$ could be also obtained by fixing a basis g_i, $i = 0, \ldots, n$, which is not the canonical one. Possible choices include Legendre, Chebyshev, Laguerre, and Hermite

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

polynomials [43]. *In particular, the numerical results presented in Section 4 refer to the specific choice of g_i given by the Lagrange polynomials centred at Chebyshev points, i.e.*

$$g_i(t) = \left(\prod_{\substack{0 \leqslant l \leqslant n \\ l \neq i}} \frac{t - t_l}{t_i - t_l} \right), \quad t_i = \frac{T}{2} + \frac{T}{2} \cos\left(\frac{2i+1}{2n+2} \pi \right), \quad i = 0, 1, \ldots, n. \tag{15}$$

Finally note that, in particular, the distance $d_n(\gamma)$ and the projection $p_n(\gamma)$ (see Remark 6) are the same for all these choices of basis, for they all generate the same finite-dimensional space $S_n = \mathcal{P}_n(\mathcal{R}^K)$. In this case, p_n in (14) is called the polynomial of best approximation of γ (see [34, Chapter 5]).

Example 4. *Other examples for the basis $\{g_i, i = 0, 1, \ldots\}$ include splines* [4] *and smooth Fourier wavelet expansions* [22].

Since the maps γ are bounded by $\|\varphi\|_\infty$, we introduce the auxiliary map $\mathbf{h} : \mathbb{R}^K \to \mathbb{R}^K$ so that we can ensure that $\mathbf{h}(\mathscr{L}a)$ is also bounded. Concretely, let $\mathbf{h}$ be a smooth bounded function $\mathbb{R}^K \ni x \mapsto \mathbf{h}(x) = \big(\mathbf{h}_1(x_1), \ldots, \mathbf{h}_K(x_K)\big)$ defined as

$$\mathbb{R}^K \ni x \mapsto \mathbf{h}(x) = \begin{cases} x, & \text{if } |x| \leq \|\varphi\|_\infty + 2d_n(\bar{\gamma}) \\ \|\varphi\|_\infty + 2d_n(\bar{\gamma}), & \text{if } |x| \geq \|\varphi\|_\infty + 2d_n(\bar{\gamma}) \end{cases}, \tag{16}$$

where it is clear that $\|\mathbf{h}\|_\infty \leq \|\varphi\|_\infty + 2d_n(\bar{\gamma})$.

Let the constant $\tau_n > 0$ be such that

$$B_{\tau_n} \supset \mathscr{L}^{-1}\big(B_{\|\varphi\|_\infty + d_n(\bar{\gamma})}\big),$$

where B_{τ_n} and $B_{\|\varphi\|_\infty + d_n(\bar{\gamma})}$ denote the balls in $\mathbb{R}^{(n+1)K}$ and S_n, respectively, with radius τ_n and $\|\varphi\|_\infty + d_n(\bar{\gamma})$, centred at zero. Note that such τ_n exists as $\mathscr{L}^{-1}$ is a linear, and thus a bounded operator (see Remark 6). Finally, we consider a non-negative real-valued function $H \in C^1(\mathbb{R}^{(n+1)K})$ such that

$$\operatorname{supp} H \subset B^c_{\tau_n} \quad \text{and} \quad \lim_{|a| \to +\infty} H(a) = +\infty. \tag{17}$$

Example 5. *A function H satisfying hypothesis (17) above is given by*

$$H(a) := \begin{cases} 0, & \text{if } a \in B_{\tau_n} \\ (|a| - \tau_n)^2, & \text{otherwise} \end{cases}, \quad \nabla H(a) := \begin{cases} 0, & \text{if } a \in B_{\tau_n} \\ \frac{2(|a| - \tau_n)}{|a|} a, & \text{otherwise} \end{cases},$$

Remark 7. Let $\bar{\gamma} \in C([0, T], \mathbb{R}^K)$ be the unique minimizer to (7) and consider the element $p_n(\bar{\gamma})$ introduced in Remark 6. Since $\|\bar{\gamma}\|_\infty \leq \|\varphi\|_\infty$, it is clear that $\|p_n(\bar{\gamma})\|_\infty \leq \|\varphi\|_\infty + d_n(\bar{\gamma})$. Therefore, by (16) and (17), we obtain that

$$\mathbf{h}\big(p_n(\bar{\gamma})\big) = p_n(\bar{\gamma}) \quad \text{and} \quad H\big(\mathscr{L}^{-1} p_n(\bar{\gamma})\big) \equiv 0.$$

In place of (7), we consider the following optimization problem:

$$\min_{a \in \mathbb{R}^{(n+1)K}} G(a) \quad \text{with} \quad G(a) := F^2(\mathbf{h} \circ \mathscr{L}a) + H(a). \tag{18}$$

The role of map $\mathbf{h}$ is to ensure that the estimate of $\bar{\gamma}$, defined in (9), remains bounded when computing the loss function F. This is reflected when we compute $F^2(\mathbf{h} \circ \mathscr{L}a)$

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

in (18). Moreover, the function H penalizes the coefficient vector a when the norm of our approximation $\mathcal{L}a$ grows beyond $\|\varphi\|_\infty + d_n(\bar{\gamma})$. This is reflected in the definition of H in (17) and also motivated in Remark 7.

Lemma 2. *Let Assumptions 1 and 2 hold true for some $R > 0$. Let $\bar{\gamma} \in C([0, T], \mathbb{R}^K)$ be the unique solution to (7) and assume that $\|\varphi\|_\infty + d_n(\bar{\gamma}) \le \kappa$ for some $\kappa > 0$. Then the minimization problem (18) admits a solution. Furthermore, for any*

$$\bar{a}_n \in \arg\min_{a \in \mathbb{R}^{(n+1)K}} G(a)$$

we have

$$F(\mathcal{L}\bar{a}_n) \le C\, d_n(\bar{\gamma}), \tag{19}$$

where $C > 0$ depends on κ, T, R, and the dimensions.

Proof. The boundedness of φ implies that $F^2(\mathbf{h} \circ \mathcal{L}a)$ is bounded. Thus, by (17), $G(a)$ tends to $+\infty$ as $|a| \to \infty$. On the other hand, Lemma 1 together with the continuity of $\mathbf{h}$ and H imply that G is continuous on $\mathbb{R}^{(n+1)K}$. Therefore, the fact that G has at least a minimum $\bar{a}_n$ on $\mathbb{R}^{(n+1)K}$ stems from Weierstrass' extreme value theorem.

Furthermore, let $\bar{a}_n$ be a minimizer of G we then have

$$\begin{aligned} F^2(\mathbf{h} \circ \mathcal{L}\bar{a}_n) \le F^2(\mathbf{h} \circ \mathcal{L}\bar{a}_n) + H_n(\mathcal{L}\bar{a}_n) &\le F^2\big(\mathbf{h} \circ p_n(\bar{\gamma})\big) + H\big(\mathscr{L}^{-1} p_n(\bar{\gamma})\big) \\ &= F^2\big(p_n(\bar{\gamma})\big) = \big|F\big(p_n(\bar{\gamma})\big) - F(\bar{\gamma})\big|^2 \\ &\le C\, \|p_n(\bar{\gamma}) - \bar{\gamma}\|_\infty^2 = C\, d_n^2(\bar{\gamma}). \end{aligned}$$

In the above, we used Remark 7 and Lemma 1. The result follows since $\|\bar{\gamma}\|_\infty \le \|\varphi\|_\infty$ and $\|p_n(\bar{\gamma})\|_\infty \le \|\varphi\|_\infty + d_n(\bar{\gamma}) \le \kappa$. □

From the above result we observe that if we can compute the distance d_n explicitly, we have an explicit upper bound for the error encountered when solving (18) instead of (7). For instance, when choosing S_n as the space of the nth-order polynomials (see Example 3), an explicit error estimate can be obtained as shown next.

Proposition 4. *Let Assumptions 1–3 hold true for some $N \in \mathbb{N}$, $R > 0$. Let $g_0, \dots, g_n$ be a family of linearly independent vectors of $C([0, T], \mathbb{R})$ generating the space of the $\mathbb{R}^K$-valued polynomials of order n, i.e. $S_n = \mathcal{P}_n(\mathcal{R}^K)$ in (12). Then the minimization problem (18) admits a solution. Furthermore, for any*

$$\bar{a}_n \in \arg\min_{a \in \mathbb{R}^{(n+1)K}} G(a)$$

we have

$$F(\mathcal{L}\bar{a}_n) \le \frac{C}{n^N},$$

where $C > 0$ depends on T,N,R, and the dimensions (and is independent of n).

Proof. By Lemma 2 the minimization problem (18) admits a solution and (19) holds true. By Proposition 2, $\bar{\gamma}$ is N-times differentiable on $[0, T]$. In particular, (4) implies that its derivatives are bounded on $[0, T]$ by a constant that only depends on N, R, and the dimensions.

A theorem of Jackson (see [49, Section 5.1.5 (p. 261)]) then yields

$$d_n(\bar{\gamma}) \le \frac{C}{n^N},$$

which concludes the proof. □

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

Remark 8. We stress that the result of Proposition 4 does not depend on the choice of the basis g_i, $i = 0, \dots, n$. In particular, it applies to all the possible choices as discussed in Remark 3. We claim that the results of this section would still hold true after lifting the boundedness assumption on α, β, and φ. Indeed, in light of Remark 2, we expect (7) to admit a unique solution $\bar{\gamma} \in C([0, T], \mathbb{R}^K)$ even if $\alpha(t, \cdot)$, $\beta(t, \cdot)$ have sub-linear growth and φ has polynomial growth (provided that the initial datum $X_0 = \xi$ has suitably many finite moments). Furthermore, the proof of Lemma 2 only relies on the estimates of Lemma 1, whose proof itself would be identical in this more general context. Similarly, by Remark 4, we expect Proposition 4 to still hold true if the derivatives of $\alpha(t, \cdot)$, $\beta(t, \cdot)$, and φ have polynomial growth.

3. SGD for MV-SDEs

We propose a stochastic SGD to solve the minimization problem (18). For the convenience of the reader we re-write the objective function G in (18) explicitly as

$$\mathbb{R}^{(n+1)K} \ni a \mapsto G(a) = \int_0^T \left|\mathrm{E}\left[\varphi(Z_t^a) - \mathbf{h}\big((\mathscr{L}a)(t)\big)\right]\right|^2 \mathrm{d}t + H(a), \tag{20}$$

where we use the shorthand notation $Z_t^a := Z_t^{\mathbf{h}\circ\mathscr{L}a}$ to denote the solution to (5) when γ is taken as $\gamma = \mathbf{h} \circ \mathscr{L}a$, i.e. Z^a takes values in $\mathbb{R}^d$ and solves

$$\mathrm{d}Z_t = \mathbf{h}\big((\mathscr{L}a)(t)\big)\big(\alpha(t, Z_t)\,\mathrm{d}t + \beta(t, Z_t)\,\mathrm{d}W_t\big), \quad Z_0 = \xi. \tag{21}$$

We point out that the objective function G only depends on the law of Z^a (and a). Thus, when weak uniqueness holds (for instance, under Assumptions 1 and 2), G can be defined in terms of any weak solution to (21). In particular, it only depends on the coefficients of (21) and on the law of ξ, but not on the specific realizations of ξ or the Brownian motion W. However, in the following, we will sometimes consider strong solutions associated to specific realizations of Brownian motion and initial datum. In these cases we will reinforce the notation by denoting $Z^a(\xi, W)$ as the strong solution to (21) with respect to a given initial random variable ξ and a given Brownian motion W.

Lastly, it is clear that within our setting, $\min_{a\in\mathbb{R}^{(n+1)K}} G(a)$ is not a convex minimization problem.

3.1. Some preliminaries

To explain the SGD algorithm we begin with some heuristics. Hereafter we write a generic element of $a \in \mathbb{R}^{(n+1)K}$ as

$$a = (a_{0,1}, \dots, a_{0,K}, \dots, a_{n,1}, \dots, a_{n,K}).$$

- For any $k = 0, \dots, n$, and $j = 1, \dots, K$, by formally differentiating (21) with respect to the element $a_{k,j}$, we obtain that the pair $(Z^a, \partial_{a_{k,j}} Z^a)$ solves the system given by (21) and

$$\begin{aligned}\mathrm{d}Y_t^{k,j} &= g_k(t)\nabla\mathbf{h}_j\big(\mathscr{L}a(t)\big)\Big(\alpha(t, Z_t)\,\mathrm{d}t + \beta(t, Z_t)\,\mathrm{d}W_t\Big) \\ &\quad + \sum_{i=1}^{d} Y_t^{k,j,i}\mathbf{h}\big(\mathscr{L}a(t)\big)\Big(\partial_{z_i}\alpha(t, Z_t)\,\mathrm{d}t + \partial_{z_i}\beta(t, Z_t)\,\mathrm{d}W_t\Big), \quad Y_0^{k,j} = 0,\end{aligned} \tag{22}$$

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

with $\mathbf{h}$ introduced in (16). Note that $Y^{k,j}$ and Z take values in $\mathbb{R}^d$, with components denoted by $Y^{k,j,i}$ and Z^i, for $i = 1, \dots, d$, respectively. In the following, we will sometimes reinforce the notation by denoting $(Z_t^a(\xi, W), Y_t^{a;k,j}(\xi, W))$ as the strong solution to (21), (22) with respect to a given initial random variable ξ and Brownian motion W.

- Suppose that we can always exchange derivatives with time integrals and expected values (we justify this exchange of operations below). For any $a \in \mathbb{R}^{(n+1)K}$, and for any $k = 0, \dots, n$ and $j = 1, \dots, K$, we obtain

$$\begin{aligned}&\partial_{a_{k,j}}\left(\int_0^T \left|\mathrm{E}\left[\varphi(Z_t^a) - \mathbf{h}\big((\mathscr{L}a)(t)\big)\right]\right|^2 \mathrm{d}t\right)\\&= 2\int_0^T \left\langle \mathrm{E}\left[\varphi(Z_t^a) - \mathbf{h}\big((\mathscr{L}a)(t)\big)\right], \mathrm{E}\left[\partial_{a_{k,j}}\Big(\varphi(Z_t^a) - \mathbf{h}\big((\mathscr{L}a)(t)\big)\Big)\right]\right\rangle \mathrm{d}t. \end{aligned} \tag{23}$$

Let now W, $\tilde{W}$ be two independent Brownian motions and ξ, $\tilde{\xi}$ be two identically distributed independent random variables, which are also independent of $(W, \tilde{W})$. Assuming weak-uniqueness for the solutions to (21) we have

$$\begin{aligned}&\left\langle \mathrm{E}\left[\varphi(Z_t^a) - \mathbf{h}\big((\mathscr{L}a)(t)\big)\right], \mathrm{E}\left[\partial_{a_{k,j}}\Big(\varphi(Z_t^a) - \mathbf{h}\big((\mathscr{L}a)(t)\big)\Big)\right]\right\rangle\\&= \left\langle \mathrm{E}\left[\varphi\big(Z_t^a(\xi, W)\big) - \mathbf{h}\big((\mathscr{L}a)(t)\big)\right], \mathrm{E}\left[\partial_{a_{k,j}}\Big(\varphi\big(Z_t^a(\tilde{\xi}, \tilde{W})\big) - \mathbf{h}\big((\mathscr{L}a)(t)\big)\Big)\right]\right\rangle\\&= \mathrm{E}\left[\left\langle \varphi\big(Z_t^a(\xi, W)\big) - \mathbf{h}\big((\mathscr{L}a)(t)\big), \partial_{a_{k,j}}\Big(\varphi\big(Z_t^a(\tilde{\xi}, \tilde{W})\big) - \mathbf{h}\big((\mathscr{L}a)(t)\big)\Big)\right\rangle\right]. \end{aligned} \tag{24}$$

- Therefore, combining (23) and (24), and since $\partial_{a_{k,j}} Z^a$ solves (22), we formally obtain

$$\nabla_a G(a) = \mathrm{E}[v(a;\xi, W;\tilde{\xi}, \tilde{W})] + \nabla_a H(a),$$

with $v(a;\xi, W;\tilde{\xi}, \tilde{W}) \in \mathbb{R}^{(n+1)K}$, whose components are given by

$$\begin{aligned}&v_{k,j}(a;\xi, W;\tilde{\xi}, \tilde{W})\\&:= 2\int_0^T \left\langle \varphi\big(Z_t^a(\xi, W)\big) - \mathbf{h}\big((\mathscr{L}a)(t)\big), \nabla_x\varphi\big(Z_t^a(\tilde{\xi}, \tilde{W})\big) Y_t^{a;k,j}(\tilde{\xi}, \tilde{W}) - \partial_{a_{k,j}} \mathbf{h}\big((\mathscr{L}a)(t)\big)\right\rangle \mathrm{d}t, \end{aligned} \tag{25}$$

where $\nabla_x\varphi$ stands for the Jacobian matrix of $\varphi : \mathbb{R}^d \to \mathbb{R}^K$.

3.2. Algorithm

In the previous section, we presented heuristics on how certain computations can be performed in order to implement the stochastic gradient algorithm. With the help of those explanations, we can now introduce the algorithm. First, let:

- $(\eta_m)_{m\in\mathbb{N}}$ be a deterministic sequence of positive scalars (*learning rates*) such that

$$\sum_{m=0}^{\infty} \eta_m = +\infty \quad \text{and} \quad \sum_{m=0}^{\infty} \eta_m^2 < +\infty; \tag{26}$$

- $(\bar{\Omega}, \bar{\mathcal{F}}, (\bar{\mathcal{F}}_m)_{m\in\mathbb{N}}, \bar{\mathrm{P}})$ be a filtered probability space rich enough to support four independent samples $(W_m)_{m\in\mathbb{N}}$, $(\tilde{W}_m)_{m\in\mathbb{N}}$, $(\xi_m)_{m\in\mathbb{N}}$, and $(\tilde{\xi}_m)_{m\in\mathbb{N}}$ such that

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

- $(W_m)_{m\in\mathbb{N}}$ and $(\tilde{W}_m)_{m\in\mathbb{N}}$ are sequences of independent q-dimensional standard Brownian motions adapted to $(\bar{\mathcal{F}}_m)_{m\in\mathbb{N}}$; namely, for any $m\in\mathbb{N}$, $W_m=(W_{m,t})_{t\in[0,T]}$ is a q-dimensional standard Brownian motion such that $\mathcal{F}_T^{W_m}\subset\bar{\mathcal{F}}_m$, and the same holds for $\tilde{W}_m$;
- $(\xi_m)_{m\in\mathbb{N}}$ and $(\tilde{\xi}_m)_{m\in\mathbb{N}}$ of ξ are sequences of independent $\mathbb{R}^d$-valued random variables distributed like ξ.

For a deterministic $\mathbf{a}_0\in\mathbb{R}^{(n+1)K}$, the estimate update is obtained iteratively as

$$\mathbf{a}_{m+1}:=\mathbf{a}_m-\eta_m\mathbf{v}_{m+1},\qquad \mathbf{v}_{m+1}=v(\mathbf{a}_m;\xi_{m+1},W_{m+1};\tilde{\xi}_{m+1},\tilde{W}_{m+1})+\nabla H(\mathbf{a}_m),\qquad m\in\mathbb{N}_0.$$

The function v is defined component-wise by (25). Note that the $\mathbb{R}^{(n+1)K}$-valued stochastic processes $(\mathbf{a}_m)_{m\in\mathbb{N}}$ and $(\mathbf{v}_m)_{m\in\mathbb{N}}$ are recursively defined, as in Algorithm 1, and are adapted to the filtration $(\bar{\mathcal{F}}_m)_{m\in\mathbb{N}}$, which then represents the history of the algorithm.

Algorithm 1: SGD-MVSDE

Input : Maximum number ofi terations $m_{\max}$, learning rates $(\eta_m)_{m\in\mathbb{N}}$;
Initialization: Deterministic $\mathbf{a}_0\in\mathbb{R}^{(n+1)K}$;
$m=0$;
while $m\le m_{\max}$ **do**
  Sample $(W_{m+1},\tilde{W}_{m+1},\xi_{m+1},\tilde{\xi}_{m+1})$;
  $\mathbf{v}_{m+1}:=v(\mathbf{a}_m;\xi_{m+1},W_{m+1};\tilde{\xi}_{m+1},\tilde{W}_{m+1})+\nabla H(\mathbf{a}_m)$;
  $\mathbf{a}_{m+1}:=\mathbf{a}_m-\eta_m\mathbf{v}_{m+1}$;
end

Figure 1 below contains the flow-chart of Algorithm 1. We can also formulate a mini-batch version of the SGD-MVSDE algorithm by computing the gradient of the loss function v via averaging its value over M replications $\left(W_{m+1}^i,\tilde{W}_{m+1}^i,\xi_{m+1}^i,\tilde{\xi}_{m+1}^i\right)_{i=1}^M$ of $(W_{m+1},\tilde{W}_{m+1},\xi_{m+1},\tilde{\xi}_{m+1})$ for a given iteration index m.

Algorithm 2: Mini-batch-SGD-MVSDE

Input : Maximum number ofi terations $m_{\max}$; learning rates $(\eta_m)_{m\in\mathbb{N}}$;
Mini-batch sample size M
Initialization: Deterministic $\mathbf{a}_0\in\mathbb{R}^{(n+1)K}$;
$m=0$;
while $m\le m_{\max}$ **do**
  Sample $\left(W_{m+1}^i,\tilde{W}_{m+1}^i,\xi_{m+1}^i,\tilde{\xi}_{m+1}^i\right)_{i=1}^M$;
  $\mathbf{v}_{m+1}:=\frac{1}{M}\sum_{i=1}^M v(\mathbf{a}_m;\xi_{m+1}^i,W_{m+1}^i;\tilde{\xi}_{m+1}^i,\tilde{W}_{m+1}^i)+\nabla H(\mathbf{a}_m)$;
  $\mathbf{a}_{m+1}:=\mathbf{a}_m-\eta_m\mathbf{v}_{m+1}$;
end

Note that Algorithm 2 for $M=1$ is equivalent to Algorithm 1.

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

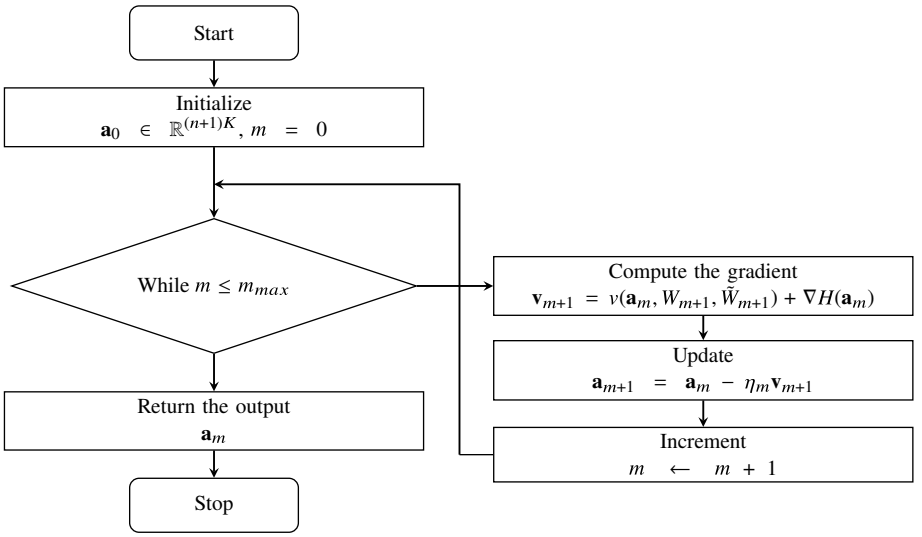

FIGURE 1. Flow-chart of Algorithm 1.

3.3. Convergence of Algorithm 1

We recall that H is a function in $C^1(\mathbb{R}^{(n+1)K})$ and that it satisfies (17). In order to obtain a convergence result for Algorithm 1, we introduce the following assumption.

Assumption 4. *The function H is such that $\nabla_a H$ is globally Lipschitz continuous on $\mathbb{R}^{(n+1)K}$.*

For example, the H specified in Example 5 satisfies Assumption 4. This requirement of Lipschitz regularity is essential to prove convergence of any iterative method with a sequential step (see Proposition 6 below).

Our main convergence result shows the convergence of the iterates $\mathbf{a}_m$ in Algorithm 1 to a stationary point of G.

Theorem 1. *Let Assumptions 1 and 4 be in force. Assume also that $\alpha(t,\cdot), \beta(t,\cdot) \in C^2(\mathbb{R}^d)$, for any $t \in [0,T]$, and $\varphi \in C^2(\mathbb{R}^K)$, with derivatives bounded by some $R > 0$. The map G is differentiable and there exists an $\mathbb{R}^{(n+1)K}$-valued random variable $\mathbf{a}_\infty$ such that, in Algorithm 1, we have*

$$\lim_{m\to+\infty} \mathbf{a}_m = \mathbf{a}_\infty \quad \bar{\mathrm{P}} - \text{almost surely} \quad \text{and} \quad \nabla_a G(\mathbf{a}_\infty) = 0.$$

To prove Theorem 1, we first prove two intermediate results. These results enable us to employ the developments in [6, 41] for the convergence of SGD algorithms in non-convex settings. The first preliminary result contains: (i) a standard consistency condition that allows us to understand the process $(\mathbf{v}_m)_{m\in\mathbb{N}}$ as a sequence of stochastic gradients of G; and (ii) a bound on the second conditional moment of the gradient noise.

Proposition 5. *Under the assumptions of Theorem 1, we have*

$$\mathrm{E}_{\bar{\mathrm{P}}}\big[\mathbf{v}_{m+1} - \nabla_a G(\mathbf{a}_m) | \bar{\mathcal{F}}_m\big] = 0, \quad m \in \mathbb{N}_0, \tag{27}$$

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

where we set $\bar{\mathcal{F}}_0 := \{\bar{\Omega}, \emptyset\}$. *Furthermore, the random variables*

$$\mathrm{E}_{\bar{\mathrm{P}}}\big[|\mathbf{v}_{m+1} - \nabla_a G(\mathbf{a}_m)|^2 \big| \bar{\mathcal{F}}_m\big]$$

are bounded, uniformly with respect to $m \in \mathbb{N}_0$.

The next proposition is a regularity result for G and its gradient.

Proposition 6. (Regularity of and $\nabla_a G$.) *Under the assumptions of Theorem 1, the function* G *is continuously differentiable and its gradient* $\nabla_a G$ *is globally Lipschitz continuous on* $\mathbb{R}^{(n+1)K}$.

We are now in the position to prove our main result in Theorem 1.

Proof of Theorem 1. Differentiability of G follows from Proposition 6. By Propositions 5 and 6, the hypothesis of [6, Proposition 3] and [41, Theorem 1] are fulfilled. In particular, Theorem 1 in [41] (see the remark on p. 5 there) implies that $\bar{\mathrm{P}}$-almost surely either $\mathbf{a}_m$ converges in $\mathbb{R}^{(n+1)K}$ or $|\mathbf{a}_m| \to +\infty$. On the other hand, by Proposition 3 in [6], either $G(\mathbf{a}_m) \to -\infty$ or $G(\mathbf{a}_m)$ converges to a finite value and $\nabla G(\mathbf{a}_m) \to 0$, $\bar{\mathrm{P}}$-almost surely. Since

$$G \geq 0 \quad \text{and} \quad \lim_{|a| \to +\infty} G(a) = +\infty,$$

we have that $\mathbf{a}_m$ converges $\bar{\mathrm{P}}$-almost surely to a stationary point of G. □

Remark 9. We claim that the convergence analysis of this section could be derived assuming $\alpha(t, \cdot)$, $\beta(t, \cdot)$, and φ with sub-linear growth (instead of bounded). Indeed, Propositions 3.2 and 3.3 (on which Theorem 1 relies) would be proved with similar, though slightly more involved, arguments. In particular, the estimates of Lemmas 4 and 5 for the solutions to the system (21), (22), which are the crux of the argument, should extend to the case of unbounded $\alpha(t, \cdot)$, $\beta(t, \cdot)$, and φ while keeping the benign boundedness assumption for their derivatives. Finally, we believe the assumptions on φ could be further relaxed by allowing for polynomial growth, provided additional integrability conditions are assumed on the initial datum.

Remark 10. The Lipschitz regularity of the derivatives of $\alpha(t, \cdot)$, $\beta(t, \cdot)$ are needed to employ standard stability estimates (see Lemma A.1) to the system (21) and (22). While it is true that dropping this assumption would allow to include popular models with Hölder-continuous coefficients (e.g. square-root diffusions), it would also result in the tangent process being the solution of an SDE with singular coefficients. This would require the development of new fundamental stability estimates on the tangent SDE (22). Therefore, this direction is left for future research.

4. Numerical study

We conduct several numerical experiments to study the numerical performance of Algorithm 2 for different values of the mini-batch sample size M. Precisely, we test it for the following models.

(i) The Kuramoto–Shinomoto–Sakaguchi model (Section 4.1), which is in the form of (2) with a low value of $K = 3$. The goal here is to test the accuracy and efficiency of the scheme in a case where the dependence on the state is truly separated from the one on the measure.

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

(ii) A model with a quadratic drift (Section 4.2), which is outside the scope of our Assumption 2. Here we show, numerically, that our SGD algorithm efficiently converges, at least for small times, even if the coefficients are super-linear in the measure variable.

(iii) An MV-SDE with a Gaussian convolution kernel (Section 4.3). The SDE has no separable coefficients, but it can be approximated by dynamics in the form of (2) with a sufficiently large K by employing a projection technique based on the generalized Fourier series (see [5]). The goal is to show that our algorithm can be successfully employed, in combination with the aforementioned projection technique, to obtain good approximations of the law of the original SDE.

The following choices are common to all our numerical experiments.

- The finite-dimensional space S_n for the domain projection of Section 2.2 is that of the n-dimensional polynomials. In particular, for a given $n \in \mathbb{N}$ and $T > 0$, the basis $g_0, \dots, g_n$ is given by the Lagrange polynomials centred at the Chebyshev nodes $t_0, \dots, t_n$ defined in (15). Therefore, recalling the lifting operator $\mathscr{L}$ defined in (13), for any $a \in \mathbb{R}^{(n+1)\times K}$ we have
$$(\mathscr{L}a)(t_j) = a_j, \quad j = 0, \dots, n.$$
- Inspired from the suggestion of learning rate in [23], the sequence of the learning rates $(\eta_m)_{m\in\mathbb{N}}$ is set as
$$\eta_m = r_0/(m+1)^\rho, \quad r_0 > 0, \ \rho \in (0.5, 1],$$
in compliance with the conditions (26).
- The functions $\mathbf{h}$ and H appearing in (18) are chosen as the identity and null functions, respectively. Although this choice does not meet the assumptions of the convergence results of Section 3.3, it does simplify the implementation. In addition, we observe that the algorithm still converges in our experiments.
- The initial iterate $\mathbf{a}_0$ (algorithm initialization) is always taken as
$$\mathbf{a}_0 = \big(\varphi(\tilde{X}_0), \dots, \varphi(\tilde{X}_0)\big),$$
with $\tilde{X}_0$ being a random realization of X_0, which yields $\mathscr{L}\mathbf{a}_0(t) \equiv \varphi(\tilde{X}_0)$. In particular, when $\mu_{X_0} = \delta_x$, then $\mathbf{a}_0 = (\varphi(x), \dots, \varphi(x))$ is deterministic.
- The benchmark value for the curve $\bar{\gamma}_\cdot = \mathrm{E}[\varphi(X_\cdot)]$ is given by
$$\bar{\gamma}_\cdot^{\mathrm{MC}} = \frac{1}{N_0}\sum_{i=1}^{N_0} \varphi(X_\cdot^{N_0,i}), \tag{28}$$
where X^{N_0} is the N_0-dimensional mean-field particle system associated to (2). The relative error after m iterations of our SGD algorithm is
$$\varepsilon_m := \frac{\|\mathscr{L}a_m - \bar{\gamma}^{\mathrm{MC}}\|_{L^2([0,T])}}{\|\bar{\gamma}^{\mathrm{MC}}\|_{L^2([0,T])}}, \tag{29}$$
with $\mathbf{a}_m$ being the mth iterate of Algorithm 2 and the $L^2([0, T])$-norm over the interval $[0, T]$ is as defined in Table 1. The algorithm is always halted when the error ε_m as computed above is below 1%.

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

- Algorithm 2 and the mean-field particle system associated with (2) make use of a time-stepping scheme to approximate the solution of the involved SDEs. In particular, the system (21), (22) is discretized by employing the Euler–Maruyama scheme with time-step $h = 10^{-2}$. Note that this further approximation is absolutely standard under the assumption of $\alpha(t, \cdot)$, $\beta(t, \cdot)$ differentiable with bounded derivatives. The discretized mean-field particle system is written as

$$X_t^{N_0,i} = X_{t_k}^{N_0,i} + \Big(\frac{1}{N_0}\sum_{i=1}^{N_0}\varphi(X_{\eta(t)}^{N_0,i})\Big)\cdot\Big(\alpha(\eta(t), X_{\eta(t)}^{N_0,i})(t-t_k) + \beta(\eta(t), X_{\eta(t)}^{N_0,i})\big(W_t^i - W_{t_k}^i\big)\Big),$$

  with $\eta(t) := t_k$ if $t \in [t_k, t_{k+1})$, and $t_{k+1} - t_k = h$. Here $\{t_k\}_k$ is a partition of $[0, T]$. Due to interactions between discretized diffusions, implementation of the above particle system will require $O(N_0^2 K h^{-1})$ arithmetic operations. On the other hand, in Algorithm 2, one sample of the mini-batch will involve $O(Kdqh^{-1})$, $O(nK^2d^2qh^{-1})$, and $O(nK^2dh^{-1})$ number of arithmetic operations for implementing discretized (21), (22) , and computing entries of v given in (25), respectively. Thus, the total number of arithmetic operations in one iteration of Algorithm 2 is $O(MnK^2d^2qh^{-1})$.

- Computational setup: All numerical experiments are conducted on the computing server of the Department of Mathematics, University of Bologna, featuring a 12th Gen Intel® Core™ i9-12900K CPU, 125 GB RAM, and running Ubuntu® 24.04.1 LTS. The computations were implemented in PYTHON™ 3.11.11 using NUMPY.

4.1. Kuramoto–Shinomoto–Sakaguchi MV-SDE

Consider the following MV-SDE, which is related to the famous Kuramoto–Shinomoto–Sakaguchi model (see [24, Section 5.3.2]):

$$\mathrm{d}X_t = \big(\mathrm{E}[\sin(X_t)]\cos(X_t) - \mathrm{E}[\cos(X_t)]\sin(X_t)\big)\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad X_0 = x_0.$$

The above MV-SDE can be seen in the form of (2) with $K = 3$, $d = 1$, $q = 1$, and

$$\varphi(x) = \big(\sin(x), \cos(x), 1\big), \qquad \alpha(t, x) = (\cos(x), -\sin(x), 0)^\top, \qquad \beta(t, x) = (0, 0, \sigma)^\top.$$

Here, $\bar\gamma(t) = \mathrm{E}[\varphi(X_t)]$ takes values on $\mathbb{R}^3$. However, as $\bar\gamma_3(t) \equiv 1$, we run the SGD algorithm with $K = 2$ to approximate $\bar\gamma_1(t) = \mathrm{E}[\sin(X_t)]$ and $\bar\gamma_2(t) = \mathrm{E}[\cos(X_t)]$, which are the real unknown functions. Note that φ, α, and β in this example satisfy the assumptions required in the convergence result of Section 3.3.

We test different choices of algorithm hyperparameters: $r_0 = 1, 5, 10$, $\rho = 0.6, 0.7, 0.8, 0.9$, $n = 3, 4, 5, 6, 7, 8$, and $M = 1, 10, 100, 1000, 10{,}000$. The benchmark $\bar\gamma^{\mathrm{MC}}$ in (28), based on the Monte Carlo particle-system approximation, is obtained with $N_0 = 10^6$ particles. For each combination of hyperparameters, we perform 1000 independent runs of the algorithm, and record the average number of iterations of the SGD algorithm required to achieve the relative error accuracy $\varepsilon_m < 1\%$.

In Tables 3–5 we report the lowest average number of iterations obtained over different (r_0, ρ) combinations for a given M and n for time horizons $T = 0.5, 1, 2$, respectively. We observe that the average number of iterations decreases with the size of the mini-batch sample M, and increases with the time horizon T. By taking into account the computational time, it is clear that $M = 1000$ gives the best performance. It is evident that our *mini-batch-SGD* algorithm converges very quickly to the Monte Carlo particle-system-based benchmark.

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

TABLE 3. Kuramoto–Shinomoto–Sakaguchi MV-SDE. Average number of iterations m (execution time in seconds) over 1000 independent runs of the algorithm to achieve relative error accuracy $\varepsilon_m < 1\%$ with the best combination of (r_0, ρ) for each pair (n,M). Here $T = 0.5$, $x_0 = 0.5$, and $\sigma = 0.5$, and the Monte Carlo benchmark exhibited an execution time of 1.98 seconds.

	$M = 1$	$M = 10$	$M = 100$	$M = 1000$	$M = 10,000$
$n = 3$	249.6 (0.23)	40.4 (0.05)	8.3 (0.01)	2.6 (0.02)	2 (0.17)
$n = 4$	368 (0.34)	48 (0.06)	8.2 (0.01)	2.6 (0.02)	2 (0.19)
$n = 5$	382.4 (0.36)	59.6 (0.06)	8.3 (0.01)	2.4 (0.02)	1.1 (0.14)
$n = 6$	405 (0.38)	53.8 (0.06)	8.3 (0.02)	2.4 (0.02)	1.1 (0.15)
$n = 7$	447.3 (0.42)	53.9 (0.06)	8.7 (0.02)	2.7 (0.02)	1.5 (0.22)
$n = 8$	451 (0.42)	55.2 (0.07)	10.3 (0.02)	2.7 (0.03)	2 (0.33)

TABLE 4. Kuramoto–Shinomoto–Sakaguchi MV-SDE. Average number of iterations m (execution time in seconds) over 1000 independent runs of the algorithm to achieve relative error accuracy $\varepsilon_m < 1\%$ with the best combination of (r_0, ρ) for each pair (n,M). Here $T = 1$, $x_0 = 0.5$, and $\sigma = 0.5$, and the Monte Carlo benchmark exhibited an execution time of 3.60 seconds.

	$M = 1$	$M = 10$	$M = 100$	$M = 1000$	$M = 10,000$
$n = 3$	444.1 (0.80)	77.5 (0.18)	16.6 (0.06)	5.1 (0.08)	4 (0.65)
$n = 4$	747 (1.35)	101.3 (0.23)	16.1 (0.06)	4.4 (0.07)	3 (0.62)
$n = 5$	709.9 (1.31)	105.8 (0.24)	16.9 (0.07)	4.5 (0.08)	3 (0.71)
$n = 6$	804.3 (1.48)	109.9 (0.25)	17.8 (0.06)	4.1 (0.08)	2.1 (0.57)
$n = 7$	899.9 (1.67)	110.9 (0.27)	16.5 (0.06)	4 (0.08)	2 (0.62)
$n = 8$	923.4 (1.70)	118.4 (0.28)	17.9 (0.07)	3.9 (0.09)	2 (0.67)

TABLE 5. Kuramoto–Shinomoto–Sakaguchi MV-SDE. Average number of iterations m (execution time in seconds) over 1000 independent runs of the algorithm to achieve relative error accuracy $\varepsilon_m < 1\%$ with the best combination of (r_0, ρ) for each pair (n,M). Here $T = 2$, $x_0 = 0.5$, and $\sigma = 0.5$, and the Monte Carlo benchmark exhibited an execution time of 6.59 seconds.

	$M = 1$	$M = 10$	$M = 100$	$M = 1000$	$M = 10,000$
$n = 3$	992.1 (3.53)	134 (0.61)	21.4 (0.14)	4.7 (0.15)	2.3 (0.86)
$n = 4$	1302.7 (4.67)	153.8 (0.70)	25.4 (0.18)	6 (0.21)	3.1 (1.33)
$n = 5$	1576.5 (5.66)	191.4 (0.87)	28.6 (0.20)	6.8 (0.26)	4.1 (1.99)
$n = 6$	1730.9 (6.22)	244.5 (1.12)	34.9 (0.25)	7.8 (0.33)	5 (2.71)
$n = 7$	1950.3 (7.02)	220.3 (1.01)	36.1 (0.27)	7.7 (0.35)	4 (2.43)
$n = 8$	2033.8 (7.32)	238.1 (1.09)	34.5 (0.26)	7.6 (0.37)	4 (2.68)

In Figures 2–4 we plot the output curves $\mathscr{L}\mathbf{a}_m = (\mathscr{L}\mathbf{a}_m^1, \mathscr{L}\mathbf{a}_m^2)$ of the SGD algorithm versus the benchmark curves $\bar{\gamma}^{\mathrm{MC}} = (\bar{\gamma}_1^{\mathrm{MC}}, \bar{\gamma}_2^{\mathrm{MC}})$, for all values of the polynomial degree $n = 3, 4, \ldots, 8$ and for time horizons $T = 0.5, 1, 2$ respectively. Note that the approximation does not deteriorate as the time horizon T increases. In particular, the slight loss of precision

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

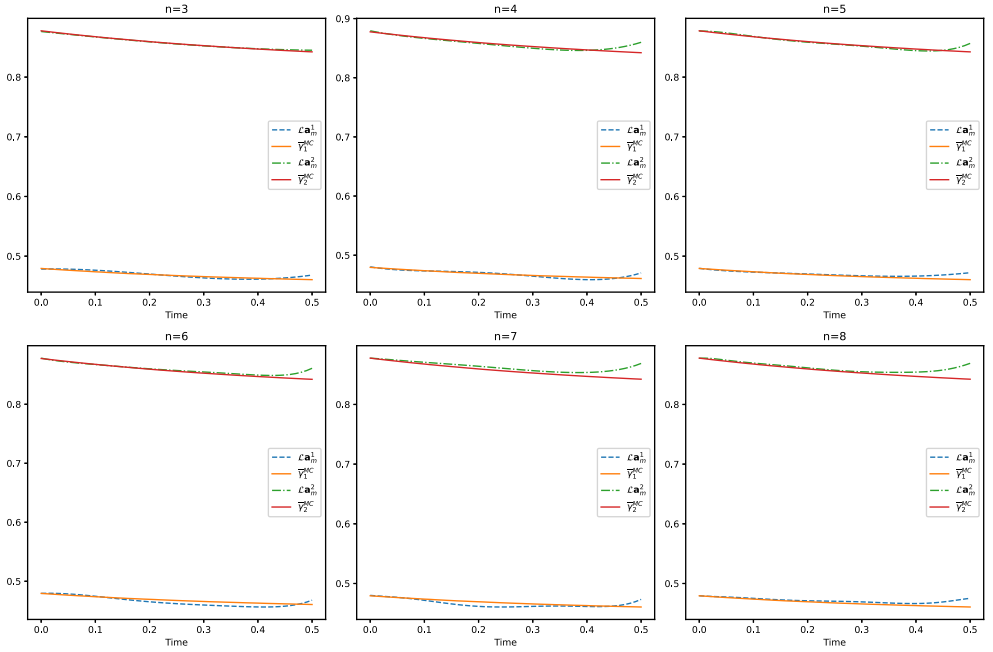

FIGURE 2. Kuramoto–Shinomoto–Sakaguchi MV-SDE. Comparison of the output curves $\mathscr{L}\mathbf{a}_m = (\mathscr{L}\mathbf{a}_m^1, \mathscr{L}\mathbf{a}_m^2)$ of the SGD algorithm versus the benchmark curves $\bar{\gamma}^{\mathrm{MC}} = (\bar{\gamma}_1^{\mathrm{MC}}, \bar{\gamma}_2^{\mathrm{MC}})$ for all values of n, for time-step size $h = 10^{-2}$, $T = 0.5$, $M = 1000$, $x_0 = 0.5$, and $\sigma = 0.5$.

that can be observed towards the right end of the time interval can be explained by the choice of polynomial approximations. For instance, the gap between $(\mathscr{L}\mathbf{a}_m)^1$ and $\bar{\gamma}_1^{\mathrm{MC}}$, which can be seen in Figure 2 in the region $t \in [0.4, 0.5]$, does not appear in the analogous plot of Figure 3 (or Figure 4) after the same number of steps. This shows that the error does not propagate over time: this is mainly due to the fact that ours is a full history-type minimization algorithm, in that it approximates the minimum in (7) over the set of the continuous curves on $[0, T]$.

Finally observe that the algorithm works best for low values of the polynomial order, namely $n = 3, 4$. Indeed, typically the number of iterations needed to reach the target relative error accuracy $\varepsilon_m < 1\%$ (cf. Tables 3–5) increase as higher values of n are considered. For the best values of the mini-batch size, namely $M = 100, 1000$, this phenomenon seems less evident. However, we can see from Figures 2–4 that the quality of the approximation deteriorates when the number of iterations does not increase with n. This can be interpreted as an overfitting phenomenon, which is expected as the benchmark curves $\bar{\gamma}^{\mathrm{MC}} = (\bar{\gamma}_1^{\mathrm{MC}}, \bar{\gamma}_2^{\mathrm{MC}})$ are relatively flat.

4.2. Polynomial drift MV-SDE

Consider the MV-SDE with polynomial drift given as follows:

$$\mathrm{d}X_t = \Big(\mathrm{E}[X_t] - X_t\mathrm{E}[X_t^2] + \delta X_t\Big)\,\mathrm{d}t + X_t\,\mathrm{d}W_t, \quad X_0 = x_0.$$

The latter can be cast in the form of (2) with $K = 3$, $d = 1$, $q = 1$, and

$$\varphi(x) = (x, x^2, 1), \quad \alpha(t, x) = (1, -x, \delta x)^T, \quad \beta(t, x) = (0, 0, x)^T.$$

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

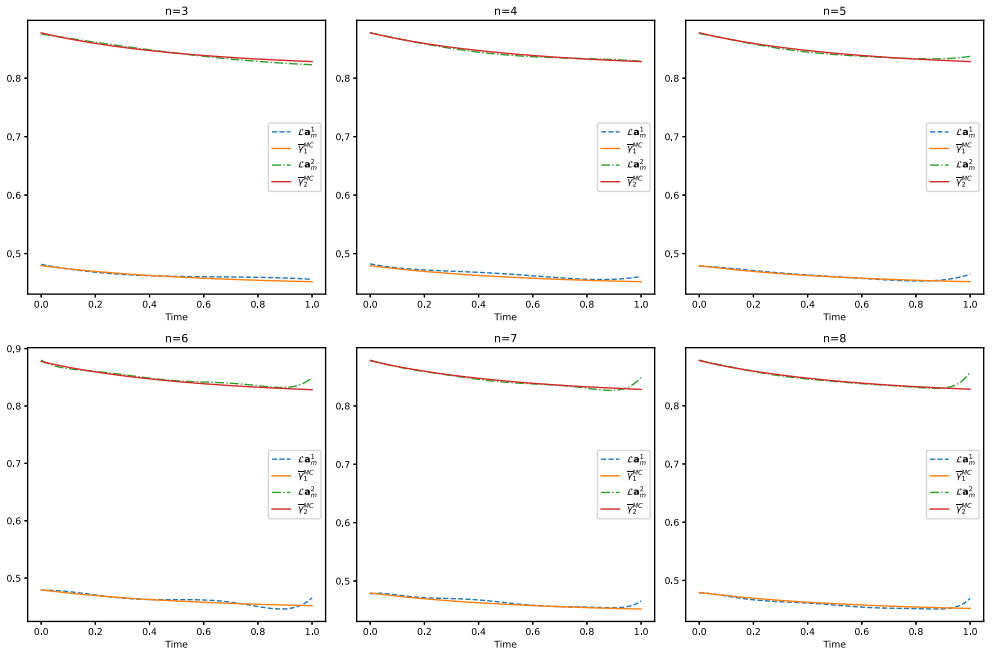

FIGURE 3. Kuramoto–Shinomoto–Sakaguchi MV-SDE. Comparison of the output curves $\mathscr{L}\mathbf{a}_m = (\mathscr{L}\mathbf{a}_m^1, \mathscr{L}\mathbf{a}_m^2)$ of the SGD algorithm versus the benchmark curves $\bar{\gamma}^{\mathrm{MC}} = (\bar{\gamma}_1^{\mathrm{MC}}, \bar{\gamma}_2^{\mathrm{MC}})$ for all values of n, for time-step size $h = 10^{-2}$, $T = 1$, $M = 1000$, $x_0 = 0.5$, and $\sigma = 0.5$

Once more, the function $\bar{\gamma}(t) = \mathrm{E}[\varphi(X_t)]$ takes values on $\mathbb{R}^3$ but $\bar{\gamma}_3(t) \equiv 1$. Thus, we run the SGD algorithm with $K = 2$ to approximate $\bar{\gamma}_1(t) = \mathrm{E}[(X_t)^2]$ and $\bar{\gamma}_2(t) = \mathrm{E}[X_t]$, which are the real unknown functions. Also note that φ, α, and β in this example are unbounded and thus do not satisfy the assumptions in the results of Section 3.3. In particular, the drift coefficient has super-linear growth in the measure component.

We test different choices of algorithm hyperparameters: $r_0 = 1, 5, 10$, $\rho = 0.6, 0.7$, and $M = 1, 10, 100, 1000$. We set the polynomial degree $n = 3$, and test the performance of the SGD algorithm for different time horizons, $T = 0.1, 0.5, 1$. The benchmark $\bar{\gamma}^{\mathrm{MC}}$ in (28) based on the Monte Carlo particle-system approximation is obtained with $N_0 = 10^7$ particles. For each combination of the hyperparameters, we perform 1000 independent runs of the algorithm, and record the average number of iterations of the SGD algorithm required to achieve the relative error accuracy $\varepsilon_m < 1\%$.

In Table 6 we report the lowest average number of iterations over different (r_0, ρ) combinations for a given M and T. Once more, as the value of the mini-batch increases the number of steps to reach convergence decreases. Compared to the model tested in Section 4.1, increasing the time horizon T has a stronger impact on the convergence, which is due to the quadratic growth of the coefficient. We stress that for this model even the MC particle method requires an increasing number of particles to obtain reliable benchmark curves for increasing values of T.

In Figure 5 we plot the output curves $\mathscr{L}\mathbf{a}_m = (\mathscr{L}\mathbf{a}_m^1, \mathscr{L}\mathbf{a}_m^2)$ of the SGD algorithm versus the benchmark curves $\bar{\gamma}^{\mathrm{MC}} = (\bar{\gamma}_1^{\mathrm{MC}}, \bar{\gamma}_2^{\mathrm{MC}})$ for all time horizons $T = 0.1$, $T = 0.5$, and $T = 1$. In the plots, particularly for $T = 0.5$, notice the considerable loss of accuracy in our approximation

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

TABLE 6. Polynomial drift MV-SDE. Average number of iterations (execution time in seconds) over 1000 independent runs of the algorithm to achieve relative error accuracy $\varepsilon_m < 1\%$ with the best combination of (r_0, ρ) for each pair (T,M). Here $x_0 = 1$, $\delta = 0.8$ and the Monte Carlo benchmark exhibited an execution time of 2.57 seconds for $T = 0.1$, 9.78 seconds for $T = 0.5$, and 19.53 seconds for $T = 1$.

	$M = 1$	$M = 10$	$M = 100$	$M = 1000$
$T = 0.1$	258 (0.06)	52.3 (0.02)	31.5 (0.01)	26.9 (0.03)
$T = 0.5$	3456.7 (3.52)	590.3 (0.72)	131.2 (0.20)	79.7 (0.39)
$T = 1.0$	4995.9 (9.91)	3582 (8.54)	1003.9 (3.07)	776.8 (8.25)

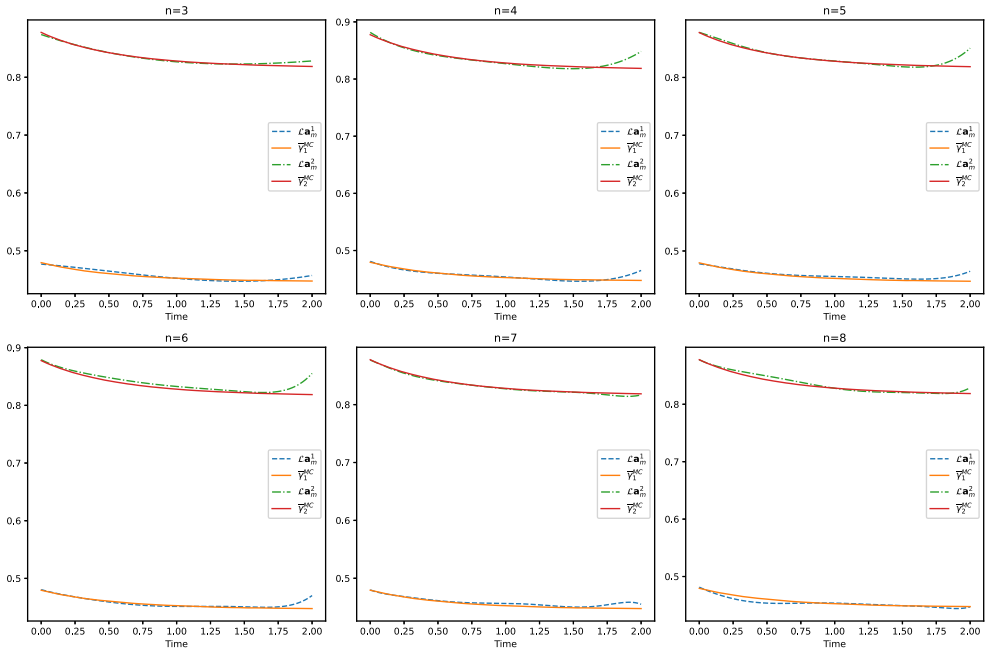

FIGURE 4. Kuramoto–Shinomoto–Sakaguchi MV-SDE. Comparison of the output curves $\mathscr{L}\mathbf{a}_m = (\mathscr{L}\mathbf{a}_m^1, \mathscr{L}\mathbf{a}_m^2)$ of the SGD algorithm versus the benchmark curves $\bar{\gamma}^{\mathrm{MC}} = (\bar{\gamma}_1^{\mathrm{MC}}, \bar{\gamma}_2^{\mathrm{MC}})$ for all values of n, for time-step size $h = 10^{-2}$, $T = 2$, $M = 1000$, $x_0 = 0.5$, and $\sigma = 0.5$

$\mathscr{L}\mathbf{a}_m^2$ of $\bar{\gamma}_2(t) = \mathrm{E}[X_t]$ near the right boundary of the interval (similar effects were visible in the plots relative to the Kuramoto model). After running additional tests, we do not believe that this effect is due to the polynomial projection, similar to Runge-type phenomena in polynomial interpolation over compact sets, but we claim that it is rather due to the norm used to define the cost function F^2 (cf. (8)), in combination with the chosen threshold level we set for ε_m in (29). Indeed, lowering this threshold from to 10^{-2} to 10^{-3} reduces considerably the error, also near the boundaries of the interval. However, one could look for strategies to distribute the error in a more uniform way across the interval, and obtain a better qualitative result while keeping the same error threshold. One possibility of improvement is to first employ the alternative objective function presented in Remark 5, i.e. using a weight function to force the SGD algorithm to be more accurate towards the end of the interval. One can also use a similar weight function to

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

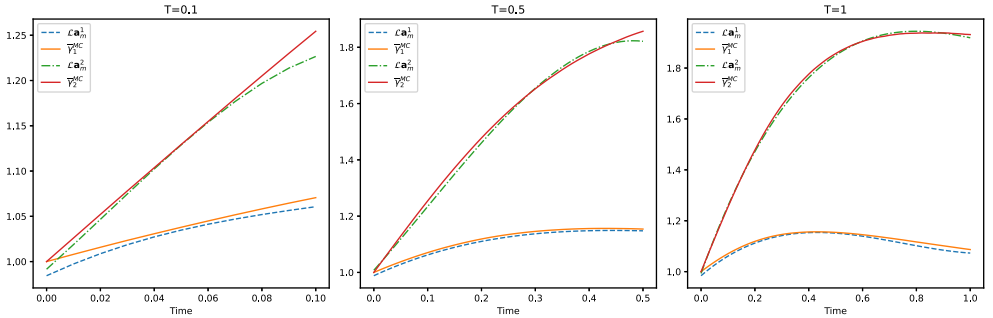

FIGURE 5. Polynomial drift MV-SDE. Comparison of the output curves $\mathscr{L}\mathbf{a}_m = (\mathscr{L}\mathbf{a}_m^1, \mathscr{L}\mathbf{a}_m^2)$ of the SGD algorithm versus the benchmark curves $\bar{\gamma}^{\mathrm{MC}} = (\bar{\gamma}_1^{\mathrm{MC}}, \bar{\gamma}_2^{\mathrm{MC}})$ for all values of T, for time-step size $h = 10^{-2}$, $n = 3$, $M = 1000$, $x_0 = 1$, and $\delta = 0.8$.

compute a weighted $L^2([0, T])$ norm for error ε_m in the stopping criterion. Another possibility is to experiment with different choices of basis functions listed in Examples 3 and 4 and choose the one with the best error performance.

4.3. Convolution-type MV-SDE

We consider the following convolution-type MV-SDE:

$$\mathrm{d}X_t = \mathrm{E}\left[\exp\left(-(X_t - x)^2/2\right)\right]\Big|_{x=X_t} \mathrm{d}t + \sigma\, \mathrm{d}W_t, \quad X_0 = \xi \sim \mathcal{N}_{(0,1)}. \tag{30}$$

As shown in [5], the equation above can be approximated by employing a projection technique based on generalized Fourier series, which we sketch out here for the reader's convenience. Noticing that the drift coefficient in (30) can be written as

$$\mathrm{E}[b(x, X_t)]|_{x=X_t}, \quad b(x, y) = \exp\left(-(y - x)^2/2\right),$$

we can expand $b(\cdot, x)$ in a suitable weighted L^2 space as follows. Setting the measure $\mu(\mathrm{d}x) = e^{-x^2/2}\,\mathrm{d}x$ on the Borel sets of $\mathbb{R}$, we consider the following orthonormal basis of $L^2(\mathbb{R}, \mu)$ of the normalized Hermite polynomials

$$\overline{H}_k(x) = c_k(-1)^k \mathrm{e}^{x^2} \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left(\mathrm{e}^{-x^2}\right), \quad c_k = \left(2^k k! \sqrt{\pi}\right)^{-1/2}, \quad k \in \mathbb{N}_0.$$

Therefore, for $K \in \mathbb{N}$ sufficiently large, $b(x, \cdot)$ can be approximated in $L^2(\mathbb{R}, \mu)$ as

$$b(x, y) \approx \sum_{k=0}^{K} \alpha_k(x)\varphi_k(y), \tag{31}$$

where

$$\varphi_k(x) = \overline{H}_k(x)\mathrm{e}^{-x^2/2}, \, quad \alpha_k(x) = \int_{\mathbb{R}} b(x, y)\varphi_k(y)\,\mathrm{d}y = \pi^{1/4}\left(\frac{1}{2}\right)^{k/2} \frac{x^k}{\sqrt{k!}} \mathrm{e}^{-x^2/4}.$$

Therefore, the MV-SDE (4.3) can be approximated by the MV-SDE with separable coefficients

$$\mathrm{d}X_t^{(K)} = \sum_{k=0}^{K} \mathbb{E}\left[\varphi_k(X_t^{(K)})\right]\alpha_k\left(X_t^{(K)}\right) \mathrm{d}t + \sigma\, \mathrm{d}W_t, \quad X_0 = \xi = \mathcal{N}_{(0,1)},$$

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

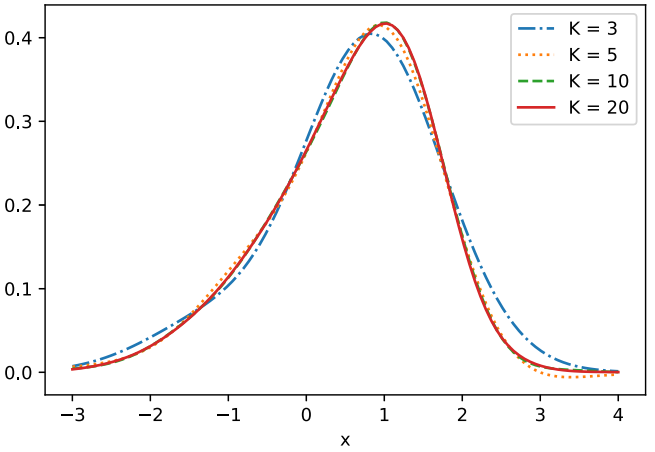

FIGURE 6. Monte Carlo method densities $\tilde{w}_T^{(K),\mathrm{MC}}(x)$ over the interval $[-3, 4]$, for $T = 1$ and $K = 3, 5, 10, 20$. The benchmark vector $\bar{\gamma}^{\mathrm{MC}}(T)$ was computed with $N_0 = 10^7$ particles. Here $X_0 \sim \mathcal{N}_{(0,1)}$ and $\sigma = 0.1$. In the model (30) we had $X_0 \sim \mathcal{N}_{(0,1)}$ and $\sigma = 0.1$.

which can be clearly cast in the form of (2). For a given $K \in \mathbb{N}$, we run the SGD algorithm in order to approximate $\bar{\gamma}_k^{(K)}(t) = \mathbb{E}[\varphi_k(X_t^{(K)})]$ for $k = 0, \ldots, K$.

For this model, we test the accuracy of the SGD algorithm by checking the quality of the density approximation of the solution X_T to Section 4.3 that can be derived from the $\bar{\gamma}_k^{(K)}$ curves as follows (see also [5]). For a fixed $T > 0$ denote by $w_T(x)$ and $w_T^{(K)}(x)$ the densities of X_T and $X_T^{(K)}$, respectively. Following the same argument that led to (31), for $K \in \mathbb{N}$ sufficiently large, $w_T(x)$ can be approximated in $L^2(\mathbb{R}, \mu)$ as

$$w_T(x) \approx w_T^{(K)}(x) \approx \sum_{k=0}^{K} \int_{\mathbb{R}} \varphi_k(y) w_T^{(K)}(y)\, \mathrm{d}y\, \varphi_k(x) = \sum_{k=0}^{K} \bar{\gamma}_k^{(K)}(T)\varphi_k(x) =: \tilde{w}_T^{(K)}(x).$$

Denote by $\tilde{w}_T^{(K),\mathrm{SGD}}$ and $\tilde{w}_T^{(K),\mathrm{MC}}$, the approximations of $\tilde{w}_T^{(K)}$ obtained by replacing the vector $\bar{\gamma}^{(K)}(T)$ with the output of the SGD algorithm, $\mathscr{L}\mathbf{a}_m(T)$, and with the particle-system benchmark, $\bar{\gamma}^{\mathrm{MC}}(T)$, respectively.

In our tests we set $\sigma = 0.1$ and $T = 1$ as in [5]. The number of particles to derive the benchmark $\bar{\gamma}^{\mathrm{MC}}(T)$ is set as $N_0 = 10^7$. Note that in [5] the authors used $N_0 = 5 \times 10^2$, which is why our density plots differ considerably from theirs.

In Figure 6 we plot the approximate density $\tilde{w}_T^{(K),\mathrm{MC}}(x)$ for $K = 3, 5, 10, 20$, with x ranging on the interval $[-3, 4]$. From this we conclude that the value $K = 10$ is high enough in order for $\tilde{w}^{(K)}$ to accurately approximate w_T over the interval $[-3, 4]$. In Figure 7 we test the accuracy of the SGD algorithm by plotting $\tilde{w}_T^{(K),\mathrm{SGD}}(x)$ versus $\tilde{w}_T^{(K),\mathrm{MC}}(x)$, with $K = 10$, over the same interval. The parameters of the SGD algorithm were chosen as $n = 3, M = 100, r_0 = 5, \rho = 0.9$, and the algorithm halted after $m = 172$ iterations (35 seconds of computation time) when the relative error reached $\varepsilon_m < 1\%$.

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

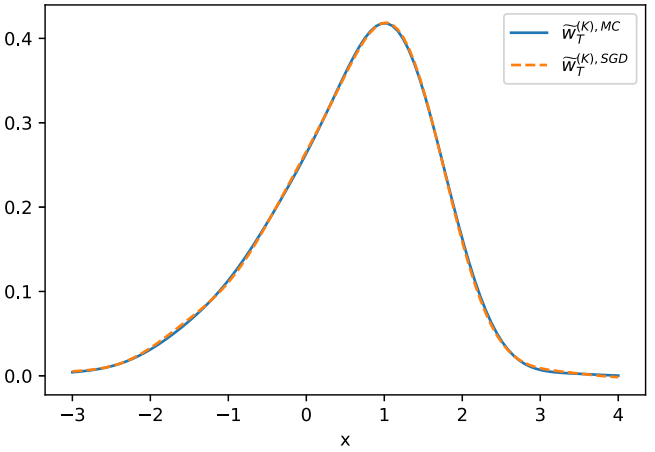

FIGURE 7. Comparison of $\tilde{w}_T^{(K),\mathrm{SGD}}(x)$ and $\tilde{w}_T^{(K),\mathrm{MC}}(x)$ over the interval $[-3, 4]$, for $T = 1$ and $K = 10$. The benchmark vector $\bar{\gamma}^{\mathrm{MC}}(T)$ was computed with $N_0 = 10^7$ particles. The SGD output $\mathscr{L}\mathbf{a}_m(T)$ was obtained after $m = 172$ iterations (35 seconds of computation time), with parameters $n = 3$, $M = 100$, $r_0 = 5$, and $\rho = 0.9$. In the model (30) we had $X_0 \sim \mathcal{N}_{(0,1)}$ and $\sigma = 0.1$.

Appendix A. SDE stability estimates

Throughout this section, fix $T > 0$ and let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathrm{P})$ be a filtered probability space, with $(\mathcal{F}_t)_{0\le t\le T}$ satisfying the usual hypothesis, supporting a q-dimensional Brownian motion $W = (W_t)_{t\in[0,T]}$. Consider the $\mathbb{R}^d$-valued SDE

$$\mathrm{d}\mathcal{Y}_t = \rho_t\big(\mu(t, \mathcal{Y}_t)\,\mathrm{d}t + \sigma(t, \mathcal{Y}_t)\,\mathrm{d}W_t\big), \quad \mathcal{Y}_0 = \zeta, \tag{A1}$$

where ζ is an $\mathbb{R}^d$-valued $\mathcal{F}_0$-measurable random variable, $\mu{:}[0, T] \times \mathbb{R}^d \to \mathbb{R}^{m\times d}$ and $\beta : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{m\times d\times q}$ are deterministic measurable coefficients, and $\rho = (\rho_t)_{t\in[0,T]}$ is a progressively measurable $\mathbb{R}^m$-valued stochastic process. Hereafter, we denote by $\mathcal{Y}^\rho$ a solution to (A1).

Lemma 3. *Assume that $\zeta \in L^2(\Omega, \mathcal{F}_0, \mathrm{P})$, ρ is bounded, and there exists a constant $\Lambda > 0$ such that*

$$|\mu(t, x) - \mu(t, y)| + |\sigma(t, x) - \sigma(t, y)| \le \Lambda|x - y|, \quad t \in [0, T],\ x, y \in \mathbb{R}^d, \tag{A2}$$

and

$$|\mu(t, x)| + |\sigma(t, x)| \le \Lambda(1 + |x|), \quad t \in [0, T],\ x \in \mathbb{R}^d. \tag{A3}$$

Then (A1) has a unique solution $\mathcal{Y}^\rho$ in the sense of indistinguishability.

Furthermore, for a given $\kappa > 0$ and for any ρ and $\hat{\rho}$ with $\|\rho\|_{\mathcal{S}^\infty_{[0,T]}}, \|\hat{\rho}\|_{\mathcal{S}^\infty_{[0,T]}} \le \kappa$ we have:

(a) if $\zeta \in L^{2p}(\Omega, \mathcal{F}_0, \mathrm{P})$ for some $p \ge 1$, then

$$\|\mathcal{Y}^\rho\|_{\mathcal{S}^{2p}_{[0,T]}} \le C_1 \tag{A4}$$

where $C_1 > 0$ only depends on κ, p, T, Λ, $\mathrm{E}[|\zeta|^{2p}]$, and the dimensions;

(b) if $\zeta \in L^{4p}(\Omega, \mathcal{F}_0, \mathrm{P})$ for some $p \ge 1$, then

$$\|\mathcal{Y}^\rho - \mathcal{Y}^{\hat{\rho}}\|_{\mathcal{S}^{2p}_{[0,T]}} \le C_2\,\|\rho - \hat{\rho}\|_{\mathbb{L}^{4p}_{[0,T]}}, \tag{A5}$$

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

where $C_2 > 0$ only depends on κ, p, T, Λ, $\mathrm{E}[|\zeta|^{4p}]$, and the dimensions;

(c) if

$$|\mu(t,x)| + |\sigma(t,x)| \le \Lambda, \quad t \in [0,T],\ x \in \mathbb{R}^d, \tag{A6}$$

then for any $p \ge 1$ we have

$$\|\mathcal{Y}^{\rho} - \mathcal{Y}^{\hat{\rho}}\|_{\mathcal{S}^{2p}_{[0,T]}} \le C_3 \, \|\rho - \hat{\rho}\|_{\mathbb{L}^{2p}_{[0,T]}} \le C_3 \, T^{\frac{1}{2p}} \|\rho - \hat{\rho}\|_{\mathcal{S}^{\infty}_{[0,T]}}, \tag{A7}$$

where $C_3 > 0$ only depends on κ, p, T, Λ, and the dimensions. Note that (A7) is true only assuming $\xi \in L^2(\Omega, \mathcal{F}_0, \mathrm{P})$ and that C_3 does not depend on $\mathrm{E}[|\zeta|^2]$.

Proof. The well-posedness of $\mathcal{Y}^{\rho}$ and part (a) of the result are particular cases of [35, Theorem 2.5.7 (p. 77)] and [35, Theorem 2.5.10 (p. 85)], respectively.

Parts (b) and (c) can be readily obtained by combining (A4) with [35, Theorem 2.5.9 (p. 83)]. Precisely, from [35, Theorem 2.5.9], there exists $C > 0$, only dependent on κ, p, T, Λ, and K, such that

$$\begin{aligned}
&\mathrm{E}\Big[\sup_{t\le T} |\mathcal{Y}^{\rho}_t - \mathcal{Y}^{\hat{\rho}}_t|^{2p}\Big] \\
&\quad \le C\,\mathrm{E}\bigg[\int_0^T \Big(\big|\rho_t\mu(t,\mathcal{Y}^{\hat{\rho}}_t) - \hat{\rho}_t\mu(t,\mathcal{Y}^{\hat{\rho}}_t)\big|^{2p} + \big|\rho_t\sigma(t,\mathcal{Y}^{\hat{\rho}}_t) - \hat{\rho}_t\sigma(t,\mathcal{Y}^{\hat{\rho}}_t)\big|^{2p}\Big)\,\mathrm{d}t\bigg] \\
&\quad \le C\,\mathrm{E}\bigg[\int_0^T |\rho_t - \hat{\rho}_t|^{2p}\Big(\big|\mu(t,\mathcal{Y}^{\hat{\rho}}_t)\big|^{2p} + \big|\sigma(t,\mathcal{Y}^{\hat{\rho}}_t)\big|^{2p}\Big)\,\mathrm{d}t\bigg] \\
&\quad \le C\Lambda\,\|\rho - \hat{\rho}\|^{2p}_{\mathbb{L}^{4p}_{[0,T]}} \bigg(\mathrm{E}\bigg[\int_0^T \big(1 + |\mathcal{Y}^{\hat{\rho}}_t|\big)^{4p}\,\mathrm{d}t\bigg]\bigg)^{\frac{1}{2}},
\end{aligned}$$

where the last inequality follows from (A3) and the Cauchy–Schwarz inequality. Combining the result with (A4), yields (A5). When (A6) holds true, we have a stronger estimate

$$\mathrm{E}\bigg[\int_0^T |\rho_t - \hat{\rho}_t|^{2p}\Big(\big|\mu(t,\mathcal{Y}^{\hat{\rho}}_t)\big|^{2p} + \big|\sigma(t,\mathcal{Y}^{\hat{\rho}}_t)\big|^{2p}\Big)\mathrm{d}t\bigg] \le \Lambda^{2p}\,\|\rho - \hat{\rho}\|^{2p}_{\mathbb{L}^{2p}_{[0,T]}},$$

which yields (A7).

Appendix B. Proofs

B.1. Proof of the time-regularity of $\bar{\gamma}$

Proof of Proposition 2. Let X be the unique solution to (2), which exists due to Proposition 1. A straightforward application of Itô's formula shows that its time-marginal laws $\mu_t(\mathrm{d}x)$ solve the Fokker–Planck equation in distributional form over $[0,T] \times \mathbb{R}^d$, which is

$$\int_0^T \int_{\mathbb{R}^d} \mathscr{K}\psi(s,x)\mu_s(\mathrm{d}x)\,\mathrm{d}s = 0, \quad \psi \in C_0^{\infty}\big([0,T] \times \mathbb{R}^d\big).$$

By Assumption 2 and Proposition 1, the coefficients of $\mathscr{K}$ are continuous. Therefore, a standard approximating procedure yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}[f(t,X_t)] = \mathrm{E}[\mathscr{K}f(t,X_t)], \quad t \in [0,T], \tag{B1}$$

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

for any test function $f \in C_b^{1,2}([0,T]\times\mathbb{R}^d)$. Now, by Assumption 3, $\varphi \in C_b^2(\mathbb{R}^d)$. Thus, (B1) with $f(t,x)=\varphi(x)$ becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\gamma}(t) = \mathrm{E}[\mathscr{K}\varphi(X_t)], \quad t\in[0,T],$$

which is (4) with $n=1$, and in particular $\bar{\gamma}\in C_b^1([0,T])$.

Let now $m\in\{1,\dots,N-1\}$, assume $\bar{\gamma}\in C_b^m([0,T])$ and (4) be true for m. By Assumption 3 we obtain

$$\mathscr{K}^m\varphi \in C_b^{1,2}([0,T]\times\mathbb{R}^d).$$

Hence, (B1) with $f=\mathscr{K}^m\varphi$ yields

$$\frac{\mathrm{d}^{m+1}}{\mathrm{d}t^{m+1}}\bar{\gamma}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}[\big(\mathscr{K}^m\varphi\big)(t,X_t)] = \mathrm{E}[\big(\mathscr{K}^{m+1}\varphi\big)(t,X_t)], \quad t\in[0,T],$$

and, in particular, $\bar{\gamma}\in C_b^{m+1}([0,T])$. This concludes the proof. □

B.2. Auxiliary stability and moment-estimate results

Hereafter, we denote by $(Z^a, Y^{a;k,j})$ any solution to (21), (22) for a given $a\in\mathbb{R}^{(n+1)K}$ and $k=0,\dots,n$, $j=1,\dots,K$. Throughout the remainder of this section we denote by C, indistinctly, any positive constant that does not depend on $a\in\mathbb{R}^{(n+1)K}$.

We start with the following stability and moment estimates for the solution to the system (21) and (22).

Lemma 4. *Let Assumption 1 be in force and assume that $\alpha(t,\cdot)$, $\beta(t,\cdot)\in C^2(\mathbb{R}^d)$, for any $t\in[0,T]$, with derivatives bounded by some $R>0$. Then, for any $a\in\mathbb{R}^{(n+1)K}$, the system (21) and (22) is strongly well-posed for any $k=0,\dots,n$ and $j=1,\dots,K$. Furthermore, we have*

$$\|Z^a\|_{\mathcal{S}^2_{[0,T]}} + \sum_{\substack{k=0,\dots,n\\ j=1,\dots,K}} \|Y^{a;k,j}\|_{\mathcal{S}^2_{[0,T]}} \le C, \qquad a\in\mathbb{R}^{(n+1)K}, \tag{B2}$$

$$\|Z^a - Z^{a'}\|_{\mathcal{S}^2_{[0,T]}} + \sum_{\substack{k=0,\dots,n\\ j=1,\dots,K}} \|Y^{a;k,j} - Y^{a';k,j}\|_{\mathcal{S}^2_{[0,T]}} \le C\,|a-a'|, \qquad a,a'\in\mathbb{R}^{(n+1)K}, \tag{B3}$$

for some constant $C>0$ independent of a.

Proof. Well-posedness and estimates (B2) and (B3) for Z follow directly from Lemma 3 using that $\mathbf{h}$ is bounded and Lipschitz continuous. In particular, Lemma 3(a) yields

$$\|Z^a\|_{\mathcal{S}^2_{[0,T]}} \le C,$$

for $C>0$ independent of a. Lemma 3(c) yields

$$\begin{aligned}\|Z^a - Z^{a'}\|_{\mathcal{S}^4_{[0,T]}} &\le C\|\mathbf{h}\circ\mathscr{L}a - \mathbf{h}\circ\mathscr{L}a'\|_\infty \\ &\le C\|\nabla\mathbf{h}\|_\infty\|\mathscr{L}a-\mathscr{L}a'\|_\infty \le C\|\nabla\mathbf{h}\|_\infty\Big(\sum_{k=0,\dots,n}\|g_k\|_\infty\Big)|a-a'|,\end{aligned} \tag{B4}$$

with C depending on $\|\mathbf{h}\|_\infty$ but not on a, and $\nabla\mathbf{h}$ stands for the Jacobian of $\mathbf{h}:\mathbb{R}^K\to\mathbb{R}^K$.

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

As for Y, to ease the notation we only consider the case $q = 1$. The general case is completely analogous. We observe that SDE (22) can be cast in the form of (A1) with $\zeta = 0$, $m = 2d(1+d)$,

$$\rho_t = \rho_t(a) = \Big(g_k(t)\nabla\mathbf{h}_j\big(\mathscr{L}a(t)\big)\alpha(t, Z_t), \mathbf{h}\big(\mathscr{L}a(t)\big)\partial_{z_1}\alpha(t, Z_t), \ldots, \mathbf{h}\big(\mathscr{L}a(t)\big)\partial_{z_d}\alpha(t, Z_t),$$
$$g_k(t)\nabla\mathbf{h}_j\big(\mathscr{L}a(t)\big)\beta(t, Z_t), \mathbf{h}\big(\mathscr{L}a(t)\big)\partial_{z_1}\beta(t, Z_t), \ldots, \mathbf{h}\big(\mathscr{L}a(t)\big)\partial_{z_d}\beta(t, Z_t)\Big),$$

and

$$\mu(t, y) = \begin{pmatrix} I_d \\ y_1 I_d \\ \vdots \\ y_d I_d \\ 0_{d(1+d),d} \end{pmatrix}, \quad \sigma(t, y) = \begin{pmatrix} 0_{d(1+d),d} \\ I_d \\ y_1 I_d \\ \vdots \\ y_d I_d \end{pmatrix}, \quad t \in [0, T],\ y = (y_1, \ldots, y_d) \in \mathbb{R}^d,$$

where I_d is the $d \times d$ identity matrix and $0_{d(1+d),d}$ is the $d(1+d) \times d$-dimensional matrix with null entries. Under our assumptions, the functions μ, σ satisfy conditions (A2) and (A3) for some $\Lambda > 0$ that is independent of a and the stochastic process ρ is bounded by a constant that is also independent of a (recall that by construction $\mathbf{h}$ is constant outside a ball – see (16)). Therefore, by Lemma 3, (22) is strongly well-posed and we have (by part (b) of Lemma 3) that for any $k = 0, \ldots, n$ and $j = 1, \ldots, K$

$$\|Y^{a;k,j}\|_{\mathcal{S}^2_{[0,T]}} \le C,$$
$$\|Y^{a;k,j} - Y^{a';k,j}\|_{\mathcal{S}^2_{[0,T]}} \le C\,\|\rho(a) - \rho(a')\|_{\mathbb{L}^4_{[0,T]}}.$$

In order to conclude, we write

$$\begin{aligned}|\rho_t(a) - \rho_t(a')| \le |g_k(t)|\Big(&\big|\nabla\mathbf{h}_j\big(\mathscr{L}a(t)\big)\alpha(t, Z^a_t) - \nabla\mathbf{h}_j\big(\mathscr{L}a'(t)\big)\alpha(t, Z^{a'}_t)\big| \\ &+ \big|\nabla\mathbf{h}_j\big(\mathscr{L}a(t)\big)\beta(t, Z^a_t) - \nabla\mathbf{h}_j\big(\mathscr{L}a'(t)\big)\beta(t, Z^{a'}_t)\big|\Big) \\ &+ \sum_{i=1}^{d}\Big(\big|\mathbf{h}\big(\mathscr{L}a(t)\big)\partial_{z_i}\alpha(t, Z^a_t) - \mathbf{h}\big(\mathscr{L}a'(t)\big)\partial_{z_i}\alpha(t, Z^{a'}_t)\big| \\ &+ \big|\mathbf{h}\big(\mathscr{L}a(t)\big)\partial_{z_i}\beta(t, Z^a_t) - \mathbf{h}\big(\mathscr{L}a'(t)\big)\partial_{z_i}\beta(t, Z^{a'}_t)\big|\Big),\end{aligned}$$

whereby from the boundedness of $\mathbf{h}$, α, β and their derivatives, we obtain

$$|\rho_t(a) - \rho_t(a')| \le C|\mathscr{L}a(t) - \mathscr{L}a'(t)| + |Z^a_t - Z^{a'}_t|.$$

This then yields, via (B4),

$$\|\rho(a) - \rho(a')\|^4_{\mathbb{L}^4_{[0,T]}} \le C|a - a'|^4,$$

and completes the proof. □

Lemma 5. *Let Assumption 1 be in force. Assume also that $\alpha(t, \cdot), \beta(t, \cdot) \in C^2(\mathbb{R}^d)$, for any $t \in [0, T]$, and $\varphi \in C^2(\mathbb{R}^K)$, with derivatives bounded by some $R > 0$. Then, the function $\mathbb{R}^{(n+1)K} \ni a \mapsto \mathrm{E}\big[\varphi(Z^a_t)\big]$ is differentiable, and*

$$\partial_{a_{k,j}}\mathrm{E}\big[\varphi(Z^a_t)\big] = \mathrm{E}\big[\nabla_x\varphi\big(Z^a_t\big)Y^{a;k,j}_t\big], \quad a \in \mathbb{R}^{(n+1)K},\ t \in [0, T], \tag{B5}$$

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

for any $k=0,\dots,n$ *and* $j=1,\dots,K$. *Furthermore, the functions* $\mathbb{R}^{(n+1)K}\ni a\mapsto \mathrm{E}\big[\nabla_x\varphi\big(Z_t^a\big)Y_t^{a;k,j}\big]$ *are bounded and Lipschitz continuous, uniformly with respect to* $t\in[0,T]$ *(for any k,j).*

Proof. To ease the notation, we prove the statement for $K=1$ and $n=0$, the general case being completely analogous.

Since $\mathbf{h}$, $\alpha(t,\cdot)$, $\beta(t,\cdot)$ and their derivatives up to order 2 are bounded, uniformly w.r.t. $t\in[0,T]$, [36, Theorem 2.3.1 (p. 218)] yields

$$\partial_a Z_t^a = Y_t^a, \quad a\in\mathbb{R}, \quad \text{a.s.}, \tag{B6}$$

for any $t\in[0,T]$. Moreover, given the differentiability of φ and the integrability of Y^a from Lemma 4, the right-hand side of (B5) is well-defined. Therefore, the identity (B5) can be obtained once the exchange of derivative and expectation operations is justified. Let us set

$$f(t;a):=\mathrm{E}\big[\varphi'\big(Z_t^a\big)Y_t^a\big],\quad t\in[0,T],\ a\in\mathbb{R}.$$

By Lemma 4 and the boundedness of φ', φ'' we have

$$\begin{aligned}|f(t;a)|&\le C\mathrm{E}\big[|Y_t^a|\big]\le C,\\ |f(t;a)-f(t;a')|&\le \mathrm{E}\big[\big|\varphi'\big(Z_t^a\big)-\varphi'\big(Z_t^{a'}\big)\big|\times\big|Y_t^a\big|\big]+\mathrm{E}\big[\big|\varphi'\big(Z_t^a\big)\big|\times\big|Y_t^a-Y_t^{a'}\big|\big]\le C|a-a'|,\end{aligned} \tag{B7}$$

for any $t\in[0,T]$, $a\in\mathbb{R}$, with $C>0$ independent of t and a. Therefore, we obtain

$$\begin{aligned}\mathrm{E}\big[\varphi(Z_t^a)\big]-\mathrm{E}\big[\varphi(Z_t^0)\big]=\mathrm{E}\big[\varphi(Z_t^a)-\varphi(Z_t^0)\big]&=\mathrm{E}\bigg[\int_0^a\partial_{\hat a}\big\{\varphi(Z_t^{\hat a})\big\}\,\mathrm{d}\hat a\bigg]\\ &=\mathrm{E}\bigg[\int_0^a\varphi'(Z_t^{\hat a})Y_t^{\hat a}\,\mathrm{d}\hat a\bigg]=\int_0^a f(t;\hat a)\,\mathrm{d}\hat a,\end{aligned}$$

where the second last equality follows by using (B6), and the last equality from Fubini's theorem (using (B7)).

Lastly, using the continuity of $f(t;\cdot)$ we conclude that

$$\partial_a\mathrm{E}\big[\varphi(Z_t^a)\big]=f(t;a),\quad a\in\mathbb{R},\ t\in[0,T],$$

which completes the proof. □

B.3. Proof of the main auxiliary propositions

We are now in the position to prove Propositions 5 and 6.

Proof of Proposition 5. To ease the notation, we prove the statement for $K=1$ and $n=0$, the general case being completely analogous. By (B5) in Lemma 5 we have

$$\partial_a\Big(\big|\mathrm{E}\big[\varphi(Z_t^a)-\mathbf{h}\big((\mathscr{L}a)(t)\big)\big]\big|^2\Big)=2\,\mathrm{E}\big[\varphi(Z_t^a)-\mathbf{h}\big((\mathscr{L}a)(t)\big)\big]\mathrm{E}\big[\varphi'\big(Z_t^a\big)Y_t^a-\partial_a\mathbf{h}\big((\mathscr{L}a)(t)\big)\big].$$

Recall the heuristics in (24) and (25). For any independent copies (ξ,W) and $(\tilde\xi,\tilde W)$, since $\mathbf{h}$, $\mathbf{h}'$, φ, φ' are bounded, we observe that

$$\begin{aligned}\mathrm{E}\big[\big|\varphi\big(Z_t^a(\bar\xi,\bar W)\big)-\mathbf{h}\big((\mathscr{L}a)(t)\big)\big|^2\big]\times\mathrm{E}\big[\big|\varphi'\big(Z_t^a(\tilde\xi,\tilde W)\big)Y_t^a(\tilde\xi,\tilde W)\big)-\partial_a\mathbf{h}\big((\mathscr{L}a)(t)\big)\big|^2\big]&\\ \le C\big(1+\|Y^a\|_{\mathcal{S}^2_{[0,T]}}\big)\le C,&\end{aligned} \tag{B8}$$

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

where the last inequality follows from (B2). Therefore, we can exchange the integration and derivative operation to obtain the following:

$$\partial_a G(a) = \mathrm{E}[v(a;\xi, W;\tilde{\xi}, \tilde{W})] + \partial_a H(a),$$

with v as defined in (25), namely

$$v(a;\xi, W;\tilde{\xi}, \tilde{W}) := 2\int_0^T \Big(\varphi\big(Z_t^a(\xi, W)\big) - \mathbf{h}\big((\mathscr{L}a)(t)\big)\Big)\Big(\varphi'\big(Z_t^a(\tilde{\xi}, \tilde{W})\big)Y_t^a(\tilde{\xi}, \tilde{W}) - \partial_a\mathbf{h}\big((\mathscr{L}a)(t)\big)\Big)\,\mathrm{d}t.$$

Moreover, Jensen's inequality and Fubini's theorem, together with (B8), prove that

$$\mathrm{E}_{\bar{\mathrm{P}}}\big[|v(a;\xi, W;\tilde{\xi}, \tilde{W})|^2\big] \le C, \quad a \in \mathbb{R},\ m \in \mathbb{N}_0, \tag{B9}$$

with C independent of a. Hence, for $m \in \mathbb{N}$ and recalling $\mathbf{v}_m$ as defined recursively in Algorithm 1, we obtain for any $m \in \mathbb{N}_0$

$$\begin{aligned}&\mathrm{E}_{\bar{\mathrm{P}}}\big[\mathbf{v}_{m+1} - \nabla_a G(\mathbf{a}_m)|\bar{\mathcal{F}}_m\big]\\ &= \mathrm{E}_{\bar{\mathrm{P}}}\Big[v(\mathbf{a}_m;\xi_{m+1}, W_{m+1};\tilde{\xi}_{m+1}, \tilde{W}_{m+1}) - \mathrm{E}_{\bar{\mathrm{P}}}\big[v(a;\xi_{m+1}, W_{m+1};\tilde{\xi}_{m+1}, \tilde{W}_{m+1})\big]_{a=\mathbf{a}_m}\Big|\bar{\mathcal{F}}_m\Big] = 0,\end{aligned}$$

which is (27). From (B9) and Assumption 4 on H we also obtain

$$\mathrm{E}_{\bar{\mathrm{P}}}\big[|\mathbf{v}_{m+1} - \nabla_a G(\mathbf{a}_m)|^2\big|\bar{\mathcal{F}}_m\big] \le C, \quad m \in \mathbb{N}_0,$$

which concludes the proof. □

Proof of Proposition 6. To ease the notation, we prove the statement for $K = 1$ and $n = 0$, the general case being completely analogous.

By the same arguments as in the proof of Proposition 5, and using Assumption 4, one obtains

$$\partial_a G(a) = 2\int_0^T \mathrm{E}\Big[\varphi\big(Z_t^a\big) - \mathbf{h}\big((\mathscr{L}a)(t)\big)\Big]\mathrm{E}\Big[\varphi'\big(Z_t^a\big)Y_t^a - \partial_a\mathbf{h}\big((\mathscr{L}a)(t)\big)\Big]\,\mathrm{d}t + \partial_a H(a).$$

From this expression we obtain the identity

$$\partial_a G(a) - \partial_a G(a') = 2\int_0^T \big(I_1(t) + I_2(t)\big)\,\mathrm{d}t + I_3,$$

where

$$\begin{aligned}I_1(t) &= \mathrm{E}\Big[\varphi\big(Z_t^a\big) - \varphi\big(Z_t^{a'}\big) + \mathbf{h}\big((\mathscr{L}a')(t)\big) - \mathbf{h}\big((\mathscr{L}a)(t)\big)\Big]\mathrm{E}\Big[\varphi'\big(Z_t^a\big)Y_t^a - \partial_a\mathbf{h}\big((\mathscr{L}a)(t)\big)\Big],\\ I_2(t) &= \mathrm{E}\Big[\varphi\big(Z_t^{a'}\big) - \mathbf{h}\big((\mathscr{L}a')(t)\big)\Big]\mathrm{E}\Big[\varphi'\big(Z_t^a\big)Y_t^a - \varphi'\big(Z_t^{a'}\big)Y_t^{a'} + \partial_a\,\mathbf{h}\big((\mathscr{L}a')(t)\big) - \partial_a\mathbf{h}\big((\mathscr{L}a)(t)\big)\Big],\\ I_3 &= \partial_a H(a) - \partial_a H(a').\end{aligned}$$

Now, by the boundedness of φ, φ', φ'', $\mathbf{h}$, $\mathbf{h}'$ we have

$$\begin{aligned}|I_1(t)| &\le C\,\mathrm{E}\big[|Z_t^a - Z_t^{a'}| + |a - a'|\big]\mathrm{E}\big[|Y_t^a| + 1\big],\\ |I_2(t)| &\le C\,\mathrm{E}\big[|a - a'|(|Y_t^a| + 1) + |Y_t^a - Y_t^{a'}|\big].\end{aligned}$$

Therefore, Lemma 4 and Assumption 4 yield the following result:

$$|\partial_a G(a) - \partial_a G(a')| \le C|a - a'|, \quad a \in \mathbb{R}.$$

□

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

Supplementary material

The supplementary material for this article can be found at https://github.com/amatoandre/Numerical-approximation-of-MKV-SDEs-via-SGD

Funding Information

The research of SP was partially supported by the PRIN22 project no. CUP_E53C23001670001. The author SP has also received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Actions Staff Exchanges (Grant Agreement No. 101183168, Call: HORIZON-MSCA-2023-SE-01). GdR acknowledges support from the *Fundação para a Ciência e a Tecnologia* (Portuguese Foundation for Science and Technology) through the project UIDB/00297/2020 and UIDP/00297/2020 (Centro de Matemática e Aplicações – NOVA Math) and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding Guarantee [Project UKRI313].

Competing Interests

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

References

[1] Agarwal, A. and Pagliarani, S. (2021). A Fourier-based Picard-iteration approach for a class of McKean-Vlasov SDEs with Lévy jumps. *Stochastics* **93**(4), 592–624.

[2] Antonelli, F. and Kohatsu-Higa, A. (2002). Rate of convergence of a particle method to the solution of the McKean-Vlasov equation. *Ann. Appl. Probab.* **12**(2), 423–476.

[3] Baladron, J., Fasoli, D., Faugeras, O. and Touboul, J. (2012). Mean-field description and propagation of chaos in networks of Hodgkin-Huxley and FitzHugh-Nagumo neurons. *J. Math. Neurosci.* 2, Art. **10**, 50.

[4] Beirão da Veiga, L., Buffa, A., Sangalli, G. and Vázquez, R. (2016). An introduction to the numerical analysis of isogeometric methods. In *Isogeometric Analysis: A New Paradigm in the Numerical Approximation of PDEs*, Lecture Notes in Mathematics, Vol. 2161. Springer, Cham, pp. 87–154.

[5] Belomestny, D. and Schoenmakers, J. (2018). Projected particle methods for solving McKean–Vlasov stochastic differential equations. *SIAM J. Numer. Anal.* **56**(6), 3169–3195.

[6] Bertsekas, D. P. and Tsitsiklis, J. N. (2000). Gradient convergence in gradient methods with errors. *SIAM J. Optim.* **10**(3), 627–642.

[7] Biswas, S., Kumar, C., Neelima, dos Reis, G. and Reisinger, C. (2024). An explicit Milstein-type scheme for interacting particle systems and McKean-Vlasov SDEs with common noise and non-differentiable drift coefficients. *Ann. Appl. Probab.* **34**(2), 2326–2363.

[8] Bossy, M. and Talay, D. (1997). A stochastic particle method for the McKean-Vlasov and the Burgers equation. *Math. Comp.* **66**(217), 157–192.

[9] Carmona, R. and Delarue, F. (2018). *Probabilistic Theory of Mean Field Games with Applications. I*, *Probability Theory and Stochastic Modelling*, Vol. **83**. Springer, Cham. Mean field FBSDEs, control, and games.

[10] Carmona, R. and Laurière, M. (2021). Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games I: The ergodic case. *SIAM J. Numer. Anal.* **59**(3), 1455–1485.

[11] Carmona, R. and Laurière, M. (2022). Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games: II—The finite horizon case. *Ann. Appl. Probab.* **32**(6), 4065–4105.

[12] Cattiaux, P., Guillin, A. and Malrieu, F. (2008). Probabilistic approach for granular media equations in the non-uniformly convex case. *Probab. Theory Relat. Fields* **140**(1–2), 19–40.

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

[13] CAZACU, N. (2025). Stochastic numerical approximation for nonlinear Fokker-Planck equations with singular kernels. arXiv preprint arXiv:2504.06132.
[14] CHASSAGNEUX, J.-F. AND PAGÈS, G. (2024). Computing the invariant distribution of McKean-Vlasov SDEs by ergodic simulation. arXiv preprint arXiv:2406.13370.
[15] CHAUDRU DE RAYNAL, P. AND GARCIA TRILLOS, C. (2015). A cubature based algorithm to solve decoupled McKean-Vlasov forward–backward stochastic differential equations. *Stoch. Process. Appl.* **125**(6), 2206–2255.
[16] CHAUDRU DE RAYNAL, P. E. (2020). Strong well posedness of McKean-Vlasov stochastic differential equations with Hölder drift. *Stoch. Process. Appl.* **130**(1), 79–107.
[17] CHEN, X., and DOS REIS, G. (2022). A flexible split-step scheme for solving McKean-Vlasov stochastic differential equations. *Appl. Math. Comput.* **427**, Paper No. 127180, 23.
[18] CHEN, X., and dos Reis, G. (2024). Euler simulation of interacting particle systems and McKean-Vlasov SDEs with fully super-linear growth drifts in space and interaction. *IMA J. Numer. Anal.* **44**(2), 751–796.
[19] CHEN, X., DOS REIS, G. and Stockinger, W. (2025). Wellposedness, exponential ergodicity and numerical approximation of fully super-linear McKean-Vlasov SDEs and associated particle systems. *Electron. J. Probab.* **30**, Paper No. 23, 50.
[20] CHEN, X., DOS REIS, G., Stockinger, W. and Wilde, Z. (2025). Improved weak convergence for the long time simulation of mean-field Langevin equations. *Electron. J. Probab.* **30**, Paper No. 89, 81.
[21] dos Reis, G., Smith, G. and Tankov, P. (2023). Importance sampling for McKean-Vlasov SDEs. *Appl. Math. Comput.* **453**, Paper No. 128078, 31.
[22] FANG, F. AND OOSTERLEE, C. W. (2008/09). A novel pricing method for European options based on Fourier-cosine series expansions. *SIAM J. Sci. Comput.* **31**(2), 826–848.
[23] FEHRMAN, B., GESS, B. AND JENTZEN, A. (2020). Convergence rates for the stochastic gradient descent method for non-convex objective functions. *J. Mach. Learn. Res.* **21**, 136.
[24] FRANK, T. D. (2005). *Nonlinear Fokker-Planck Equations: Fundamentals and Applications*. Springer Science & Business Media, Berlin Heidelberg.
[25] GOBET, E. AND PAGLIARANI, S. (2018). Analytical approximations of non-linear SDEs of McKean-Vlasov type. *J. Math. Anal. Appl.* **466**(1), 71–106.
[26] GODDARD, B. D., GOODING, B., SHORT, H. AND PAVLIOTIS, G. A. (2022). Noisy bounded confidence models for opinion dynamics: The effect of boundary conditions on phase transitions. *IMA J. Appl. Math.* **87**(1), 80–110.
[27] GOLSE, F., MOUHOT, C. AND PAUL, T. (2016). On the mean field and classical limits of quantum mechanics. *Comm. Math. Phys.* **343**(1), 165–205.
[28] HAJI-ALI, A.-L. AND TEMPONE, R. (2018). Multilevel and multi-index Monte Carlo methods for the McKean–Vlasov equation. *Stat. Comput.* **28**(4), 923–935.
[29] HAN, J., JENTZEN, A. AND E, W. (2018). Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci.* **115**(34), 8505–8510.
[30] HUTZENTHALER, M., JENTZEN, A., KRUSE, T., ANH NGUYEN, T. AND VON WURSTEMBERGER, P. (2020). Overcoming the curse of dimensionality in the numerical approximation of semilinear parabolic partial differential equations. *Proc. R. Soc. A* **476**(2244), 20190630.
[31] HUTZENTHALER, M., JENTZEN, A., KRUSE, T. AND NGUYEN, T. A. (2020). A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations. *SN Partial Differ. Equations Appl.* **1**, 1–34.
[32] HUTZENTHALER, M., KRUSE, T. AND NGUYEN, T. A. (2022). Multilevel Picard approximations for McKean-Vlasov stochastic differential equations. *J. Math. Anal. Appl.* **507**(1), Paper No. 125761, 14.
[33] IOANNOU, E., KLUS, S. AND DOS REIS, G. (2025). Data-driven approximation of transfer operators for mean-field stochastic differential equations. arXiv preprint arXiv:2509.09891.
[34] ISAACSON, E., and KELLER, H. B. (2012). *Analysis of Numerical Methods*. Courier Corporation.
[35] KRYLOV, N. V. (2008). *Controlled Diffusion Processes*, Vol. **14**. Springer Science & Business Media, New York.
[36] KUNITA, H. (1984). Stochastic differential equations and stochastic flows of diffeomorphisms. In *Ecole d'été de probabilités de Saint-Flour XII-1982*. Springer, Berlin Heidelberg, pp. 143–303.
[37] MAHATO, N. K., KLAR, A. AND TIWARI, S. (2018). Particle methods for multi-group pedestrian flow. *Appl. Math. Model.* **53**, 447–461.
[38] MALRIEU, F. (2001). Logarithmic Sobolev inequalities for some nonlinear PDE's. *Stoch. Process. Appl.* **95**(1), 109–132.
[39] MAURY, B. AND FAURE, S. (2019). *Crowds in Equations*, Advanced Textbooks in Mathematics. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ. An introduction to the microscopic modeling of crowds, With a foreword by Laure Saint-Raymond.

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press

[40] MÉLÉARD, S. (1996). Asymptotic behaviour of some interacting particle systems; McKean-Vlasov and Boltzmann models. In *Probabilistic Models for Nonlinear Partial Differential Equations*. Springer, Berlin Heidelberg, pp. 42–95.
[41] PATEL, V., and ZHANG, S. (2021). Stochastic gradient descent on nonconvex functions with general noise models. arXiv preprint arXiv:2104.00423.
[42] SONI, S., NEELIMA, KUMAR, C. and DOS REIS, G. (2025). Tamed Euler approximation for fully super-linear growth McKean-Vlasov SDE and their particle systems: Sharp rates for strong propagation of chaos, convergence and ergodicity. arXiv preprint arXiv:2510.16427.
[43] SÜLI, E., and MAYERS, D. F. (2003). *An Introduction to Numerical Analysis*. Cambridge University Press, Cambridge.
[44] SUN, Y., YANG, J. AND ZHAO, W. (2017). Itô-Taylor schemes for solving mean-field stochastic differential equations. *Numer. Math. Theory Methods Appl.* **10**(4), 798–828.
[45] SUZUKI, T. (2005). *Free Energy and Self-Interacting Particles*, *Progress in Nonlinear Differential Equations and their Applications*, Vol. **62**. Birkhäuser Boston, Inc., Boston, MA.
[46] SZNITMAN, A.-S. (1991). Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX—1989*, Lecture Notes in Mathematics, Vol. 1464. Springer, Berlin, pp. 165–251.
[47] SZPRUCH, L., TAN, S. AND TSE, A. (2019). Iterative particle approximation for McKean–Vlasov SDEs with application to multilevel Monte Carlo estimation. *Ann. Appl. Probab.* **29**(4), 2230–2265.
[48] TALAY, D. AND VAILLANT, O. (2003). A stochastic particle method with random weights for the computation of statistical solutions of McKean-Vlasov equations. *Ann. Appl. Probab.* **13**(1), 140–180.
[49] TIMAN, A. F. (1994). *Theory of Approximation of Functions of a Real Variable*. Dover Publications, Inc., New York. Translated from the Russian by J. Berry, Translation edited and with a preface by J. Cossar, Reprint of the 1963 English translation.
[50] YUANPING, C., XIAOYUE, L. AND YI, L. (2024). Explicit numerical approximations for McKean-Vlasov stochastic differential equations in finite and infinite time. arXiv preprint arXiv:2401.02878.

https://doi.org/10.1017/apr.2026.10057 Published online by Cambridge University Press