

# NON-PARAMETRIC LEARNING OF STOCHASTIC DIFFERENTIAL EQUATIONS WITH NON-ASYMPTOTIC FAST RATES OF CONVERGENCE\*

RICCARDO BONALLI<sup>†</sup> AND ALESSANDRO RUDI<sup>‡</sup>

**Abstract.** We propose a novel non-parametric learning paradigm for the identification of drift and diffusion coefficients of multi-dimensional non-linear stochastic differential equations, which relies upon discrete-time observations of the state. The key idea essentially consists of fitting a RKHS-based approximation of the corresponding Fokker-Planck equation to such observations, yielding theoretical estimates of non-asymptotic learning rates which, unlike previous works, become increasingly tighter when the regularity of the unknown drift and diffusion coefficients becomes higher. Our method being kernel-based, offline pre-processing may be profitably leveraged to enable efficient numerical implementation, offering excellent balance between precision and computational complexity.

**Key words.** Non-parametric system identification with guarantees, non-linear stochastic differential equations, discrete-time observation, Reproducing Kernel Hilbert Space.

**1. Introduction.** Consider non-linear *stochastic differential equations* of type

$$(1.1) \quad dX(t) = b(t, X(t)) dt + \sqrt{a(t, X(t))} dW_t, \quad t \in [0, T].$$

Here,  $X(t) \in \mathbb{R}^n$  denotes the (stochastic) state of dimension  $n \in \mathbb{N}$  of the system at time  $t \in [0, T]$ , with  $T > 0$  a fixed time horizon,  $(W_t)_{t \in [0, T]}$  is a  $n$ -dimensional Wiener process, whereas  $b : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  and  $a : [0, T] \times \mathbb{R}^n \rightarrow \text{Sym}_{++}(n)$  are regular enough *drift* and *diffusion* coefficients, with  $\text{Sym}_{++}(n) \subseteq \mathbb{R}^{n \times n}$  the subset of symmetric positive-definite matrices. Equations such as (1.1) may be profitably leveraged to accurately model complex phenomena in a wide range of applications, such as aerospace, finance, and robotics to name a few [38, 50, 7, 6, 47].

In practice, both coefficients  $a$  and  $b$  might be completely unknown, e.g.,  $a$  might model external perturbations due to many different physical phenomena which affect the motion of the system, as it occurs in aerospace and robotics. Therefore, appropriate *stochastic system identification* procedures for both  $a$  and  $b$  must be devised.

**1.1. Related Work.** The problem of stochastic system identification has been investigated for several decades. The earliest methods mainly address either discrete-time models [23, 33] or continuous-time models that are however linear in the state variable [18], which thus do not fit the identification setting introduced by (1.1).

More recently, identification of non-linear stochastic differential equations such as (1.1) have seen an important surge of interest. More specifically, besides some likelihood-based [27] or Kalman filtering-based estimation methods [44, 5, 41], which assume continuous-time observation of the stochastic state, the vast majority of the existing methods leverage more realistic discrete-time observations of  $X$ . Efficient paradigms range from 1. non-parametric estimation [12, 19, 11, 21] and Bayesian estimation [46, 36], to 2. maximum likelihood and quasi-likelihood methods [15, 51, 30, 45], generalized methods of moments [17, 16, 37], and online gradient descent-based methods [20, 32], among others. On the one hand, although Group 1. of these works

\*Submitted for review on April 24, 2024.

<sup>†</sup>Laboratoire des Signaux et Systèmes, Université Paris-Saclay, CNRS, CentraleSupélec, Bât. Bréguet, 3 Rue Joliot Curie, 91190 Gif-sur-Yvette, France. *E-mail:* riccardo.bonalli@cnrs.fr.

<sup>‡</sup>Inria, École Normale Supérieure, CNRS, PSL Research University, Paris, France. *E-mail:* alessandro.rudi@inria.fr.

provides learning rates which become increasingly tighter when the number of observations of the state grows, estimates of the error between the unknown drift and diffusion coefficients and the learned ones are obtained by assuming diffusion coefficients are either known or have specific parametric structures. On the other hand, although Group 2. of the aforementioned works rely on generally milder assumptions to be implemented, they require working with given families of finite-dimensional parametric drift and diffusion coefficients, which might hinder the identification process in the case the family of parametric coefficients is not rich enough.

Importantly, none of the above works has investigated how to leverage the regularity, i.e., high order continuous differentiability of drift and diffusion coefficients to improve learning rates. Although such analysis has been recently undertaken in, e.g., [1, 35, 2, 28], these studies often work with SDE that are either scalar or perturbed by constant diffusion coefficients, and they offer learning rates that are often only asymptotic. Therefore, designing methods for the identification of multi-dimensional non-linear SDE with non-constant diffusion coefficients under discrete-time observations of the state remains a challenging open question (note that multivariate settings are known to generally require very different techniques, see, e.g., [35, 2]). In particular, a key part of the challenge lies in developing non-asymptotic rates of convergence that become tighter not only when the number of observations of the state grows, but also when the regularity of drift and diffusion coefficients is higher.

Finally, from a numerical standpoint, learning-based methods, such as scalable gradient methods [24] and infinitely deep Bayesian neural networks [49], have shown remarkable performance on complex stochastic differential equations. Unfortunately, it seems extremely challenging to endow these latter paradigms with theoretical guarantees of convergence, motivating investigation of numerically efficient identification methods for (1.1) which enjoy guarantees of accuracy under mild assumptions.

**1.2. Outline and Contributions.** We propose a non-parametric, *Reproducing Kernel Hilbert Space* (RKHS)-based learning paradigm for the identification of drift and diffusion coefficients of the multi-dimensional non-linear SDE (1.1), which relies upon discrete-time observation of the state. In particular, motivated by classical likelihood-based methods [37] we propose a two-step, discrete-time observation-based scheme which entails fitting the *Fokker-Planck equation* related to (1.1).

Under assumptions of smoothness for the unknown drift and diffusion coefficients, we provide theoretical estimates of non-asymptotic learning rates which become increasingly tighter when the number of observations of the state grows. In particular, given the nature of our data set, which essentially depends on observations of the state process, we consider the error between solutions to (1.1) generated with the learned drift and diffusion coefficients and the unknown trajectories of (1.1) is a “good metric” to test the accuracy of our identification method. Importantly, under this accuracy metric we can additionally prove our learning rates become tighter when the regularity (in a Sobolev sense) of the unknown drift and diffusion coefficients is higher. Finally, from a numerical standpoint, our method being kernel-based, offline pre-processing may be successfully leveraged to enable efficient implementations, offering excellent balance between precision and computational complexity (details are in Section 6).

Our method is composed of two steps which are informally summarized below.

#### A. Learning the laws of the stochastic differential equation through independent discrete-time observation of the state.

We assume there exist regular enough, i.e. essentially  $C^{2m+1}$ ,  $m \in \mathbb{N}$ , drift

$b : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  and diffusion  $a : [0, T] \times \mathbb{R}^n \rightarrow \text{Sym}_{++}(n)$ , and a stochastic process  $X : [0, T] \times \Omega \rightarrow \mathbb{R}^n$  in some filtered probability space  $(\Omega, \mathcal{G}, \mathcal{F}, \mathbb{P})$ , such that  $X$  solves (1.1) with drift  $b$  and diffusion  $a$ . In addition, we may assume we can sample  $X$  at  $M \in \mathbb{N}$  times  $0 = t_1 < \dots < t_M = T$ , and more specifically that at every time  $t_\ell$ ,  $\ell = 1, \dots, M$ , we have access to  $N \in \mathbb{N}$  samples  $X_{\ell,1} \triangleq X(t_\ell, \omega_1), \dots, X_{\ell,N} \triangleq X(t_\ell, \omega_N)$  of the solution  $X$  which have been independently drawn from  $\mathbb{P}_{X(t_\ell, \cdot)}$ .

In such setting, as first step we propose to approximate the unknown densities  $p : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$  of the laws of  $X$  through the (random) RKHS-based model:

$$\hat{p}(t, x) \triangleq \sum_{\ell=1}^M c_\ell(t) \hat{g}_\ell(x), \quad \text{where} \quad \hat{g}_\ell(x) \triangleq \frac{1}{N} \sum_{j=1}^N \rho_R(x - X_{\ell,j}),$$

for appropriate coefficients  $c_\ell : [0, T] \rightarrow \mathbb{R}$  and radial mappings  $\rho_R$ ,  $R > 0$ . Let  $\mu_X$  denote the probability measure which is generated by the process  $X$ . For every tuple of precision parameters  $0 < \varepsilon, \delta < 1$ , if the unknown drift  $b$  and diffusion  $a$  are regular enough, by leveraging RKHS approximation theory we can show appropriate values for  $M$ ,  $N$ , and  $R$  (which depend on  $\varepsilon$  and  $m$  uniquely) may be selected so that the following learning rate holds with probability  $\mu_X$  at least  $1 - \delta$ :

$$(1.2) \quad \int_0^T \left\| \frac{\partial \hat{p}}{\partial t}(t, \cdot) - \frac{\partial p}{\partial t}(t, \cdot) \right\|_{L^2}^2 + \|\hat{p}(t, \cdot) - p(t, \cdot)\|_{H^2}^2 dt = O \left( \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^2 \right),$$

where  $\|\cdot\|_{L^2}$  and  $\|\cdot\|_{H^2}$  denote the norms of the Hilbert spaces  $L^2(\mathbb{R}^n, \mathbb{R})$  and  $H^2(\mathbb{R}^n, \mathbb{R})$ , respectively. Moreover, the higher the degree of smoothness of  $b$  and  $a$  is, the lower the values of  $M$ ,  $N$ , and  $R$  needed to achieve this precision become.

The main benefit which comes with the model  $\hat{p}$  consists of computing accurate finite-dimensional approximations of the laws of solutions to (1.1) without a priori resorting to conservative families of parametric densities. In particular, the model  $\hat{p}$  being kernel-based, one may considerably reduce the computational effort by resorting to prior offline computations, which essentially boil down to simply inverting a  $M \times M$  matrix. As a byproduct, (1.2) provides a quantitative estimate of the approximation error which is key to derive theoretical guarantees for the accuracy of our identification method in the next step.

## B. Learning finite-dimensional models for the drift and diffusion coefficients by fitting approximated solutions to the Fokker-Planck equation.

If the unknown drift  $b$  and diffusion  $a$  are regular enough, the unknown densities  $p$  satisfy the following Fokker-Planck equation:

$$(1.3) \quad \frac{\partial p}{\partial t}(t, y) = (\mathcal{L}_t^{a,b})^* p(t, y), \quad (t, y) \in [0, T] \times \mathbb{R}^n,$$

where  $(\mathcal{L}_t^{a,b})^*$  is the dual operator of the Kolmogorov generator

$$\mathcal{L}_t^{a,b} \varphi(y) \triangleq \frac{1}{2} \sum_{i,j=1}^n a_{ij}(t, y) \frac{\partial^2 \varphi}{\partial y_i \partial y_j}(y) + \sum_{i=1}^n b_i(t, y) \frac{\partial \varphi}{\partial y_i}(y), \quad \varphi \in C^2(\mathbb{R}^n, \mathbb{R}).$$

Given the results at the previous step, it is then natural to learn models of the drift  $\hat{b}_Q : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  and the diffusion  $\hat{a}_Q : [0, T] \times \mathbb{R}^n \rightarrow \text{Sym}_{++}(n)$  which “best”

match the Fokker-Planck equation when evaluated at  $\widehat{p}$ , that is as solutions to the following finite-dimensional convex optimization problem:

$$(1.4) \quad \min_{(\widehat{a}_Q, \widehat{b}_Q) \in \mathcal{H}_Q} \int_0^T \left\| \frac{\partial \widehat{p}}{\partial t}(t, \cdot) - (\mathcal{L}_t^{\widehat{a}_Q, \widehat{b}_Q})^* \widehat{p}(t, \cdot) \right\|_{L^2}^2 dt + \lambda \|(\widehat{a}_Q, \widehat{b}_Q)\|_{\mathcal{H}}^2,$$

where  $\mathcal{H}_Q$  is an appropriate finite-dimensional subspace of a RKHS  $\mathcal{H}$  with norm  $\|\cdot\|_{\mathcal{H}}$ , in which the unknown drift and diffusion coefficients lie, whereas  $\lambda > 0$  is a regularization weight to be appropriately selected. As detailed in Section 6, problem (1.4) admits a finite dimensional characterization that can be solved exactly, i.e., no approximation or integral discretization is needed, with reduced computational cost.

Thanks to estimate (1.2), one shows that, for appropriate choices of the subspace  $\mathcal{H}_Q$  and the regularization weight  $\lambda$ , the solution  $(\widehat{a}_Q, \widehat{b}_Q) \in \mathcal{H}_Q$  to problem (1.4) satisfies the following, with probability  $\mu_X$  at least  $1 - \delta$ :

$$(1.5) \quad \int_0^T \left\| \frac{\partial \widehat{p}}{\partial t}(t, \cdot) - (\mathcal{L}_t^{\widehat{a}_Q, \widehat{b}_Q})^* \widehat{p}(t, \cdot) \right\|_{L^2}^2 dt + \lambda \|(\widehat{a}_Q, \widehat{b}_Q)\|_{\mathcal{H}}^2 = O \left( \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^2 \right).$$

By combining estimate (1.5) with energy-type estimates for parabolic partial differential equations, we may infer theoretical estimates of learning rates for the identification of drift and diffusion coefficients of non-linear stochastic differential equations, which we informally summarize as follows.

As we mentioned previously, since our data consists of observation of the state process, the error between the unknown densities and the densities stemming from the learned coefficients is a “good metric” with which the convergence of an identification algorithm for stochastic differential equations which leverage observations of the state process may be tested. To better formalize this metric, let  $X^{\widehat{a}_Q, \widehat{b}_Q}$  and  $p_{\widehat{a}_Q, \widehat{b}_Q}$  respectively denote the solutions to (1.1) and (1.3) with coefficients  $(\widehat{a}_Q, \widehat{b}_Q) \in \mathcal{H}_H$ . We thus define the following metric to test the accuracy:

$$E(\widehat{a}_Q, \widehat{b}_Q) \triangleq \|p_{\widehat{a}_Q, \widehat{b}_Q} - p\|_{L^2}^2 = \int_0^T \|p_{\widehat{a}_Q, \widehat{b}_Q}(t, \cdot) - p(t, \cdot)\|_{L^2}^2 dt.$$

By adopting this metric, our main result on the accuracy of our learning method may be summarized in the following meta-theorem:

**META-THEOREM 1.1.** *Assume the unknown drift  $b$  and diffusion  $a$  coefficients are regular enough. The following estimate holds true with probability  $\mu_X$  at least  $1 - \delta$ :*

$$E(\widehat{a}_Q, \widehat{b}_Q) = O \left( \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^2 \right)$$

*In particular, for every regular enough function  $f : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$ :*

$$(1.6) \quad \mathbb{E} \left[ \int_0^T f(t, X(t)) dt \right] = \mathbb{E} \left[ \int_0^T f(t, X^{\widehat{a}_Q, \widehat{b}_Q}(t)) dt \right] + O \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right).$$

Although the estimates provided in this meta-theorem are informal and need some clarification (see Section 5 for more rigorous statements), they show the aforementioned two-step identification method enjoys practical theoretical guarantees of accuracy, i.e., estimate (1.6): an observation/regulation metric  $f$  computed at the unknown state solution to (1.1) with unknown drift  $b$  and diffusion  $a$  may be rather observed through the process solution to (1.1) with model drift  $\hat{b}_Q$  and diffusion  $\hat{a}_Q$  up to an error  $O(\log(\frac{1}{\delta\varepsilon})\varepsilon)$ . Such result has important implications in observation and regulation of stochastic differential equations [21], and it may represent a good starting result to develop paradigms for the identification of *controlled stochastic differential equations*, which are crucial for the control of complex autonomous systems.

**1.3. Paper Organization.** The paper is organized as follows. After gathering basic notation and preliminary results in Section 2, in Section 3 we summarize both classical and less classical results about stochastic differential equations and corresponding relationships with the Fokker-Planck equation. To ease the reading of this section, we moved a more detailed description of the aforementioned results to Section 8 and their technical proofs to Appendix A. Similarly, in Section 4 we expose the results in RKHS theory which we leverage in this work. Our main contributions are contained in Section 5, which in particular details the methodologies we exposed at the previous steps 1) and 2) and corresponding learning rates. Precision and computational complexity of our approach are discussed in Section 6. Finally, in Section 7 we provide concluding remarks and some perspectives.

**2. Notation and Preliminary Results.** We fix the dimension  $n \in \mathbb{N}$  of the state space and a time horizon  $T > 0$  for the identification process. We denote by  $\text{Sym}_+(n)$  and  $\text{Sym}_{++}(n)$  the subsets of  $\mathbb{R}^{n \times n}$  of symmetric semi-positive-definite and symmetric positive-definite matrices, respectively.

We assume stochastic differential equations are defined on a given filtered probability space  $(\Omega, \mathcal{G} \triangleq \mathcal{F}_T, \mathcal{F} \triangleq (\mathcal{F}_t)_{t \in [0, T]}, \mathbb{P})$ , which is complete, and the noise is generated by a  $\mathcal{F}$ -adapted Wiener process  $W : [0, T] \times \Omega \rightarrow \mathbb{R}^n$  (e.g., we may consider the canonical process in the space  $\Omega = C([0, T], \mathbb{R}^n)$ , equipped with the Wiener measure). Moreover, we introduce the complete metric space  $(S, \mathfrak{d})$ , where

$$S \triangleq C([0, T], \mathbb{R}^n), \quad \mathfrak{d}(w_1, w_2) \triangleq \sup_{t \in [0, T]} \|w_1(t) - w_2(t)\|,$$

and equip it with the Borel sigma-algebra  $\mathcal{B}(S)$  induced by the metric  $\mathfrak{d}$ .

For any  $\ell \in \mathbb{N}$ ,  $r, s > 0$ , and any  $A \subseteq \mathbb{R}^\ell$ , we denote by  $H^{r,s}([0, T] \times \mathbb{R}^n, A)$  the Sobolev (Hilbert) space of functions whose image is in  $A$ , and whose weak time derivatives are defined up to order  $r$  and whose weak space derivatives are defined up to order  $s$ ; in particular, we denote  $H^r([0, T] \times \mathbb{R}^n, A) \triangleq H^{r,r}([0, T] \times \mathbb{R}^n, A)$ . Finally, we introduce the following Hilbert space and corresponding positive convex cone, in which we will assume the unknown drift and diffusion coefficients lie (this latter requirement will be made more explicit shortly):

$$\begin{aligned} \mathcal{H}_m &\triangleq H^{d(m)}([0, T] \times \mathbb{R}^n, \mathbb{R}^{n \times n}) \times H^{d(m)}([0, T] \times \mathbb{R}^n, \mathbb{R}^n), \\ \mathcal{H}_m^+ &\triangleq \left\{ (a, b) : [0, T] \times \mathbb{R}^n \rightarrow \text{Sym}_+(n) \times \mathbb{R}^n : (a, b) \in \mathcal{H}_m \right\}, \end{aligned}$$

for every  $m \in \mathbb{N}$ , where  $d(m) \triangleq 2(m+1) + \lfloor \frac{n}{2} \rfloor \in \mathbb{N}$  is the unique integer greater than  $2m+1$  such that  $\lfloor d(m) - \frac{n}{2} \rfloor = 2m+1$ . The choice of taking the same exponent  $d(m)$  for both the drift and the diffusion coefficients has been made without loss of generality for the sake of conciseness. The following regularity result will be crucial:

**THEOREM 2.1.** *There exists a constant  $C > 0$ , which depends on  $m$  uniquely, such that every  $(a, b) \in \mathcal{H}_m^+$  satisfies the following properties:*

1. *Differentiability:*

$$a \in C^{2m+1}([0, T] \times \mathbb{R}^n, \mathbb{R}^{n \times n}), \quad b \in C^{2m+1}([0, T] \times \mathbb{R}^n, \mathbb{R}^n).$$

2. *Boundness of functions and their derivatives:*

$$\sum_{i=0}^{2m+1} \left( \|D_{(t,y)}^i a\|_{L^\infty} + \|D_{(t,y)}^i b\|_{L^\infty} \right) \leq C \|(a, b)\|_{\mathcal{H}_m},$$

where  $D_{(t,y)}^i$  denotes the differential of order  $i$  with respect to  $(t, y)$ . In particular, the following refined bound holds:

$$\sum_{i=0}^2 \left( \|D_{(t,y)}^i a\|_{L^\infty} + \|D_{(t,y)}^i b\|_{L^\infty} \right) \leq C \|(a, b)\|_{H^{d(m)-2m+1}}.$$

The proof of this result makes use of classical embedding arguments and it is reported in Appendix A for the sake of completeness.

**3. Stochastic Differential and Fokker-Planck Equations.** In this section, we summarize both classical and less classical results about stochastic differential equations, and corresponding relationships with the Fokker-Planck equation. In particular, we propose a minimally detailed discussion for the sake of conciseness, reporting a more structured exposition in Section 8 for the sake of completeness.

**3.1. The Fokker-Planck equation.** From now on, we fix  $m \in \mathbb{N}$  (to be specified later) and a constant  $\alpha > 0$ . Since we will need to work with diffusion coefficients which are never trivial, for every  $(a, b) \in \mathcal{H}_m^+$  we will rather replace  $a$  with the mapping

$$a + \alpha I : [0, T] \times \mathbb{R}^n \rightarrow \text{Sym}_{++}(n) \quad \text{with} \quad a + \alpha I = \sqrt{a + \alpha I} \sqrt{a + \alpha I}.$$

At this step, we fix a non-negative density  $p_0 \in L^2(\mathbb{R}^n, \mathbb{R})$  which will serve as appropriate initial condition, and we denote by  $\mu_0 \in \mathcal{P}(\mathbb{R}^n)$  the associate probability measure. Motivated by [43, 14], we recall the following notions of stochastic differential equation and its solutions (though they differ from the ones in [43, 14]):

**THEOREM-DEFINITION 3.1.** *There exists a measurable mapping  $X : \mathbb{R}^n \times \Omega \rightarrow S$  such that each process  $X_x(t, \omega) \triangleq X(x, \omega)(t)$  is  $\mathcal{F}$ -progressively measurable for every  $x \in \mathbb{R}^n$ , and such that the following Stochastic Differential Equation (SDE)*

$$\text{SDE}_x \quad \begin{cases} dX_x(t) = b(t, X_x(t)) dt + \sqrt{(a + \alpha I)(t, X_x(t))} dW_t, \\ \mathbb{P}(X_x(0) = x) = 1, \end{cases}$$

holds in  $(\Omega, \mathcal{G}, \mathcal{F}, \mathbb{P})$  for  $\mu_0$ -almost every  $x \in \mathbb{R}^n$ . We then say that  $X$  solves or is solution to SDE (with coefficients  $(a, b) \in \mathcal{H}_m^+$ ). The solution  $X$  to SDE is unique in the following sense: if a measurable mapping  $Y : \mathbb{R}^n \times \Omega \rightarrow S$  solves SDE, then it holds that  $X(x, \cdot) = Y(x, \cdot)$  a.s., for  $\mu_0$ -almost every  $x \in \mathbb{R}^n$ .

For the well-posedness of Theorem-Definition 3.1, see Section 8. Solutions to SDE share a close relationship with the solutions to the Fokker-Planck equation, which we introduce next. For every  $(a, b) \in \mathcal{H}_m^+$ , we denote the Kolmogorov generator by

$$\mathcal{L}_t^{a,b} \varphi(y) \triangleq \frac{1}{2} \sum_{i,j=1}^n (a + \alpha I)_{ij}(t, y) \frac{\partial^2 \varphi}{\partial y_i \partial y_j}(y) + \sum_{i=1}^n b_i(t, y) \frac{\partial \varphi}{\partial y_i}(y), \quad \varphi \in C^2(\mathbb{R}^n, \mathbb{R}).$$

Let  $X : \mathbb{R}^n \times \Omega \rightarrow S$  be solution of SDE with coefficients  $(a, b) \in \mathcal{H}_m^+$ . Thanks to specific regularity properties of  $X$  (we refer to Section 8 for these latter), we may define the curve  $\mu : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  of probability measures

$$(3.1) \quad \mu_t(A) \triangleq \int_{\mathbb{R}^n} \int_{\Omega} \mathbf{1}_{\{X_x(t) \in A\}} d\mathbb{P} \mu_0(dx), \quad A \in \mathcal{B}(\mathbb{R}^n),$$

and note that  $\mu$  is narrowly continuous, i.e., for every  $\varphi \in C_b(\mathbb{R}^n, \mathbb{R})$ , the mapping

$$t \in [0, T] \mapsto \int_{\mathbb{R}^n} \varphi(y) \mu_t(dy) = \int_{\mathbb{R}^n} \int_{\Omega} \varphi(X_x(t)) d\mathbb{P} \mu_0(dx) \in \mathbb{R}$$

is continuous. By combining this latter property with the results in [14], we introduce the following notions of Fokker-Planck equation, its solutions, and additional relationship between these solutions and the solutions to SDE (see also Section 8):

**THEOREM-DEFINITION 3.2.** *The curve  $\mu : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  defined through (3.1) is the unique narrowly continuous curve satisfying the Fokker-Planck Equation*

$$\text{FPE} \quad \begin{cases} \frac{d}{dt} \int_{\mathbb{R}^n} \varphi(y) \mu_t(dy) = \int_{\mathbb{R}^n} \mathcal{L}_t^{a,b} \varphi(y) \mu_t(dy), & \varphi \in C_c^\infty(\mathbb{R}^n, \mathbb{R}), \\ \mu_{t=0} = \mu_0. \end{cases}$$

We then say that  $\mu$  solves or is solution to FPE (with coefficients  $(a, b) \in \mathcal{H}_m^+$ ). If  $X$  denotes the solution to SDE, the following representation formula holds:

$$(3.2) \quad \int_{\mathbb{R}^n} \varphi(y) \mu_t(dy) = \int_{\mathbb{R}^n} \int_{\Omega} \varphi(X_x(t)) d\mathbb{P} \mu_0(dx), \quad \text{for } t \in [0, T], \varphi \in C_c(\mathbb{R}^n, \mathbb{R}).$$

### 3.2. Absolutely continuous solutions to the Fokker-Planck equation.

Thanks to the regularity of the coefficients  $(a, b) \in \mathcal{H}_m^+$  which is offered through Theorem 2.1, the solution  $\mu$  to FPE is absolutely continuous, i.e., it takes the form

$$\mu_t(A) = \int_A p(t, y) dy, \quad A \in \mathcal{B}(\mathbb{R}^n),$$

for an appropriate  $p : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$ . To elucidate this property, we first introduce broader definitions of FPE and corresponding solutions, which encompass Theorem-Definition 3.2 as a sub-case (see Theorem 3.4 below) and will be useful in our analysis:

**DEFINITION 3.3.** *Let  $f \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$  and  $\bar{p} \in L^2(\mathbb{R}^n, \mathbb{R})$ . A (regular enough) function  $p : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$  is said to solve or be solution to the non-homogeneous Fokker-Planck Equation (with coefficients  $(a, b) \in \mathcal{H}_m^+$ ), if*

$$\text{FPE}_f \quad \begin{cases} \frac{d}{dt} \int_{\mathbb{R}^n} \varphi(y) p(t, y) dy = \\ \quad = \int_{\mathbb{R}^n} \left( \mathcal{L}_t^{a,b} \varphi(y) p(t, y) + f(t, y) \varphi(y) \right) dy, & \varphi \in C_c^\infty(\mathbb{R}^n, \mathbb{R}), \\ p(0, \cdot) = \bar{p}(\cdot). \end{cases}$$

The next theorem gathers important properties of the solution to  $\text{FPE}_f$ , additionally showing the solution to FPE is absolutely continuous. We refer the reader to Section 8 for exhaustive presentation and proof of these results.



**THEOREM 3.4.** *For every  $(a, b) \in \mathcal{H}_m^+$ ,  $f \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$ , and  $\bar{p} \in L^2(\mathbb{R}^n, \mathbb{R})$ , there exists a unique mapping  $p \in C(0, T; L^2(\mathbb{R}^n, \mathbb{R})) \cap H^{0,1}([0, T] \times \mathbb{R}^n, \mathbb{R})$  which solves  $FPE_f$  with coefficients  $(a, b) \in \mathcal{H}_m^+$ . In addition, the following estimate holds:*

$$(3.3) \quad \begin{aligned} & \|p(t, \cdot)\|_{L^2}^2 + \int_0^t \|p(s, \cdot)\|_{H^1}^2 \, ds \leq \\ & \leq C(\|(a, b)\|_{\mathcal{H}_m}) \left( \|\bar{p}\|_{L^2}^2 + \int_0^t \|f(s, \cdot)\|_{L^2}^2 \, ds \right), \quad t \in [0, T], \end{aligned}$$

where  $C(\|(a, b)\|_{\mathcal{H}_m}) > 0$  is a constant which continuously depends on  $\|(a, b)\|_{\mathcal{H}_m^+}$ . Assume  $f = 0$  and  $\bar{p} = p_0$ . Then, it holds that

$$p(t, \cdot) \geq 0, \quad \int_{\mathbb{R}^n} p(t, y) \, dy = 1, \quad t \in [0, T],$$

and therefore, if for every  $t \in [0, T]$  we define

$$\mu_t(A) \triangleq \int_A p(t, y) \, dy, \quad A \in \mathcal{B}(\mathbb{R}^n),$$

then the curve  $\mu : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  is narrowly continuous and solves FPE. Finally, if in addition  $p_0 \in H^{2m+1}(\mathbb{R}^n, \mathbb{R})$ , then the function  $p$  satisfies

$$p \in H^{m+1, 2(m+1)}([0, T] \times \mathbb{R}^n, \mathbb{R})$$

and the following Strong Fokker-Planck Equations (with coefficients  $(a, b) \in \mathcal{H}_m^+$ ):

$$\text{SFPE} \quad \begin{cases} \frac{\partial p}{\partial t}(t, y) = (\mathcal{L}_t^{a,b})^* p(t, y), & \text{a.e. } (t, y) \in [0, T] \times \mathbb{R}^n, \\ p(0, \cdot) = p_0(\cdot), \end{cases}$$

where  $(\mathcal{L}_t^{a,b})^*$  denotes the dual operator of the Kolmogorov generator  $\mathcal{L}_t^{a,b}$ .

Combining Theorem 3.4 with the representation formula (3.2) allows us to introduce criteria to establish satisfactory guarantees for our learning approach in the context of observation and regulation of stochastic differential equations (a more exhaustive presentation is provided in Section 8). More specifically, let  $f \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$  be an observation/regulation integral metric and  $(a, b) \in \mathcal{H}_m^+$ . By denoting  $X^{a,b}$  and  $p_{a,b}$  respectively the solution to SDE and to SFPE with coefficients  $(a, b) \in \mathcal{H}_m^+$ , we are interested in studying the accuracy with which the following observation/regulation metric is approximated by our learning approach:

$$(3.4) \quad \mathbb{E}_{\mu_0 \times \mathbb{P}} \left[ \int_0^T f(t, X_x^{a,b}(t)) \, dt \right] = \int_0^T \int_{\mathbb{R}^n} f(t, y) p_{a,b}(t, y) \, dy \, dt.$$

To give precise estimates of the approximation error for (3.4), the following corollary of Theorem 3.4 will be crucial (see Section 8 for a proof):

**COROLLARY 3.5.** *For every  $(a_1, b_1), (a_2, b_2) \in \mathcal{H}_m^+$ , it holds that*

$$(3.5) \quad \begin{aligned} & \left| \mathbb{E}_{\mu_0 \times \mathbb{P}} \left[ \int_0^T f(t, X_x^{a_1, b_1}(t)) \, dt \right] - \mathbb{E}_{\mu_0 \times \mathbb{P}} \left[ \int_0^T f(t, X_x^{a_2, b_2}(t)) \, dt \right] \right| \\ & \leq \|f\|_{L^2} \|p_{a_1, b_1} - p_{a_2, b_2}\|_{L^2}. \end{aligned}$$



**4. Useful Results from RKHS Theory.** In this section, we list classical results about Sobolev Spaces of functions with scattered zeros and Reproducing Kernel Hilbert Spaces (RKHS), which will be extensively leveraged in the following sections and for which we mainly refer to [3, 48, 42].

Let  $r, s \in \mathbb{N}$  such that  $r > s/2$ . Every function  $u \in H^r(\mathbb{R}^s, \mathbb{R})$  satisfies

$$(4.1) \quad C\|u\|_{L^\infty} \leq \|u\|_{H^r} = \|J_r(\cdot)(\mathcal{F}u)(\cdot)\|_{L^2},$$

where  $J_r(z) \triangleq (1 + \|z\|^2)^{r/2}/(2\pi)^{s/2}$ ,  $z \in \mathbb{R}^s$  denotes the Bessel potential and  $\mathcal{F}u$  the Fourier transform of  $u$ , whereas the constant  $C > 0$  depends on  $s$  uniquely.

Let  $\ell \in \mathbb{N}$ ,  $D \subseteq \mathbb{R}^s$  be an open and bounded domain, and  $u \in H^r(D, \mathbb{R})$  be a function such that  $u|_{\widehat{X}} = 0$ , where  $\widehat{X}_\ell \triangleq \{x_1, \dots, x_\ell\} \subseteq \mathbb{R}^s$  is a finite set of  $\ell$  given points. For the *fill distance*, which is defined by

$$h_{\widehat{X}_\ell, D} \triangleq \sup_{x \in D} \min_{i=1, \dots, \ell} \|x - x_i\|,$$

the following inequality holds true, e.g., [48, Theorem 11.32],

$$(4.2) \quad \|u\|_{H^\nu} \leq Ch_{\widehat{X}_\ell, D}^{r-\nu} \|u\|_{H^r}, \quad 0 \leq \nu \leq r,$$

where the constant  $C > 0$  depends on  $\nu$ ,  $r$ , and  $D$  uniquely. Inequality 4.2 may be interpreted as follows: if  $u$  is zero on a well distributed set of points over  $D$ , i.e.,  $h_{\widehat{X}_\ell, D}$  is small, and is very regular, i.e.,  $r$  is large, then any norm  $\|\cdot\|_{H^\nu}$ ,  $0 \leq \nu \leq r$ , e.g., the  $L^2$  norm for  $\nu = 0$ , is small over the whole domain  $D$ .

Given a set  $D$ , a RKHS  $\mathcal{H}_D$  is a separable Hilbert space of functions  $u : D \rightarrow \mathbb{R}$  such that the following *reproducing property* holds true:

**DEFINITION 4.1** (Reproducing property [3]). *For every point  $x \in D$  there exists a reproducing function  $k_x \in \mathcal{H}_D$  such that*

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}_D}, \quad f \in \mathcal{H}_D.$$

We denote  $K_D : D \times D \rightarrow \mathbb{R}$ , the *reproducing kernel* associated to  $\mathcal{H}_D$ , i.e.,

$$K_D(x_1, x_2) \triangleq \langle k_{x_1}, k_{x_2} \rangle_{\mathcal{H}_D}, \quad x_1, x_2 \in D.$$

The following crucial result holds true, see, e.g., [42]:

**THEOREM 4.2.** *Given a reproducing kernel for a RKHS  $\mathcal{H}_D$ , by the reproducing property, the reproducing function  $k_x \in \mathcal{H}_D$  for any  $x \in D$  corresponds to*

$$k_x = K_D(x, \cdot) \in \mathcal{H}_D.$$

We recall that the Sobolev space  $H^r(D, \mathbb{R})$  is a RKHS for  $r > s/2$ , and  $D \subseteq \mathbb{R}^s$  be either  $\mathbb{R}^s$  or an open domain with Lipschitz boundary and  $s \in \mathbb{N}$ . In this case, the associated reproducing kernel has a known closed form, see, e.g., [48].

**5. Learning Stochastic Differential Equations.** In this section, we finally introduce and study the problem of learning drift and diffusion coefficients of a stochastic differential equation. For this, from now on we assume the following hypothesis, which naturally stems from our setting, to hold true (see also Definition 3.1):

- (A) Let  $m \in \mathbb{N}$  with  $m \geq 1$ ,  $\alpha > 0$ , and  $p_0 \in H^{2m+1}(\mathbb{R}^n, \mathbb{R})$ . There exist  $(a_*, b_*) \in \mathcal{H}_m^+$  and a (unique) solution  $X$  to SDE with coefficients  $(a_*, b_*) \in \mathcal{H}_m^+$ .

Thanks to the results we gathered in Section 3, one readily checks that Assumption (A) yields the following characterization of the mapping  $X$ :

**COROLLARY 5.1.** *With the notation  $\mathcal{L}_t \triangleq \mathcal{L}_t^{a_*, b_*}$ , there exists a unique function  $p \in H^{m+1, 2(m+1)}([0, T] \times \mathbb{R}^n, \mathbb{R})$  which is non-negative and such that*

$$\int_A p(t, x) \, dx = \mu_t(A) \triangleq \int_{\mathbb{R}^n} \int_{\Omega} \mathbb{1}_{\{X_x(t) \in A\}} \, d\mathbb{P} \, \mu_0(dx), \quad \text{for all } A \in \mathcal{B}(\mathbb{R}^n),$$

and which satisfies:

$$\frac{\partial p}{\partial t}(t, y) = \mathcal{L}_t^* p(t, y), \quad \text{a.e. } (t, y) \in [0, T] \times \mathbb{R}^n.$$

To achieve our goal, we proceed along three successive steps:

1. First, we approximate the unknown curve of densities  $p$  from given samples  $\{X_{x_j}(t_\ell, \omega_j)\}_{\ell=1, \dots, M, j=1, \dots, N}$  of the unknown solution  $X$  to SDE, through a finite-dimensional RKHS-based model  $\hat{p}$ . Thanks to this, all the quantities appearing in the learning problems we define in the next steps can be actually (tractably) computed. As is customary in many applications, we assume our data are collected by independently sampling  $X$  at pre-defined times.
2. Second, we approximate the unknown coefficients  $(a_*, b_*) \in \mathcal{H}_m^+$  through additional coefficients  $(\hat{a}, \hat{b}) \in \mathcal{H}_m^+$  which “best” match SFPE when evaluated at  $\hat{p}$ , through an appropriately well-posed infinite-dimensional learning problem. We then combine appropriate estimates with (3.3) to show the solution to SFPE with coefficients  $(\hat{a}, \hat{b})$  well-approximates the unknown  $p$ .
3. Third, we make the learning problem at the second step “tractable” by approximating the coefficients  $(\hat{a}, \hat{b})$  through a finite-dimensional RKHS-based model  $(\hat{a}_Q, \hat{b}_Q)$ . We then combine the estimates we obtained at the second step with (3.3) to show the solution to SFPE with coefficients  $(\hat{a}_Q, \hat{b}_Q)$  well-approximates the unknown  $p$ . Finally, this latter property is combined with (3.5) to show theoretical error bounds in the context of observation and regulation of stochastic differential equations.

**5.1. Approximating solutions to SDE via RKHS-based models.** We start by defining our data set. We assume sampling happens at  $M \in \mathbb{N}$  fixed times  $0 = t_1 < \dots < t_M = T$ , which are equally spaced for simplicity, i.e.,  $t_\ell \triangleq T(\ell-1)/(M-1)$ . Then, we assume at each time  $t_\ell$ ,  $\ell = 1, \dots, M$ , we have access to  $N \in \mathbb{N}$  samples  $X_{\ell,1} \triangleq X_{x_1}(t_\ell, \omega_1), \dots, X_{\ell,N} \triangleq X_{x_N}(t_\ell, \omega_N)$  of the solution  $X$  to SDE, which have been independently drawn from the same probability  $\mu_{t_\ell}$ , with density  $p(t_\ell, \cdot)$  (see Corollary 5.1). Now, for every  $t \in [0, T]$  the family of probability measures given by

$$\mu_{t,k} \triangleq \underbrace{\mu_t \otimes \dots \otimes \mu_t}_{k\text{-times}} : \underbrace{\mathcal{B}(\mathbb{R}^n \times \dots \times \mathbb{R}^n)}_{k\text{-times}} \rightarrow [0, 1], \quad k \in \mathbb{N}$$

may be extended to a unique probability measure  $\mu_t^{\mathbb{N}} : \mathcal{B}^{\mathbb{N}} \rightarrow [0, 1]$  thanks to Kolmogorov (extension) theorem, and by definition, for every  $\ell = 1, \dots, M$  and  $N \in \mathbb{N}$  the samples  $\{X_{\ell,j}\}_{j=1, \dots, N}$  may be seen as independent random variables in the probability space  $((\mathbb{R}^n)^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, \mu_{t_\ell}^{\mathbb{N}})$ , with equal density  $p(t_\ell, \cdot)$ . Finally, by one further application of Kolmogorov theorem, we extend the family of probability measures  $\{\mu_{s_1}^{\mathbb{N}} \otimes \dots \otimes \mu_{s_k}^{\mathbb{N}}\}_{0 \leq s_1 \leq \dots \leq s_k \leq T}$  to a unique probability measure  $\mu_X : (\mathcal{B}^{\mathbb{N}})^{[0, T]} \rightarrow [0, 1]$ , so that, with an abuse of notation,  $\mu_X|_{t_1, \dots, t_M} = \mu_{t_1}^{\mathbb{N}} \otimes \dots \otimes \mu_{t_M}^{\mathbb{N}}$  for every  $M \in \mathbb{N}$ ,

and for every  $M, N \in \mathbb{N}$  the samples  $\{X_{\ell,j}\}_{\ell=1,\dots,M,j=1,\dots,N}$  may be seen as (not necessarily i.i.d.) random variables in the probability space  $((\mathbb{R}^n)^{\mathbb{N}})^{[0,T]}, (\mathcal{B}^{\mathbb{N}})^{[0,T]}, \mu_X$ . Below, the assessment “with probability at least” will be meant with respect to  $\mu_X$ .

At this step, fix  $M, N \in \mathbb{N}$  and define the following model (random) density

$$\hat{p}(t, x) \triangleq \sum_{\ell=1}^M c_{\ell}(t) \hat{g}_{\ell}(x), \quad \text{where} \quad \hat{g}_{\ell}(x) \triangleq \frac{1}{N} \sum_{j=1}^N \rho_R(x - X_{\ell,j})$$

with  $\rho_R(x) \triangleq R^{n/2} \|x\|^{-n/2} B_{n/2}(2\pi R \|x\|)$ ,  $R > 0$  and  $B_{n/2}$  is the Bessel  $J$  function of order  $n/2$ , while  $c_{\ell}(t) \triangleq e_{\ell}^{\top} G^{-1} v(t)$  with  $\{e_1, \dots, e_M\}$  the canonical basis of  $\mathbb{R}^M$  and we define  $v(t) \triangleq (K_{m+1}(t, t_1), \dots, K_{m+1}(t, t_M))$ . The notation  $G$  is used for the Gram matrix with elements  $G_{j,k} \triangleq K_{m+1}(t_j, t_k)$ , where  $K_{m+1}$  denotes the Sobolev kernel of smoothness degree  $m+1$  (see, e.g., [48, Page 133] for an explicit formula). Finally, for every  $u \in H^{1,2}([0, T] \times \mathbb{R}^n, \mathbb{R})$  we denote

$$L(u) \triangleq \int_0^T \left( \left\| \frac{\partial p}{\partial t}(t, \cdot) - \frac{\partial u}{\partial t}(t, \cdot) \right\|_{L^2}^2 + \|p(t, \cdot) - u(t, \cdot)\|_{H^2}^2 \right) dt.$$

Our result on the approximation of  $p$  via  $\hat{p}$  is as follows:

**THEOREM 5.2.** *There exists a constant  $C > 0$ , which depends on  $n$ ,  $m$ , and  $T$  uniquely, such that the following learning rate for the random model  $\hat{p}$  holds with probability at least  $1 - \delta$ , for every  $M, N \in \mathbb{N}$  such that  $M \geq 2T$ :*

$$L(\hat{p}) \leq C \left( \|p\|_{H^{m+1,2}}^2 + \|p\|_{H^{1,2(m+1)}}^2 \right) \times \left( M^{-2m} + R^{-4m} + R^n \log \left( \frac{4M}{\delta} \right) N^{-1} \right).$$

In particular, up to overloading  $C$ , for every  $\varepsilon > 0$  small enough we have that

$$\begin{aligned} L(\hat{p}) &\leq C \left( \|p\|_{H^{m+1,2}}^2 + \|p\|_{H^{1,2(m+1)}}^2 \right) \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^2 \\ &\leq C \left( \|p\|_{H^{m+1,2}}^2 + \|p\|_{H^{1,2(m+1)}}^2 \right) \left( \log \left( \frac{MN}{\delta} \right)^{\frac{1}{2}} (MN)^{-\frac{2m}{n+2(2m+1)}} \right)^2, \end{aligned}$$

if  $M = \varepsilon^{-1/m}/4$ ,  $N = \varepsilon^{-(2+n/(2m))}$  (as closest integers), and  $R = \varepsilon^{-1/(2m)}$ .

*Proof.* For the sake of clarity, we divide the proof in several steps.

**1) Preliminaries.** With obvious notation, fix  $r, s \geq 1$  and define the operator

$$P_M : H^r([0, T], \mathbb{R}) \rightarrow H^r([0, T], \mathbb{R}), \quad \text{via} \quad (P_M u)(t) = \sum_{\ell=1}^M c_{\ell}(t) u(t_{\ell}).$$

For  $u \in H^{r,s}([0, T] \times \mathbb{R}^n, \mathbb{R})$ , we denote by  $u_x$  the function  $u_x(\cdot) = u(\cdot, x)$ ,  $x \in \mathbb{R}^n$ , and by  $\tilde{u}$  the function  $\tilde{u}(t, x) = (P_M u_x)(t)$ . Now, note that the function  $v_x := u_x - P_M u_x$  satisfies  $v_x(t_{\ell}) = 0$  for any  $\ell = 1, \dots, M$ . From the results we recalled in Section 4, we can bound the norm of functions with scattered zeros as follows: for every  $0 \leq \sigma \leq r$  there exists a constant  $C_{r,\sigma} > 0$  such that

$$(5.1) \quad \|u_x - P_M u_x\|_{H^{\sigma}} \leq C_{r,\sigma} M^{\sigma-r} \|u_x\|_{H^r},$$

for almost every  $x \in \mathbb{R}^n$ . Below, we will often implicitly overload the constant  $C_{r,\sigma}$ . Moreover, we recall that, for  $\alpha, \beta \geq 0$ ,

$$D_t^\alpha D_x^\beta \tilde{u}(t, x) = D_t^\alpha D_x^\beta (P_M u_x)(t) = D_t^\alpha P_M (D_x^\beta u)(t, x),$$

so that by leveraging (5.1) with  $\sigma = \alpha$ , for  $0 \leq \alpha \leq r$  and  $0 \leq \nu \leq s$ , yields

$$\begin{aligned} (5.2) \quad & \int_0^T \|D_t^\alpha u(t, \cdot) - D_t^\alpha \tilde{u}(t, \cdot)\|_{H^\nu}^2 dt = \\ & = \sum_{|\beta| \leq \nu} \int_0^T \int_{\mathbb{R}^n} (D_t^\alpha D_x^\beta u(t, x) - D_t^\alpha D_x^\beta \tilde{u}(t, x))^2 dx dt \\ & = \sum_{|\beta| \leq \nu} \int_0^T \int_{\mathbb{R}^n} (D_t^\alpha D_x^\beta u(t, x) - D_t^\alpha P_M D_x^\beta u(t, x))^2 dx dt \\ & \leq \sum_{|\beta| \leq \nu} \int_{\mathbb{R}^n} \|D_x^\beta u(\cdot, x) - P_M D_x^\beta u(\cdot, x)\|_{H^\alpha}^2 dx \\ & \leq \sum_{|\beta| \leq \nu} \int_{\mathbb{R}^n} C_{r,\alpha}^2 M^{2(\alpha-r)} \|D_x^\beta u(\cdot, x)\|_{H^r}^2 dx = C_{r,\alpha}^2 M^{2(\alpha-r)} \|u\|_{H^{r,\nu}}^2. \end{aligned}$$

At this step, by decomposing  $p - \hat{p} = (p - \tilde{p}) + (\tilde{p} - \hat{p})$  and applying the triangular inequality to  $L(\hat{p})$ , we obtain that

$$\begin{aligned} L(\hat{p})^{1/2} & \leq L(\tilde{p})^{1/2} + A^{1/2}, \\ A & \triangleq \int_0^T \left( \left\| \frac{\partial \tilde{p}}{\partial t}(t, \cdot) - \frac{\partial \hat{p}}{\partial t}(t, \cdot) \right\|_{L^2}^2 + \|\tilde{p}(t, \cdot) - \hat{p}(t, \cdot)\|_{H^2}^2 \right) dt. \end{aligned}$$

Thanks to (5.2) and  $p \in H^{m+1,2(m+1)}([0, T] \times \mathbb{R}^n, \mathbb{R})$ , we may bound  $L(\tilde{p})$  by

$$L(\tilde{p})^{1/2} \leq C_{m+1,1} M^{-m} \|p\|_{H^{m+1,0}} + C_{m+1,0} M^{-(m+1)} \|p\|_{H^{m+1,2}}.$$

The rest of the proof is devoted to appropriately bounding  $A$ .

**2) Further decomposing the term  $A$ .** For this, from the definition of both  $\tilde{p}$  and  $\hat{p}$ , for every  $0 \leq \alpha \leq m+1$  and  $0 \leq \nu \leq 2(m+1)$  we have that

$$\begin{aligned} & \left( \int_0^T \|D_t^\alpha \tilde{p}(t, \cdot) - D_t^\alpha \hat{p}(t, \cdot)\|_{H^\nu}^2 dt \right)^{1/2} = \\ & = \left( \int_0^T \left\| \sum_{\ell=1}^M D_t^\alpha c_\ell(t) (p(t_\ell, \cdot) - \hat{g}_\ell(\cdot)) \right\|_{H^\nu}^2 dt \right)^{1/2} \\ & \leq \sum_{\ell=1}^M \|c_\ell\|_{H^\alpha} \|p(t_\ell, \cdot) - \hat{g}_\ell(\cdot)\|_{H^\nu}. \end{aligned}$$

In, particular, by applying this bound to the two terms in  $A$ , we infer that

$$A^{1/2} \leq \sum_{\ell=1}^M \left( \|c_\ell\|_{H^1} \|p(t_\ell, \cdot) - \hat{g}_\ell(\cdot)\|_{L^2} + \|c_\ell\|_{L^2} \|p(t_\ell, \cdot) - \hat{g}_\ell(\cdot)\|_{H^2} \right).$$

**3) Bounding each term  $\|p(t_\ell, \cdot) - \widehat{g}_\ell(\cdot)\|_{H^\nu}$ .** Fix  $\ell = 1, \dots, M$  and  $0 \leq \nu \leq 2(m+1)$ . As recalled in Section 4, the norm of any  $u \in H^\nu(\mathbb{R}^n, \mathbb{R})$ , with  $\nu \in \mathbb{N}$  writes

$$\|u\|_{H^\nu} = \|J_\nu(\cdot)(\mathcal{F}u)(\cdot)\|_{L^2},$$

where  $\mathcal{F}u$  denotes the Fourier transform of  $u$ , and  $J_\nu(z) \triangleq (1 + \|z\|^2)^{\nu/2}/(2\pi)^{n/2}$ ,  $z \in \mathbb{R}^n$ . In particular, since the Fourier transform of  $\rho_R$  is  $\mathbb{1}_{B_R^{\mathbb{R}^n}(0)}$ , it holds that

$$(5.3) \quad \mathcal{F}\widehat{g}_\ell(z) = \sum_{j=1}^N \mathbb{1}_{B_R^{\mathbb{R}^n}(0)}(z) e^{2\pi i z^\top X_{\ell,j}} = \mathbb{1}_{B_R^{\mathbb{R}^n}(0)}(z) \mathcal{F}\widehat{g}_\ell(z), \quad z \in \mathbb{R}^n,$$

so that, thanks to the fact that  $(1 - \mathbb{1}_{B_R^{\mathbb{R}^n}(0)})\mathbb{1}_{B_R^{\mathbb{R}^n}(0)} = 0$  and  $\mathbb{1}_{B_R^{\mathbb{R}^n}(0)}^2 = \mathbb{1}_{B_R^{\mathbb{R}^n}(0)}$ , by denoting the function  $p_\ell(\cdot) \triangleq p(t_\ell, \cdot)$  we may compute

$$\begin{aligned} \|p(t_\ell, \cdot) - \widehat{g}_\ell(\cdot)\|_{H^\nu} &= \|J_\nu(\mathcal{F}p_\ell - \mathcal{F}\widehat{g}_\ell)\|_{L^2} \\ &= \left\| J_\nu(1 - \mathbb{1}_{B_R^{\mathbb{R}^n}(0)})\mathcal{F}p_\ell + J_\nu\mathbb{1}_{B_R^{\mathbb{R}^n}(0)}(\mathcal{F}p_\ell - \mathcal{F}\widehat{g}_\ell) \right\|_{L^2} \\ &\leq \left\| J_\nu(1 - \mathbb{1}_{B_R^{\mathbb{R}^n}(0)})J_{2(m+1)}^{-1} \right\|_{L^\infty} \|J_{2(m+1)}\mathcal{F}p_\ell\|_{L^2} \\ &\quad + \left\| J_\nu\mathbb{1}_{B_R^{\mathbb{R}^n}(0)} \right\|_{L^\infty} \left\| \mathbb{1}_{B_R^{\mathbb{R}^n}(0)}(\mathcal{F}p_\ell - \mathcal{F}\widehat{g}_\ell) \right\|_{L^2} \\ &\leq R^{\nu-2(m+1)}\|p_\ell\|_{H^{2(m+1)}} + 2^\nu R^\nu \left\| \mathbb{1}_{B_R^{\mathbb{R}^n}(0)}(\mathcal{F}p_\ell - \mathcal{F}\widehat{g}_\ell) \right\|_{L^2}, \end{aligned}$$

where we used the fact that  $J_{2(m+1)}$ ,  $J_\nu$ , and  $\mathbb{1}_{B_R^{\mathbb{R}^n}(0)}$  are radial functions, and thus

$$\begin{aligned} \left\| J_\nu(1 - \mathbb{1}_{B_R^{\mathbb{R}^n}(0)})J_{2(m+1)}^{-1} \right\|_{L^\infty} &= \sup_{r>R} (1 + R^2)^{-(2(m+1)+\nu)/2} \leq R^{\nu-2(m+1)}, \\ \left\| J_\nu\mathbb{1}_{B_R^{\mathbb{R}^n}(0)} \right\|_{L^\infty} &= \sup_{0<r\leq R} (1 + R^2)^{\nu/2} \leq 2^\nu R^\nu, \quad \text{as soon as } R \geq 1. \end{aligned}$$

**4) Bounding each term  $\left\| \mathbb{1}_{B_R^{\mathbb{R}^n}(0)}(\mathcal{F}p_\ell - \mathcal{F}\widehat{g}_\ell) \right\|_{L^2}$ .** For this, for every integers  $\ell = 1, \dots, M$  and  $j = 1, \dots, N$  we define a random mapping  $\zeta_{\ell,j} : (\mathbb{R}^n)^\mathbb{N} \rightarrow L^2(\mathbb{R}^n, \mathbb{C})$  through the following expression, which holds for every  $x^\mathbb{N} \in (\mathbb{R}^n)^\mathbb{N}$ ,

$$\zeta_{\ell,j}(x^\mathbb{N})(z) \triangleq \mathbb{1}_{B_R^{\mathbb{R}^n}(0)}(z) e^{2\pi i z^\top X_{\ell,j}(x^\mathbb{N})}, \quad z \in \mathbb{R}^n.$$

Note that, for every fixed  $\ell = 1, \dots, M$ , since they depend deterministically from  $X_{\ell,1}, \dots, X_{\ell,M}$ , the random mappings  $\zeta_{\ell,1}, \dots, \zeta_{\ell,N}$  are independent random mappings in the probability space  $((\mathbb{R}^n)^\mathbb{N}, \mathcal{B}^\mathbb{N}, \mu_{t_\ell}^\mathbb{N})$ , and they are additionally equally distributed with respect to  $p_\ell(\cdot) \triangleq p(t_\ell, \cdot)$ . In particular, for every  $j = 1, \dots, N$  we may compute

$$\mathbb{E}_{\mu_{t_\ell}^\mathbb{N}}[\zeta_{\ell,j}](z) = \int_{\mathbb{R}^n} \mathbb{1}_{B_R^{\mathbb{R}^n}(0)}(z) e^{2\pi i z^\top x} p(t_\ell, x) dx = \mathbb{1}_{B_R^{\mathbb{R}^n}(0)}(z) (\mathcal{F}p_\ell)(z), \quad z \in \mathbb{R}^n.$$

We may also compute

$$\text{ess sup}_{x^\mathbb{N} \in (\mathbb{R}^n)^\mathbb{N}} \left\| \frac{\zeta_{\ell,j}(x^\mathbb{N}) - \mathbb{E}_{\mu_{t_\ell}^\mathbb{N}}[\zeta_{\ell,j}]}{N} \right\|_{L^2}^2 \leq \frac{4}{N^2} \int_{\mathbb{R}^n} \mathbb{1}_{B_R^{\mathbb{R}^n}(0)}(z) dz = \frac{4}{N^2} V_n R^n,$$

where  $V_n$  is the volume of the  $n$ -dimensional unit ball. At this step, thanks to Pinelis inequality, for every  $\eta > 0$  and  $\ell = 1, \dots, M$ , we may compute

$$\begin{aligned} & \mu_X \left( \left\| \mathbb{1}_{B_R^{\mathbb{R}^n}(0)} (\mathcal{F}p_\ell - \mathcal{F}\widehat{g}_\ell) \right\|_{L^2} > \frac{3(V_n R)^{n/2} \log\left(\frac{4}{\eta}\right)}{\sqrt{N}} \right) \leq \\ & \leq \mu_X \left( \left\| \mathbb{1}_{B_R^{\mathbb{R}^n}(0)} (\mathcal{F}p_\ell - \mathcal{F}\widehat{g}_\ell) \right\|_{L^2} > \frac{2\sqrt{2}(V_n R)^{n/2} \sqrt{\log\left(\frac{2}{\eta}\right)}}{\sqrt{N}} \right) \\ & = \mu_{t_\ell}^{\mathbb{N}} \left( \left\| \sum_{j=1}^N \frac{\zeta_{\ell,j} - \mathbb{E}_{\mu_{t_\ell}^{\mathbb{N}}}[\zeta_{\ell,j}]}{N} \right\|_{L^2} > \frac{2\sqrt{2}(V_n R)^{n/2} \sqrt{\log\left(\frac{2}{\eta}\right)}}{\sqrt{N}} \right) \leq \eta. \end{aligned}$$

Therefore, for every  $\delta > 0$ , by taking the union bound with  $\eta = \delta/M$ , with probability at least  $1 - \delta$  it finally holds that

$$\max_{\ell=1,\dots,M} \left\| \mathbb{1}_{B_R^{\mathbb{R}^n}(0)} (\mathcal{F}p_\ell - \mathcal{F}\widehat{g}_\ell) \right\|_{L^2} \leq \frac{3(V_n R)^{n/2} \log\left(\frac{4M}{\delta}\right)^{\frac{1}{2}}}{\sqrt{N}}.$$

**5) Bounding each term  $\|c_\ell\|_{H^\nu}$ .** By construction, each  $c_\ell \in H^{m+1}([0, T], \mathbb{R})$  is the function with minimum norm that satisfies  $c_{\ell_1}(t_{\ell_2}) = \delta_{\ell_1, \ell_2}$ , for  $\ell_1, \ell_2 = 1, \dots, M$ . In particular, since the function  $z_\ell(t) \triangleq \text{sinc}(M(t - t_\ell)/T)$  is analytic, and thus  $z_\ell \in H^{m+1}([0, T], \mathbb{R})$ , and satisfies  $z_{\ell_1}(t_{\ell_2}) = \delta_{\ell_1, \ell_2}$ , for  $\ell_1, \ell_2 = 1, \dots, M$ , it must hold that  $\|c_\ell\|_{H^{m+1}} \leq \|z_\ell\|_{H^{m+1}}$ , for every  $\ell = 1, \dots, M$ . Therefore, since

$$\max_{\ell_1=1,\dots,M} |c_{\ell_2}(t_{\ell_1})| = 1, \quad \ell_2 = 1, \dots, M,$$

by applying the bound for Sobolev functions with scattered zeros recalled in 4 with  $h = T/M$ , we obtain that, for every  $0 \leq \nu \leq m+1$  there exists a constant  $C_{m+1,\nu} > 0$  such that

$$\|c_\ell\|_{H^\nu} \leq C_{m+1,\nu} \left( (T/M)^{-\nu} + (T/M)^{m+1-\nu} \|z_\ell\|_{H^{m+1}} \right), \quad \ell = 1, \dots, M.$$

To conclude, we note that the Fourier transform of the extension of each  $z_\ell$  to  $H^{m+1}(\mathbb{R}, \mathbb{R})$  is the function  $(T/M)\mathbb{1}_{\{M/T\}}$ . Combining this latter result with the Fourier characterization of the norm  $\|\cdot\|_{H^{m+1}}$  yields, for every  $\ell = 1, \dots, M$ ,

$$\|z_\ell\|_{H^{m+1}}^2 \leq \frac{T}{M} \int_{-\frac{M}{2T}}^{\frac{M}{2T}} (1 + \tau^2)^{m+1} d\tau \leq \left(\frac{M}{T}\right)^{2(m+1)},$$

where the last step follows from the assumption  $M \geq 2T^1$ . Therefore, for every  $\ell = 1, \dots, M$  we finally obtain that (we implicitly overload the constant  $C_{m+1,\nu}$ )

$$\|c_\ell\|_{H^\nu} \leq C_{m+1,\nu} \left(\frac{M}{T}\right)^\nu, \quad \nu = 0, \dots, m+1,$$

---

<sup>1</sup>Indeed, under this assumption,  $1 \leq 3/4M^2/T^2$ , and thus  $\max_{0 \leq \tau \leq M/(2T)} (1 + \tau^2) \leq M^2/T^2$ .

**6) Gathering the previous bounds together and end of the proof.** Recalling from the results in Section 4 that there exists a constant  $C > 0$  such that

$$\|u\|_{L^\infty} \leq C\|u\|_{H^1}, \quad \text{for every } u \in H^{m+1}([0, T], \mathbb{R}),$$

we may compute, for every  $\ell = 1, \dots, M$ ,

$$\|p(t_\ell, \cdot)\|_{H^{2(m+1)}(\mathbb{R}^n, \mathbb{R})} \leq \|p\|_{L^\infty([0, T], \mathbb{R}) \otimes H^{2(m+1)}(\mathbb{R}^n, \mathbb{R})} \leq C\|p\|_{H^{1, 2(m+1)}}.$$

Combining this latter inequality with all the previous bounds finally yields

$$\begin{aligned} L(\widehat{p})^{1/2} &\leq L(\widehat{p})^{1/2} + A^{1/2} \\ &\leq C_{m+1,1} M^{-m} \|p\|_{H^{m+1,0}} + C_{m+1,0} M^{-(m+1)} \|p\|_{H^{m+1,2}} + A^{1/2} \\ &\leq C(\|p\|_{H^{m+1,2}} + \|p\|_{H^{1,2(m+1)}}) \times \\ &\quad \times \left( M^{-m} + R^{-2m} + R^{n/2} \log \left( \frac{4M}{\delta} \right)^{\frac{1}{2}} N^{-1/2} \right), \end{aligned}$$

for some appropriate constant  $C > 0$ , which holds with probability at least  $1 - \delta$ , and the sought conclusion may be easily inferred.  $\square$

The main benefit offered by replacing the unknown density  $p$  with the model density  $\widehat{p}$  consists of the fact that integrals of this latter mapping and of its derivatives can be actually (easily) computed, enabling to correctly instantiate our first learning problem for  $(a_*, b_*) \in \mathcal{H}_m^+$ , which is our next step.

**5.2. The infinite-dimensional learning problem and its fidelity.** In this section, our goal consists of instantiating and analyzing the learning problem to identify the coefficients  $(a_*, b_*) \in \mathcal{H}_m^+$ . This problem computes coefficients which “best” matches SFPE when evaluated at the model density  $\widehat{p}$ . Below, we seek those coefficients in the whole set  $\mathcal{H}_m^+$ , deferring to a later section the problem of learning  $(a_*, b_*) \in \mathcal{H}_m^+$  through finite-dimensional models. In particular, the main result contained herein consists of appropriate learning error estimates which pave the way to obtaining learning error estimates for the problem of learning  $(a_*, b_*) \in \mathcal{H}_m^+$  through finite-dimensional models. However, for pedagogical purposes, we present complete learning rates for the infinite-dimensional learning problem as well.

From now on, we fix  $0 < \varepsilon, \delta < 1$ , and select  $M, N \in \mathbb{N}$  and  $R > 0$  as claimed in Theorem 5.2, so that, with probability at least  $1 - \delta$ , it holds that:

$$\begin{aligned} (5.4) \quad L(\widehat{p}) &\leq C(a_*, b_*) \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^2 \\ &\leq C(a_*, b_*) \left( \log \left( \frac{MN}{\delta} \right)^{\frac{1}{2}} (MN)^{-\frac{2m}{n+2(2m+1)}} \right)^2, \end{aligned}$$

where the constant  $C(a_*, b_*) > 0$  depends on  $a_*$  and  $b_*$  uniquely. To compute accurate learning rates, we leverage classical RKHS approximation theory, which is essentially well-posed for mappings which are defined on bounded domains. For this, we will make use of the following additional assumption:

(B) There exists  $R_* > 0$  such that the coefficients  $(a_*, b_*) \in \mathcal{H}_m^+$  satisfy:

$$\text{supp}(a_*(t, \cdot), b_*(t, \cdot)) \subseteq \overline{B_{R_*}^{\mathbb{R}^n}(0)}, \quad \text{for every } t \in [0, T].$$



*Remark 5.3.* Assumption (B) plays a key role in computing learning rates which, on the one hand, leverage classical RKHS approximation theory, and which, on the other hand, depend explicitly on the parameters defining the learning problem. In particular, the latter property can not be generally obtained by just (smoothly) restricting to balls quantities which are defined on unbounded domains. That said, our result can be shown to hold under less restricting assumptions, such as assuming the coefficients  $(a_*, b_*) \in \mathcal{H}_m^+$  decay to zero at infinity under specific rates [48], although we opted for Assumption (B) to avoid excessively tedious computations, in turn fostering a smooth exposition. In addition, it is worth mentioning Assumption (B) is often naturally verified in many applications ranging from biology to robotics, where the state space is a bounded domain.

Thanks to Assumption (B), we may restrict ourselves to the closed subspace:

$$\mathcal{H}_{m,R_*}^+ \triangleq \left\{ (a, b) \in \mathcal{H}_m^+ : \text{supp}(a(t, \cdot), b(t, \cdot)) \subseteq \overline{B_{R_*}^{\mathbb{R}^n}(0)}, \text{ for every } t \in [0, T] \right\} \subseteq \mathcal{H}_m^+,$$

We are now ready to define our infinite-dimensional problem for learning the coefficients  $(a_*, b_*) \in \mathcal{H}_{m,R_*}^+$ , which writes as follows:

**DEFINITION 5.4.** *For every positive real  $\lambda > 0$ , the (random) infinite-dimensional Learning Problem to learn stochastic differential equations is defined as:*

$$\text{LP} \quad \min_{(a,b) \in \mathcal{H}_{m,R_*}^+} L_\lambda(a, b) \triangleq \int_0^T \left\| \frac{\partial \widehat{p}}{\partial t}(t, \cdot) - (\mathcal{L}_t^{a,b})^* \widehat{p}(t, \cdot) \right\|_{L^2}^2 dt + \lambda \|(a, b)\|_{\mathcal{H}_m}^2.$$

The well-posedness of problem LP is proven in the following proposition:

**PROPOSITION 5.5.** *For every  $\lambda > 0$ , problem LP is well-posed and has a unique solution, which is denoted by  $(\widehat{a}, \widehat{b}) \in \mathcal{H}_{m,R_*}^+$ .*

*Proof.* Since the mapping

$$(5.5) \quad (a, b) \in \mathcal{H}_m \mapsto \int_0^T \left\| \frac{\partial \widehat{p}}{\partial t}(t, \cdot) - (\mathcal{L}_t^{a,b})^* \widehat{p}(t, \cdot) \right\|_{L^2}^2 dt + \lambda \|(a, b)\|_{\mathcal{H}_m}^2$$

is strictly convex when restricted to the closed convex subset  $\mathcal{H}_{m,R_*}^+ \subseteq \mathcal{H}_m$ , we just need to prove the existence of a solution to LP. For this, if  $(a_k, b_k)_{k \in \mathbb{N}} \in \mathcal{H}_{m,R_*}^+$  is any minimizing sequence for LP, there must exist some constant  $C > 0$  such that  $\|(a_k, b_k)\|_{\mathcal{H}_m^+} \leq C$  for every  $k \in \mathbb{N}$ , and therefore, since  $\mathcal{H}_{m,R_*}^+$  is in particular a closed and convex subset of the Hilbert space  $\mathcal{H}_m$ , up to extracting a subsequence there exists  $(\widehat{a}, \widehat{b}) \in \mathcal{H}_{m,R_*}^+$  such that  $(a_k, b_k)_{k \in \mathbb{N}}$  converges to  $(\widehat{a}, \widehat{b})$  for the weak topology of  $\mathcal{H}_m$ . Now, thanks to Theorem 2.1, it is clear the mapping (5.5) is in addition continuous for the strong topology of  $\mathcal{H}_m$ . Thus, we infer that the mapping (5.5) is in particular weakly lower semi-continuous, and therefore we finally obtain that

$$\begin{aligned} & \int_0^T \left\| \frac{\partial \widehat{p}}{\partial t}(t, \cdot) - (\mathcal{L}_t^{\widehat{a}, \widehat{b}})^* \widehat{p}(t, \cdot) \right\|_{L^2}^2 dt + \lambda \|(\widehat{a}, \widehat{b})\|_{\mathcal{H}_m}^2 \leq \\ & \leq \liminf_{k \rightarrow \infty} \int_0^T \left\| \frac{\partial \widehat{p}}{\partial t}(t, \cdot) - (\mathcal{L}_t^{a_k, b_k})^* \widehat{p}(t, \cdot) \right\|_{L^2}^2 dt + \lambda \|(a_k, b_k)\|_{\mathcal{H}_m}^2 \\ & = \min_{(a,b) \in \mathcal{H}_{m,R_*}^+} \int_0^T \left\| \frac{\partial \widehat{p}}{\partial t}(t, \cdot) - (\mathcal{L}_t^{a,b})^* \widehat{p}(t, \cdot) \right\|_{L^2}^2 dt + \lambda \|(a, b)\|_{\mathcal{H}_m}^2, \end{aligned}$$

and the sought conclusion follows.  $\square$

We now investigate the fidelity of the learning problem LP. More specifically, through the estimates of Section 3, under appropriate norms we compute bounds for the approximation error between solutions to SFPE, associated with the drift and diffusion coefficients solutions to LP, and the unknown mapping  $p$ . In turn, these bounds will endow our learning procedure with well-posedness and high fidelity.

We start with the following technical lemma:

LEMMA 5.6. *For every  $\lambda > 0$ , with probability at least  $1 - \delta$ , it holds that:*

$$\begin{aligned} L_\lambda(\hat{a}, \hat{b}) &\leq C(a_*, b_*) \left( \lambda + \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^2 \right) \\ &\leq C(a_*, b_*) \left( \lambda + \left( \log \left( \frac{MN}{\delta} \right)^{\frac{1}{2}} (MN)^{-\frac{2m}{n+2(2m+1)}} \right)^2 \right), \end{aligned}$$

where the constant  $C(a_*, b_*) > 0$  depends on  $a_*$  and  $b_*$  uniquely.

*Proof.* Thanks to Theorem 2.1 and Corollary 5.1, a routine use of Hölder and Young inequalities allows us to compute

$$\begin{aligned} L_\lambda(\hat{a}, \hat{b}) - \lambda \| (a_*, b_*) \|_{\mathcal{H}_m}^2 &\leq \\ &\leq \int_0^T \left\| \frac{\partial \hat{p}}{\partial t}(t, \cdot) - \mathcal{L}_t^* \hat{p}(t, \cdot) \right\|_{L^2}^2 dt \\ &\leq 2 \int_0^T \left\| \frac{\partial \hat{p}}{\partial t}(t, \cdot) - \frac{\partial p}{\partial t}(t, \cdot) \right\|_{L^2}^2 dt + 2 \int_0^T \left\| \mathcal{L}_t^* (\hat{p} - p)(t, \cdot) \right\|_{L^2}^2 dt \\ &\leq 2 \int_0^T \left\| \frac{\partial \hat{p}}{\partial t}(t, \cdot) - \frac{\partial p}{\partial t}(t, \cdot) \right\|_{L^2}^2 dt + 2C(a_*, b_*) \int_0^T \left\| \hat{p}(t, \cdot) - p(t, \cdot) \right\|_{H^2}^2 dt \\ &\leq C(a_*, b_*) L(\hat{p}) \leq C(a_*, b_*) \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^2, \end{aligned}$$

where the (overloaded) constant  $C(a_*, b_*) > 0$ , which depends on  $a_*$  and  $b_*$  uniquely, comes from the constant in (5.4), and the conclusion follows.  $\square$

We are now ready to compute error bounds between solutions to SFPE, associated with the drift and diffusion coefficients solutions to LP, and the unknown mappings  $p$ . It is important to note that, given the nature of our data set, which essentially depends on observations of the law of the state process, the error between the unknown densities and the densities stemming from the learned coefficients is a “good metric” with which the convergence of an identification algorithm for stochastic differential equations which leverage observations of the state process may be tested.

We better formalize this metric as follows. For every  $(a, b) \in \mathcal{H}_m^+$ , let  $p_{a,b} \in H^{m+1, 2(m+1)}([0, T] \times \mathbb{R}^n, \mathbb{R})$  denote the unique solution to SFPE with coefficients  $(\hat{a}, \hat{b}) \in \mathcal{H}_m^+$ . Note that the existence and uniqueness of the regular mapping  $p_{a,b} \in H^{m+1, 2(m+1)}([0, T] \times \mathbb{R}^n, \mathbb{R})$  as non-negative solution to SFPE, of unitary mass and with coefficients  $(a, b) \in \mathcal{H}_m^+$  is immediate consequence of Theorem 3.4. We define the following metric to test the accuracy of our method:

$$E(a, b) \triangleq \| p_{a,b} - p \|_{L^2}^2 = \int_0^T \| p_{a,b}(t, \cdot) - p(t, \cdot) \|_{L^2}^2 dt, \quad (a, b) \in \mathcal{H}_m^+.$$

Among other benefits, we will see this metric is also particularly well-suited to estimate the error which is done when computing the observation/regulation metric (3.4).

Our main result on the accuracy of our learning method writes as follows:

**THEOREM 5.7.** *Let the coefficients  $(\widehat{a}, \widehat{b}) \in \mathcal{H}_{m, R_*}^+$  be the unique solution to LP with  $\lambda = \left(\log\left(\frac{1}{\delta\varepsilon}\right)^{\frac{1}{2}} \varepsilon\right)^2$ . With probability at least  $1 - \delta$ , it holds that:*

$$\begin{aligned} E(\widehat{a}, \widehat{b}) &\leq C(a_*, b_*) \left(\log\left(\frac{1}{\delta\varepsilon}\right)^{\frac{1}{2}} \varepsilon\right)^2 \\ &\leq C(a_*, b_*) \left(\log\left(\frac{MN}{\delta}\right)^{\frac{1}{2}} (MN)^{-\frac{2m}{n+2(2m+1)}}\right)^2, \end{aligned}$$

where the constant  $C(a_*, b_*) > 0$  depends on  $a_*$  and  $b_*$  uniquely.

*Proof.* We define

$$\rho \triangleq p_{\widehat{a}, \widehat{b}} - \widehat{p} \in C(0, T; L^2(\mathbb{R}^n, \mathbb{R})) \cap H^{0,1}([0, T] \times \mathbb{R}^n, \mathbb{R}),$$

and

$$f(t, \cdot) \triangleq - \left( \frac{\partial \widehat{p}}{\partial t}(t, \cdot) - (\mathcal{L}_t^{\widehat{a}, \widehat{b}})^* \widehat{p}(t, \cdot) \right) \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R}).$$

It is readily seen that

$$\begin{cases} \frac{d}{dt} \int_{\mathbb{R}^n} \varphi(y) \rho(t, y) dy = \\ \quad = \int_{\mathbb{R}^n} \left( \mathcal{L}_t^{\widehat{a}, \widehat{b}} \varphi(y) \rho(t, y) + f(t, y) \varphi(y) \right) dy, & \varphi \in C_c^\infty(\mathbb{R}^n, \mathbb{R}), \\ \rho(0, \cdot) = 0, \end{cases}$$

and therefore, thanks to Theorem 3.4, we may apply the estimate (3.3) to  $\rho$ , which in combination with Lemma 5.6 with the choice  $\lambda = \left(\log\left(\frac{1}{\delta\varepsilon}\right)^{\frac{1}{2}} \varepsilon\right)^2$  yields

$$\sup_{t \in [0, T]} \|p_{\widehat{a}, \widehat{b}}(t, \cdot) - \widehat{p}(t, \cdot)\|_{L^2}^2 \leq C(\|(\widehat{a}, \widehat{b})\|_{\mathcal{H}_m}) C(a_*, b_*) \left(\log\left(\frac{1}{\delta\varepsilon}\right)^{\frac{1}{2}} \varepsilon\right)^2,$$

where the constant  $C(a_*, b_*) > 0$  depends on  $a_*$  and  $b_*$  uniquely, whereas the constant  $C(\|(\widehat{a}, \widehat{b})\|_{\mathcal{H}_m}) > 0$  continuously depends on  $\|(\widehat{a}, \widehat{b})\|_{\mathcal{H}_m}$  uniquely. Up to overloading these constants, combined with (5.4) this latter inequality readily yields

$$(5.6) \quad \int_0^T \|p_{\widehat{a}, \widehat{b}}(t, \cdot) - p(t, \cdot)\|_{L^2}^2 dt \leq C(\|(\widehat{a}, \widehat{b})\|_{\mathcal{H}_m}) C(a_*, b_*) \left(\log\left(\frac{1}{\delta\varepsilon}\right)^{\frac{1}{2}} \varepsilon\right)^2.$$

At this step, from Lemma 5.6 in particular we obtain that

$$(5.7) \quad \|(\widehat{a}, \widehat{b})\|_{\mathcal{H}_m}^2 \leq \frac{C(a_*, b_*)}{\lambda} \left( \lambda + \left(\log\left(\frac{1}{\delta\varepsilon}\right)^{\frac{1}{2}} \varepsilon\right)^2 \right) = 2C(a_*, b_*),$$

as soon as  $\lambda = \left(\log\left(\frac{1}{\delta\varepsilon}\right)^{\frac{1}{2}} \varepsilon\right)^2$ . Therefore, up to overloading the constant  $C(a_*, b_*)$ , the conclusion follows from combining (5.6) with (5.7).  $\square$

**5.3. The finite-dimensional learning problem and its fidelity.** The learning problem we introduced in Section 5.2 (see Definition 5.4) remains difficult to numerically solve. Here, we discuss appropriate finite-dimensional approximations of LP and error bounds ensuring the fidelity of this latter approximation, ultimately making our learning approach for stochastic differential equations accurate and tractable.

We start by recalling and adapting the approximation tools we introduced in Section 4 to our framework. Let  $D = [0, T] \times \overline{B_{R_*}^{\mathbb{R}^n}(0)}$ , and consider the RKHS  $\mathcal{H}_D = H^{d(m)}(D, \mathbb{R})$  with associated kernel  $K_D$ . Without loss of generality, we may assume  $K_D$  and all its derivatives equal zero on  $[0, T] \times \partial B_{R_*}^{\mathbb{R}^n}(0)$ . For any set of  $Q \in \mathbb{N}$  points

$$\tilde{X}_D \triangleq \{(t_1, x_1), \dots, (t_Q, x_Q)\},$$

we consider the following coordinate-wise finite dimensional models to approximate the candidate drift and diffusion coefficient solutions  $(\hat{a}, \hat{b})$ :

$$\begin{aligned} \tilde{a}(t, y) &\triangleq \sum_{\ell, \ell'=1}^Q A_{\ell, \ell'} K_D((t, y), (t_\ell, x_\ell)) K_D((t, y), (t_{\ell'}, x_{\ell'})), \\ &\quad (t, y) \in D, \quad i, j \in \{1, \dots, n\}, \\ \tilde{b}(t, x) &\triangleq \sum_{\ell=1}^Q B_\ell K_D((t, y), (t_\ell, x_\ell)), \quad (t, y) \in D, \quad i \in \{1, \dots, n\}, \end{aligned}$$

with  $A_{\ell, \ell'} \in \mathbb{R}^{n \times n}$ ,  $B_\ell \in \mathbb{R}^n$ . Note in particular that, by defining  $\Phi_Q : D \rightarrow \mathbb{R}^{n \times Qn}$  to be the map  $\Phi_Q(t, y) = (K_D((t, x), (t_1, x_1))I_{n \times n} | \dots | K_D((t, y), (t_Q, x_Q))I_{n \times n})$ , then

$$(5.8) \quad \tilde{a}(t, y) = \Phi_Q(t, y) A \Phi_Q(t, y)^\top, \quad \tilde{b}(t, y) = \Phi_Q(t, y) B,$$

with  $B \triangleq (B_1, \dots, B_Q)^\top \in \mathbb{R}^{Qn}$ , and  $A \in \mathbb{R}^{Qn \times Qn}$  is the  $Q \times Q$  block matrix with block entries  $A_{\ell, \ell'}$ . Therefore, we define the finite dimensional convex subset of  $\mathcal{H}_{m, R_*}^+$  in which the tuple  $(\hat{a}, \hat{b})$  is approximated to be (zero is imposed as value outside  $D$ )

$$\mathcal{H}_{m, R_*}^{+, Q} \triangleq \text{span} \left\{ (\tilde{a}, \tilde{b}) : A \in \mathbb{R}^{Qn \times Qn}, A \succeq 0 \text{ and } B \in \mathbb{R}^{Qn} \right\} \subseteq \mathcal{H}_{m, R_*}^+.$$

Before moving to the core of this section, by leveraging the facts we recalled in Section 4, we provide a crucial approximation results for coefficients  $(a, b) \in \mathcal{H}_{m, R_*}^+$ . Specifically, by combining the bounds for Sobolev functions with scattered zeros we listed in Section 4 with Theorem 2.1, we obtain the following:

**THEOREM 5.8.** *Denote the fill distance between  $\tilde{X}_D$  and  $D$  with*

$$h_Q \triangleq \sup_{(t, y) \in D} \min_{\ell=1, \dots, Q} \|(t_\ell, x_\ell) - (t, y)\|.$$

*There exists a constant  $C > 0$  such that, for every  $(a, b) \in \mathcal{H}_{m, R_*}^+$  with  $a(t, y) \succ \beta I$ ,  $(t, y) \in D$ , for some constant  $\beta > 0$ , there exists a tuple  $(\mathcal{P}_Q(a), \mathcal{P}_Q(b)) \in \mathcal{H}_{m, R_*}^{+, Q}$ :*

$$\|(a, b) - (\mathcal{P}_Q(a), \mathcal{P}_Q(b))\|_{W^{2, \infty}} \leq C(\|\sqrt{a}\|_{H^{d(m)}}^2 + \|b\|_{H^{d(m)}})h_Q^{2m-1},$$

$$\|(\mathcal{P}_Q(a), \mathcal{P}_Q(b))\|_{\mathcal{H}_m} \leq C(\|\sqrt{a}\|_{H^{d(m)}}^2 + \|b\|_{H^{d(m)}}).$$

*Proof.* First, by combining the bounds for Sobolev functions with scattered zeros (4.2) in Section 4 with Theorem 2.1, we may readily claim the existence of a constant  $C > 0$  (implicitly overloaded below) and  $\mathcal{P}_Q(b) \triangleq \tilde{b}$  in the form (5.8) such that

$$\|b - \tilde{b}\|_{W^{2,\infty}} \leq C_1 \|b\|_{H^{d(m)}} h_Q^{2m-1}.$$

We now turn to the construction of  $\mathcal{P}_Q(a) \triangleq \tilde{a}$ . This is slightly more elaborated and is taken from [40, 31, 39, 29], in particular, adapting the construction of [31]. Note that the matrix square root  $\sqrt{\cdot}$  has uniformly bounded derivatives of any order on the set  $\{A \in \mathbb{R}^{n \times n} | A \succ \beta I\}$ . By assumption, it thus follows that  $\sqrt{a} \in H^{d(m)}(D, \mathbb{R}^{n \times n})$ , and in particular that the element-wise functions  $(\sqrt{a})_{i,j} \triangleq e_i \sqrt{a} e_j \in H^{d(m)}(D, \mathbb{R})$ . Now, we can build our approximation  $\tilde{a}$  as follows

$$\tilde{a}(t, x) \triangleq v(t, y)^\top v(t, y) \succeq 0, \quad (t, y) \in D,$$

where  $v : D \rightarrow \mathbb{R}^{n \times n}$  is obtained by element-wise applying the bound for Sobolev functions with scattered zeros (4.2) in Section 4. In particular, we obtain that

$$\|(\sqrt{a})_{i,j} - v_{i,j}\|_{W^{2,\infty}} \leq C \|\sqrt{a}\|_{H^{d(m)}} h_Q^{2m-1}, \quad i, j \in \{1, \dots, n\}.$$

Therefore, since  $\sqrt{a}^\top \sqrt{a} - v^\top v = \sqrt{a}^\top (\sqrt{a} - v) + (\sqrt{a} - v)^\top \sqrt{a} - (\sqrt{a} - v)^\top (\sqrt{a} - v)$ , and  $W^{2,\infty}$  is in particular a Banach algebra, we may infer that

$$\|a - \tilde{a}\|_{W^{2,\infty}} \leq 2 \|\sqrt{a} - v\|_{W^{2,\infty}} \|\sqrt{a}\|_{W^{2,\infty}} + \|\sqrt{a} - v\|_{W^{2,\infty}}^2.$$

At this step, the inequalities

$$\|\sqrt{a} - v\|_{W^{2,\infty}}^2 \leq \sum_{i,j=1}^n \|(\sqrt{a})_{i,j} - v_{i,j}\|_{W^{2,\infty}}^2$$

and  $\|\cdot\|_{W^{2,\infty}} \leq C \|\cdot\|_{H^{d(m)}}$  may be combined to readily derive that

$$\|a - \tilde{a}\|_{W^{2,\infty}} \leq C n^2 \|\sqrt{a}\|_{H^{d(m)}}^2 h_Q^{2m-1},$$

that is the first inequality claimed in the theorem. In particular, note that

$$v(t, y) = \sum_{\ell=1}^Q R_\ell K_D((t, y), (t_\ell, x_\ell)), \quad (t, y) \in D,$$

for some  $R_1, \dots, R_Q \in \mathbb{R}^{n \times n}$ . By denoting  $R = (R_1, \dots, R_Q) \in \mathbb{R}^{n \times Qn}$ , we have that

$$\tilde{a}(t, y) = \Phi_Q(t, y) A \Phi_Q(t, y)^\top, \quad (t, y) \in D,$$

with  $A \triangleq R R^\top \in \mathbb{R}^{Qn \times Qn}$  semi-positive-definite. We thus infer that  $(\tilde{a}, \tilde{b}) \in \mathcal{H}_{m, R_*}^{+, Q}$ .

To conclude, according to the results we gathered in Section 4, it holds that  $\tilde{b} = \Pi_Q(b)$  and  $v = \Pi_Q(\sqrt{a})$ , where  $\Pi_Q$  denotes the corresponding projection operator of each involved RKHS, and thus  $\tilde{a} = \Pi_Q(\sqrt{a})^\top \Pi_Q(\sqrt{a})$ . This implies that

$$\|\Pi_Q(b)\|_{H^{d(m)}} \leq \|b\|_{H^{d(m)}} \quad \text{and} \quad \|\Pi_Q(\sqrt{a})\|_{H^{d(m)}} \leq \|\sqrt{a}\|_{H^{d(m)}}.$$

The proof is concluded considering that  $H^{d(m)}$  is in particular a Banach algebra.  $\square$

In the rest of the manuscript, we adopt the notation we introduced in Theorem 5.8. We are now ready to define our finite-dimensional problem for learning the coefficients  $(a_*, b_*) \in \mathcal{H}_{m, R_*}^+$ , which writes as follows:

DEFINITION 5.9. *For every positive real  $\lambda > 0$ , the (random) finite-dimensional Learning Problem to learn stochastic differential equations is defined as:*

$$\text{LP}_Q \quad \min_{(a, b) \in \mathcal{H}_{m, R_*}^{+, Q}} L_\lambda(a, b) \triangleq \int_0^T \left\| \frac{\partial \hat{p}}{\partial t}(t, \cdot) - (\mathcal{L}_t^{a, b})^* \hat{p}(t, \cdot) \right\|_{L^2}^2 dt + \lambda \|(a, b)\|_{\mathcal{H}_m}^2.$$

Since  $\mathcal{H}_{m, R_*}^{+, Q}$  is a finite-dimensional convex set, the well-posedness of  $\text{LP}_Q$  may be proven similarly to the well-posedness of the original problem  $\text{LP}$ , i.e., by replicating the proof of Proposition 5.5. We thus report this result below without proof:

PROPOSITION 5.10. *For every  $\lambda > 0$ , problem  $\text{LP}_Q$  is well-posed and has a unique solution, which is denoted by  $(\hat{a}_Q, \hat{b}_Q) \in \mathcal{H}_{m, R_*}^{+, Q}$ .*

Remark 5.11. Thanks to Assumption (A), up to re-scaling  $\alpha$  and overloading  $a_*$ , one may always assume that  $a_*(t, y) \succ \alpha I$ ,  $(t, y) \in D$ . This property would enable applying Theorem 5.8 directly to  $(a_*, b_*)$ . Unfortunately, the setting  $a_*(t, y) \succ \alpha I$ ,  $(t, y) \in D$ , does not fit Assumption (B), and in turn the problem formulation of  $\text{LP}_Q$ . Nevertheless, one may appropriately tailor the definition of  $\mathcal{H}_{m, R_*}^{+, Q}$  to allow for  $a_*(t, y) \succ \alpha I$ ,  $(t, y) \in D$ . Indeed, in such setting Assumption (B) would rather require that  $\text{supp}(a(t, \cdot) - \alpha I, b(t, \cdot)) \subseteq \overline{B_{R_*}^{\mathbb{R}^n}(0)}$ , for every  $t \in [0, T]$ . Therefore, we should consider a larger domain  $D$  in which learning is pursued, e.g.,  $D \triangleq [0, T] \times \overline{B_{R_*+1}^{\mathbb{R}^n}}$ , selecting  $K_D$  to equal 1 on  $[0, T] \times \partial B_{R_*+1}^{\mathbb{R}^n}(0)$  and such that all its derivatives equal zero on  $[0, T] \times \partial B_{R_*+1}^{\mathbb{R}^n}(0)$  (this selection is always possible under mild assumptions). Thanks to this choice,  $\mathcal{H}_{m, R_*}^{+, Q}$  may be replaced by the following subset

$$\text{span} \left\{ (\chi \tilde{a} + (1 - \chi)\alpha I, \tilde{b}) : A \in \mathbb{R}^{Qn \times Qn}, A \succeq \alpha I, \text{ and } B \in \mathbb{R}^{Qn} \right\} \subseteq \mathcal{H}_{m, R_*+1}^+,$$

where the definition of  $\mathcal{H}_{m, R_*+1}^+$  corresponds to  $\mathcal{H}_{m, R_*}^+$ 's with updated requirement  $\text{supp}(a(t, \cdot) - \alpha I, b(t, \cdot)) \subseteq \overline{B_{R_*+1}^{\mathbb{R}^n}(0)}$ . Here,  $\chi \in C^\infty([0, T] \times \mathbb{R}^n, [0, 1])$  is some fixed cut-off function satisfying  $\chi|_{\overline{B_{R_*}^{\mathbb{R}^n}(0)}} = 1$  and  $\text{supp } \chi \subseteq \overline{B_{R_*+1}^{\mathbb{R}^n}(0)}$ . One readily check this updated definition does not affect the validity of Theorem 5.8, and in turn of the rates of convergence we will develop shortly. From a numerical point of view, the computational cost of solving  $\text{LP}_Q$  with this updated definition of  $\mathcal{H}_{m, R_*}^{+, Q}$  remains unchanged (see Section 6). Therefore, without loss of generality, from now on we assume we may apply Theorem 5.8 to  $(a_*, b_*) \in \mathcal{H}_{m, R_*}^+$  without further ado.

The next result is a natural extension of Lemma 5.6 to the setting of problem  $\text{LP}_Q$ , and it represents the main result of this section.

LEMMA 5.12. *For  $Q \in \mathbb{N}$  with  $h_Q \leq 1$ , with probability at least  $1 - \delta$ , it holds that:*

$$L_\lambda(\hat{a}_Q, \hat{b}_Q) \leq C(a_*, b_*) \left( \lambda + \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^2 + h_Q^{2(2m-1)} \right),$$

where the constant  $C(a_*, b_*) > 0$  depends on  $a_*$  and  $b_*$  uniquely.

*Proof.* Thanks to Lemma 5.6 (actually, its proof), we may compute

$$\begin{aligned}
L_\lambda(\widehat{a}_Q, \widehat{b}_Q) - \lambda \|(\mathcal{P}_Q(a_*), \mathcal{P}_Q(b_*))\|_{\mathcal{H}_m}^2 &\leq \\
&\leq \int_0^T \left\| \frac{\partial \widehat{p}}{\partial t}(t, \cdot) - (\mathcal{L}_t^{\mathcal{P}_Q(a_*), \mathcal{P}_Q(b_*)})^* \widehat{p}(t, \cdot) \right\|_{L^2}^2 dt \\
&\leq 2 \int_0^T \left\| \frac{\partial \widehat{p}}{\partial t}(t, \cdot) - \mathcal{L}_t^* \widehat{p}(t, \cdot) \right\|_{L^2}^2 dt \\
&\quad + 2 \int_0^T \left\| \left( (\mathcal{L}_t^{\mathcal{P}_Q(a_*), \mathcal{P}_Q(b_*)})^* - \mathcal{L}_t^* \right) \widehat{p}(t, \cdot) \right\|_{L^2}^2 dt \\
&\leq C(a_*, b_*) \left( \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^2 \right. \\
&\quad \left. + \|(a_*, b_*) - (\mathcal{P}_Q(a_*), \mathcal{P}_Q(b_*))\|_{W^{2,\infty}}^2 \int_0^T \|\widehat{p}(t, \cdot)\|_{H^2}^2 dt \right),
\end{aligned}$$

where the constant  $C(a_*, b_*) > 0$ , which we will overload below, depends on  $a_*$  and  $b_*$  uniquely. Theorem 5.2 and the choice  $0 < \varepsilon, \delta < 1$  readily yield

$$\int_0^T \|\widehat{p}(t, \cdot)\|_{H^2}^2 dt \leq C(a_*, b_*).$$

By combining this inequality with Theorem 5.8 (and Remark 5.11), we infer that

$$\|(a_*, b_*) - (\mathcal{P}_Q(a_*), \mathcal{P}_Q(b_*))\|_{W^{2,\infty}}^2 \int_0^T \|\widehat{p}(t, \cdot)\|_{H^2}^2 dt \leq C(a_*, b_*) h_Q^{2(2m-1)},$$

yielding that

$$\begin{aligned}
(5.9) \quad L_\lambda(\widehat{a}_Q, \widehat{b}_Q) - \lambda \|(\mathcal{P}_Q(a_*), \mathcal{P}_Q(b_*))\|_{\mathcal{H}_m}^2 &\leq \\
&\leq C(a_*, b_*) \left( \lambda + \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^2 + h_Q^{2(2m-1)} \right).
\end{aligned}$$

Finally, observe that Theorem 5.8 additionally provides that

$$\|(\mathcal{P}_Q(a_*), \mathcal{P}_Q(b_*))\|_{\mathcal{H}_m}^2 \leq C(a_*, b_*),$$

which combined with (5.9) leads to the conclusion.  $\square$

Thanks to Lemma 5.12, Theorem 5.7 may be straightforwardly extended to the context of the finite-dimensional learning problem  $\text{LP}_Q$  (just by replicating its proof). We thus report this result without proof in the proposition below.

**THEOREM 5.13.** *Let the coefficients  $(\widehat{a}_Q, \widehat{b}_Q) \in \mathcal{H}_{m,R_*}^{+,Q}$  be the unique solution to  $\text{LP}_Q$  with  $\lambda = \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^2$  and  $Q \in \mathbb{N}$  so that  $h_Q = \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^{\frac{1}{2m-1}} \leq 1$ . With probability at least  $1 - \delta$ , it holds that:*

$$\begin{aligned}
E(\widehat{a}_Q, \widehat{b}_Q) &\leq C(a_*, b_*) \left( \log \left( \frac{1}{\delta \varepsilon} \right)^{\frac{1}{2}} \varepsilon \right)^2 \\
&\leq C(a_*, b_*) \left( \log \left( \frac{MN}{\delta} \right)^{\frac{1}{2}} (MN)^{-\frac{2m}{n+2(2m+1)}} \right)^2,
\end{aligned}$$



where the constant  $C(a_*, b_*) > 0$  depends on  $a_*$  and  $b_*$  uniquely.

We conclude this section with a result which summarizes our essential contributions, showing how our learning approach may be leveraged for efficient observation/regulation of stochastic differential equations. First, we recall the setting introduced at the end of Section 3. Specifically, for every coefficients  $(a, b) \in \mathcal{H}_m^+$ , we denote by  $X^{a,b}$  and  $p_{a,b}$  respectively the (unique) solutions to SDE and SFPE with coefficients  $(a, b) \in \mathcal{H}_m^+$ . By combining Theorem 5.13 with Corollary 3.5 (in particular, compare with (3.4); more details about this remark are provided in Section 8), we readily obtain the following summarizing result:

**THEOREM 5.14.** *Let  $m \in \mathbb{N}$ ,  $\alpha > 0$ ,  $R_* > 0$ ,  $p_0 \in H^{2m+1}(\mathbb{R}^n, \mathbb{R})$ , and  $(a_*, b_*) \in \mathcal{H}_{m, R_*}^+$  satisfy Assumptions (A) and (B), and denote by  $X$  and  $p$  respectively the (unique) solutions to SDE and SFPE with coefficients  $(a_*, b_*) \in \mathcal{H}_{m, R_*}^+$ . There exists a constant  $C(a_*, b_*) > 0$  which only depends on  $a_*$  and  $b_*$ , such that by choosing the following learning parameters for the fixed precision parameters  $0 < \varepsilon, \delta < 1$ :*

- $M = \varepsilon^{-1/m}/4$ ,  $N = \varepsilon^{-(2+n/(2m))}$  (as closest integers), and  $R = \varepsilon^{-1/(2m)}$ ,
- $\lambda = \left(\log\left(\frac{1}{\delta\varepsilon}\right)^{\frac{1}{2}}\varepsilon\right)^2$ , and  $Q \in \mathbb{N}$  so that  $h_Q = \min\left(1, \left(\log\left(\frac{1}{\delta\varepsilon}\right)^{\frac{1}{2}}\varepsilon\right)^{\frac{1}{2m-1}}\right)$ ,

with probability at least  $1 - \delta$ , the following estimate holds:

$$(5.10) \quad \begin{aligned} \|p - p_{\hat{a}_Q, \hat{b}_Q}\|_{L^2} &\leq C(a_*, b_*) \log\left(\frac{1}{\delta\varepsilon}\right)^{\frac{1}{2}} \varepsilon \\ &\leq C(a_*, b_*) \log\left(\frac{MN}{\delta}\right)^{\frac{1}{2}} (MN)^{-\frac{2m}{n+2(2m+1)}}, \end{aligned}$$

where  $(\hat{a}_Q, \hat{b}_Q) \in \mathcal{H}_{m, R_*}^{+, Q}$  is the unique solution to the finite-dimensional learning problem  $\text{LP}_Q$ , where the learning parameters have been selected as above. Therefore, for every  $f \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$  the following observation/regulation estimate holds:

$$\begin{aligned} &\left| \mathbb{E}_{\mu_0 \times \mathbb{P}} \left[ \int_0^T f(t, X_x(t)) \, dt \right] - \mathbb{E}_{\mu_0 \times \mathbb{P}} \left[ \int_0^T f(t, X_x^{\hat{a}_Q, \hat{b}_Q}(t)) \, dt \right] \right| \leq \\ &\leq C(a_*, b_*) \log\left(\frac{1}{\delta\varepsilon}\right)^{\frac{1}{2}} \varepsilon \leq C(a_*, b_*) \log\left(\frac{MN}{\delta}\right)^{\frac{1}{2}} (MN)^{-\frac{2m}{n+2(2m+1)}}. \end{aligned}$$

**Remark 5.15** (On the optimality of the learning rates (5.10)). To the best of our knowledge, the learning rates (5.10) have not yet appeared in the literature on non-parametric learning of SDE. Specifically, although in recent works, see, e.g., [1, 35, 2, 28], learning rates that improve as the regularity of the drift and diffusion coefficients increases are proposed, for the first time our method offers non-asymptotic learning rates when considering estimation of multi-dimensional SDE with non-constant diffusion coefficient, under discrete-time observations of  $X$ .

For the sake of completeness, one may want to compare (5.10) with existing learning rates, e.g., the rates offered by [1, 35], though as we mentioned the estimation methods provided in these works apply to less general settings. If we were to do such comparison, we would infer that, up to a log factor, (5.10) would be “almost optimal” in that their would be sub-optimal by a factor  $(NM)^{\frac{1}{n+2(2m+1)}}$ . However, this factor exponentially converges to 1 when  $m$  tends to infinity.

**6. Computational Considerations.** From a numerical viewpoint, we can solve problem  $\text{LP}_Q$  exactly in closed form. Indeed, such a problem amounts to a finite semi-definite program in the variables  $A \in \mathbb{R}^{Qn \times Qn}$ ,  $B \in \mathbb{R}^{Qn \times 1}$ , subject to  $A \succeq 0$ .

**Problem reformulation.** Recall that we are naturally, though implicitly extending the mapping  $\Phi_Q$  to zero outside  $D$ , smoothly. Thanks to the characterization form (5.8), which is valid for every  $(a, b) \in \mathcal{H}_{m, R_*}^{+, Q}$ , one easily shows that

$$\mathcal{L}_t^{a, b} \hat{p}(t, y) = \text{tr}(V(t, y) A) - U(t, y) B + r(t, y), \quad (t, y) \in [0, T] \times \mathbb{R}^n,$$

where

$$r(t, y) \triangleq \frac{\alpha}{2} \Delta \hat{p}(t, y),$$

and  $U : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^{1 \times Qn}$  and  $V : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^{Qn \times Qn}$  are defined as

$$\begin{aligned} U_\ell(t, y) &\triangleq \sum_{i=1}^n \frac{\partial}{\partial x_i} [\hat{p}(t, y) e_i^\top K_D((t, y), (t_\ell, x_\ell))] \in \mathbb{R}^n, \quad \ell, \ell' = 1, \dots, Q, \\ V_{\ell, \ell'}(t, x) &\triangleq \frac{1}{2} \sum_{i, j=1}^n \frac{\partial^2}{\partial y_j \partial y_i} [\hat{p}(t, y) K_D((t, y), (t_\ell, x_\ell)) \times \\ &\quad \times K_D((t, y), (t_{\ell'}, x_{\ell'})) e_i e_j^\top] \in \mathbb{R}^{n \times n}. \end{aligned}$$

Therefore, by denoting  $q \triangleq \frac{\partial \hat{p}}{\partial t}$ , the cost in  $\text{LP}_Q$  thus writes

$$\begin{aligned} L_\lambda(a, b) &= \int_{[0, T] \times \mathbb{R}^n} (q(t, y) - \text{tr}(V(t, y) A) + U(t, y) B - r(t, y))^2 dt dy \\ &= \int_{[0, T] \times \mathbb{R}^n} (1, B^\top, \text{vec}(A)^\top) W(t, y) (1, B^\top, \text{vec}(A)^\top)^\top dt dy \\ &= (1, B^\top, \text{vec}(A)^\top)^\top H (1, B^\top, \text{vec}(A)^\top). \end{aligned}$$

where, by denoting  $\tilde{q} \triangleq q - r$ , we define  $W : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^{1+Qn+Qn^2 \times 1+Qn+Qn^2}$  as

$$W(t, y) \triangleq \begin{pmatrix} \tilde{q}^2(t, y) & -\tilde{q}(t, y) U(t, y) & -\tilde{q}(t, y) \text{vec}(V(t, y))^\top \\ -\tilde{q}(t, y) U(t, y) & U(t, y) U(t, y)^\top & U(t, y)^\top \text{vec}(V(t, y))^\top \\ -\tilde{q}(t, y) \text{vec}(V(t, y)) & \text{vec}(V(t, y)) U(t, y) & \text{vec}(V(t, y)) \text{vec}(V(t, y))^\top \end{pmatrix}$$

whereas  $H \in \mathbb{R}^{1+Qn+Q^2 n^2 \times 1+Qn+Q^2 n^2}$  is defined as

$$H \triangleq \int_{[0, T] \times \mathbb{R}^n} W(t, y) dt dy.$$

Summing up,  $\text{LP}_Q$  equals the following semi-definite program

$$\begin{aligned} &\min && v^\top H v. \\ &\begin{cases} A \in \mathbb{R}^{Qn \times Qn}, B \in \mathbb{R}^{Qn} \\ v \triangleq (1, B, \text{vec}(A)) \in \mathbb{R}^{1+Qn+Q^2 n^2} \\ A \succeq 0 \end{cases} \end{aligned}$$

Such semi-definite programs enjoy a particularly simple form. They can be efficiently solved via damped Newton methods, with computational cost  $O((Qn)^{3.5})$  [34].

**Computing the integral  $H$ .** The computation of each entry of the matrix  $H$  can be done in closed form, given that they correspond to the integral of products of functions, that are linear combinations of elementary functions. As a matter of example, below we explicitly compute the first element of  $H$ , i.e.,  $H_{11}$ , the same argument holding for the other elements. We have that

$$\int_{[0,T] \times \mathbb{R}^n} W_{11}(t, y) dt dy = \int_{[0,T] \times \mathbb{R}^n} (q(t, y)^2 - 2q(t, y)r(t, y) + r(t, y)^2) dt dy.$$

Due to their similarity, we proceed with computations only for the integral of the first term in the right-hand side of this equality. For this, we may compute

$$\begin{aligned} & \int_{[0,T] \times \mathbb{R}^n} q(t, y)^2 dt dy \\ &= \sum_{\ell, \ell'=1}^M \sum_{j, j'=1}^N \frac{1}{N^2} \int_{[0,T] \times \mathbb{R}^n} \dot{c}_\ell(t) \dot{c}_{\ell'}(t) \rho_R(y - x_{\ell, j}) \rho_R(y - x_{\ell', j'}) dt dy \\ &= \sum_{i, i', \ell, \ell'=1}^M \sum_{j, j'=1}^N \frac{\alpha_{i\ell} \alpha_{i'\ell'}}{N^2} \int_0^T \dot{k}_m(t - t_i) \dot{k}_m(t - t_{i'}) dt \times \\ & \quad \times \int_{\mathbb{R}^n} \rho_R(y - x_{\ell, j}) \rho_R(y - x_{\ell', j'}) dy. \end{aligned}$$

Above we used the definition of  $\widehat{p}$ , and the fact that  $c_\ell$  are minimum-norm kernel interpolators [48], i.e., they are given by  $c_\ell(t) = \sum_{i=1}^N \alpha_{i,\ell} k_m(t - t_i)$ , where  $\alpha_{i,\ell} = (K^{-1})_{i,\ell}$ , whereas  $K \in \mathbb{R}^{N \times N}$  is given by  $K_{i,i'} = k_m(t_i, t_{i'})$ . By denoting

$$R_{\ell, \ell'} \triangleq \int_0^T \dot{k}_m(t - t_i) \dot{k}_m(t - t_{i'}) dt, \quad S_{\ell, j, \ell', j'} \triangleq \int_{\mathbb{R}^n} \rho_R(y - x_{\ell, j}) \rho_R(y - x_{\ell', j'}) dy,$$

the computations above thus amount to

$$\int_{[0,T] \times \mathbb{R}^n} q(t, y)^2 dt dy = \sum_{\ell, \ell'=1}^M \sum_{j, j'=1}^N (K^{-1} R K^{-1})_{\ell, \ell'} S_{\ell, j, \ell', j'}.$$

Assuming that the computational cost of analytically computing such integrals of elementary functions is  $O(1)$  per integral, the total computational cost to compute  $H_{11}$  is thus  $O(M^3 + M^2 N^2)$ . Under the same assumption and iterating the same argument for all the elements of  $H$ , since  $U$  is the linear combination of  $Q n^2$  elementary functions, whereas  $V$  is the linear combination of  $Q^2 n^4$  elementary functions, the total computational cost to compute the whole matrix  $H$  is thus  $O(M^3 + M^2 N^2 + Q^4 n^8)$ .

**Considerations on the computational complexity.** Since to achieve a desired value for the fill distance  $Q = O(h_Q^{-(n+1)})$  number of centers are required, taken, e.g. on a grid of stepsize  $h_Q$  [40], we thus need  $Q = O\left(\varepsilon^{-\frac{n+1}{m-(n-1)/4}}\right)$  to meet the rates of Theorem 5.13. In particular, by dimensioning  $M$  and  $N$  according to Theorem 5.13 to achieve the desired precision  $\varepsilon$ , it is required that  $Q^4 \leq M^2 N^2$ . Therefore, the proposed algorithm enjoy the following performance:

1. Approximation error of  $(a^*, b^*)$ :  $O(\varepsilon)$ ,
2. Total number of required samples:  $O(\varepsilon^{-2 - \frac{n+1/2}{2m}})$ ,
3. Total computational cost:  $O(\varepsilon^{-4 - \frac{2n+1}{2m}} n^8)$ .

In particular, under regular enough settings, i.e.,  $m \geq n + 1/2$ , we infer that our identification method achieves approximation error  $O(\varepsilon)$  by requiring  $O(\varepsilon^{-2.5})$  samples and a total computational cost of  $O(\varepsilon^{-5}n^8)$ .

**7. Conclusion and Perspectives.** In this paper, we propose a Reproducing Kernel Hilbert Space-based learning paradigm for the identification of drift and diffusion coefficients of non-linear stochastic differential equations, which relies upon discrete-time observation of the state. Under assumptions of smoothness for the unknown drift and diffusion coefficients, we provide theoretical estimates of learning rates which become increasingly tighter when both the number of observations of the state and the regularity of the unknown drift and diffusion coefficients grow.

Some possible improvements and perspectives are in order. Since our learning rates essentially apply to the laws of the state process, it would be interesting to understand whether and under what conditions our method may be extended to derive stronger  $L^p$  norm-based learning rates. Finally, it would be interesting to investigate extensions of our work for the identification of controlled stochastic differential equations: although these models are crucial for the control of complex systems, e.g., in aerospace and robotics, methods which offer relevant guarantees of accuracy and efficiency of the identification process still require extensive investigation.

**8. Details on Stochastic Differential and Fokker-Planck Equations.** In this section, we provide a more structured exposition of the concepts we previously introduced in Section 3. For this, we chronologically retrace in more details every definition and result of Section 3 step by step.

**8.1. The Fokker-Planck equation.** We recall we fixed  $m \in \mathbb{N}$  and a constant  $\alpha > 0$ , and since we worked with diffusion coefficients which are never trivial, for every  $(a, b) \in \mathcal{H}_m^+$ , the coefficient  $a$  has been replaced with the mapping

$$a + \alpha I : [0, T] \times \mathbb{R}^n \rightarrow \text{Sym}_{++}(n) \quad \text{with} \quad a + \alpha I = \sqrt{a + \alpha I} \sqrt{a + \alpha I}.$$

Also, we fixed a non-negative density  $p_0 \in L^2(\mathbb{R}^n, \mathbb{R})$  which served as appropriate initial condition, and we denoted by  $\mu_0 \in \mathcal{P}(\mathbb{R}^n)$  the associate probability measure. We recall the following notions of stochastic differential equation and its solutions:

**DEFINITION 8.1.** *A measurable mapping  $X : \mathbb{R}^n \times \Omega \rightarrow S$  solves the Stochastic Differential Equation with coefficients  $(a, b) \in \mathcal{H}_m^+$  if each process  $X_x(t, \omega) \triangleq X(x, \omega)(t)$  is  $\mathcal{F}$ -progressively measurable for every  $x \in \mathbb{R}^n$ , and*

$$\text{SDE}_x \quad \begin{cases} dX_x(t) = b(t, X_x(t)) dt + \sqrt{(a + \alpha I)(t, X_x(t))} dW_t, \\ \mathbb{P}(X_x(0) = x) = 1, \end{cases}$$

*holds in  $(\Omega, \mathcal{G}, \mathcal{F}, \mathbb{P})$  for  $\mu_0$ -almost every  $x \in \mathbb{R}^n$ . A solution  $X$  to SDE is unique if, for every measurable mapping  $Y : \mathbb{R}^n \times \Omega \rightarrow S$  which solves SDE with coefficients  $(a, b) \in \mathcal{H}_m^+$ , it holds that  $X(x, \cdot) = Y(x, \cdot)$  a.s., for  $\mu_0$ -almost every  $x \in \mathbb{R}^n$ .*

The well-posedness of Definition 8.1 is contained in the following theorem, together with other useful properties on solutions to SDE:

**THEOREM 8.2.** *For every  $(a, b) \in \mathcal{H}_m^+$ , there exists a unique measurable mapping  $X : \mathbb{R}^n \times \Omega \rightarrow S$  which solves SDE with coefficients  $(a, b) \in \mathcal{H}_m^+$ . In addition, the following properties hold true for the mapping  $X$ :*

1. *The following mapping is measurable:*

$$(x, \omega, t) \in \mathbb{R}^n \times \Omega \times [0, T] \mapsto X_x(t, \omega) \in \mathbb{R}^n.$$

2. For every  $\varphi \in C_b([0, T] \times \mathbb{R}^n, \mathbb{R})$ , the following mapping is continuous:

$$t \in [0, T] \mapsto \int_{\mathbb{R}^n} \int_{\Omega} \varphi(t, X_x(t)) \, d\mathbb{P} \, \mu_0(dx) \in \mathbb{R}.$$

*Proof.* Thanks to Theorem 2.1, for every  $x \in \mathbb{R}^n$  there exists a unique (up to stochastic indistinguishability)  $\mathcal{F}$ -adapted process  $X_x : [0, T] \times \Omega \rightarrow \mathbb{R}^n$  with continuous sample paths which solves  $\text{SDE}_x$ , and which satisfies the following inequality

$$(8.1) \quad \mathbb{E}[\mathfrak{d}(X_{x_1}, X_{x_2})^2] \leq C(\|(a, b)\|_{\mathcal{H}_m^+}) \|x_1 - x_2\|^2,$$

where the constant  $C(\|(a, b)\|_{\mathcal{H}_m^+}) > 0$  depends on  $(a, b) \in \mathcal{H}_m^+$  uniquely (see, e.g., [22]).<sup>2</sup> Therefore, we define the mapping  $X : \mathbb{R}^n \times \Omega \rightarrow S$  by  $X(x, \omega)(t) \triangleq X_x(t, \omega)$ . From the continuity of the sample paths of each  $X_x$ , one may show that

$$X(x, \cdot)^{-1}(\overline{B_\varepsilon^S(w_0)}) = \bigcap_{t \in [0, T] \cap \mathbb{Q}} X_x(t)^{-1}(\overline{B_\varepsilon^{\mathbb{R}^n}(w_0(t))}), \quad x \in \mathbb{R}^n,$$

from which, by leveraging a routine monotone class argument, we easily infer the measurability of the process  $(X(x, \cdot))_{x \in \mathbb{R}^n}$ .

To prove the measurability of  $X : \mathbb{R}^n \times \Omega \rightarrow S$ , we rather build a measurable mapping  $\tilde{X} : \mathbb{R}^n \times \Omega \rightarrow S$  such that the process  $(\tilde{X}(x, \cdot))_{x \in \mathbb{R}^n}$  is a modification of  $(X(x, \cdot))_{x \in \mathbb{R}^n}$ . In particular, such property would imply for every  $x \in \mathbb{R}^n$  the existence of a subset  $N_x \in \mathcal{F}_T$  with  $\mathbb{P}(N_x) = 1$ , and such that

$$(8.2) \quad \tilde{X}_x(t, \omega) = X_x(t, \omega), \quad t \in [0, T], \quad \omega \in N_x.$$

From (8.2), together with the completeness of  $\mathcal{F}$ , we would infer each process  $\tilde{X}_x$  is  $\mathcal{F}$ -adapted and has continuous sample paths (the latter property being trivially true by definition). Moreover, thanks to Theorem 2.1, a routine application of Burkholder-Davis-Gundy inequality, and the fact that each process  $X_x$  satisfies  $\text{SDE}_x$ , one may straightforwardly compute, for every  $x \in \mathbb{R}^n$ ,

$$\begin{aligned} & \mathbb{E} \left[ \sup_{t \in [0, T]} \left\| \tilde{X}_x(t) - x - \int_0^t b(s, \tilde{X}_x(s)) \, ds - \int_0^t \sqrt{(a + \alpha I)(s, \tilde{X}_x(s))} \, dW_s \right\|^2 \right] \leq \\ & \leq C\mathbb{E} \left[ \sup_{t \in [0, T]} \|\tilde{X}_x(t) - X_x(t)\|^2 \right] + C\mathbb{E} \left[ \left( \int_0^T \|b(t, \tilde{X}_x(t)) - b(t, X_x(t))\| \, dt \right)^2 \right] \\ & \quad + C\mathbb{E} \left[ \int_0^T \left\| \sqrt{(a + \alpha I)(t, \tilde{X}_x(t))} - \sqrt{(a + \alpha I)(t, X_x(t))} \right\|^2 \, dt \right] \\ & \leq C\mathbb{E}[\mathfrak{d}(\tilde{X}_x, X_x)^2], \end{aligned}$$

for some (overloaded) constant  $C > 0$ , and therefore by combining (8.1) with (8.2) yields that each process  $\tilde{X}_x$  satisfies  $\text{SDE}_x$ ,  $x \in \mathbb{R}^n$ . In turn, we showed the existence of

<sup>2</sup>Here, we use the fact that the mapping  $y \mapsto \sqrt{a(t, y) + \alpha I}$  is Lipschitz, for every  $t \in [0, T]$ . This is a straightforward consequence of Theorem 2.1 and the fact that the mapping  $A \in \text{Sym}_{++}(n) \mapsto \sqrt{A}$  is Lipschitz on  $\{A \in \text{Sym}_{++}(n) : \|A\| \leq \|a\|_{L^\infty} + \alpha, \, y^\top A y \geq \alpha \|y\|^2, \, y \in \mathbb{R}^n\}$ . Indeed, the latter set is compact and the mapping  $A \in \text{Sym}_{++}(n) \mapsto \sqrt{A}$  is continuously differentiable.

a measurable mapping  $X : \mathbb{R}^n \times \Omega \rightarrow S$  which solves SDE with coefficients  $(a, b) \in \mathcal{H}_m^+$ . By leveraging the same argument (and a routine application of Gronwall's inequality), one also shows the uniqueness of this mapping as in Definition 8.1.

At this step, we build the aforementioned mapping  $\tilde{X} : \mathbb{R}^n \times \Omega \rightarrow S$  via Kolmogorov's lemma. More precisely, combining (8.1) with Kolmogorov's lemma yields the existence of a modification  $(\tilde{X}(x, \cdot))_{x \in \mathbb{R}^n}$  of  $(X(x, \cdot))_{x \in \mathbb{R}^n}$  such that, for every  $\omega \in \Omega$ , the mapping  $x \in \mathbb{R}^n \mapsto \tilde{X}_x(\cdot, \omega) \in S$  is continuous. Therefore, the mapping  $\tilde{X} : \mathbb{R}^n \times \Omega \rightarrow S$  is Caratheodory, and thus measurable (see, e.g., [4, Lemma 8.2.6]).

At this step, note that the mapping

$$\pi : [0, T] \times S \rightarrow \mathbb{R}^n : (t, w) \mapsto w(t)$$

satisfies

$$\|\pi(t_1, w_1) - \pi(t_2, w_2)\| \leq \|w_1(t_1) - w_1(t_2)\| + \mathfrak{d}(w_1, w_2),$$

for every  $t_1, t_2 \in [0, T]$ ,  $w_1, w_2 \in S$ . Hence,  $\pi$  is continuous and property 1. follows from  $X_x(t, \omega) = \pi(t, X(x, \omega))$ , for  $(x, \omega, t) \in \mathbb{R}^n \times \Omega \times [0, T]$ . Moreover, if  $\varphi \in C_b([0, T] \times \mathbb{R}^n, \mathbb{R})$  and  $(t_k)_{k \in \mathbb{N}} \subseteq [0, T]$  satisfies  $t_k \rightarrow t$ , from what we just proved:

$$\varphi(t_n, \pi(t_n, X(x, \omega))) \rightarrow \varphi(t, \pi(t, X(x, \omega))), \quad \text{a.e. in } \mathbb{R}^n \times \Omega,$$

and a routine application of the dominated convergence theorem yields property 2.  $\square$

At this step, for every  $(a, b) \in \mathcal{H}_m^+$ , we denote the Kolmogorov generator by

$$\mathcal{L}_t^{a,b} \varphi(y) \triangleq \frac{1}{2} \sum_{i,j=1}^n (a + \alpha I)_{ij}(t, y) \frac{\partial^2 \varphi}{\partial y_i \partial y_j}(y) + \sum_{i=1}^n b_i(t, y) \frac{\partial \varphi}{\partial y_i}(y), \quad \varphi \in C^2(\mathbb{R}^n, \mathbb{R}).$$

Fix  $(a, b) \in \mathcal{H}_m^+$ , and assume we are given a measurable mapping  $X : \mathbb{R}^n \times \Omega \rightarrow S$  which solves SDE with coefficients  $(a, b) \in \mathcal{H}_m^+$ . For  $\varphi \in C_c^\infty(\mathbb{R}^n, \mathbb{R})$ , a straightforward application of Itô's formula to  $\text{SDE}_x$  yields

$$(8.3) \quad \int_{\Omega} \varphi(X_x(t)) \, d\mathbb{P} = \varphi(x) + \int_{\Omega} \int_0^t \mathcal{L}_s^{a,b} \varphi(X_x(s)) \, ds \, d\mathbb{P}, \quad t \in [0, T],$$

which holds for  $\mu_0$ -almost every  $x \in \mathbb{R}^n$ . Thanks to Theorem 8.2, we may define the curve  $\mu : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  of probability measures

$$\mu_t(A) \triangleq \int_{\mathbb{R}^n} \int_{\Omega} \mathbf{1}_{\{X_x(t) \in A\}} \, d\mathbb{P} \, \mu_0(dx), \quad A \in \mathcal{B}(\mathbb{R}^n),$$

and note that  $\mu$  is narrowly continuous, i.e., for every  $\varphi \in C_b(\mathbb{R}^n, \mathbb{R})$ , the mapping

$$t \in [0, T] \mapsto \int_{\mathbb{R}^n} \varphi(y) \mu_t(dy) = \int_{\mathbb{R}^n} \int_{\Omega} \varphi(X_x(t)) \, d\mathbb{P} \, \mu_0(dx) \in \mathbb{R}$$

is continuous. By combining this latter property with (8.3) and Theorem 8.2, one readily checks that the curve of probabilities  $\mu$  satisfies:

**DEFINITION 8.3.** *A narrowly continuous curve  $\mu : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  is said to solve the Fokker-Planck Equation with coefficients  $(a, b) \in \mathcal{H}_m^+$  if*

$$\text{FPE} \quad \begin{cases} \frac{d}{dt} \int_{\mathbb{R}^n} \varphi(y) \mu_t(dy) = \int_{\mathbb{R}^n} \mathcal{L}_t^{a,b} \varphi(y) \mu_t(dy), & \varphi \in C_c^\infty(\mathbb{R}^n, \mathbb{R}), \\ \mu_{t=0} = \mu_0. \end{cases}$$

Importantly, thanks to the regularity of the coefficients  $(a, b) \in \mathcal{H}_m^+$ , FPE can have one narrowly continuous solution  $\mu$  at most, as we state in the following:

**PROPOSITION 8.4** (Propositions 4.1 and 4.2 in [14]). *Given any tuple of coefficients  $(a, b) \in \mathcal{H}_m^+$ , at most one narrowly continuous curve  $\mu : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  can solve FPE with coefficients  $(a, b) \in \mathcal{H}_m^+$ .*

Our previous computations show that solutions to FPE may be obtained from solutions to SDE, and we now establish this process may be inverted. Specifically, motivated by the results in [43, 14], we prove that any narrowly continuous curve  $\mu : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  solution to FPE with coefficients  $(a, b) \in \mathcal{H}_m^+$  is associated with a unique (in the sense of Definition 8.1) measurable mapping  $X : \mathbb{R}^n \times \Omega \rightarrow S$  which solves SDE with coefficients  $(a, b) \in \mathcal{H}_m^+$ . We gather such result in the following theorem, which is a natural extension of [14, Theorem 2.6] to our setting:

**THEOREM 8.5.** *Let a narrowly continuous curve  $\mu : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  be solution to FPE with coefficients  $(a, b) \in \mathcal{H}_m^+$  (which is unique thanks to Proposition 8.4). There exists a unique measurable mapping  $X : \mathbb{R}^n \times \Omega \rightarrow S$  which solves SDE with coefficients  $(a, b) \in \mathcal{H}_m^+$ , and which satisfies the representation formula:*

$$\int_{\mathbb{R}^n} \varphi(y) \mu_t(dy) = \int_{\mathbb{R}^n} \int_{\Omega} \varphi(X_x(t)) \, d\mathbb{P} \, \mu_0(dx), \quad \text{for } t \in [0, T], \quad \varphi \in C_c(\mathbb{R}^n, \mathbb{R}).$$

*Proof.* Thanks to Theorem 8.2, we already know there exists a unique measurable mapping  $X : \mathbb{R}^n \times \Omega \rightarrow S$  which solves SDE with coefficients  $(a, b) \in \mathcal{H}_m^+$ .

To conclude we just need to show the representation formula holds true. For this, we define the curve  $\bar{\mu} : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  of probability measures

$$\bar{\mu}_t(A) \triangleq \int_{\mathbb{R}^n} \int_{\Omega} \mathbf{1}_{\{X_x(t) \in A\}} \, d\mathbb{P} \, \mu_0(dx), \quad A \in \mathcal{B}(\mathbb{R}^n),$$

which is well-defined and narrowly continuous thanks to Theorem 8.2. In addition, by combining this latter property with (8.3) and Theorem 8.2, we see that the curve  $\bar{\mu} : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  solves FPE, and thus Proposition 8.4 yields

$$\int_{\mathbb{R}^n} \varphi(y) \mu_t(dy) = \int_{\mathbb{R}^n} \varphi(y) \bar{\mu}_t(dy) = \int_{\mathbb{R}^n} \int_{\Omega} \varphi(X_x(t)) \, d\mathbb{P} \, \mu_0(dx),$$

for every  $t \in [0, T]$  and  $\varphi \in C_c(\mathbb{R}^n, \mathbb{R})$ , and the conclusion follows.  $\square$

### 8.2. Absolutely continuous solutions to the Fokker-Planck equation.

From what we showed, solutions to SDE may be found by solving FPE. In this section, we show the existence of narrowly continuous curves  $\mu : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  of type

$$\mu_t(A) = \int_A p(t, y) \, dy, \quad A \in \mathcal{B}(\mathbb{R}^n),$$

for appropriate densities  $p : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$ , which are solutions to FPE. Note that, if such solutions to FPE exist, then they are unique thanks to Proposition 8.4. Although such existence result is classic (see, e.g., [14, 8]), we retrace its proof in Appendix A to characterize the constants appearing in some appropriate estimates and regularity properties which have been paramount to derive the results in Section 5.

For this, let us first recall the broader definition of solution to FPE, which in particular encompasses Definition 8.3 as a sub-case (see also Theorem 8.7 below):



DEFINITION 8.6. Let  $f \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$  and  $\bar{p} \in L^2(\mathbb{R}^n, \mathbb{R})$ . A (regular enough) function  $p : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$  is said to solve the non-homogeneous Fokker-Planck Equation with coefficients  $(a, b) \in \mathcal{H}_m^+$ , if

$$\text{FPE}_f \quad \begin{cases} \frac{d}{dt} \int_{\mathbb{R}^n} \varphi(y) p(t, y) dy = \\ = \int_{\mathbb{R}^n} \left( \mathcal{L}_t^{a,b} \varphi(y) p(t, y) + f(t, y) \varphi(y) \right) dy, & \varphi \in C_c^\infty(\mathbb{R}^n, \mathbb{R}), \\ p(0, \cdot) = \bar{p}(\cdot). \end{cases}$$

We gather results on the existence, uniqueness, and energy-type estimates for solutions to  $\text{FPE}_f$ , and therefore for solutions to FPE, in the following:

THEOREM 8.7. For every  $f \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$  and every  $\bar{p} \in L^2(\mathbb{R}^n, \mathbb{R})$ , there exists a unique mapping  $p \in C(0, T; L^2(\mathbb{R}^n, \mathbb{R})) \cap L^2(0, T; H^1(\mathbb{R}^n, \mathbb{R}))$  which solves  $\text{FPE}_f$  with coefficients  $(a, b) \in \mathcal{H}_m^+$ , and with  $\frac{\partial p}{\partial t} \in L^2(0, T; H^{-1}(\mathbb{R}^n, \mathbb{R}))$ . In addition, the following first parabolic estimate holds:

$$(8.4) \quad \begin{aligned} \|p(t, \cdot)\|_{L^2}^2 + \int_0^t \|p(s, \cdot)\|_{H^1}^2 ds \leq \\ \leq C(\|(a, b)\|_{\mathcal{H}_m}) \left( \|\bar{p}\|_{L^2}^2 + \int_0^t \|f(s, \cdot)\|_{L^2}^2 ds \right), \quad t \in [0, T], \end{aligned}$$

where  $C(\|(a, b)\|_{\mathcal{H}_m}) > 0$  is a constant which continuously depends on  $\|(a, b)\|_{\mathcal{H}_m}$ . Finally, if  $f = 0$  and  $\bar{p}$  is a non-negative density in  $L^2(\mathbb{R}^n, \mathbb{R})$ , then

$$p(t, \cdot) \geq 0, \quad \int_{\mathbb{R}^n} p(t, y) dy = 1, \quad t \in [0, T],$$

and therefore, if for every  $t \in [0, T]$  we define

$$\mu_t(A) \triangleq \int_A p(t, y) dy, \quad A \in \mathcal{B}(\mathbb{R}^n),$$

then the curve  $\mu : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  is narrowly continuous and solves FPE.

Remark 8.8. The regularity of the mapping  $f$  may be weakened (see, e.g., [25, 8]), although the requirement  $f \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$  already fits our purpose.

The proof of Theorem 8.7 is reported in Appendix A and is based on the classical Lions scheme (see, e.g., [25, 10, 13]). Among straightforward benefits, we recall Theorem 8.7 enables introducing rigorous criteria to establish satisfactory guarantees for our learning approach in many circumstances. Let us better introduce this concept below. As a matter of example, when dealing with stochastic differential equations in applications such as observation and regulation, one must often manipulate metrics

$$(8.5) \quad \mathbb{E}_{\mu_0 \times \mathbb{P}} \left[ \int_0^T f(t, X_x(t)) dt \right] = \int_0^T \left( \int_{\mathbb{R}^{2n}} f(t, y) \mathbb{P}_{X_x(t)}(dy) \mu_0(dx) \right) dt,$$

where the measurable mapping  $X : \mathbb{R}^n \times \Omega \rightarrow S$  solves SDE for some coefficients  $(a, b) \in \mathcal{H}_m^+$ , and the mapping  $f : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$  is regular enough. In the case the coefficients  $(a, b) \in \mathcal{H}_m^+$  of SDE are to be learned, one does not have perfect knowledge of (8.5), and therefore the error between (8.5) and its counterpart in which  $(X_x)_{x \in \mathbb{R}^n}$  is replaced with the solution to SDE stemming from rather learned coefficients must be estimated. This gap is filled with the following:

COROLLARY 8.9. *For any coefficients  $(a, b) \in \mathcal{H}_m^+$ , let  $X^{a,b} : \mathbb{R}^n \times \Omega \rightarrow S$  denote the unique measurable mapping which solves SDE with coefficients  $(a, b) \in \mathcal{H}_m^+$  (see Theorem 8.5). There exists an operator*

$$\begin{aligned} \mathcal{O} : \mathcal{H}_m^+ &\rightarrow L^\infty(0, T; L^2(\mathbb{R}^n, \mathbb{R})^*) \cong L^\infty(0, T; L^2(\mathbb{R}^n, \mathbb{R})) \\ (a, b) &\mapsto \mathcal{O}^{a,b}(\cdot) \end{aligned}$$

which, for every  $t \in [0, T]$  and every  $\varphi \in C_c(\mathbb{R}^n, \mathbb{R})$ , satisfies

$$\mathcal{O}_t^{a,b}(\varphi) = \int_{\mathbb{R}^{2n}} \varphi(y) \mathbb{P}_{X_x^{a,b}(t)}(dy) \mu_0(dx).$$

In addition, if  $p_{a,b} \in C(0, T; L^2(\mathbb{R}^n, \mathbb{R}))$  denotes the unique solution to FDE<sub>0</sub> with coefficients  $(a, b) \in \mathcal{H}_m^+$  and  $\bar{p} = p_0$  (which uniquely exists thanks to Theorem 8.7), for every  $f \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$  the mapping

$$t \in [0, T] \mapsto \mathcal{O}_t^{a,b}(f(t, \cdot)) \in \mathbb{R}$$

is measurable and in  $L^1([0, T], \mathbb{R})$ , and it satisfies, for every  $(a_1, b_1), (a_2, b_2) \in \mathcal{H}_m^+$ ,

$$(8.6) \quad \int_0^T |\mathcal{O}_t^{a_1, b_1}(f(t, \cdot)) - \mathcal{O}_t^{a_2, b_2}(f(t, \cdot))| dt \leq \|f\|_{L^2} \|p_{a_1, b_1} - p_{a_2, b_2}\|_{L^2}.$$

Remark 8.10. Thanks to the previous Corollary, we may provide (8.5) with a rigorous meaning by defining, for every  $f \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$ ,

$$\mathbb{E}_{\mu_0 \times \mathbb{P}} \left[ \int_0^T f(t, X_x^{a,b}(t)) dt \right] \triangleq \int_0^T \mathcal{O}_t^{a,b}(f(t, \cdot)) dt.$$

As we showed in Section 5, the estimate (8.6) enables the control of any estimation error occurring during the computation of the metric (8.5). Note that more refined estimates than (8.6) may be easily obtained by requiring more regularity on the function  $f$  (e.g., see the proof of Corollary 8.9 below).

*Proof.* Fix any  $(a, b) \in \mathcal{H}_m^+$ . For  $t \in [0, T]$  and  $\varphi \in L^2(\mathbb{R}^n, \mathbb{R})$ , we define

$$\mathcal{O}_t^{a,b}(\varphi) = \lim_{k \rightarrow \infty} \int_{\mathbb{R}^{2n}} \varphi_k(y) \mathbb{P}_{X_x^{a,b}(t)}(dy) \mu_0(dx),$$

where  $(\varphi_k)_{k \in \mathbb{N}} \subseteq C_c(\mathbb{R}^n, \mathbb{R})$  is any sequence which converges to  $\varphi$  for the strong topology of  $L^2(\mathbb{R}^n, \mathbb{R})$ . This definition is well-posed thanks to Theorem 8.5: indeed, the limit above uniquely exists, given that if  $(\varphi_k^1)_{k \in \mathbb{N}}, (\varphi_k^2)_{k \in \mathbb{N}} \subseteq C_c(\mathbb{R}^n, \mathbb{R})$  are two sequences which converge to  $\varphi$  for the strong topology of  $L^2(\mathbb{R}^n, \mathbb{R})$ , we may compute

$$\begin{aligned} \left| \int_{\mathbb{R}^{2n}} \varphi_k^1(y) \mathbb{P}_{X_x^{a,b}(t)}(dy) \mu_0(dx) - \int_{\mathbb{R}^{2n}} \varphi_k^2(y) \mathbb{P}_{X_x^{a,b}(t)}(dy) \mu_0(dx) \right| &\leq \\ &\leq \|p_{a,b}(t, \cdot)\|_{L^2} \|\varphi_k^1 - \varphi_k^2\|_{L^2} \rightarrow 0, \quad k \rightarrow \infty. \end{aligned}$$

Similarly, one may easily prove that  $\mathcal{O}_t^{a,b}(\cdot) \in L^2(\mathbb{R}^n, \mathbb{R})^* \cong L^2(\mathbb{R}^n, \mathbb{R})$  for every  $t \in [0, T]$ . In particular, note that thanks to (8.4), for  $t \in [0, T]$  and  $\varphi \in L^2(\mathbb{R}^n, \mathbb{R})$ ,

$$(8.7) \quad |\mathcal{O}_t^{a,b}(\varphi)| \leq C(a, b) \|p_0\|_{L^2} \|\varphi\|_{L^2},$$

for some appropriate constant  $C(a, b) > 0$ .

Finally, let  $g \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$  and let  $(g_k)_{k \in \mathbb{N}} \subseteq C_c([0, T] \times \mathbb{R}^n, \mathbb{R})$  converge to  $g$  for the strong topology of  $L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$ . In particular, up to some subsequence, we infer that each sequence  $(g_k(t, \cdot))_{k \in \mathbb{N}} \subseteq C_c([0, T] \times \mathbb{R}^n, \mathbb{R})$  converges to  $g(t, \cdot) \in L^2(\mathbb{R}^n, \mathbb{R})$  for the strong topology of  $L^2(\mathbb{R}^n, \mathbb{R})$ , for almost every  $t \in [0, T]$ . Hence, the definition of  $\mathcal{O}_t^{a,b}(\cdot)$  and Theorem 8.5 yield

$$(8.8) \quad \mathcal{O}_t^{a,b}(g(t, \cdot)) = \lim_{k \rightarrow \infty} \int_{\mathbb{R}^n} g_k(t, y) p_{a,b}(t, y) dy, \quad \text{for almost every } t \in [0, T].$$

On the one hand, combined with Theorem 8.7 and Pettis' theorem, (8.8) yields the Bochner-measurability of the mapping  $t \in [0, T] \mapsto \mathcal{O}_t^{a,b}(\cdot) \in L^2(\mathbb{R}^n, \mathbb{R})^*$ , and therefore from (8.7) we deduce that  $\mathcal{O}^{a,b}(\cdot) \in L^\infty(0, T; L^2(\mathbb{R}^n, \mathbb{R})^*)$ . On the other hand, (8.8) yields the measurability of the mapping  $t \in [0, T] \mapsto \mathcal{O}_t^{a,b}(f(t, \cdot)) \in \mathbb{R}$ , which together with (8.7) makes the mapping  $t \in [0, T] \mapsto \mathcal{O}_t^{a,b}(f(t, \cdot)) \in \mathbb{R}$  measurable and in  $L^1([0, T], \mathbb{R})$ . The conclusion of the proof follows from the fact that (8.6) is a straightforward consequence of (8.8) and Fatou's lemma.  $\square$

### 8.3. Additional regularity of solutions to the Fokker-Planck equation.

In Section 5, we dealt with solutions  $p : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$  of  $\text{FPE}_0$  which enjoy higher regularity properties, and specifically for which  $p \in H^{m+1,2(m+1)}([0, T] \times \mathbb{R}^n, \mathbb{R})$ , and for which  $\text{FPE}_0$  holds pointwise. Below, we are going to show such additional attributes stem from the properties of coefficients  $(a, b) \in \mathcal{H}_m^+$  listed in Theorem 2.1.

First, for every  $(a, b) \in \mathcal{H}_m^+$  and almost every  $t \in [0, T]$  we introduce the notation:

$$(\mathcal{L}_t^{a,b})^* u(t, y) \triangleq \frac{1}{2} \sum_{i,j=1}^n \frac{\partial^2 (a_{ij}^\alpha u)}{\partial y_j \partial y_i}(t, y) - \sum_{i=1}^n \frac{\partial (b_i u)}{\partial y_i}(t, y), \quad u \in L^2(0, T; H^2(\mathbb{R}^n, \mathbb{R})),$$

which denotes the dual operator of  $\mathcal{L}_t^{a,b}$ . This operator is well-defined with image in  $L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$ . Indeed, thanks to Theorem 2.1, one readily shows that

$$\begin{aligned} b_i u &\in L^2(0, T; H^1(\mathbb{R}^n, \mathbb{R})), \quad \text{with} \quad \frac{\partial (b_i u)}{\partial y_r} = b_i \frac{\partial u}{\partial y_r} + u \frac{\partial b_i}{\partial y_r} \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R}), \\ a_{ij} u &\in L^2(0, T; H^2(\mathbb{R}^n, \mathbb{R})), \quad \text{with} \quad \frac{\partial (a_{ij} u)}{\partial y_r} = a_{ij} \frac{\partial u}{\partial y_r} + u \frac{\partial a_{ij}}{\partial y_r} \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R}), \\ &\text{and} \quad \frac{\partial^2 (a_{ij}^\alpha u)}{\partial y_r \partial y_s} = a_{ij}^\alpha \frac{\partial^2 u}{\partial y_r \partial y_s} + \frac{\partial a_{ij}^\alpha}{\partial y_s} \frac{\partial u}{\partial y_r} \\ &\quad + \frac{\partial a_{ij}^\alpha}{\partial y_r} \frac{\partial u}{\partial y_s} + u \frac{\partial^2 a_{ij}^\alpha}{\partial y_r \partial y_s} \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R}), \end{aligned}$$

for  $u \in L^2(0, T; H^2(\mathbb{R}^n, \mathbb{R}))$ ,  $i, j, r, s = 1, \dots, n$ , thus the well-posedness of  $(\mathcal{L}_t^{a,b})^*$ .

Thanks to Theorem 2.1, the following higher regularity result holds:

**THEOREM 8.11.** *Under the setting and notation of Theorem 8.7, if furthermore  $\bar{p} \in H^{2m+1}(\mathbb{R}^n, \mathbb{R})$ , then the unique solution  $p : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$  to  $\text{FPE}_0$  with coefficients  $(a, b) \in \mathcal{H}_m^+$  is additionally such that*

$$p \in H^{m+1}(0, T; H^{2(m+1)}(\mathbb{R}^n, \mathbb{R})).$$

In particular, the mapping  $p$  additionally satisfies the following Strong Fokker-Planck Equations with coefficients  $(a, b) \in \mathcal{H}_m^+$ :

$$\text{SFPE} \quad \begin{cases} \frac{\partial p}{\partial t}(t, y) = (\mathcal{L}_t^{a, b})^* p(t, y), & \text{a.e. } (t, y) \in [0, T] \times \mathbb{R}^n, \\ p(0, \cdot) = \bar{p}(\cdot). \end{cases}$$

*Remark 8.12.* This result is classically proved under either time-independent coefficients or bounded domains (e.g., [13, 26]). Although in classical reference which consider time-dependent coefficients such as [26] it is explicitly mentioned therein that their results may be extended to unbounded domains, we found such extension non-straightforward. We therefore decided to revisit and extend the regularity results for parabolic equations which are contained in [13, Section 6.3] to time-dependent coefficients and unbounded domains, and to report the proof in Appendix A.

## Appendix A. Proofs of Sections 2 and 8.

### A.1. Proofs of Section 2.

*Proof of Theorem 2.1.* The first and second statements in Theorem 2.1 stem from Morrey theorem as soon as we show the existence of a linear and bounded operator:

$$E : H^1([0, T] \times \mathbb{R}^n, \mathbb{R}) \rightarrow H^1(\mathbb{R}^{n+1}, \mathbb{R}),$$

for which there exists a constant  $C > 0$  such that:

- $Eu|_{[0, T] \times \mathbb{R}^n} = u$ , for every  $u \in H^1([0, T] \times \mathbb{R}^n, \mathbb{R})$ ,
- $\|Eu\|_{L^2} \leq C\|u\|_{L^2}$  and  $\|Eu\|_{H^1} \leq C\|u\|_{H^1}$ , for every  $u \in H^1([0, T] \times \mathbb{R}^n, \mathbb{R})$ ,

for instance, see [9]. For this, we may follow the proof of [9, Theorem 8.6]. More specifically, let  $\eta \in C^\infty(\mathbb{R}, [0, 1])$  be such that

$$\eta(t) = \begin{cases} 1, & t < \frac{T}{4}, \\ 0, & t > \frac{3}{4}T, \end{cases}$$

and for every  $u : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$  define the mappings

$$u^-(t, y) \triangleq \begin{cases} u(t, y), & 0 < t \leq T, \\ 0, & t \leq 0 \end{cases} \quad \text{and} \quad u^+(t, y) \triangleq \begin{cases} u(t, y), & 0 \leq t < T, \\ 0, & t \geq T. \end{cases}$$

If  $u \in H^1([0, T] \times \mathbb{R}^n, \mathbb{R})$ , one sees that  $\eta u^+ \in H^1([0, \infty) \times \mathbb{R}^n, \mathbb{R})$  with

$$\frac{\partial \eta u^+}{\partial t} = \eta \left( \frac{\partial u}{\partial t} \right)^+ + u^+ \frac{\partial \eta}{\partial t} \quad \text{and} \quad \frac{\partial \eta u^+}{\partial y_i} = \eta \left( \frac{\partial u}{\partial y_i} \right)^+, \quad i = 1, \dots, n,$$

and that  $(1 - \eta)u^- \in H^1((-\infty, T] \times \mathbb{R}^n, \mathbb{R})$  with, for every  $i = 1, \dots, n$ ,

$$\frac{\partial (1 - \eta)u^-}{\partial t} = (1 - \eta) \left( \frac{\partial u}{\partial t} \right)^- + u^- \frac{\partial \eta}{\partial t} \quad \text{and} \quad \frac{\partial (1 - \eta)u^-}{\partial y_i} = (1 - \eta) \left( \frac{\partial u}{\partial y_i} \right)^-.$$

At this step, since  $u = \eta u^+ + (1 - \eta)u^-$ , we first extend  $\eta u^+$  by reflection at  $t = 0$  through a mapping  $v^1 \in H^1(\mathbb{R}^{n+1}, \mathbb{R})$  which thus satisfies

$$\|v^1\|_{L^2} \leq C\|u\|_{L^2} \quad \text{and} \quad \|v^1\|_{H^1} \leq C\|u\|_{H^1},$$

and then we extend  $(1 - \eta)u^-$  by reflection at  $t = T$  through another mapping  $v^2 \in H^1(\mathbb{R}^{n+1}, \mathbb{R})$  which thus satisfies

$$\|v^2\|_{L^2} \leq C\|u\|_{L^2} \quad \text{and} \quad \|v^2\|_{H^1} \leq C\|u\|_{H^1},$$

where the constant  $C > 0$  only depends on  $\eta$ . The conclusion easily follows if we define the linear and bounded operator  $E$  as  $Eu \triangleq v^1 + v^2$ .  $\square$

**A.2. Proofs of Section 8.** Below, we will make use of the notation  $a^\alpha \triangleq a + \alpha I$ .

*Proof of Theorem 8.7.* The proof is standard and based on the classical Lions scheme (see, e.g., [25, 10, 13]), therefore we mainly focus on deriving the constant  $C(\|(a, b)\|_{\mathcal{H}_m}) = C(\alpha, \|(a, b)\|_{\mathcal{H}_m})$  in (8.4) and the properties of  $p$  when  $f = 0$  and  $\bar{p}$  is a non-negative density in  $L^2(\mathbb{R}^n, \mathbb{R})$ . For this, consider the Gelfand triple

$$V \triangleq H^1(\mathbb{R}^n, \mathbb{R}) = H_0^1(\mathbb{R}^n, \mathbb{R}) \hookrightarrow H \triangleq L^2(\mathbb{R}^n, \mathbb{R}) \hookrightarrow V^* = H^{-1}(\mathbb{R}^n, \mathbb{R})$$

and define the  $t$ -measurable bilinear form

$$\begin{aligned} A^{\alpha, a, b} : [0, T] \times V \times V &\rightarrow \mathbb{R} \\ (t; u, v) &\mapsto \sum_{i, j=1}^n \int_{\mathbb{R}^n} \frac{1}{2} a_{ij}^\alpha(t, y) \frac{\partial u}{\partial y_j}(y) \frac{\partial v}{\partial y_i}(y) \, dy \\ &\quad + \sum_{i=1}^n \int_{\mathbb{R}^n} \left( \sum_{j=1}^n \frac{1}{2} \frac{\partial a_{ij}^\alpha}{\partial y_j}(t, y) - b_i(t, y) \right) \frac{\partial v}{\partial y_i}(y) u(y) \, dy. \end{aligned}$$

Thanks to Theorem 2.1, one may prove that the form  $A^{\alpha, a, b}$  is continuous and semi-coercive (see also the computations for the estimate (8.4) below). Therefore, thanks to a straightforward modification of the proof of [10, Theorem 11.7], there exists a unique solution  $p \in C(0, T; H) \cap L^2(0, T; V)$  to the variational problem

$$(A.1) \quad \begin{cases} \frac{d}{dt}(p(t, \cdot), \varphi)_H + A^{\alpha, a, b}(t; p(t, \cdot), \varphi) = (f(t, \cdot), \varphi)_H, & \varphi \in V, \\ p(0, \cdot) = \bar{p}(\cdot), \end{cases}$$

which in addition satisfies  $\frac{\partial p}{\partial t} \in L^2(0, T; V^*)$ . At this step, thanks to Theorem 2.1

$$a_{ij}^\alpha(t, \cdot) p(t, \cdot) \in H^1(\mathbb{R}^n, \mathbb{R}), \quad \text{with} \quad \frac{\partial(a_{ij}^\alpha p)}{\partial y_j}(t, \cdot) = p(t, \cdot) \frac{\partial a_{ij}^\alpha}{\partial y_j}(t, \cdot) + a_{ij}^\alpha(t, \cdot) \frac{\partial p}{\partial y_j}(t, \cdot),$$

for every  $i, j = 1, \dots, n$  and almost every  $t \in [0, T]$ . Therefore, for every  $\varphi \in C_c^\infty(\mathbb{R}^n; \mathbb{R})$ ,  $i, j = 1, \dots, n$ , and almost every  $t \in [0, T]$ , it holds that

$$\begin{aligned} \int_{\mathbb{R}^n} a_{ij}^\alpha(t, y) p(t, y) \frac{\partial^2 \varphi}{\partial y_j \partial y_i}(y) \, dy &= \\ &= - \int_{\mathbb{R}^n} \left( a_{ij}^\alpha(t, y) \frac{\partial p}{\partial y_j}(t, y) + p(t, y) \frac{\partial a_{ij}^\alpha}{\partial y_j}(t, y) \right) \frac{\partial \varphi}{\partial y_i}(y) \, dy, \end{aligned}$$

showing that  $p$  solves  $\text{FPE}_f$  with coefficients  $(a, b) \in \mathcal{H}_m^+$ ,  $\alpha > 0$ .

Next, since  $p \in L^2(0, T; V)$  and  $\frac{\partial p}{\partial t} \in L^2(0, T; V^*)$ , from (A.1) we may compute

$$\frac{d}{dt} \frac{\|p(t, \cdot)\|_H^2}{2} = \left\langle \frac{\partial p}{\partial t}(t, \cdot), p(t, \cdot) \right\rangle_{V^*} = (f(t, \cdot), p(t, \cdot))_H - A^{\alpha, a, b}(t; p(t, \cdot), p(t, \cdot)),$$

and therefore Young's inequality yields

$$\begin{aligned} \frac{d}{dt} \frac{\|p(t, \cdot)\|_H^2}{2} &\leq \|f(t, \cdot)\|_H \|p(t, \cdot)\|_H - \frac{\alpha}{2} \|\nabla p(t, \cdot)\|_H^2 \\ &\quad + \bar{C}(a, b) \sum_{i=1}^n \int_{\mathbb{R}^n} \left| \frac{\partial p(t, y)}{\partial y_i} p(t, y) \right| dy \\ &\leq \frac{\|f(t, \cdot)\|_H^2}{2} + \frac{\|p(t, \cdot)\|_H^2}{2} - \frac{\alpha}{2} \|\nabla p(t, \cdot)\|_H^2 \\ &\quad + nM\bar{C}(a, b)^2 \frac{\|p(t, \cdot)\|_{L^2}^2}{2} + \frac{\|\nabla p(t, \cdot)\|_H^2}{2M} \\ &\leq (1 + nM\bar{C}(a, b)^2) \frac{\|p(t, \cdot)\|_H^2}{2} + \frac{\|f(t, \cdot)\|_H^2}{2} - \frac{\alpha}{4} \|\nabla p(t, \cdot)\|_H^2, \end{aligned}$$

where  $M > 2/\alpha$  is some large enough constant which stems from applying Young's inequality, whereas  $\bar{C}(a, b) > 0$  is a constant which stems from Theorem 2.1 and continuously depends on the  $L^\infty$  norms of  $b$  and of the derivatives of  $a$  uniquely, and a routine application of Gronwall's inequality together with Theorem 2.1 yield (8.4).

Finally, assume  $f = 0$  and that  $\bar{p}$  is a non-negative density in  $L^2(\mathbb{R}^n, \mathbb{R})$ . Thanks to the fact that  $u^+, u^- \in H^1(\mathbb{R}^n, \mathbb{R})$  for every  $u \in H^1(\mathbb{R}^n, \mathbb{R})$ , with  $\nabla(u^+) = \nabla u \mathbb{1}_{\{u > 0\}}$  and  $\nabla(u^-) = \nabla u \mathbb{1}_{\{u < 0\}}$ , and that  $u^- = (-u)^+$ , leveraging the notation we introduced previously and applying [10, Lemma 11.2] with  $u = -p$  yield

$$\begin{aligned} \frac{d}{dt} \frac{\|p(t, \cdot)^-\|_H^2}{2} &= - \left\langle \frac{\partial p}{\partial t}(t, \cdot), p(t, \cdot)^- \right\rangle_{V^*} \\ &= - \sum_{i,j=1}^n \int_{\mathbb{R}^n} \frac{1}{2} a_{ij}^\alpha(t, y) \frac{\partial(p^+ - p^-)}{\partial y_j}(t, y) \frac{\partial(-p^-)}{\partial y_i}(t, y) dy \\ &\quad + \sum_{i=1}^n \int_{\mathbb{R}^n} \left( \sum_{j=1}^n \frac{1}{2} \frac{\partial a_{ij}^\alpha}{\partial y_j}(t, y) - b_i(t, y) \right) \frac{\partial(p^+ - p^-)}{\partial y_i}(t, y) p(t, y)^- dy \\ &\leq -\frac{\alpha}{2} \|\nabla(p^-)(t, \cdot)\|_H^2 + \bar{C}(a, b) \sum_{i=1}^n \int_{\mathbb{R}^n} \left| \frac{\partial p^-}{\partial y_i}(t, y) p(t, y)^- \right| dy \\ &\leq -\frac{\alpha}{2} \|\nabla(p^-)(t, \cdot)\|_H^2 + nM\bar{C}(a, b)^2 \frac{\|p(t, \cdot)^-\|_H^2}{2} + \frac{\|\nabla(p^-)(t, \cdot)\|_H^2}{2M} \\ &\leq nM\bar{C}(a, b)^2 \frac{\|p(t, \cdot)^-\|_H^2}{2}, \end{aligned}$$

and thanks to a routine application of Gronwall's inequality, the fact that  $\bar{p}(x) \geq 0$  almost everywhere yields  $p(t, x) \geq 0$  for every  $t \in [0, T]$  and almost every  $x \in \mathbb{R}^n$ .

At this step, for every  $k \in \mathbb{N}$  choose a cut-off function  $\varphi_k \in C_c^\infty(\mathbb{R}^n, [0, 1])$  such that  $\varphi_k(x) = 1$  for every  $x \in \overline{B_k^{\mathbb{R}^n}}(0)$ ,  $\text{supp}(\varphi_k) \subseteq B_{2k}^{\mathbb{R}^n}(0)$ , and whose first and second derivatives are uniformly bounded. For every  $t \in [0, T]$ , from the definition

of  $\text{FPE}_f$  and the monotone and dominated convergence theorems (here, we leverage both Theorem 2.1 and the fact that  $p(t, \cdot) \in L^2(\mathbb{R}^n, \mathbb{R})$ ,  $t \in [0, T]$ ) we may compute

$$\begin{aligned}
\int_{\mathbb{R}^n} p(t, y) dy &= \lim_{k \rightarrow \infty} \int_{\mathbb{R}^n} p(t, y) \varphi_k(y) dy \\
&= \lim_{k \rightarrow \infty} \left( \int_{\mathbb{R}^n} \bar{p}(y) \varphi_k(y) dy + \sum_{i,j=1}^n \int_{\mathbb{R}^n} \frac{1}{2} a_{ij}^\alpha(t, y) \frac{\partial^2 \varphi_k}{\partial y_i \partial y_j}(y) p(t, y) dy \right. \\
&\quad \left. + \sum_{i=1}^n \int_{\mathbb{R}^n} b_i(t, y) \frac{\partial \varphi_k}{\partial y_i}(y) p(t, y) dy \right) \\
&= \int_{\mathbb{R}^n} \bar{p}(y) dy = 1.
\end{aligned}$$

To conclude, it is clear that we just need to prove that the curve  $\mu : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^n)$  defined in the statement of the theorem is narrowly continuous. For this, let  $t \in [0, T]$  and  $(t_k)_{k \in \mathbb{N}} \subseteq [0, T]$  such that  $t_k \rightarrow t$  for  $k \rightarrow \infty$ . Note that  $p \in C(0, T; H)$  implies that  $\|p(t_n, \cdot) - p(t, \cdot)\|_{L^2} \rightarrow 0$  for  $k \rightarrow \infty$ . In particular, for any function  $\varphi \in C_c(\mathbb{R}^n, \mathbb{R})$ , as soon as  $k \rightarrow \infty$  we infer that

$$\left| \int_{\mathbb{R}^n} \varphi(y) p(t_n, y) dy - \int_{\mathbb{R}^n} \varphi(y) p(t, y) dy \right| \leq \|\varphi\|_{L^2} \|p(t_n, \cdot) - p(t, \cdot)\|_{L^2} \rightarrow 0.$$

We conclude from the fact that the narrow and weak\* topologies coincide in  $\mathcal{P}(\mathbb{R}^n)$ .  $\square$

*Proof of Theorem 8.11.* Given that the proof is substantially long, for the sake of clarity we divide it in several step. Below, we adopt the notation we introduced and used in the proof of Theorem 8.7.

**1) A second parabolic estimate.** In this section, we provide computations by rather considering  $\text{SFPE}_f$  with  $0 \neq f \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$ . Let  $(v_i)_{i \in \mathbb{N}}$  be a countable basis of  $H^{2(m+1)}(\mathbb{R}^n, \mathbb{R})$  (and in turn of  $H^\ell(\mathbb{R}^n, \mathbb{R})$ , for  $\ell = 0, \dots, 2m+1$ ), such that  $(v_i)_{i \in \mathbb{N}}$  is orthonormal in  $H^1(\mathbb{R}^n, \mathbb{R})$ . Therefore, there is  $(\bar{x}_i)_{i \in \mathbb{N}} \subseteq \mathbb{R}$  such that

$$(\text{A.2}) \quad \lim_{j \rightarrow \infty} \sum_{i=1}^j \bar{x}_i v_i = \bar{p}, \quad \text{in } H^1(\mathbb{R}^n, \mathbb{R}).$$

Moreover, by revisiting the proof of [10, Theorem 11.7], one easily see that for every  $k \in \mathbb{N}$  there exists  $x^k \in AC([0, T], \mathbb{R}^k)$  such that the function

$$p_k(t, \cdot) \triangleq \sum_{i=1}^k x_i^k(t) v_i \in C(0, T; L^2(\mathbb{R}^n, \mathbb{R})) \cap L^2(0, T; H^1(\mathbb{R}^n, \mathbb{R}))$$

is the unique solution to the variational problem

$$(\text{A.3}) \quad \begin{cases} \frac{d}{dt}(p_k(t, \cdot), v_i)_{L^2} + A^{\alpha, a, b}(t; p_k(t, \cdot), v_i) = (f(t, \cdot), v_i)_H, & i = 1, \dots, k, \\ x_i^k(0) = \bar{x}_i, & i = 1, \dots, k, \end{cases}$$



which in addition satisfies (see also [10, Page 197])

$$(A.4) \quad \begin{cases} \frac{\partial p_k}{\partial t}(t, \cdot) = \sum_{i=1}^k \frac{d}{dt} x_i^k(t) v_i \in L^2(0, T; H^1(\mathbb{R}^n, \mathbb{R})), \\ \frac{\partial p_k}{\partial y_j}(t, \cdot) = \sum_{i=1}^k x_i^k(t) \frac{\partial v_i}{\partial y_j} \in AC(0, T; L^2(\mathbb{R}^n, \mathbb{R})), \quad j = 1, \dots, n, \\ p_k \rightarrow p \quad \text{strongly in } L^2(0, T; H^1(\mathbb{R}^n, \mathbb{R})). \end{cases}$$

Moreover, from the proof of Theorem 8.7, it is clear that each  $p_k$  satisfies (8.4).

At this step, by multiplying each equation in (A.3) by  $\frac{d}{dt} x_i^k(t)$  and summing those over  $i = 1, \dots, k$ , thanks to (A.4) for almost every  $t \in [0, T]$  we may compute

$$(A.5) \quad \begin{aligned} \left\| \frac{\partial p_k}{\partial t}(t, \cdot) \right\|_{L^2}^2 &= \sum_{i=1}^k \left( \frac{d}{dt} x_i^k(t) \right) \frac{d}{ds} \Big|_{s=t} (p_k(s, \cdot), v_i)_{L^2} \\ &= \sum_{i=1}^k \left( \frac{d}{dt} x_i^k(t) \right) \left( (f(t, \cdot), v_i)_{L^2} - A^{\alpha, a, b}(t; p_k(t, \cdot), v_i) \right) \\ &\leq \|f(t, \cdot)\|_{L^2} \left\| \frac{\partial p_k}{\partial t}(t, \cdot) \right\|_{L^2} - \sum_{i,j=1}^n \int_{\mathbb{R}^n} \frac{1}{2} a_{ij}^\alpha(t, y) \frac{\partial p_k}{\partial y_j}(t, y) \frac{d}{dt} \left( \frac{\partial p_k}{\partial y_i} \right)(t, y) dy \\ &\quad - \sum_{i=1}^n \int_{\mathbb{R}^n} \left( \sum_{j=1}^n \frac{1}{2} \frac{\partial a_{ij}^\alpha}{\partial y_j}(t, y) - b_i(t, y) \right) \frac{\partial}{\partial y_i} \left( \frac{\partial p_k}{\partial t} \right)(t, y) p_k(t, y) dy. \end{aligned}$$

We are going to bound the last two terms in (A.5). Specifically, for the first to the last term, thanks to Theorem 2.1 and (A.4) for every  $t \in [0, T]$  we obtain that

$$\begin{aligned} & - \int_0^t \int_{\mathbb{R}^n} \sum_{i,j=1}^n \frac{1}{2} a_{ij}^\alpha(s, y) \frac{\partial p_k}{\partial y_j}(s, y) \frac{d}{ds} \left( \frac{\partial p_k}{\partial y_i} \right)(s, y) dy ds = \\ & = - \int_0^t \int_{\mathbb{R}^n} \sum_{i,j=1}^n \frac{1}{2} \left( \frac{d}{ds} \left( a_{ij}^\alpha \frac{\partial p_k}{\partial y_j} \frac{\partial p_k}{\partial y_i} \right)(s, y) - \dot{a}_{ij}^\alpha(s, y) \frac{\partial p_k}{\partial y_j}(s, y) \frac{\partial p_k}{\partial y_i}(s, y) \right. \\ & \quad \left. - a_{ij}^\alpha(s, y) \frac{d}{ds} \left( \frac{\partial p_k}{\partial y_j} \right)(s, y) \frac{\partial p_k}{\partial y_i}(s, y) \right) dy ds, \end{aligned}$$

and since  $a_{ji}^\alpha = a_{ij}^\alpha$  for every  $i, j = 1, \dots, n$ , we may compute

$$\begin{aligned} & - \int_0^t \int_{\mathbb{R}^n} \sum_{i,j=1}^n a_{ij}^\alpha(s, y) \frac{\partial p_k}{\partial y_j}(s, y) \frac{d}{ds} \left( \frac{\partial p_k}{\partial y_i} \right)(s, y) dy ds \leq \\ & \leq - \sum_{i,j=1}^n \frac{1}{2} \int_{\mathbb{R}^n} \left( a_{ij}^\alpha(t, y) \frac{\partial p_k}{\partial y_j}(t, y) \frac{\partial p_k}{\partial y_i}(t, y) - a_{ij}^\alpha(0, y) \frac{\partial p_k}{\partial y_j}(0, y) \frac{\partial p_k}{\partial y_i}(0, y) \right) dy ds \\ & \quad + \frac{\|\dot{a}\|_{L^\infty}}{2} \int_0^t \int_{\mathbb{R}^n} \sum_{i,j=1}^n \left| \frac{\partial p_k}{\partial y_j}(s, y) \right| \left| \frac{\partial p_k}{\partial y_i}(s, y) \right| dy ds. \end{aligned}$$

Therefore, the convergence (A.2) and (8.4) finally yield

$$\begin{aligned}
 & - \int_0^t \int_{\mathbb{R}^n} \sum_{i,j=1}^n a_{ij}^\alpha(s, y) \frac{\partial p_k}{\partial y_j}(s, y) \frac{d}{ds} \left( \frac{\partial p_k}{\partial y_i}(s, y) \right) dy ds \leq \\
 (A.6) \quad & \leq -\frac{\alpha}{2} \|\nabla p_k(t, \cdot)\|_{L^2}^2 + C(\alpha, \|(a, b)\|_{\mathcal{H}_m}) \left( \|\bar{p}\|_{H^1}^2 + \int_0^t \|f(s, \cdot)\|_{L^2}^2 ds \right),
 \end{aligned}$$

where  $C(\alpha, \|(a, b)\|_{\mathcal{H}_m}) > 0$  is a constant which continuously depends on  $\alpha$  and  $\|(a, b)\|_{\mathcal{H}_m}$  uniquely. Below, we will implicitly overload the constant  $C(\alpha, \|(a, b)\|_{\mathcal{H}_m})$ .

We now focus on the last term in (A.5). For this, first thanks to Theorem 2.1 one obtains that, for every  $i = 1, \dots, n$ ,  $\varphi \in C_c^\infty(\mathbb{R}^n, \mathbb{R})$ , and almost every  $t \in [0, T]$ ,

$$\begin{aligned}
 & - \int_{\mathbb{R}^n} \left( \sum_{j=1}^n \frac{1}{2} \frac{\partial a_{ij}^\alpha}{\partial y_j}(t, y) - b_i(t, y) \right) \frac{\partial \varphi}{\partial y_i}(y) p_k(t, y) dy = \\
 & = \int_{\mathbb{R}^n} \frac{\partial}{\partial y_i} \left( \left( \sum_{j=1}^n \frac{1}{2} \frac{\partial a_{ij}^\alpha}{\partial y_j} - b_i \right) p_k \right) (t, y) \varphi(y) dy.
 \end{aligned}$$

Since from (A.4) we infer the existence of a sequence  $(\varphi_\ell)_{\ell \in \mathbb{N}} \subseteq C_c^\infty(\mathbb{R}^n, \mathbb{R})$  such that

$$\left\| \frac{\partial p_k}{\partial t}(t, \cdot) - \varphi_\ell \right\|_{H^1} \rightarrow 0, \text{ for } \ell \rightarrow \infty, \text{ again Theorem 2.1 and (8.4) yield}$$

$$\begin{aligned}
 & - \int_0^t \sum_{i=1}^n \int_{\mathbb{R}^n} \left( \sum_{j=1}^n \frac{1}{2} \frac{\partial a_{ij}^\alpha}{\partial y_j}(s, y) - b_i(s, y) \right) \frac{\partial}{\partial y_i} \left( \frac{\partial p_k}{\partial s}(s, y) \right) p_k(s, y) dy ds = \\
 & = \int_0^t \sum_{i=1}^n \int_{\mathbb{R}^n} \left( \sum_{j=1}^n \frac{1}{2} \frac{\partial^2 a_{ij}^\alpha}{\partial y_i \partial y_j}(s, y) - \frac{\partial b_i}{\partial y_i}(s, y) \right) \frac{\partial p_k}{\partial s}(s, y) p_k(s, y) dy ds \\
 & \quad + \int_0^t \sum_{i=1}^n \int_{\mathbb{R}^n} \left( \sum_{j=1}^n \frac{1}{2} \frac{\partial a_{ij}^\alpha}{\partial y_j}(s, y) - b_i(s, y) \right) \frac{\partial p_k}{\partial y_i}(s, y) \frac{\partial p_k}{\partial s}(s, y) dy ds \\
 & \leq C(\alpha, \|(a, b)\|_{\mathcal{H}_m}) \int_0^t \int_{\mathbb{R}^n} \left( |p_k(s, y)| + \sum_{i=1}^n \left| \frac{\partial p_k}{\partial y_i}(s, y) \right| \right) \left| \frac{\partial p_k}{\partial s}(s, y) \right| dy ds \\
 & \leq \frac{1}{2} \int_0^t \left\| \frac{\partial p_k}{\partial s}(s, \cdot) \right\|_{L^2}^2 ds + C(\alpha, \|(a, b)\|_{\mathcal{H}_m}) \left( \|\bar{p}\|_{H^1}^2 + \int_0^t \|f(s, \cdot)\|_{L^2}^2 ds \right),
 \end{aligned}$$

for every  $t \in [0, T]$ . By summing up this latter inequality with (A.6) and (A.5), via a routine Granwäll's inequality argument we infer that, for  $k \in \mathbb{N}$  and  $t \in [0, T]$ ,

$$(A.7) \quad \int_0^t \left\| \frac{\partial p_k}{\partial s}(s, \cdot) \right\|_{L^2}^2 ds \leq C(\alpha, \|(a, b)\|_{\mathcal{H}_m^+}) \left( \|\bar{p}\|_{H^1}^2 + \int_0^t \|f(s, \cdot)\|_{L^2}^2 ds \right).$$

**2) First-order-in-time and second-order-in-space regularity.** In this section, we provide computations by rather considering SFPE $_f$  with  $0 \neq f \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$ .

We first show the inclusion  $\frac{\partial p}{\partial t} \in L^2(0, T; L^2(\mathbb{R}^n, \mathbb{R})) \cong L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$ , this latter

identification being true since  $L^2(\mathbb{R}^n, \mathbb{R})$  is a separable Hilbert space. For this, we show the existence of  $\xi \in L^2(0, T; L^2(\mathbb{R}^n, \mathbb{R}))$  such that the following holds in  $L^2(\mathbb{R}^n, \mathbb{R})$ :

$$(A.8) \quad \int_0^T \xi(t, \cdot) \psi(t) dt = - \int_0^T p(t, \cdot) \frac{d\psi}{dt}(t) dt, \quad \text{for all } \psi \in C_c^\infty([0, T], \mathbb{R}).$$

For this, we first note that, thanks to (A.7) there exists  $\xi \in L^2(0, T; L^2(\mathbb{R}^n, \mathbb{R}))$  such that  $\left(\frac{\partial p_k}{\partial t}\right)_{k \in \mathbb{N}} \subseteq L^2(0, T; L^2(\mathbb{R}^n, \mathbb{R}))$  weakly converges to  $\xi$ , up to a subsequence. Now, for any  $\psi \in C_c^\infty([0, T], \mathbb{R})$  and any  $\varphi \in L^2(\mathbb{R}^n, \mathbb{R})$ , thanks to the last (convergence) property in (A.4) we may compute

$$\begin{aligned} \left( \int_0^T \xi(t, \cdot) \psi(t) dt, \varphi \right)_{L^2} &= \int_0^T (\xi(t, \cdot), \psi(t) \varphi)_{L^2} dt = \\ &= \lim_{k \rightarrow \infty} \int_0^T \left( \frac{\partial p_k}{\partial t}(t, \cdot), \psi(t) \varphi \right)_{L^2} dt = \lim_{k \rightarrow \infty} \left( \int_0^T \frac{\partial p_k}{\partial t}(t, \cdot) \psi(t) dt, \varphi \right)_{L^2} \\ &= - \lim_{k \rightarrow \infty} \left( \int_0^T p_k(t, \cdot) \frac{d\psi}{dt}(t) dt, \varphi \right)_{L^2} = - \lim_{k \rightarrow \infty} \int_0^T \left( p_k(t, \cdot), \frac{d\psi}{dt}(t) \varphi \right)_{L^2} dt \\ &= - \int_0^T \left( p(t, \cdot), \frac{d\psi}{dt}(t) \varphi \right)_{L^2} dt = \left( - \int_0^T p(t, \cdot) \frac{d\psi}{dt}(t) dt, \varphi \right)_{L^2}, \end{aligned}$$

and the equivalence (A.8) readily follows.

Next, we prove that  $p \in L^2(0, T; H^2(\mathbb{R}^n, \mathbb{R}))$ . For this, we first note that, from Theorem 2.1 and the variational problem (A.1), thanks to our previous computations the following variational equality holds true for almost every  $t \in [0, T]$ :

$$(A.9) \quad \int_{\mathbb{R}^n} \sum_{i,j=1}^n \frac{1}{2} a_{ij}^\alpha(t, y) \frac{\partial p}{\partial y_j}(t, y) \frac{\partial \varphi}{\partial y_i}(y) dy \quad \text{for } \varphi \in H^1(\mathbb{R}^n, \mathbb{R}) \\ + \sum_{i=1}^n \int_{\mathbb{R}^n} \left( \sum_{j=1}^n \frac{1}{2} \frac{\partial a_{ij}^\alpha}{\partial y_j}(t, y) - b_i(t, y) \right) \frac{\partial \varphi}{\partial y_i}(y) p(t, y) dy = \int_{\mathbb{R}^n} g(t, y) \varphi(y) dy,$$

where

$$(A.10) \quad g \triangleq f - \frac{\partial p}{\partial t} \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R}).$$

Therefore, Theorem 2.1 and the classical elliptic regularity theory (see, e.g., [13, Section 6.3]) imply, for almost every  $t \in [0, T]$ , both that  $p(t, \cdot) \in H^2(\mathbb{R}^n, \mathbb{R})$  and that

$$(A.11) \quad \|p(t, \cdot)\|_{H^2}^2 \leq C(\alpha, \|(a, b)\|_{\mathcal{H}_m}) \left( \|p(t, \cdot)\|_{L^2}^2 + \|g(t, \cdot)\|_{L^2}^2 \right).$$

At this step, from our choice for the countable basis  $(v_i)_{i \in \mathbb{N}} \subseteq H^{2(m+1)}(\mathbb{R}^n, \mathbb{R})$ , for almost every  $t \in [0, T]$  and every  $i \in \mathbb{N}$ , in particular there exists  $z_i(t) \in \mathbb{R}$  such that

$$(A.12) \quad p(t, \cdot) = \lim_{k \rightarrow \infty} \sum_{i=1}^k z_i(t) v_i(\cdot), \quad \text{in } H^2(\mathbb{R}^n, \mathbb{R}),$$

for almost every  $t \in [0, T]$ . But, up to extracting a subsequence, (A.4) yields

$$p(t, \cdot) = \lim_{k \rightarrow \infty} \sum_{i=1}^k x_i^k(t) v_i(\cdot), \quad \text{in } H^1(\mathbb{R}^n, \mathbb{R}),$$

for almost every  $t \in [0, T]$ , and from the orthonormality of  $(v_i)_{i \in \mathbb{N}}$  in  $H^1(\mathbb{R}^n, \mathbb{R})$  we obtain that  $z_i(t) = x_i^k(t)$ , for every  $k \in \mathbb{N}$ ,  $i = 1, \dots, k$  and almost every  $t \in [0, T]$ . Hence, (A.12) yields that the mapping  $p : [0, T] \rightarrow H^2(\mathbb{R}^n, \mathbb{R})$  is strongly Bochner measurable. In addition, combining the elliptic estimate (A.11) together with the estimates (8.4) and (A.7) finally provides that  $p \in L^2(0, T; H^2(\mathbb{R}^n, \mathbb{R}))$  with

$$(A.13) \quad \int_0^t \|p(s, \cdot)\|_{H^2}^2 ds \leq C(\alpha, \|(a, b)\|_{\mathcal{H}_m}) \left( \|\bar{p}\|_{H^1}^2 + \int_0^t \|f(s, \cdot)\|_{L^2}^2 ds \right), \quad t \in [0, T].$$

Before moving on with additional regularity properties, we note that, thanks to Theorem 2.1 and the fact that  $p \in L^2(0, T; H^2(\mathbb{R}^n, \mathbb{R}))$ , integrating by parts the variational problem (A.1) with  $f = 0$  readily yields

$$\int_0^T \int_{\mathbb{R}^n} \left( \frac{\partial p}{\partial t}(t, y) - (\mathcal{L}_t^{a,b})^* p(t, y) \right) \varphi(y) dy dt = 0, \quad \varphi \in C_c^\infty(\mathbb{R}^n, \mathbb{R}),$$

and SFPE<sub>0</sub> follows from the previous regularity properties and a density argument.

**3) Second-order-in-time and fourth-order-in-space regularity.** We now turn to the original setting SFPE<sub>0</sub>, i.e., SFPE<sub>f</sub> with  $f = 0$ . The first step to further improve the regularity of  $p$  consists of formally differentiating (A.1) with respect to time and studying solutions  $q : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$  to the new variational problem

$$(A.14) \quad \begin{cases} \frac{d}{dt}(q(t, \cdot), \varphi)_{L^2} + A^{\alpha, a, b}(t; q(t, \cdot), \varphi) + \\ \quad + A^{0, \dot{a}, \dot{b}}(t; p(t, \cdot), \varphi) = 0, \quad \varphi \in H^1(\mathbb{R}^n, \mathbb{R}), \\ q(0, \cdot) = (\mathcal{L}_0^{a,b})^* \bar{p}(\cdot). \end{cases}$$

Problem (A.14) is well-posed. Indeed, since  $p \in L^2(0, T; H^2(\mathbb{R}^n, \mathbb{R}))$ , thanks to Theorem 2.1 one easily shows that

$$A^{0, \dot{a}, \dot{b}}(t; p(t, \cdot), \varphi) = - \int_{\mathbb{R}^n} (\mathcal{L}_t^{0, \dot{a}, \dot{b}})^* p(t, y) \varphi(y) dy,$$

for every  $\varphi \in H^1(\mathbb{R}^n, \mathbb{R})$  and almost every  $t \in [0, T]$ . Therefore, since  $(\mathcal{L}_t^{0, \dot{a}, \dot{b}})^* p \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$  and  $(\mathcal{L}_0^{a,b})^* \bar{p} \in H^1(\mathbb{R}^n, \mathbb{R})$ , problem (A.14) fits the setting of Theorem 8.7, thus from our previous computations there exists a unique solution  $q \in L^2(0, T; H^2(\mathbb{R}^n, \mathbb{R}))$  to (A.14), which additionally satisfies  $\frac{\partial q}{\partial t} \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$ .

At this step, we define the function

$$w : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R} \\ (t, y) \mapsto \bar{p}(y) + \left( \int_0^t q(s, \cdot) ds \right)(y) = \bar{p}(y) + \int_0^t q(s, y) ds,$$

where the Lebesgue integral in the last equality stems from the properties of the Bochner integral for functions in  $L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$ . Clearly,  $w \in C^0(0, T; H^2(\mathbb{R}^n, \mathbb{R}))$ , and Fubini theorem yields, for every  $\varphi_1 \in L^2(\mathbb{R}^n, \mathbb{R})$  and  $\psi \in C_c^\infty([0, T], \mathbb{R})$ ,

$$\int_{\mathbb{R}^n} \left( \int_0^T \left( \int_0^t q(s, y) \, ds \right) \frac{d\psi}{dt}(t) \, dt + \int_0^T q(t, y) \psi(t) \, dt \right) \varphi_1(y) \, dy = 0,$$

and, for almost every  $t \in [0, T]$  and every  $\varphi_2 \in C_c^\infty(\mathbb{R}^n, \mathbb{R})$ ,

$$\int_{\mathbb{R}^n} w(t, y) \frac{\partial \varphi_2}{\partial y_j}(y) \, dy = - \int_{\mathbb{R}^n} \left( \frac{\partial \bar{p}}{\partial y_j}(y) + \int_0^t \frac{\partial q}{\partial y_j}(s, y) \, ds \right) \varphi_2(y) \, dy, \quad j = 1, \dots, n.$$

In particular, from such properties we readily deduce that, for every  $j = 1, \dots, n$ ,

$$(A.15) \quad \frac{\partial w}{\partial t} = q \in L^2(0, T; H^2(\mathbb{R}^n, \mathbb{R})), \quad \text{with} \quad \frac{\partial}{\partial y_j} \left( \frac{\partial w}{\partial t} \right) = \frac{\partial}{\partial t} \left( \frac{\partial w}{\partial y_j} \right).$$

Thanks to Theorem 2.1 and the properties listed in (A.15), for every  $\varphi \in C_c^\infty(\mathbb{R}^n, \mathbb{R})$  and almost every  $t \in [0, T]$ , we may compute

$$\begin{aligned} \frac{d}{dt} (w(t, \cdot), \varphi)_{L^2} &= (q(t, \cdot), \varphi)_{L^2} \\ &= ((\mathcal{L}_0^{a,b})^* \bar{p}, \varphi)_{L^2} - \int_0^t A^{\alpha, a, b}(s; q(s, \cdot), \varphi) \, ds - \int_0^t A^{0, \dot{a}, \dot{b}}(s; p(s, \cdot) \pm w(s, \cdot), \varphi) \, ds \\ &= \int_{\mathbb{R}^n} (\mathcal{L}_0^{a,b})^* \bar{p}(y) \varphi(y) \, dy - \\ &\quad - \sum_{i,j=1}^n \int_{\mathbb{R}^n} \frac{1}{2} \left( \int_0^t \left( a_{ij}^\alpha(s, y) \frac{\partial}{\partial s} \left( \frac{\partial w}{\partial y_j} \right)(s, y) + \dot{a}_{ij}^\alpha(s, y) \frac{\partial w}{\partial y_j}(s, y) \right) ds \right) \frac{\partial \varphi}{\partial y_i}(y) \, dy \\ &\quad - \sum_{j=1}^n \int_{\mathbb{R}^n} \left( \int_0^t \left( \left( \sum_{j=1}^n \frac{\partial \dot{a}_{ij}^\alpha}{\partial y_j}(s, y) - \dot{b}_i(s, y) \right) w(s, y) \right. \right. \\ &\quad \left. \left. + \left( \sum_{j=1}^n \frac{\partial a_{ij}^\alpha}{\partial y_j}(s, y) - b_i(s, y) \right) \frac{\partial w}{\partial s}(s, y) \right) ds \right) \frac{\partial \varphi}{\partial y_i}(y) \, dy \\ &\quad - \int_0^t A^{0, \dot{a}, \dot{b}}(s; (p - w)(s, \cdot), \varphi) \, ds. \end{aligned}$$

Therefore, from [10, Theorem 11.5] and the fact that

$$\int_{\mathbb{R}^n} (\mathcal{L}_0^{a,b})^* \bar{p}(y) \varphi(y) \, dy = -A^{\alpha, a, b}(0; \bar{p}, \varphi),$$

we conclude that  $w$  solves the following integro-differential variational problem

$$(A.16) \quad \begin{cases} \frac{d}{dt} (w(t, \cdot), \varphi)_{L^2} + A^{\alpha, a, b}(t; w(t, \cdot), \varphi) + \\ \quad + \int_0^t A^{0, \dot{a}, \dot{b}}(s; (p - w)(s, \cdot), \varphi) \, ds = 0, & \varphi \in H^1(\mathbb{R}^n, \mathbb{R}), \\ w(0, \cdot) = \bar{p}(\cdot). \end{cases}$$

Clearly,  $p = w$  is solution to (A.16). In particular, if so we would obtain that  $\frac{\partial p}{\partial t} \in L^2(0, T; H^2(\mathbb{R}^n, \mathbb{R}))$  and  $\frac{\partial^2 p}{\partial t^2} \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$ , and thanks to the choice of the basis  $(v_i)_{i \in \mathbb{N}} \subseteq H^{2(m+1)}(\mathbb{R}^n, \mathbb{R})$  and to a higher order elliptic regularity argument applied to (A.9)–(A.10) with  $f = 0$ , we would also obtain that  $p \in L^2(0, T; H^4(\mathbb{R}^n, \mathbb{R}))$ , that is the sought second-order-in-time and fourth-order-in-space regularity.

Hence, to conclude we need to prove that the integro-differential variational problem (A.16) may have one solution at most. For this, let  $z_1, z_2 \in L^2(0, T; H^2(\mathbb{R}^n, \mathbb{R}))$ , such that  $\frac{\partial z_1}{\partial t}, \frac{\partial z_2}{\partial t} \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$ , satisfy (A.16). By defining the mapping

$$h : [0, T] \rightarrow L^2(\mathbb{R}^n, \mathbb{R})$$

$$t \mapsto \int_0^t (\mathcal{L}_s^{0, \dot{a}, \dot{b}})^*(z_1 - z_2)(s, \cdot) \, ds,$$

it is readily checked that  $h \in L^2([0, T] \times \mathbb{R}^n, \mathbb{R})$ , in particular satisfying

$$(A.17) \quad \int_0^t \|h(s, \cdot)\|_{L^2}^2 \, ds \leq C(\alpha, \|(a, b)\|_{\mathcal{H}_m}) \int_0^t \left( \int_0^s \|(z_1 - z_2)(r, \cdot)\|_{H^2}^2 \, dr \right) \, ds, \quad t \in [0, T],$$

and that

$$\begin{cases} \frac{d}{dt}((z_1 - z_2)(t, \cdot), \varphi)_{L^2} + A^{\alpha, a, b}(t; (z_1 - z_2)(t, \cdot), \varphi) = \\ \quad = (h(t, \cdot), \varphi)_{L^2}, \quad \varphi \in H^1(\mathbb{R}^n, \mathbb{R}), \\ (z_1 - z_2)(0, \cdot) = 0. \end{cases}$$

Therefore, we are in the setting of Theorem 8.7 and of our previous computations. In particular, we may combine (A.13) with (A.17) to obtain that, for every  $t \in [0, T]$ ,

$$\begin{aligned} \int_0^t \|(z_1 - z_2)(s, \cdot)\|_{H^2}^2 \, ds &\leq C(\alpha, \|(a, b)\|_{\mathcal{H}_m}) \int_0^t \|h(s, \cdot)\|_{L^2}^2 \, ds \\ &\leq C(\alpha, \|(a, b)\|_{\mathcal{H}_m}) \int_0^t \left( \int_0^s \|(z_1 - z_2)(r, \cdot)\|_{H^2}^2 \, dr \right) \, ds, \end{aligned}$$

and a routine application of Gronwall's inequality allows us to conclude that  $z_1 = z_2$ .

**4) Conclusion.** Given Theorem 2.1, the regularity of the initial condition  $\bar{p} \in H^{2m+1}(\mathbb{R}^n, \mathbb{R})$ , and the choice of the basis  $(v_i)_{i \in \mathbb{N}} \subseteq H^{2(m+1)}(\mathbb{R}^n, \mathbb{R})$ , one may easily iterate the previous computations by induction to define successive time derivatives of  $p$  via iteratively considering formal differentiation-in-time of the variational problem (A.14). In particular, it is clear how to extend properties (A.15) and define updated integro-differential variational problems (A.16) by induction, which, thanks to iteratively higher order elliptic regularity arguments applied to (A.9)–(A.10) and the elliptic estimate (A.13), provide that  $\frac{\partial^\ell p}{\partial t^\ell} \in L^2(0, T; H^{2(m+1-\ell)}(\mathbb{R}^n, \mathbb{R}))$ , for  $\ell = 1, \dots, m+1$ , and  $p \in L^2(0, T; H^{2(m+1)}(\mathbb{R}^n, \mathbb{R}))$ . The conclusion follows.  $\square$

## REFERENCES

- [1] K. ABRAHAM, *Nonparametric Bayesian Posterior Contraction Rates for Scalar Diffusions with High-frequency Data*, Bernoulli, 25 (2019), pp. 2696–2728.
- [2] C. AECKERLE-WILLEMS AND C. STRAUCH, *Sup-norm Adaptive Drift Estimation for Multivariate Nonreversible Diffusions*, The Annals of Statistics, 50 (2022), pp. 3484–3509.
- [3] N. ARONSZAJN, *Theory of reproducing kernels*, Transactions of the American mathematical society, 68 (1950), pp. 337–404.
- [4] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser, 1990.
- [5] T. BHUDISAKSANG AND A. CARTEA, *Online Drift Estimation for Jump-Diffusion Processes*, Bernoulli, 27 (2021), pp. 2494–2518.
- [6] R. BONALLI AND B. BONNET-WEILL, *First-Order Pontryagin Maximum Principle for Risk-Averse Stochastic Optimal Control Problems*, SIAM Journal on Control and Optimization, 61 (2023), pp. 1881–1909.
- [7] R. BONALLI, T. LEW, AND M. PAVONE, *Sequential convex programming for non-linear stochastic optimal control*, ESAIM: Control, Optimisation and Calculus of Variations, 28 (2022), p. 64.
- [8] T. BREITEN, K. KUNISCH, AND L. PFEIFFER, *Control Strategies for the Fokker–Planck Equation*, ESAIM: Control, Optimisation and Calculus of Variations, 24 (2018).
- [9] H. BREZIS, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, 2011.
- [10] M. CHIPOT, *Elements of Nonlinear Analysis*, Springer, 2000.
- [11] F. COMTE AND V. GENON-CATALOT, *Drift Estimation on Non Compact Support for Diffusion Models*, Stochastic Processes and their Applications, 134 (2021), pp. 174–207.
- [12] D. DACUNHA-CASTELLE AND D. FLORENS-ZMIROU, *Estimation of the Coefficients of a Diffusion from Discrete Observations*, Stochastics: An International Journal of Probability and Stochastic Processes, 19 (1986), pp. 263–284.
- [13] L. C. EVANS, *Partial Differential Equations*, American Mathematical Society, 2010.
- [14] A. FIGALLI, *Existence and uniqueness of martingale solutions for SDEs with rough or degenerate coefficients*, Journal of Functional Analysis, 254 (2008).
- [15] D. FLORENS-ZMIROU, *Approximate Discrete-Time Schemes for Statistics of Diffusion Processes*, Statistics, 20 (1989), pp. 547–557.
- [16] A. R. GALLANT, *The Relative Efficiency of Method of Moments Estimators*, Journal of Econometrics, 92 (1999), pp. 149–172.
- [17] A. R. GALLANT, D. HSIEH, AND G. E. TAUCHEN, *Estimation of Stochastic Volatility Models with Diagnostics*, Journal of Econometrics, 81 (1997), pp. 159–192.
- [18] H. GARNIER AND L. WANG, *Identification of Continuous-Time Models from Sampled Data*, Springer, 2008.
- [19] V. GENON-CATALOT AND C. LAREDO, *Asymptotic Equivalence of Nonparametric Diffusion and Euler Scheme Experiments*, The Annals of Statistics, 42 (2014), pp. 1145–1165.
- [20] H. J. KUSHNER AND G. G. YIN, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, 2001.
- [21] H. LAVENANT ET AL., *Towards a Mathematical Theory of Trajectory Inference*. arXiv:2102.09204, 2021.
- [22] J.-F. LE GALL, *Brownian Motion, Martingales, and Stochastic Calculus*, Springer, 2016.
- [23] L. LENNART, *System Identification – Theory For the User*, PTR Prentice Hall, 1998.
- [24] X. LI ET AL., *Scalable Gradients for Stochastic Differential Equations*, in Int. Conference on Artificial Intelligence and Statistics, 2020, pp. 3870–3882.
- [25] J.-L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, 1968.
- [26] J.-L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications: Vol. 2*, Springer, 2012.
- [27] R. S. LIPTSER AND A. N. SHIRIAEV, *Statistics of Random Processes: General Theory*, Springer, 1977.
- [28] N. MARIE AND A. ROSIER, *Nadaraya Watson Estimator for IID Paths of Diffusion Processes*, Scandinavian Journal of Statistics, 50 (2023), pp. 589–637.
- [29] U. MARTEAU-FEREY, F. BACH, AND A. RUDI, *Non-parametric models for non-negative functions*, Advances in neural information processing systems, 33 (2020), pp. 12816–12826.
- [30] H. MASUDA, *Convergence of Gaussian Quasi-Likelihood Random Fields for Ergodic Lévy driven SDE Observed at High Frequency*, The Annals of Statistics, 41 (2013), pp. 1593–1641.
- [31] B. MUZELLEC, F. BACH, AND A. RUDI, *Learning psd-valued functions using kernel sums-of-squares*, arXiv preprint arXiv:2111.11306, (2021).
- [32] S. NAKAKITA, *Parametric Estimation of Stochastic Differential Equations via Online Gradient*

- Descent*. arXiv:2210.08800, 2022.
- [33] O. NELLES, *Nonlinear System Identification*, Springer, 2001.
  - [34] Y. NESTEROV AND A. NEMIROVSKII, *Interior-point polynomial algorithms in convex programming*, SIAM, 1994.
  - [35] R. NICKL AND K. RAY, *Nonparametric Statistical Inference for Drift Vector Fields of Multidimensional Diffusions*, The Annals of Statistics, 48 (2020), pp. 1383–1408.
  - [36] R. NICKL AND J. SÖHL, *Nonparametric Bayesian Posterior Contraction Rates for Discretely Observed Scalar Diffusions*, The Annals of Statistics, 45 (2017), pp. 1664–1693.
  - [37] J. N. NIELSEN, H. MADSEN, AND P. C. YOUNG, *Parameter Estimation in Stochastic Differential Equations: an Overview*, Annual Reviews in Control, 24 (2000), pp. 83–94.
  - [38] J. RIDDERHOF, K. OKAMOTO, AND P. TSIOTRAS, *Nonlinear Uncertainty Control with Iterative Covariance Steering*, in Int. Conference on Decision and Control, 2019, pp. 3484–3490.
  - [39] A. RUDI AND C. CILIBERTO, *Psd representations for effective probability models*, in Advances in Neural Information Processing Systems, 2021.
  - [40] A. RUDI, U. MARTEAU-FEREY, AND F. BACH, *Finding global minima via kernel approximations*, Mathematical Programming, (2024), pp. 1–82.
  - [41] L. SHARROCK AND N. KANTAS, *Joint Online Parameter Estimation and Optimal Sensor Placement for the Partially Observed Stochastic Advection-Diffusion Equation*, SIAM Journal on Uncertainty Quantification, 10 (2022), pp. 55–95.
  - [42] I. STEINWART AND A. CHRISTMANN, *Support vector machines*, Springer Science & Business Media, 2008.
  - [43] D. W. STROOCK AND S. VARADHAN, *Multidimensional Diffusion Processes*, Springer Science & Business Media, 1997.
  - [44] S. C. SURACE AND J.-P. PFISTER, *Online Maximum-Likelihood Estimation of the Parameters of Partially Observed Diffusion Processes*, IEEE Transactions on Automatic Control, 64 (2018), pp. 2814–2829.
  - [45] M. UCHIDA AND N. YOSHIDA, *Quasi Likelihood Analysis of Volatility and Nondegeneracy of Statistical Random Field*, Stochastic Processes and their Applications, 123 (2013), pp. 2851–2876.
  - [46] F. VAN DER MEULEN AND H. VAN ZANTEN, *Consistent Nonparametric Bayesian Inference for Discretely Observed Scalar Diffusions*, Bernulli, 19 (2013), pp. 44–63.
  - [47] G. VELHO, J. AURIOL, AND R. BONALLI, *A Gradient Descent-Ascent Method for Continuous-Time Risk-Averse Optimal Control*. arxiv:2306.12878, 2024.
  - [48] H. WENDLAND, *Scattered Data Approximation*, Cambridge University Press, 2004.
  - [49] W. XU ET AL., *Infinitely Deep Bayesian Neural Networks with Stochastic Differential Equations*. arXiv:2102.06559, 2022.
  - [50] J. YONG AND X. Y. ZHOU, *Stochastic Controls: Hamiltonian Systems and HJB equations*, Springer Science and Business Media, 1999.
  - [51] N. YOSHIDA, *Polynomial Type Large Deviation Inequalities and Quasi-Likelihood Analysis for Stochastic Differential Equations*, Annals of the Institute of Statistical Mathematics, 63 (2011), pp. 431–479.