A TRANSITION-DENSITY-BASED OPERATOR LEARNING METHOD FOR FOKKER-PLANCK EQUATIONS WITH VARIOUS INITIAL CONDITIONS*

LI ZENG†, XIAOLIANG WAN‡, YAOBIN WANG§, FABIO NOBILE¶, AND TAO ZHOU‖

Abstract. Solving Fokker-Planck equations (FPEs) for multiple initial conditions typically requires repeated computations, leading to substantial computational costs. In this work, we propose a transition-density-based operator learning method to efficiently approximate the solution operator of FPEs with various initial conditions. The core idea is to learn the transition probability density function (PDF) of the underlying stochastic differential equation (SDE), from which the solution associated with a new initial distribution can be obtained through the Chapman–Kolmogorov equation without retraining the model. A major challenge in learning the transition PDF lies in the singular behavior induced by the Dirac initial condition. To address it, we introduce a conditional normalizing flow whose base distribution is given by the explicit transition PDF of a linearized SDE. This base distribution captures the short-time behavior of the target transition PDF and allows the normalizing flow to learn a near-identity transformation at small times. We further incorporate a time-weighted loss function to stabilize training near the initial time and develop an importance-sampling strategy for evaluating solutions associated with general initial conditions. A variety of numerical experiments are presented to illustrate the effectiveness and robustness of the proposed method.

Key words. Conditional Normalizing flow, Fokker-Planck equation, Operator learning

MSC codes. 68T07, 62G07, 65M99

1. Introduction. The FPE describes the time evolution of the PDF associated with a stochastic dynamical system of diffusion type, serving as a deterministic representation of the stochastic process. Numerically solving the FPE is of high interest in many fields, such as statistical physics, chemistry, biology, finance, and data science, as it enables uncertainty quantification and statistical inference. Moreover, in many of these applications, such as ensemble forecasting [24] and data assimilation [30], the governing stochastic dynamic is fixed, whereas the initial distribution varies across scenarios or posterior updates. Solving the FPE repeatedly for each initial condition can therefore lead to substantial computational costs.

Classical numerical discretization schemes for the FPE include the finite element method [9], the finite difference method [22], the path integral method [45], and the Jordan-Kinderlehrer-Otto (JKO) scheme [21], among others. Nevertheless, these traditional approaches incur high computational cost and suffer from poor scalability in high-dimensional settings. To overcome these challenges, deep learning methods [11, 33, 37] bring a new paradigm. In general, deep learning–based methods for solving

*Submitted to the editor's DATE.

Funding: The first author was supported by the NSF of China (under grant 12401566). The second author was supported by NSF grant DMS-2513234. The last author was supported by the NSF of China (under grants 12288201 and 12461160275).

†School of Mathematics and Statistics, Fuzhou University, Fuzhou, China (lzeng@fzu.edu.cn).

‡Department of Mathematics and Center for Computation and Technology, Louisiana State University, Baton Rouge 70803, USA (xlwan@lsu.edu).

§Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai 519087, China (wangyaobin@bnbu.edu.cn).

¶CSQI, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland (Fabio.Nobile@epfl.ch).

‖Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China (tzhou@lsec.cc.ac.cn).

the FPE can be categorized into two main classes: those that directly represent the solution with neural networks [7, 31, 42, 46, 47, 51], and those that employ generative modeling techniques, such as normalizing flows [12, 13, 29, 39, 49, 50] and score-based approaches [3, 19, 52]. Alternatively, Bruna et al. [6, 46] approximate the solution of the interacting-particle FPE using a Gaussian mixture model, and numerically solve the ordinary differential equation (ODE) system for the mixture coefficients using the Neural Galerkin method, subsequently normalizing the solution at each time step to enforce the PDF constraints. In parallel, several techniques have been developed for some special FPEs that exploit the reinterpretation as the L^2-Wasserstein gradient flow of the Kullback-Leibler divergence [23, 26].

While a variety of numerical methods have shown promising performance in solving FPEs with prescribed initial conditions, there remains a lack of methods capable of efficiently handling multiple initial conditions. The aforementioned methods have to solve the FPE anew every time a new initial condition arrives. This challenge can be viewed from the perspective of operator learning, where the objective is to learn a mapping from an initial distribution to the corresponding time-dependent solution of the FPE. Recently, Wang et al. [43] proposed a Gaussian-mixture latent surrogate that directly learns the evolution from initial to corresponding transient PDF representation, while focusing on Gaussian-mixture initial distributions and a truncated computational domain. Yang et al. [48] proposed a pseudoreversible normalizing flow and leveraged samples generated from an SDE to learn the terminal solution of the corresponding transition FPE. In the general data-driven setting, generative models are also widely used to learn the conditional PDF, such as conditional Glow [28], monotone generative adversarial networks [2], the gradient of a partially input-convex neural network, and Neural ODE [44], to name a few. For general parametric PDEs, some techniques have been proposed to learn the solution operator, such as the deep operator network (DeepONet) [27], the Fourier neural operator (FNO) [25], and their variants. Gaby et al. [15] learned the solution operator of the evolution PDEs for various initial conditions by learning a control vector field in the parameter space and solving the ODE system of the neural network parameters. There are also works that learn the Green's function to construct the solution operator, leveraging the explicit representation of the solution in terms of the Green's function [1, 5, 40]. However, direct application of these methods to FPEs requires additional treatments of boundary conditions and PDF constraints. In this work, we take a different route: rather than learning the dependence of the FPE solution on the initial distribution directly, we learn the transition PDF of the underlying stochastic dynamics. Solutions for new initial distributions are then recovered through the Chapman-Kolmogorov equation. Our framework is built on the following ideas.

First, we reformulate operator learning for the FPE as the problem of approximating the transition PDF, i.e., the solution of the FPE starting at initial time from a Dirac mass at an arbitrary location $\boldsymbol{x}_0$. We develop a conditional normalizing flow parametric in time and $\boldsymbol{x}_0$ to learn the transition PDF. Benefiting from the inherent property of a normalizing flow, the proposed approach naturally satisfies the PDF constraints without requiring additional enforcement. We mention that the work [35, 36] developed in parallel to ours, also proposes a conditional normalizing flow to approximate the transition PDF, however, in a Neural Galerkin framework.

Second, we propose a time- and initial state-dependent base distribution to improve the conditioning of the conditional normalizing flow. In contrast to the commonly used standard normal distribution, our base distribution is given by the explicit solution of a linearized SDE parametric with respect to the initial state $\boldsymbol{x}_0$. The lin-

earized SDE is obtained by applying a first-order Taylor expansion to the drift term and a zero-order expansion to the diffusion term of the target SDE. Using the solution of the linearized SDE as the base random process not only respects the causality for evolution PDEs [8] while bypassing the singular initial condition, but also alleviates training difficulties. We show, in particular, that the total variation distance between the transition PDF of the linearized process and that of the true process is of order $\mathcal{O}(t)$ while the Kullback–Leibler divergence scales as $\mathcal{O}(t^2)$ as $t \to 0$. The corresponding transformation tends to the identity map as $t \to 0$, making it easier to learn than transformation based on a fixed standard normal base distribution at all times.

Third, we propose a time-weighted loss function to mitigate numerical instabilities arising at small times when the transition PDF approaches a Dirac measure. The weight function takes the form t^{d+2} or $t^{\frac{d}{2}+2}$ depending on the distribution of the training points. It is designed to capture the temporal scaling of the residual when a small perturbation around the base PDF is considered. Introducing this weighting strategy can be interpreted as a trade-off between maintaining causality and alleviating optimization difficulties.

Fourth, we incorporate the learned conditional normalizing flow and an importance-sampling Monte Carlo estimator to approximate the final solution of the original FPE.

The remainder of this paper is organized as follows. Section 2 formalizes the operator learning problem for the FPE under different initial conditions. In section 3, we introduce the conditional normalizing flow with base distribution provided by the linearized SDE. In section 4, we analyze the behavior of the loss function as $t \to 0$, and we propose a time-weighted loss and an adaptive sampling algorithm. Importance sampling is discussed in section 5 to efficiently evaluate $p(\boldsymbol{x}, t)$. Section 6 presents a series of numerical experiments that illustrate the effectiveness and robustness of the proposed method, followed by concluding remarks in section 7. Additional details of certain proofs and calculations are provided in the Appendix.

2. Problem setup and reformulation. Consider the state variable $\boldsymbol{X}_t$ modeled by the following non-degenerate SDE,

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t)\,\mathrm{d}t + \boldsymbol{g}(\boldsymbol{X}_t)\,\mathrm{d}\boldsymbol{W}_t, \quad t > 0, \tag{2.1}$$

where $\boldsymbol{f}(\boldsymbol{X}_t) : \mathbb{R}^d \to \mathbb{R}^d$ and $\boldsymbol{g}(\boldsymbol{X}_t) : \mathbb{R}^d \to \mathbb{R}^{d\times m}, m \geq d$ denote the drift and diffusion terms, respectively, and $\boldsymbol{W}_t$ is an m-dimensional standard Wiener process. The PDF $p(\boldsymbol{x}, t)$ for $\boldsymbol{X}_t$ satisfies the corresponding time-dependent FPE:

$$\frac{\partial p}{\partial t} = \mathcal{L}^*_{\boldsymbol{f},\boldsymbol{g}} p \coloneqq -\nabla \cdot (p\boldsymbol{f}) + \frac{1}{2}\nabla \cdot \nabla \cdot (\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}} p). \tag{2.2}$$

In general, equation (2.2) is defined on $\mathbb{R}^d$ with the following boundary condition

$$p(\boldsymbol{x}) \to 0 \quad \text{as} \quad |\boldsymbol{x}|_2 \to \infty. \tag{2.3}$$

Meanwhile, the solution as a probability density function should be conservative and non-negative, i.e.,

$$\int_{\mathbb{R}^d} p(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x} \equiv 1, \quad \text{and} \quad p(\boldsymbol{x}, t) \geq 0, \quad \forall\, t \geq 0. \tag{2.4}$$

In this work, we address scenarios with varying initial conditions and develop a transition-density-based operator learning method that generalizes across them,

thereby eliminating the need to solve the corresponding FPE repeatedly for different initial conditions, i.e., we consider the problem

$$\begin{cases} \dfrac{\partial p(\boldsymbol{x},t)}{\partial t} = \mathcal{L}^*_{\boldsymbol{f},\boldsymbol{g}} p(\boldsymbol{x},t), & \boldsymbol{x}\in\mathbb{R}^d, t>0, \\ p(\boldsymbol{x},0) = p_0(\boldsymbol{x}), \quad p_0\in\mathcal{G}, & \boldsymbol{x}\in\mathbb{R}^d \end{cases} \tag{2.5}$$

where $\mathcal{G}$ is a functional class including all possible initial distributions of interest.

To achieve this, we turn to the transition PDF $p(\boldsymbol{x},t|\boldsymbol{x}_0)$ given an initial state. The corresponding transition PDF satisfies the same governing equation as the original FPE (2.5), but with a Dirac Delta function as the initial condition,

$$\begin{cases} \dfrac{\partial p(\boldsymbol{x},t|\boldsymbol{x}_0)}{\partial t} = \mathcal{L}^*_{\boldsymbol{f},\boldsymbol{g}} p(\boldsymbol{x},t|\boldsymbol{x}_0), & \boldsymbol{x}\in\mathbb{R}^d, t>0, \\ p(\boldsymbol{x},0|\boldsymbol{x}_0) = \delta(\boldsymbol{x}-\boldsymbol{x}_0), & \boldsymbol{x}\in\mathbb{R}^d. \end{cases} \tag{2.6}$$

Applying the Chapman-Kolmogorov equation yields the solution to the FPE (2.5),

$$p(\boldsymbol{x},t) = \int_{\Omega_0} p(\boldsymbol{x},t|\boldsymbol{x}_0) p_0(\boldsymbol{x}_0) \mathrm{d}\boldsymbol{x}_0. \tag{2.7}$$

This motivates us to solve the FPE for different initial conditions by approximating the transition PDF $p(\boldsymbol{x}|\boldsymbol{x}_0,t)$. In this work, we extend the normalizing flow framework [13, 49] to handle the transition FPE with Dirac delta functions centered at various points as initial conditions.

2.1. Notation. For convenience, we introduce here some notations that will be used throughout the paper. For a vector $x\in\mathbb{R}^d$, $|x|$ denotes its Euclidean norm. For a matrix $A\in\mathbb{R}^{d\times d}$, $\|A\|$ denotes its spectral norm induced by the Euclidean norm. Likewise, for a bilinear map $B:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}^d$, $\|B\|$ denotes its operator norm $\|B\| \coloneqq \sup_{x,y\in\mathbb{R}^d,\, x\neq 0,\, y\neq 0} \frac{|B(x,y)|}{|x|\,|y|}$.

Given two probability measures ν and μ on $\mathbb{R}^d$, $\nu \ll \mu$, $H(\nu \mid \mu)$ denotes the relative entropy of ν w.r.t. μ, and $\|\nu-\mu\|_{\mathrm{TV}}$ denotes the total variation distance between ν and μ.

$$H(\nu\mid\mu) \coloneqq \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right) \mathrm{d}\nu, \quad \|\nu-\mu\|_{\mathrm{TV}} \coloneqq \sup_{A\in\mathcal{B}(\mathbb{R}^d)} |\nu(A)-\mu(A)|.$$

Furthermore, if $\mu\ll\rho$, and $\mathrm{d}\nu = p_\nu \mathrm{d}\rho$, $\mathrm{d}\mu = p_\mu \mathrm{d}\rho$, then

$$H(\nu\mid\mu) = \int_{\mathbb{R}^d} p_\nu(\boldsymbol{x}) \log\left(\frac{p_\nu(\boldsymbol{x})}{p_\mu(\boldsymbol{x})}\right) \mathrm{d}\rho(\boldsymbol{x}), \quad \|\nu-\mu\|_{\mathrm{TV}} = \frac{1}{2}\int_{\mathbb{R}^d} |p_\nu(\boldsymbol{x}) - p_\mu(\boldsymbol{x})| \mathrm{d}\rho(\boldsymbol{x}).$$

Moreover, given a measurable map $T:\mathbb{R}^d\to\mathbb{R}^d$, $T_\#\nu$ denotes the pushforward measure of ν through T, i.e., $(T_\#\nu)(A) \coloneqq \nu\big(T^{-1}(A)\big) = \nu\left(\{\boldsymbol{x}\in\mathbb{R}^d : T(\boldsymbol{x})\in A\}\right)$ for any $A\in\mathcal{B}(\mathbb{R}^d)$. By $\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ we denote the PDF of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For $\boldsymbol{f}\in L^2(\mathbb{R}^d,\nu;\mathbb{R}^m)$, $\|\boldsymbol{f}\|_{L^2(\mathbb{R}^d,\nu;\mathbb{R}^m)}$ denotes its L^2-norm w.r.t. the measure ν, i.e., $\|\boldsymbol{f}\|_{L^2(\mathbb{R}^d,\nu;\mathbb{R}^m)} = \left(\int_{\mathbb{R}^d} |\boldsymbol{f}(\boldsymbol{x})|^2 \mathrm{d}\nu(\boldsymbol{x})\right)^{1/2}$.

3. The conditional normalizing flow. In this section, we introduce a conditional normalizing flow to approximate the transition PDF $p(\boldsymbol{x}, t|\boldsymbol{x}_0)$ given the initial state $\boldsymbol{x}_0$. Before proceeding to the conditional normalizing flow, we briefly review the idea of normalizing flows.

A normalizing flow seeks an invertible transformation T from the unknown random vector $\boldsymbol{X} \in \mathbb{R}^d$ to a given random vector $\boldsymbol{Z} \in \mathbb{R}^d$ with a well-known and easy to sample distribution. According to the change of variables formula, the PDF of $\boldsymbol{X} = T^{-1}(\boldsymbol{Z})$ can be expressed by,

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = p_{\boldsymbol{Z}}(T(\boldsymbol{x}))\left|\det \nabla_{\boldsymbol{x}} T(\boldsymbol{x})\right|. \tag{3.1}$$

Furthermore, by allowing the distribution of the given random variable $\boldsymbol{Z}$ and the transformation T to depend on some parameters, the initial state $\boldsymbol{x}_0$ and time t in our setting, we can derive a conditional normalizing flow that maps the current state $\boldsymbol{X}_t$ of (2.1) to a base random variable $\boldsymbol{Z}_t(\boldsymbol{x}_0)$ conditioned on an initial state $\boldsymbol{X}_0 = \boldsymbol{x}_0$, i.e.,

$$\begin{aligned} &\boldsymbol{X}_t|_{\boldsymbol{X}_0=\boldsymbol{x}_0} = T^{-1}(\boldsymbol{Z}_t(\boldsymbol{x}_0); \theta(\boldsymbol{x}_0, t)), \\ &p_{\boldsymbol{X}_t|\boldsymbol{X}=\boldsymbol{x}_0}(\boldsymbol{x}) = p_{\boldsymbol{Z}_t(\boldsymbol{x}_0)}\left(T(\boldsymbol{x}; \theta(\boldsymbol{x}_0, t))\right)\left|\det \nabla_{\boldsymbol{x}} T(\boldsymbol{x}; \theta(\boldsymbol{x}_0, t))\right|, \end{aligned} \tag{3.2}$$

where $\theta(\boldsymbol{x}_0, t)$ represents the parameter set of the invertible transformation T conditioned on the initial state $\boldsymbol{x}_0$ and time t. This transformation T maps an unknown stochastic process $\boldsymbol{X}_t$ to another stochastic process $\boldsymbol{Z}_t$ given the initial state $\boldsymbol{X}_0 = \boldsymbol{x}_0$.

3.1. What makes a good base stochastic process for the conditional normalizing flow? The transition PDF provides a way to approximate the solution of the FPE under different initial conditions. However, directly approximating the solution of the FPE with a Dirac delta function as the initial condition is challenging. The challenge lies in the fact that the Dirac delta function lacks regularity, and there is no invertible transformation that maps it to a regular distribution (e.g., normal or uniform distributions) used in normalizing flows. To address this challenge, we linearize the original SDE and use the solution of the resulting linear SDE as the base stochastic process of the normalizing flow. The linearization of the original SDE is based on the assumption that the drift term $\boldsymbol{f}$ and diffusion term $\boldsymbol{g}$ are smooth functions. This allows us to approximate the original SDE by a linear SDE, which admits an analytical solution. Specifically, we approximate the original drift term by the first-order Taylor expansion and the diffusion term by the zero-order Taylor expansion at the initial state $\boldsymbol{x}_0$, i.e.,

$$\begin{cases} \mathrm{d}\widehat{\boldsymbol{X}}_t = (\boldsymbol{f}(\boldsymbol{x}_0) + \nabla \boldsymbol{f}(\boldsymbol{x}_0)(\widehat{\boldsymbol{X}}_t - \boldsymbol{x}_0))\mathrm{d}t + \boldsymbol{g}(\boldsymbol{x}_0)\mathrm{d}\boldsymbol{W}_t, & t > 0, \\ \widehat{\boldsymbol{X}}_0 = \boldsymbol{x}_0. \end{cases} \tag{3.3}$$

The corresponding solution admits an analytical form,

$$\widehat{\boldsymbol{X}}_t = \boldsymbol{x}_0 + \int_0^t e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)(t-s)} \boldsymbol{f}(\boldsymbol{x}_0)\mathrm{d}s + \int_0^t e^{\nabla \boldsymbol{f}(\boldsymbol{x_0})(t-s)} \boldsymbol{g}(\boldsymbol{x}_0)\mathrm{d}\boldsymbol{W}_s, \tag{3.4}$$

which is a multivariate Gaussian process with mean and covariance given by

$$\begin{aligned} &\mathbb{E}\left[\widehat{\boldsymbol{X}}_t|\widehat{\boldsymbol{X}}_0 = \boldsymbol{x}_0\right] = \boldsymbol{x}_0 + \int_0^t e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)(t-s)} \boldsymbol{f}(\boldsymbol{x}_0)\mathrm{d}s, \\ &\boldsymbol{\Sigma}\left[\widehat{\boldsymbol{X}}_t|\widehat{\boldsymbol{X}}_0 = \boldsymbol{x}_0\right] = \int_0^t e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)(t-s)} \boldsymbol{g}(\boldsymbol{x}_0)\boldsymbol{g}(\boldsymbol{x}_0)^\top e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)^\top (t-s)}\mathrm{d}s. \end{aligned} \tag{3.5}$$

The two integrals with respect to time t in equation (3.5) can be effectively approximated by Gauss-Legendre quadrature, thus,

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\boldsymbol{X}}_t|\widehat{\boldsymbol{X}}_0=\boldsymbol{x}_0\right]\approx \boldsymbol{x}_0+\sum_{i=1}^m w_i e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)(t-s_i)}\boldsymbol{f}(\boldsymbol{x}_0),\\
&\boldsymbol{\Sigma}\left[\widehat{\boldsymbol{X}}_t|\widehat{\boldsymbol{X}}_0=\boldsymbol{x}_0\right]\approx \sum_{i=1}^m w_i e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)(t-s_i)}\boldsymbol{g}(\boldsymbol{x}_0)\boldsymbol{g}(\boldsymbol{x}_0)^\top e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)^\top(t-s_i)},\\
&w_i=\frac{t}{2}\hat{w}_i,\quad s_i=\frac{t}{2}(\hat{s}_i+1),
\end{aligned}\tag{3.6}$$

where $\{(\hat{w}_i,\hat{s}_i)\}$ are the weights and nodes associated with the Legendre polynomial defined on $[-1,1]$. Moreover, if $\nabla \boldsymbol{f}(\boldsymbol{x}_0)$ is invertible, we can rewrite the solution as

$$\widehat{\boldsymbol{X}}_t=\boldsymbol{x}_0+(\nabla \boldsymbol{f}(\boldsymbol{x}_0))^{-1}\left(e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)t}-\boldsymbol{I}\right)\boldsymbol{f}(\boldsymbol{x}_0)+\int_0^t e^{\nabla \boldsymbol{f}(\boldsymbol{x_0})(t-s)}\boldsymbol{g}(\boldsymbol{x}_0)\mathrm{d}\boldsymbol{W}_s.\tag{3.7}$$

The corresponding conditional mean admits the equivalent expression,

$$\mathbb{E}\left[\widehat{\boldsymbol{X}}_t|\widehat{\boldsymbol{X}}_0=\boldsymbol{x}_0\right]=\boldsymbol{x}_0+(\nabla \boldsymbol{f}(\boldsymbol{x}_0))^{-1}\left(e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)t}-\boldsymbol{I}\right)\boldsymbol{f}(\boldsymbol{x}_0).$$

Thus, instead of approximating the integral, one may compute the inverse of $\nabla f(\boldsymbol{x}_0)$.

We note that the linear SDE system (3.3) provides a close pathwise approximation to the original SDE (2.1) for small t. The corresponding transition PDF associated with the linear SDE provides a good approximation of the target transition PDF in the short-time regime, as verified by Proposition 3.1 below.

PROPOSITION 3.1. *Suppose that the drift term $\boldsymbol{f}=\boldsymbol{f}_\nu$ in the SDE (2.1) is differentiable, and the diffusion term $\boldsymbol{g}$ is constant, $\boldsymbol{G}=(G^{ij})_{i,j\le d}=\frac{1}{2}\boldsymbol{g}\boldsymbol{g}^\top$ is non-degenerate, $\{p_{\nu_t}\}_{t\in[0,t_f]}$ is the PDF solution on the time interval $[0,t_f]$ to the corresponding FPE with a Dirac delta function $\delta_{\boldsymbol{x}_0}$ as the initial condition, $\{p_{\sigma_t}\}_{t\in[0,t_f]}$ is the PDF solution to the FPE corresponding to the linearized SDE (3.3) with drift term $\boldsymbol{f}_\sigma$ and the same initial condition $\delta_{\boldsymbol{x}_0}$. Denote ν_t and σ_t the probability measures corresponding to p_{ν_t}, p_{σ_t}, respectively. If $\|\nabla \boldsymbol{f}_\nu(\boldsymbol{x})\|\le C_{\boldsymbol{f}}$, $\forall \boldsymbol{x}\in\mathbb{R}^d$ for some $C_{\boldsymbol{f}}\ge 0$, then for any $t\in(0,t_f)$,*

$$\begin{aligned}
H\left(\sigma_t\mid\nu_t\right)&\le\frac{1}{2}\int_0^t\int_{\mathbb{R}^d}\left|\boldsymbol{G}^{-1/2}(\boldsymbol{f}_\sigma(\boldsymbol{x})-\boldsymbol{f}_\nu(\boldsymbol{x}))\right|^2\mathrm{d}\sigma_s\mathrm{d}s,\\
\|\nu_t-\sigma_t\|_{\mathrm{TV}}^2&\le\frac{1}{4}\int_0^t\int_{\mathbb{R}^d}\left|\boldsymbol{G}^{-1/2}\left(\boldsymbol{f}_\sigma(\boldsymbol{x})-\boldsymbol{f}_\nu(\boldsymbol{x})\right)\right|^2\mathrm{d}\sigma_s\mathrm{d}s.
\end{aligned}$$

Both the relative entropy $H(\sigma_t|\nu_t)$ and the square of total variation $\|\nu_t-\sigma_t\|_{\mathrm{TV}}^2$ are $\mathcal{O}(t^2)$ as $t\to 0$. Moreover, if $\|\nabla^2\boldsymbol{f}_\nu(\boldsymbol{x})\|\le H_{\boldsymbol{f}}$, $\forall\boldsymbol{x}\in\mathbb{R}^d$ for some $H_{\boldsymbol{f}}\ge 0$, then $H(\sigma_t|\nu_t)$, $\|\nu_t-\sigma_t\|_{\mathrm{TV}}^2$ are both $\mathcal{O}(t^3)$ as $t\to 0$.

Proof. See Appendix A.

Statistical quantities evaluated under the target measure can be approximated by those under the approximate measure, with an error bounded by the total variation distance between the two measures, as established in the next proposition.

PROPOSITION 3.2. *Suppose that ν,σ are two probability measures on $\mathbb{R}^d$. Then for every function $\boldsymbol{f}\in L^2(\mathbb{R}^d,\nu;\mathbb{R}^m)\cap L^2(\mathbb{R}^d,\sigma;\mathbb{R}^m)$, we have*

$$\left|\int_{\mathbb{R}^d}\boldsymbol{f}(\boldsymbol{x})\mathrm{d}\nu(\boldsymbol{x})-\int_{\mathbb{R}^d}\boldsymbol{f}(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x})\right|\le\sqrt{2\|\nu-\sigma\|_{\mathrm{TV}}}\left(\|\boldsymbol{f}\|_{L^2(\mathbb{R}^d,\nu;\mathbb{R}^m)}+\|\boldsymbol{f}\|_{L^2(\mathbb{R}^d,\sigma;\mathbb{R}^m)}\right).$$

Proof. See Appendix B.

In particular, if we take $\sigma = T_{\#}\nu$ for some measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$ and $f(\boldsymbol{x}) = \boldsymbol{x}$, assuming ν, σ have finite second moments, we have

$$\left| \int_{\mathbb{R}^d} \boldsymbol{x} \mathrm{d}\nu(\boldsymbol{x}) - \int_{\mathbb{R}^d} T(\boldsymbol{x}) \mathrm{d}\nu(\boldsymbol{x}) \right| \leq \sqrt{2\|\nu - \sigma\|_{\mathrm{TV}}} \left(\|\boldsymbol{x}\|_{L^2(\mathbb{R}^d,\nu;\mathbb{R}^d)} + \|\boldsymbol{x}\|_{L^2(\mathbb{R}^d,\sigma;\mathbb{R}^d)} \right).$$

We also state that if $\nu_i \to \mu$ in total variation and their second moments converge, then the Knothe-Rosenblatt (KR) map T_i satisfying $(T_i)_{\#}\mu = \nu_i$ converges to the identity map in $L_2(\mu)$.

PROPOSITION 3.3. *Let $\{\nu_i\}_{i=1}^{\infty}$ be a family of absolutely continuous probability measures on $\mathbb{R}^d$ w.r.t. the Lebesgue measure, each with finite second moments, and assume that μ is an absolutely continuous measure, $\|\nu_i - \mu\|_{\mathrm{TV}} \to 0$, $\|\boldsymbol{x}\|_{L^2(\mathbb{R}^d,\nu_i)} \to \|\boldsymbol{x}\|_{L^2(\mathbb{R}^d,\mu)}$ as $i \to \infty$. Then, let T_i be the KR map between μ and ν_i. We have $T_i \to \mathrm{Id}$ in $L^2(\mu)$ as $i \to \infty$.*

Proof. See Appendix C.

Proposition 3.1 shows that, for small t, the transition PDF of the linearized SDE is close to that of the original SDE in total variation. Proposition 3.3 implies that, under certain conditions, KR maps associated with a sequence of measures and its limiting measure converge to the identity map. This observation motivates our design choice: the conditional normalizing flow is constrained to be the identity at $t = 0$ and to deform gradually as t increases. In this way, we avoid directly mapping from a Dirac delta initial condition and reduce the training difficulty near $t = 0$.

3.2. Architecture of the conditional normalizing flow. The transition-density-based method reformulates the original problem of finding $p(\boldsymbol{x}, t|\boldsymbol{x}_0)$ as the task of identifying an invertible transformation $T(\cdot; \theta(\boldsymbol{x}_0, t))$, mapping a stochastic process $\boldsymbol{X}_t$ to $\boldsymbol{Z}_t$ given initial state $\boldsymbol{X}_0 = \boldsymbol{x}_0$. In practice, we can construct such a complex invertible mapping by stacking a sequence of simple bijections, i.e.,

$$\boldsymbol{Z}_t|_{\boldsymbol{Z}_0=\boldsymbol{Z}_0(\boldsymbol{x}_0)} = T(\boldsymbol{X}_t|_{\boldsymbol{X}_0=\boldsymbol{x}_0}; \theta(\boldsymbol{x}_0, t)) = T_{[K]} \circ T_{[K-1]} \circ \cdots \circ T_{[1]}(\boldsymbol{X}_t|_{\boldsymbol{X}_0=\boldsymbol{x}_0}),$$

where $T_{[i]} = T_{[i]}(\cdot; \theta_{[i]}(\boldsymbol{x}_0, t))$ is based on shallow neural networks. The inverse and Jacobian determinant are given as

$$\begin{aligned}
&\boldsymbol{X}_t|_{\boldsymbol{X}_0=\boldsymbol{x}_0} = T^{-1}\left(\boldsymbol{Z}_t|_{\boldsymbol{Z}_0=\boldsymbol{z}_0}; \theta(\boldsymbol{x}_0, t)\right) = T_{[1]}^{-1} \circ \cdots \circ T_{[K-1]}^{-1} \circ T_{[K]}^{-1}(\boldsymbol{Z}_t|_{\boldsymbol{Z}_0=\boldsymbol{Z}_0(\boldsymbol{x}_0)}),\\
&|\det\left(\nabla_{\boldsymbol{x}} T(\cdot; \theta(\boldsymbol{x}_0, t))\right)| = \prod_{i=1}^{L} |\det\left(\nabla_{\boldsymbol{x}_{[i-1]}} T_{[i]}(\cdot)\right)|,
\end{aligned}$$

which motivates the use of techniques that enable efficient computation of both the inverse and the Jacobian matrix of $T_{[i]}(\cdot)$. In this work, we adopt the structure with descending active transformation dimensions in the standard KRnet [38], which is inspired by the KR rearrangement [34] and affine coupling layers [10], to enhance accuracy and reduce model complexity in high-dimensional settings.

Let $\boldsymbol{x}^\top = \left(\left(\boldsymbol{x}^{(1)}\right)^\top, \left(\boldsymbol{x}^{(2)}\right)^\top, \ldots, \left(\boldsymbol{x}^{(K)}\right)^\top\right)$ be a partition of $\boldsymbol{x}^\top$, where $\boldsymbol{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_{d_i}^{(i)}\right)^\top$ with $1 \leq K \leq d$, $1 \leq d_i \leq d$ and $\sum_{i=1}^{K} d_i = d$. Denote $\boldsymbol{x}^{(i:j)} = \left(\left(\boldsymbol{x}^{(i)}\right)^\top, \left(\boldsymbol{x}^{(i+1)}\right)^\top, \ldots, \left(\boldsymbol{x}^{(j)}\right)^\top\right)^\top$. Our conditional normalizing flow $T_{\text{C-KRnet}}$ is

defined as follows,

$$T_{[1]} = \tilde{T}_K, \quad T_{[k]} = \begin{pmatrix} \tilde{T}_{K+1-k} \\ \mathrm{Id}_{K+2-k:K} \end{pmatrix}, \quad k = 2, \ldots, K,$$

$$\boldsymbol{Z}_t|_{\boldsymbol{Z}_0=\boldsymbol{z}_0} = T_{\text{C-KRnet}}(\boldsymbol{X}_t|_{\boldsymbol{X}_0=\boldsymbol{x}_0}; \theta(\boldsymbol{x}_0, t)) = \begin{pmatrix} \tilde{T}_1 \\ \mathrm{Id}_{(2:K)} \end{pmatrix} \circ \cdots \circ \begin{pmatrix} \tilde{T}_{K-1} \\ \mathrm{Id}_{(K)} \end{pmatrix} \circ \tilde{T}_K(\boldsymbol{X}_t|_{\boldsymbol{X}_0=\boldsymbol{x}_0}),$$

where $\mathrm{Id}_{(i:j)}(\boldsymbol{x}) = \boldsymbol{x}^{(i:j)}$ denotes the projection operator, $\tilde{T}_k : [-1,1]^{\sum_{i=1}^k d_i} \to [-1,1]^{\sum_{i=1}^k d_i}$ is an invertible mapping of $\boldsymbol{x}^{(1:k)}$ that is constructed by the composition of conditional affine coupling layers defined in the Section 3.2.1.

The conditional normalizing flow $T_{\text{C-KRnet}}$ is primarily constructed using two nested loops: an outer loop and an inner loop. The outer loop consists of K stages, each corresponding to a mapping $\tilde{T}_k$. The inner loop $\tilde{T}_k$ contains l_k stages, representing the number of coupling layers. Following the application of $\tilde{T}_k$, the kth partition is kept fixed in the subsequent outer stages. As the outer loop progresses, the active dimension gradually decreases, which motivates reducing the depth l_k for later stages.

3.2.1. Conditional affine coupling layer. In this section, we focus on the construction of the inner loop $\tilde{T}_k = \tilde{T}_{k,l_k} \circ \cdots \circ \tilde{T}_{k,2} \circ \tilde{T}_{k,1}$. To this end, we first introduce a conditional affine coupling layer, given some information $\boldsymbol{\xi}$ and time t, applied to a vector $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2) \in \mathbb{R}^m$, where $\boldsymbol{y}_1 \in \mathbb{R}^{m_1}$, and $\boldsymbol{y}_2 \in \mathbb{R}^{m-m_1}$ as follows:

$$\begin{cases} \hat{\boldsymbol{y}}_1 = \boldsymbol{y}_1, \\ \hat{\boldsymbol{y}}_2 = \boldsymbol{y}_2 \odot \left(\mathbf{1}_{m-m_1} + \beta \tanh(t \cdot \boldsymbol{s})\right) + e^{\boldsymbol{\zeta}} \odot \tanh(t \cdot \boldsymbol{q}), \end{cases} \tag{3.8}$$

where $\beta \in (0,1)$ is user defined, $\boldsymbol{\zeta}$ is trainable and $\boldsymbol{s}$ and $\boldsymbol{q}$ are outputs of a neural network that takes as input the unchanged portion $\boldsymbol{y}_1$, the given information $\boldsymbol{\xi}$ and time t, i.e.,

$$(\boldsymbol{s}, \boldsymbol{q}) = \mathrm{NN}(\boldsymbol{y}_1, \boldsymbol{\xi}, t). \tag{3.9}$$

To emphasize the influence of the initial state, we adopt the approach proposed in [18], applying a random Fourier feature transformation [41] before the fully connected layers of the neural network used in equation (3.9). The resulting neural network is

$$\begin{aligned} &\boldsymbol{h}_0 = [\boldsymbol{y}_1, \boldsymbol{\xi}, t]^\top, \quad \boldsymbol{h}_1 = \left[\sin\left(\frac{1}{e^\gamma}\boldsymbol{F}\boldsymbol{h}_0 + \boldsymbol{b}_0\right), \cos\left(\frac{1}{e^\gamma}\boldsymbol{F}\boldsymbol{h}_0 + \boldsymbol{b}_0\right), \quad \boldsymbol{h}_0\right]^\top, \\ &\boldsymbol{h}_j = \mathrm{SiLU}(\boldsymbol{W}_{j-1}\boldsymbol{h}_{j-1} + \boldsymbol{b}_{j-1}), \text{ for } j = 2, \ldots, M, \quad (\boldsymbol{s}, \boldsymbol{q}) = \boldsymbol{W}_M\boldsymbol{h}_M + \boldsymbol{b}_M, \end{aligned} \tag{3.10}$$

where $\boldsymbol{F} \in \mathbb{R}^{r_h/2 \times \dim(h_0)}$, $\boldsymbol{b}_0 \in \mathbb{R}^{r_h/2}$ are sampled from a standard normal distribution and a uniform distribution over $[0, 2\pi]^{r_h/2}$, respectively and r_h denotes the dimension of the random features. Both $\boldsymbol{F}$ and $\boldsymbol{b}_0$ are fixed after initialization. $\gamma, \{\boldsymbol{W}_j, \boldsymbol{b}_j\}_{j=1}^M$ are trainable parameters, and the sigmoid linear unit (SiLU) function is used as the activation function.

Since the conditional affine coupling layer (3.8) updates only $\boldsymbol{y}_2$, we alternate the roles of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ in subsequent layers to ensure a full update of $\boldsymbol{y}$. Note that the transformation defined by the conditional affine coupling layer (3.8) reduces to the identity when $t = 0$, which is consistent with using the PDF corresponding to the linearized SDE as the base distribution of the conditional normalizing flow.

Now we are ready to construct $\tilde{T}_k$, which consists of l_k conditional affine coupling layers. Denote $\boldsymbol{x}_{[0]} = \boldsymbol{x}, \boldsymbol{x}_{[k]} = T_{[k]}(\boldsymbol{x}_{[k-1]})$, $\boldsymbol{\xi} = \boldsymbol{x}_0$, and $\boldsymbol{y}_{[k]} = \boldsymbol{x}_{[k]}^{(1:K-k)}$ the active part of $\boldsymbol{x}_{[k]}$ for $k = 1, \ldots, K-1$. We apply a sequence of conditional affine coupling transformations (3.8) to $\boldsymbol{y}_{[k]}$ using a half-half split.

4. Deep adaptive algorithm. Given an SDE, we solve the corresponding transition FPE with Dirac delta functions centered at various points as initial conditions, which is considered as offline training. Let $p(\boldsymbol{x}, t|\boldsymbol{x}_0)$ be a conditional probability density function defined on $\mathbb{R}^d \times [0, t_f] \times \Omega_0$. It satisfies the transition FPE (2.6). By using the solution of the linearized SDE as the base stochastic process and enforcing the transformation to be the identity mapping at $t = 0$, the initial condition is naturally satisfied. Applying PINNs yields

(4.1)
$$\mathcal{L}(\boldsymbol{\theta}) = \int_{\mathbb{R}^d \times (0,t_f] \times \Omega_0} |r(\boldsymbol{x}, \boldsymbol{x}_0, t; p_{\boldsymbol{\theta}})|^2 \mathrm{d}\rho(\boldsymbol{x}, \boldsymbol{x}_0, t), \quad r(\boldsymbol{x}, \boldsymbol{x}_0, t; p_{\boldsymbol{\theta}}) = \frac{\partial p_{\boldsymbol{\theta}}}{\partial t} - \mathcal{L}^*_{\boldsymbol{f},\boldsymbol{g}}[p_{\boldsymbol{\theta}}],$$

where $\rho(\boldsymbol{x}, \boldsymbol{x}_0, t) = \rho(\boldsymbol{x}|\boldsymbol{x}_0, t)\rho(t|\boldsymbol{x}_0)\rho(\boldsymbol{x}_0)$ is the sampling measure to be used in the training process. For the case where $\boldsymbol{x}_0$ lies within a bounded domain Ω_0 and t_f is finite, we define $\rho(t|\boldsymbol{x}_0)\rho(\boldsymbol{x}_0)$ to be the uniform measure over $\Omega_0 \times [0, t_f]$. For the remaining term, $\rho(\boldsymbol{x}|\boldsymbol{x}_0, t)$, we employ an adaptive sampling strategy similar to that used for specific FPEs in our previous work [13, 49]. The only difference is that, instead of sampling solely from the uniform distribution at the start, we are motivated to sample from the base stochastic process, as it approximates the target transition PDF particularly well when the time t is small. After some training epochs, we obtain a conditional normalizing flow that approximates the target solution and can be used to generate samples from it. These samples are then used to update the collocation points. The procedure offers an adaptive strategy for designing the sampling measure $\rho_k(\boldsymbol{x}|\boldsymbol{x}_0, t)$ at stage k, by combining the uniform measure, the measure from the previous stage, and the one induced by the current normalizing flow, i.e.,

(4.2)
$$\begin{aligned} &\rho_0(\boldsymbol{x}|\boldsymbol{x}_0, t) = \frac{\gamma_1}{\gamma_1 + \gamma_3}\mu(\boldsymbol{x}) + \frac{\gamma_3}{\gamma_1 + \gamma_3}\rho_{\boldsymbol{Z}}(\boldsymbol{x}|\boldsymbol{x}_0, t), \quad \gamma_1 + \gamma_2 + \gamma_3 = 1, \\ &\rho_{k+1}(\boldsymbol{x}|\boldsymbol{x}_0, t) = \gamma_1\mu(\boldsymbol{x}) + \gamma_2\rho_k(\boldsymbol{x}|\boldsymbol{x}_0, t) + \gamma_3\rho_{\text{C-KRnet},\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{x}_0, t), \end{aligned}$$

where μ denotes the uniform measure over a user-defined bounded domain, $\rho_{\boldsymbol{Z}}(\boldsymbol{x}, t|\boldsymbol{x}_0)$ denotes the measure induced by the base random process, $\rho_{\text{C-KRnet},\boldsymbol{\theta}}(\boldsymbol{x}, t|\boldsymbol{x}_0)$ denotes the measure induced by the current conditional normalizing flow and $\gamma_1, \gamma_2, \gamma_3$ are hyperparameters used to tune the sampling measure.

The transition PDF of the base stochastic process serves as a good approximation of the target transition PDF when t is small. However, as $t \to 0$, the target solution approaches a Dirac delta function, and even small perturbations may lead to the divergence of the residual. To illustrate this, we consider a small perturbation around the transition PDF of the linearized SDE and analyze the asymptotic behavior of the corresponding residual as $t \to 0$.

PROPOSITION 4.1. *Suppose the drift term $\boldsymbol{f}$ in the SDE (2.1) is differentiable, and the diffusion term $\boldsymbol{g}$ is constant, $\boldsymbol{G} = (G^{ij})_{i,j\le d} = \frac{1}{2}\boldsymbol{g}\boldsymbol{g}^\top$ is non-degenerate. Assume that $\|\nabla \boldsymbol{f}(\boldsymbol{x})\| \le C_{\boldsymbol{f}}$, $\forall \boldsymbol{x} \in \mathbb{R}^d$ for some $C_{\boldsymbol{f}} \ge 0$. Given an initial state $\boldsymbol{x}_0$ and time t, let $\tilde{p}(\boldsymbol{x}, t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}(t), \boldsymbol{\Sigma}(t))$ where $\boldsymbol{\Sigma}(t) \sim t\boldsymbol{G}$, and $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{m}(t)$ is*

bounded. Then,

$$\int_{\mathbb{R}^d} |r(\boldsymbol{x}, \boldsymbol{x}_0, t; \tilde{p})|^2 \mathrm{d}\boldsymbol{x} = \mathcal{O}(t^{-(\frac{d}{2}+2)}), \tag{4.3}$$

$$\mathbb{E}_{\boldsymbol{x}\sim\tilde{p}}\left[|r(\boldsymbol{x}, \boldsymbol{x}_0, t; \tilde{p})|^2\right] = \int_{\mathbb{R}^d} |r(\boldsymbol{x}, \boldsymbol{x}_0, t; \tilde{p})|^2 \tilde{p}(\boldsymbol{x}, t|\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x} = \mathcal{O}(t^{-(d+2)}). \tag{4.4}$$

Proof. See Appendix D.

According to Proposition 4.1, although the approximation may be close to the target transition PDF for small t, the integral of the residual over the spatial domain can scale as $t^{-(d+2)}$ or $t^{-(\frac{d}{2}+2)}$ as $t \to 0$ depending on the sampling measure $\rho(\boldsymbol{x}|\boldsymbol{x}_0, t)$, which potentially leads to a blow-up. This motivates us to define a time-weighted loss function. Incorporating the adaptive sampling strategy, we rewrite the loss function at the $k-$th stage as a sum of three parts:

$$\begin{aligned}\mathcal{L}^k(\boldsymbol{\theta}) =& \gamma_1 \int_{\Omega\times\Omega_0\times(0,t_f]} |r(\boldsymbol{x}, \boldsymbol{x}_0, t; p_{\boldsymbol{\theta}})|^2 \mathrm{d}\mu(\boldsymbol{x})\mathrm{d}\rho(t|\boldsymbol{x}_0)\mathrm{d}\rho(\boldsymbol{x}_0) \\ &+ \gamma_2 \int_{\Omega\times\Omega_0\times(0,t_f]} |r(\boldsymbol{x}, \boldsymbol{x}_0, t; p_{\boldsymbol{\theta}})|^2 \mathrm{d}\rho_{k-1}(\boldsymbol{x}|\boldsymbol{x}_0, t)\mathrm{d}\rho(t|\boldsymbol{x}_0)\mathrm{d}\rho(\boldsymbol{x}_0) \\ &+ \gamma_3 \int_{\Omega\times\Omega_0\times(0,t_f]} |r(\boldsymbol{x}, \boldsymbol{x}_0, t; p_{\boldsymbol{\theta}})|^2 \mathrm{d}\rho_{\text{C-KRnet},\boldsymbol{\theta}}(\boldsymbol{x}, t|\boldsymbol{x}_0)\mathrm{d}\rho(t|\boldsymbol{x}_0)\mathrm{d}\rho(\boldsymbol{x}_0).\end{aligned}$$

We apply weighting functions $w_1(t) = t^{\frac{d}{2}+2}$ to the first part, and $w_2(t) = t^{d+2}$ to the last two parts. The resulting weighted loss function is then defined as

$$\begin{aligned}\mathcal{L}_w^k(\boldsymbol{\theta}) =& \gamma_1 \int_{\Omega\times\Omega_0\times(0,t_f]} |r(\boldsymbol{x}, \boldsymbol{x}_0, t; p_{\boldsymbol{\theta}})|^2 t^{(\frac{d}{2}+2)} \mathrm{d}\mu(\boldsymbol{x})\mathrm{d}\rho(t|\boldsymbol{x}_0)\mathrm{d}\rho(\boldsymbol{x}_0) \\ &+ \gamma_2 \int_{\Omega\times\Omega_0\times(0,t_f]} |r(\boldsymbol{x}, \boldsymbol{x}_0, t; p_{\boldsymbol{\theta}})|^2 t^{(d+2)} \mathrm{d}\rho_{k-1}(\boldsymbol{x}|\boldsymbol{x}_0, t)\mathrm{d}\rho(t|\boldsymbol{x}_0)\mathrm{d}\rho(\boldsymbol{x}_0) \\ &+ \gamma_3 \int_{\Omega\times\Omega_0\times(0,t_f]} |r(\boldsymbol{x}, \boldsymbol{x}_0, t; p_{\boldsymbol{\theta}})|^2 t^{(d+2)} \mathrm{d}\rho_{\text{C-KRnet}}(\boldsymbol{x}|\boldsymbol{x}_0, t)\mathrm{d}\rho(t|\boldsymbol{x}_0)\mathrm{d}\rho(\boldsymbol{x}_0).\end{aligned}$$

We emphasize that the usage of the linearized SDE respects the causality, and the weighting functions introduced here not only help prevent blow-up but also assign greater weight to more challenging training regions, making our method well-suited for the transition FPE.

We introduce an indicator variable to distinguish samples drawn from different distributions. Specifically, let $N = N_1 + N_2 + N_3$ with proportions $N_1 : N_2 : N_3 = \gamma_1 : \gamma_2 : \gamma_3$. Define $S_1 = \left\{\left(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i, \eta^i = 0\right)\right\}_{i=1}^{N_1}$, where $\left(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i\right)$ are sampled from $\mu(\boldsymbol{x})\rho(t|\boldsymbol{x}_0)\rho(\boldsymbol{x}_0)$. Similarly, define $S_2 = \left\{\left(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i, \eta^i = 1\right)\right\}_{i=N_1+1}^{N_1+N_2}$ where $\left\{\left(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i\right)\right\}$ are drawn from $\rho_{k-1}(\boldsymbol{x}|\boldsymbol{x}_0, t)\rho(t|\boldsymbol{x}_0)\rho(\boldsymbol{x}_0)$. Define $S_3 = \left\{\left(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i, \eta^i = 1\right)\right\}_{i=N_1+N_2+1}^{N}$ where $\left\{\left(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i\right)\right\}$ are sampled from $\rho_{\text{C-KRnet}}(\boldsymbol{x}|\boldsymbol{x}_0, t)\rho(t|\boldsymbol{x}_0)\rho(\boldsymbol{x}_0)$. Let $S = S_1 \cup S_2 \cup S_3$. The empirical time-weighted loss function is then defined as:

$$\widehat{\mathcal{L}}_w^k(\boldsymbol{\theta}; S) \coloneqq \frac{1}{N}\sum_{i=1}^{N} |r(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i; p_{\boldsymbol{\theta}})|^2 \left(\eta^i t^{d+2} + (1-\eta^i) t^{\frac{d}{2}+2}\right). \tag{4.5}$$

The optimal parameter $\boldsymbol{\theta}^*$ can be obtained via solving the optimization problem:

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}} \widehat{\mathcal{L}}_w(\boldsymbol{\theta}; \tilde{S}). \tag{4.6}$$

The adaptive training process is summarized in Algorithm 4.1. The training set is divided into several mini-batches, and the Adam optimizer is used to update the parameters $\boldsymbol{\theta}$. The training set S is updated according to the strategy (4.2).

Algorithm 4.1 Solving the transition FPE with various initial states

Input: maximum epoch number N_e, maximum iteration number N_{adaptive}, rate $\gamma_1, \gamma_2, \gamma_3$, initial learning rate l_r, decay rate η, step size n_s, the number of the training samples N.
for $k = 0, \cdots, N_{\text{adaptive}}$ **do**
  if $k = 0$ **then**
    $N_1 = \lfloor \frac{\gamma_1}{\gamma_1+\gamma_3} * N \rfloor$, $N_3 = N - N_1$,
    Generate $S_1 = \{(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i, 0)\}_{i=1}^{N_1}$ with $(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i) \sim \mu(\boldsymbol{x})\rho(t|\boldsymbol{x}_0)\rho(\boldsymbol{x}_0)$,
    $S_3 = \{(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i, 1)\}_{i=N_1+1}^{N}$ with $(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i) \sim \rho_{\text{C-KRnet},\boldsymbol{\theta}}(\boldsymbol{x}, t|\boldsymbol{x}_0)\rho(t|\boldsymbol{x}_0)\rho(\boldsymbol{x}_0)$.
    Initialize the training set $S = S_1 \cup S_3$.
  else
    $N_1 = \lfloor \gamma_1 * N \rfloor$, $N_2 = \lfloor \gamma_2 * N \rfloor$, $N_3 = N - N_1 - N_2$,
    Generate $S_1 = \{(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i, 0)\}_{i=1}^{N_1}$ with $(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i) \sim \mu(\boldsymbol{x})\rho(t|\boldsymbol{x}_0)\rho(\boldsymbol{x}_0)$,
    S_2 consists of N_2 random samples drawn from S.
    $S_3 = \{(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i, 1)\}_{i=N_1+N_2+1}^{N}$, $(\boldsymbol{x}^i, \boldsymbol{x}_0^i, t^i) \sim \rho_{\text{C-KRnet},\boldsymbol{\theta}}(\boldsymbol{x}, t|\boldsymbol{x}_0)\rho(t|\boldsymbol{x}_0)\rho(\boldsymbol{x}_0)$.
    Update the training set: $S = S_1 \cup S_2 \cup S_3$.
  end if
  for $j = 1, \cdots, N_e$ **do**
    if $((k-1)N_e + j)\%n_s == 0$ **then**
      $l_r = \eta * l_r$.
    end if
    Divide S into n mini-batches $\{S^{ib}\}_{ib=1}^{n}$ randomly.
    for $ib = 1, \cdots, n$ **do**
      Compute the loss function (4.5) $\widehat{\mathcal{L}}_w(\boldsymbol{\theta}; S^{ib})$.
      Update $\boldsymbol{\theta}$ by using the Adam optimizer.
    end for
  end for
end for
Output: The predicted solution $p_{\text{C-KRnet},\boldsymbol{\theta}}(\boldsymbol{x}, t|\boldsymbol{x}_0)$.

5. Importance sampling-based approximation of $p(\cdot, t)$. Once the transition PDF is available, a natural way to approximate the solution of equation (2.5) is by approximating the integral (2.7) through the Monte Carlo method,

$$p(\boldsymbol{x}, t) \approx \frac{1}{M}\sum_{i=1}^{M} p(\boldsymbol{x}, t|\boldsymbol{x}_0^i), \quad \boldsymbol{x}_0^i \sim p_0(\boldsymbol{x}), \tag{5.1}$$

where $\{\boldsymbol{x}_0^i\}_{i=1}^{M}$ are samples from the initial distribution p_0. However, the transition PDF approximates the Dirac delta function as $t \to 0$. Consequently, a direct Monte Carlo method may incur large errors due to high variance. Alternatively, we apply importance sampling, leveraging the knowledge of the transition PDF.

$$\begin{aligned} p(\boldsymbol{x}, t) &= \int_{\Omega_0} \frac{p(\boldsymbol{x}, t|\boldsymbol{x}_0)p_0(\boldsymbol{x}_0)}{q(\boldsymbol{x}_0|\boldsymbol{x}, t)} q(\boldsymbol{x}_0|\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}_0 \\ &\approx \hat{p}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) \coloneqq \frac{1}{M}\sum_{i=1}^{M} \frac{p(\boldsymbol{x}, t|\boldsymbol{x}_0^i)p_0(\boldsymbol{x}_0^i)}{q(\boldsymbol{x}_0^i|\boldsymbol{x}, t)}, \quad \boldsymbol{x}_0^i \sim q(\cdot|\boldsymbol{x}, t). \end{aligned} \tag{5.2}$$

Recall that given an initial state $\boldsymbol{x}_0$ and time t, $\boldsymbol{x}$ concentrates around the initial state $\boldsymbol{x}_0$ for small time. The corresponding base distribution used in our conditional normalizing flow is

$$p_{\widehat{X}_t}(\boldsymbol{x}, t|\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{x_0} + \int_0^t e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)(t-s)} \boldsymbol{f}(\boldsymbol{x}_0)\mathrm{d}s, \boldsymbol{\Sigma}[\widehat{\boldsymbol{X}}_t|\widehat{\boldsymbol{X}}_0 = \boldsymbol{x}_0]\right).$$

Let

$$q_1(\boldsymbol{x}_0|\boldsymbol{x}, t) = \mathcal{N}(\boldsymbol{x}_0; m(\boldsymbol{x}, t), \tilde{\boldsymbol{\Sigma}}(\boldsymbol{x}, t)), \tag{5.3}$$

where $m(\boldsymbol{x}, t) = \boldsymbol{x} - \int_0^t e^{\nabla \boldsymbol{f}(\boldsymbol{x})s} \boldsymbol{f}(\boldsymbol{x})\mathrm{d}s$, $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{x}, t) = \int_0^t e^{\nabla \boldsymbol{f}(\boldsymbol{x})s} \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top e^{\nabla \boldsymbol{f}(\boldsymbol{x})^\top s}\mathrm{d}s$. The distribution q_1 is primarily designed to capture the local concentration of the transition PDF and to mitigate the effect of large variance when t is small. However, the approximation accuracy of q_1 deteriorates as t increases. As t grows, the transition PDF becomes increasingly smooth and therefore does not hinder the approximation of $p(\boldsymbol{x}, t)$. This motivates incorporating the initial distribution into the proposal. Accordingly, we propose to use a mixture of $q_1(\cdot|\boldsymbol{x}, t)$ and $p_0(\cdot)$ as the proposal distribution in the importance sampling scheme. Define

$$q_2(\boldsymbol{x}_0|\boldsymbol{x}, t) = \alpha(t)q_1(\boldsymbol{x}_0|\boldsymbol{x}, t) + (1 - \alpha(t))p_0(\boldsymbol{x}_0), \tag{5.4}$$

where $\alpha(t) \in (0, 1)$ is decreasing with respect to t, e.g., $\alpha(t) = \exp(-at), a > 0$. The set of samples $\{\boldsymbol{x}_0^i\}_{i=1}^M \sim q_2(\boldsymbol{x}_0|\boldsymbol{x}, t)$ can be generated by drawing $\alpha(t)M$ samples from q_1 and $(1 - \alpha(t))M$ samples from p_0.

6. Numerical experiments. In this section, we present three numerical experiments to demonstrate the effectiveness of the proposed approach. The conditional normalizing flow is applied to solve the transition FPE under various initial states. Once obtained, the importance sampling method is used to evaluate $p(\boldsymbol{x}, t)$ given an initial distribution. We first consider a benchmark problem where the analytical solution of the transition PDF is available. We next apply our method to an SDE with nonlinear drift and constant diffusion. To further demonstrate the robustness, we investigate the case of state-dependent diffusion.

The experiments are implemented in PyTorch with an initial learning rate of 0.001. The number of quadrature points in the equation (3.6) is set to 10. To quantify the effectiveness of the proposed method, we consider several metrics. For some time t, we compute the relative L^2 error of the approximation for the transition PDF $p(\boldsymbol{x}, t|\boldsymbol{x}_0)$ if the exact transition PDF is known via

$$\mathrm{Rel}(p_{\text{C-KRnet}}(\cdot, t|\cdot)) = \sqrt{\frac{\sum_{i=1}^{N_v}(p(\boldsymbol{x}^i, t|\boldsymbol{x}_0^i) - p_{\text{C-KRnet}}(\boldsymbol{x}^i, t|\boldsymbol{x}_0^i))^2}{\sum_{i=1}^{N_v} p(\boldsymbol{x}^i, t|\boldsymbol{x}_0^i)^2}}, \tag{6.1}$$

where $\{(\boldsymbol{x}_0^i, \boldsymbol{x}^i)\}_{i=1}^{N_v} \subset \Omega_0 \times \Omega$ denotes the validation dataset. When approximating the solution of the FPE given an initial distribution $\rho_0(\boldsymbol{x})$, we consider the relative L^2 error of the approximation $\hat{p}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ of $p(\boldsymbol{x}, t)$, as well as the maximum mean discrepancy (MMD) [17]. The relative L^2 error reads

$$\mathrm{Rel}(\hat{p}_{\boldsymbol{\theta}}(\cdot, t)) = \sqrt{\frac{\sum_{i=1}^{N_v}(p(\boldsymbol{x}^i, t) - \hat{p}_{\boldsymbol{\theta}}(\boldsymbol{x}^i, t))^2}{\sum_{i=1}^{N_v} p(\boldsymbol{x}^i, t)^2}}, \tag{6.2}$$

where $\{\boldsymbol{x}^i\}_{i=1}^{N_v} \subset \Omega$ denotes the validation dataset. When the analytical form of the transition PDF is available, we employ Gauss–Kronrod quadrature rule to obtain the reference solution $p(\boldsymbol{x}, t)$ for a given initial distribution. In the absence of an analytical solution, the alternating-direction implicit (ADI) finite difference scheme [32] with a time step of 0.001 and a spatial discretization size of 0.01 is used to compute the reference solution for cases where the initial distributions are continuous over the entire domain $\mathbb{R}^2$. An MMD is introduced in particular to assess the effectiveness of the proposed method for discontinuous initial distributions.

$$\text{MMD}^2(t) = \frac{1}{M_1^2}\sum_{i,j=1}^{M_1} K(\boldsymbol{x}_t^i, \boldsymbol{x}_t^j) - \frac{2}{M_1 M_2}\sum_{i,j=1}^{M_1,M_2} K(\boldsymbol{x}_t^i, \boldsymbol{y}_t^j) + \frac{1}{M_2^2}\sum_{i,j=1}^{M_2} K(\boldsymbol{y}_t^i, \boldsymbol{y}_t^j), \tag{6.3}$$

$$K(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{3}\exp\left(-\frac{4\|\boldsymbol{x}-\boldsymbol{y}\|^2}{\hat{\sigma}^2}\right) + \frac{1}{3}\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{\hat{\sigma}^2}\right) + \frac{1}{3}\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{4\hat{\sigma}^2}\right).$$

$\{\boldsymbol{x}_t^i\}$ and $\{\boldsymbol{y}_t^i\}$ are samples at time t obtained using the Euler-Maruyama scheme with a time step of 0.001 and generated by the normalizing flow, respectively, with initial distribution $p_0(\boldsymbol{x})$. $\hat{\sigma}$ is the median distance between $\{\boldsymbol{x}_t^i\}$ and $\{\boldsymbol{y}_t^i\}$.

6.1. Beneš SDE. We start with demonstrating the performance of our method on the Beneš SDE,

$$\begin{cases} \mathrm{d}\boldsymbol{X}_t = \tanh \boldsymbol{X}_t \mathrm{d}t + \boldsymbol{I}_d \mathrm{d}\boldsymbol{W}_t, \\ \boldsymbol{X}_0 = \boldsymbol{x}_0, \quad \boldsymbol{x}_0 \in [-1, 1]^d, \end{cases} \tag{6.4}$$

for which the corresponding transition PDF admits an analytical solution,

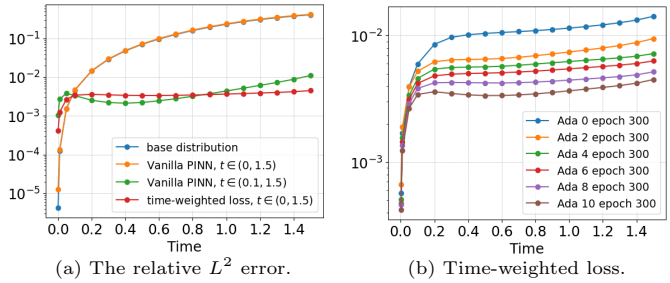
$$p(\boldsymbol{x}, t|\boldsymbol{x}_0) = \frac{1}{(2\pi t)^{d/2}}\exp\left(-\frac{d}{2}t\right)\exp\left(-\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{x}_0\|^2\right)\prod_{i=1}^{d}\frac{\cosh x_i}{\cosh x_{0,i}}.$$

6.1.1. Two-dimensional case. We first consider the two-dimensional case. The conditional normalizing flow is composed of eight coupling layers. The parameters of each coupling layer are generated by a neural network with 32 random Fourier features and two hidden layers with 32 neurons. Training is performed for 10 adaptive iterations, each with 300 epochs. The Adam optimizer is used with an initial learning rate of 0.001, which is halved every 2000 epochs. 2×10^5 training points are employed with a batch size of 10^4 each epoch. A validation dataset of 10^6 samples is used to compute relative errors throughout the training process.

We first investigate the approximation performance of the transition PDF for various initial states, using the following five types of loss functions:

1. Vanilla PINN, $t \in (0, 1.5)$, $\gamma_1 = 0.2, \gamma_2 = 0.6, \gamma_3 = 0.2$;
2. Vanilla PINN, $t \in (0.1, 1.5)$, $\gamma_1 = 0.2, \gamma_2 = 0.6, \gamma_3 = 0.2$;
3. Time-weighted loss function, $t \in (0, 1.5)$, $\gamma_1 = 0.2, \gamma_2 = 0.6, \gamma_3 = 0.2$.

We present the $\text{Rel}(p_{\text{C-KRnet}}(\cdot, t|\cdot))$ (6.1) in Figure 1a, where the validation dataset $\{\boldsymbol{x}_0, \boldsymbol{x}\}$ consists of 10^5 samples drawn uniformly from $[-1, 1]^2 \times [-5, 5]^2$. The blue curve, which overlaps with the orange one, represents the relative error of the base distribution. While it achieves high accuracy as $t \to 0$, its approximation quality deteriorates as t increases. A similar trend is observed when applying vanilla PINN. Although the base distribution stays close to the ground truth and the conditional normalizing flow remains close to the identity at small times, the corresponding residual at small times may still dominate, so the approximation accuracy is nearly the



(a) The relative L^2 error. (b) Time-weighted loss.

Fig. 1: Left: $\mathrm{Rel}(p_{\text{C-KRnet}}(\cdot, t|\cdot))$ for different loss functions. Right: the decay of $\mathrm{Rel}(p_{\text{C-KRnet}}(\cdot, t|\cdot))$ against the adaptive iterations for time-weighted loss.

same as that of the base distribution. If we exclude small times by setting the minimum training time to $t = 0.1$, the PDE constraints are no longer imposed near the initial condition. In this case, the approximation relies primarily on the continuity of the flow, and the accuracy improves for $t > 0.1$. By contrast, our method exploits the information from the base distribution and provides accurate approximations even as t increases, owing to the expressiveness of the conditional normalizing flow and the more balanced PINN's loss. Notably, we do not observe the typical error growth with time that often occurs in time-dependent problems. This highlights the effectiveness of the time-weighted loss function in mitigating error accumulation over time. Figure 1b further illustrates the decay of the $\mathrm{Rel}(p_{\text{C-KRnet}}(\cdot, t|\cdot))$ across adaptive iterations for the time-weighted loss function. Leveraging samples from the solution itself allows the model to capture the local behavior of the solution and enhances training efficiency.

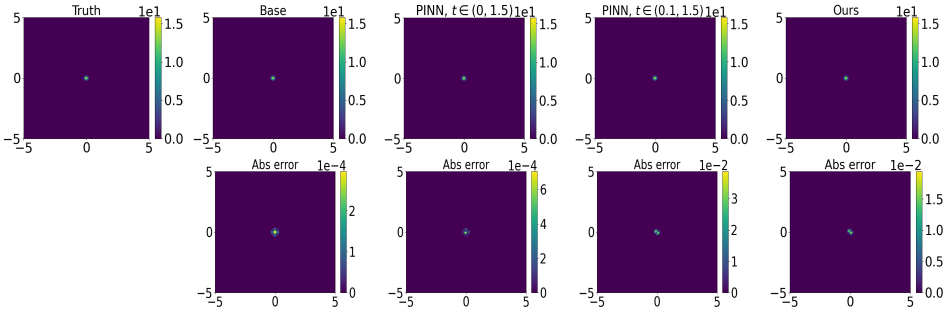


Fig. 2: Solutions (top row) and the absolute errors (bottom row) at $t = 0.01$. The first column shows the ground-truth solution. Columns 2–5 display numerical approximations obtained by, from left to right, the base distribution, Vanilla PINN with $t \in (0, 1.5)$, Vanilla PINN with $t \in (0.1, 1.5)$, time-weighted loss function with $t \in (0, 1.5)$.

Figure 2, Figure 3, and Figure 4 present the solutions at different times with the initial state $\boldsymbol{x}_0 = (0, 0)$ computed with the different loss functions. They all align closely to the ground truth except for the vanilla PINN's loss for $t \in (0, 1.5)$. While the base distribution and vanilla PINNs yield much more accurate approximations for small t, their performance degrades progressively as t increases. Especially, Figure 4 demonstrates that the conditional normalizing flow is capable of capturing multimodal solutions. These results confirm the effectiveness of the proposed approach.

We next assess how well the solution of the FPE $p(\boldsymbol{x}, t)$ can be approximated under various initial conditions. To this end, we employ a variety of Beta distributions on $[-1, 1]^d$ as initial conditions, thereby evaluating the effectiveness of the conditional

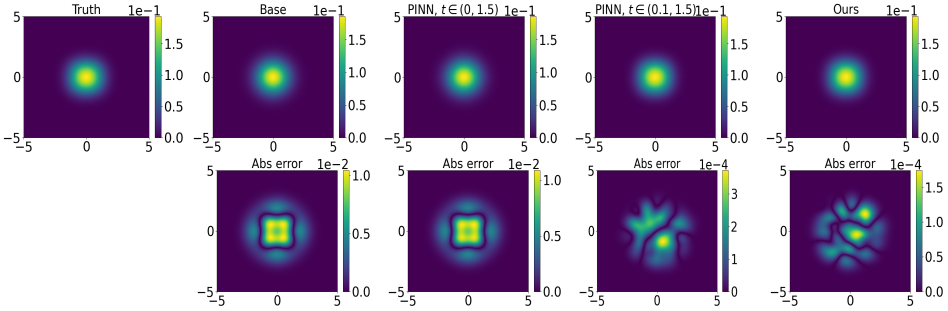


Fig. 3: Same as Figure 2 for $t = 0.5$.

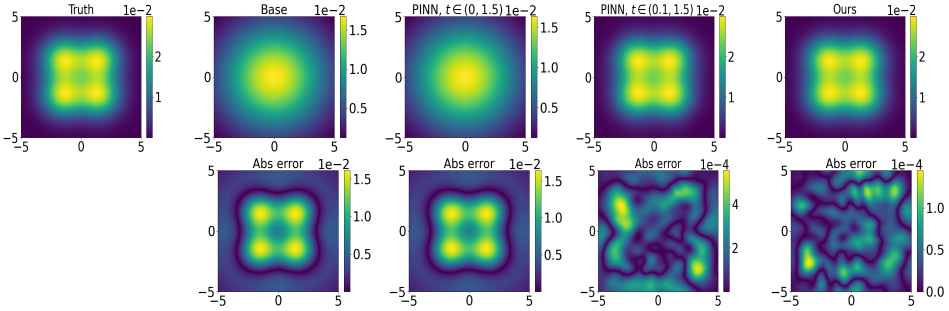


Fig. 4: Same as Figure 2 for $t = 1.5$.

normalizing flow for operator learning problems. More precisely,

$$\mathbf{Beta}(\boldsymbol{x}; \boldsymbol{\zeta}, \boldsymbol{\eta}) = \prod_{i=1}^{d} \frac{1}{\boldsymbol{B}(\zeta_i, \eta_i)} \left(\frac{1+x_i}{2}\right)^{\zeta_i - 1} \left(\frac{1-x_i}{2}\right)^{\eta_i - 1}, \boldsymbol{x} \in [-1, 1]^d,$$

where $\boldsymbol{B}(\zeta_i, \eta_i)$ denotes a beta function. We first explore the uniform case where $\boldsymbol{\zeta} = \boldsymbol{\eta} = (1, 1)$. We use the conditional normalizing flow trained with the time-weighted loss as an approximation of the transition PDF. Figure 5 reports the $\mathrm{Rel}(p_{\boldsymbol{\theta}}(\cdot, t))$ (6.2) with a validation data set consisting 10^4 grid points over $[-5, 5]^2$ for different numbers of Monte Carlo samples used in evaluating $p_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ (5.2). Using q_1 as the proposal distribution yields the most accurate approximation for $t < 0.3$, whereas drawing samples directly from the initial distribution p_0 provides more reliable approximations for $t \geq 0.4$. By contrast, adopting q_2, a mixture of q_1 and p_0 with $\alpha(t) = \exp(-6t)$, offers a balanced compromise, delivering acceptable accuracy across both small and large times. Figure 6 further illustrates the error decay with respect to the number of Monte Carlo samples. The relative error follows the Monte Carlo error scaling for $M \leq 10^5$. For a larger Monte Carlo sample size, however, the error in approximating the transition PDF becomes dominant. We also present the variance of the importance-sampling Monte Carlo estimator $\hat{p}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ in (5.2) as a function of t in Figure 7 using 10^4 samples. The proposal q_1 substantially reduces the variance as $t \to 0$, but its variance increases as t grows. However, q_2 with different $\alpha(t)$, which combines q_1 with the initial distribution p_0, yields a more balanced variance behavior over time. We note that traditional methods lose accuracy when the initial distribution is uniform, owing to the discontinuity of the initial distribution over the entire domain, if no special treatment of the discontinuity is taken. In contrast, our proposed methods remain unaffected by this discontinuity, thereby demonstrating their robustness in handling discontinuous initial data.

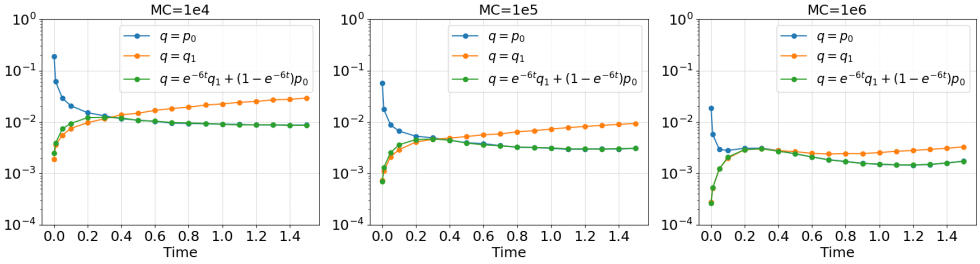
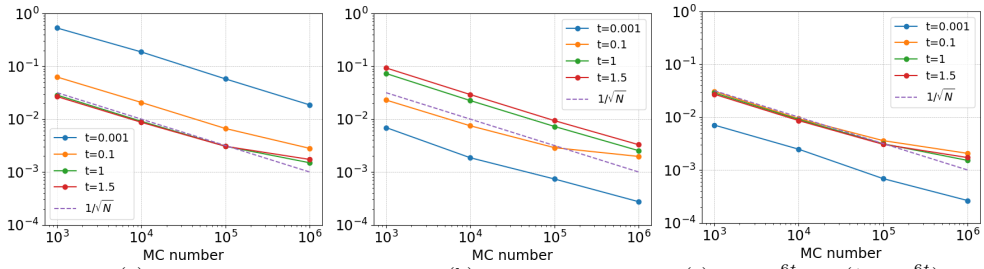


Fig. 5: Uniform case: Evolution of $\mathrm{Rel}(\hat{p}_{\boldsymbol{\theta}}(\cdot, t))$ over time. The subplots are arranged from left to right with $M = 10^4, 10^5, 10^6$.



(a) $q = p_0$ (b) $q = q_1$ (c) $q = e^{-6t}q_1 + (1 - e^{-6t})p_0$

Fig. 6: Uniform case: The decay of $\mathrm{Rel}(\hat{p}_{\boldsymbol{\theta}}(\cdot, t))$ with respect to the number of Monte Carlo samples at different times. Left: Draw samples from p_0; Middle: Draw samples from q_1; Right: Draw samples from the mixture of p_0 and q_1.

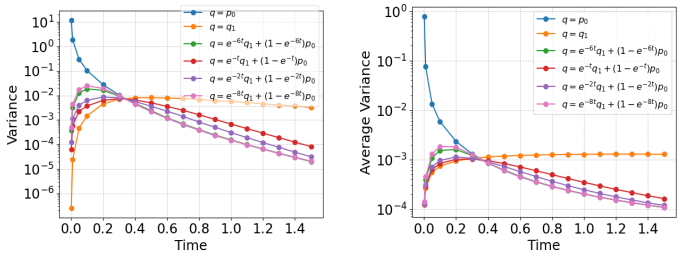


Fig. 7: The variance of $\hat{p}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ over time. Left: $\boldsymbol{x} = (0, 0)$. Right: Average variance of the set $\{\boldsymbol{x}^i\}$ consisting 10^4 grid points in $[-5, 5]^2$.

Moreover, we report in Figure 8 the time evolution of the $\mathrm{Rel}(p_{\boldsymbol{\theta}}(\cdot, t))$ for other Beta distributions chosen as initial conditions using q_2 with $\alpha(t) = \exp(-6t)$ as the proposal distribution. The numbers of Monte Carlo samples used in evaluating $\hat{p}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ are set to be 10^4 and 10^6. The results confirm that the proposed method yields reliable approximations.

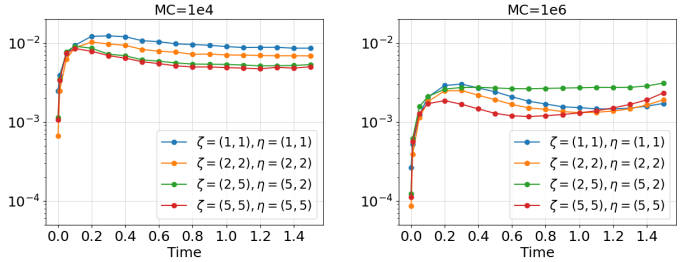
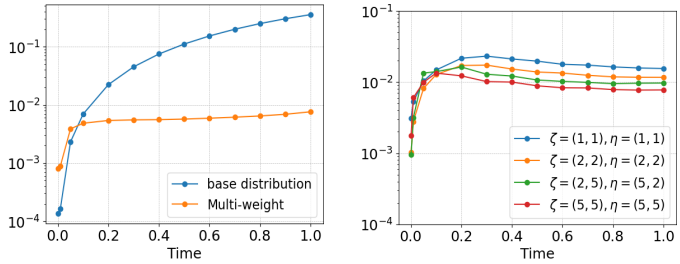


Fig. 8: Time evolution of the $\mathrm{Rel}(\hat{p}_{\boldsymbol{\theta}}(\cdot, t))$ for various initial distributions. Left: $M = 10^4$; Right: $M = 10^6$.

6.1.2. Four-dimensional case. We then solve the four-dimensional FPE with various initial distributions using the proposed time-weighted loss function. Moreover, when evaluating the interior residual, we scale the solution by a constant factor, set to 1000 in the four-dimensional case, to avoid numerical underflow, since the solution of the FPE becomes very small in higher dimensions. The conditional normalizing flow follows the KRnet structure with descending active transformation dimensions. It consists of three stages with $l_1 = 10, l_2 = 8$, and $l_3 = 6$. The parameters of each coupling layer are determined by a neural network with the same structure as in the two-dimensional case. We apply 20 adaptive iterations, each comprising 300 epochs. The updating rates are set to $\gamma_1 = 0.2$, $\gamma_2 = 0.4$, and $\gamma_3 = 0.4$. The initial learning rate is set to 0.001 and is halved every 4000 epochs. 2×10^5 training samples are employed with a batch size of 2×10^4 per epoch.



(a) $p(\boldsymbol{x}, t|\boldsymbol{x}_0)$ (b) $p(\boldsymbol{x}, t)$

Fig. 9: The relative L^2 errors in the four-dimensional case. Left: $\mathrm{Rel}(p_{\text{C-KRnet}}(\cdot, t|\cdot))$; Right: $\mathrm{Rel}(\hat{p}_{\boldsymbol{\theta}}(\cdot, t))$ for various initial distributions with $M = 10^4$.

Figure 9a reports the $\mathrm{Rel}(p_{\text{C-KRnet}}(\cdot, t|\cdot))$, where the validation dataset $\{\boldsymbol{x}_0, \boldsymbol{x}\}$ is uniformly sampled from $[-1, 1]^4 \times [-5, 5]^4$ with a total of 2×10^6 samples. It shows that the conditional normalizing flow achieves close agreement with the ground truth. Figure 9b presents $\mathrm{Rel}(\hat{p}_{\boldsymbol{\theta}}(\cdot, t))$ for different Beta distributions as initial conditions. For each initial condition, the numerical solution is obtained using the equation (5.2), where $q = q_2$ is defined in the equation (5.4) with $\alpha(t) = e^{-6t}$ and $M = 10^4$. A total of 10^5 test points are uniformly drawn from $[-5, 5]^4$. Together, these results demonstrate the effectiveness of the proposed method for moderately high-dimensional FPEs.

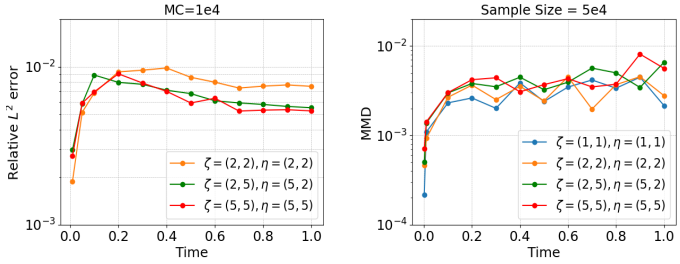
6.2. SDE with nonlinear drift and constant diffusion. In this section, we consider the following SDE with various initial distributions over $[-1, 1]^2$,

$$\mathrm{d}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 2X_2 \\ 2X_1 - 0.8X_2 - 0.2X_1^3 \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \sqrt{0.4} & \\ & \sqrt{0.8} \end{pmatrix} \mathrm{d}\boldsymbol{W}_t. \tag{6.5}$$

Both the structure of the conditional normalizing flow and the training setting are the same as in the two-dimensional Beneš equation. After obtaining the transition PDF, we approximate the PDF corresponding to various Beta initial distributions using the importance sampling (5.2) where $q = q_2$ (5.4) with $\alpha(t) = \exp(-6t)$ and $M = 10^4$.

Figure 10a presents the $\mathrm{Rel}(\hat{p}_{\boldsymbol{\theta}}(\cdot, t))$ (6.2) for different initial conditions, which is evaluated using a validation dataset consisting of 10^4 uniformly distributed grid points over $[-6, 6]^2$. We also report the MMD (6.3) with a sample size of 5×10^4 in Figure 10b. These results demonstrate that the conditional normalizing flow yields highly accurate approximations for small t, as the base distribution is quite close to the ground truth. It also provides reliable approximations as t increases.

6.3. SDE with nonlinear drift and state-dependent diffusion. For our final example, we validate the robustness of the proposed method in approximating



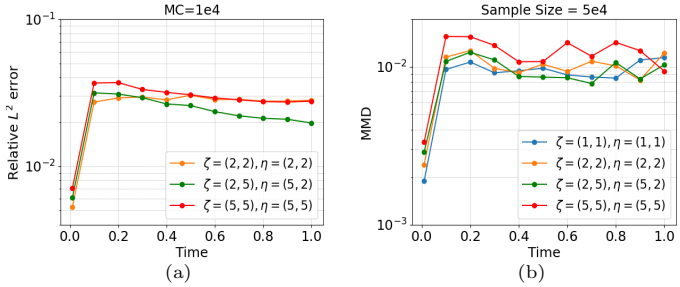
(a) $p(\boldsymbol{x},t)$ (b) $p(\boldsymbol{x},t)$

Fig. 10: Errors for various initial distributions in the nonlinear-drift, constant-diffusion case. Left: $\mathrm{Rel}(\hat{p}_{\boldsymbol{\theta}}(\cdot,t))$. Right: MMD.

the PDF of the following SDE with a state-dependent diffusion,

$$\begin{cases} \mathrm{d}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_2 \\ -0.5X_1 - 0.3X_1^3 + \sin X_2 \end{pmatrix}\mathrm{d}t + \begin{pmatrix} 0.5+0.3X_1 & \\ & 0.4+0.1\sin X_2 \end{pmatrix}\mathrm{d}\boldsymbol{W}_t, \\ \boldsymbol{X}_0 = \boldsymbol{x}_0, \quad \boldsymbol{x}_0 \in [-1,1]^2. \end{cases}$$

The conditional normalizing flow adopts the same structure as in the two-dimensional Beneš equation. Training is conducted over 4 adaptive iterations, each consisting of 1000 epochs. 4×10^5 training points are used, with a batch size of 5×10^4 per epoch.



(a) (b)

Fig. 11: Errors for various initial distributions in the nonlinear-drift, state-dependent-diffusion case. Left: $\mathrm{Rel}(\hat{p}_{\boldsymbol{\theta}}(\cdot,t))$. Right: MMD.

Figure 11a shows the $\mathrm{Rel}(\hat{p}_{\boldsymbol{\theta}}(\cdot,t))$ (6.2) for some initial conditions. The PDF $\hat{p}_\theta(\cdot,t)$ is evaluated using the importance sampling estimator (5.2) where $q=q_2$ with $\alpha(t)=\exp(-6t)$ and 10^4 Monte Carlo samples, while the $\mathrm{Rel}(\hat{p}_\theta(\cdot,t))$ is computed on a validation dataset consisting of 10^4 uniformly distributed grid points over $[-5,5]^2$, with the solution obtained by the ADI scheme used as the reference. We report the MMD (6.3) computed using 5×10^4 samples, as a function of time in the Figure 11b.

7. Conclusion. In this work, we proposed a transition-density-based operator learning method for solving the FPE under various initial conditions. The conditional normalizing flow is employed to approximate the transition PDF given any initial state, effectively constructing a mapping from the target stochastic process to a simple base process. The solution of the linearized SDE serves as a suitable candidate for the base stochastic process. We quantified the discrepancy between the PDFs of the linearized and original SDEs, which is characterized by differences in their drift and diffusion terms. By employing the solution of the linearized SDE and enforcing the identity mapping at $t=0$, the initial condition is naturally satisfied by our conditional normalizing flow, thereby bypassing the explicit handling of the

Dirac delta initial condition. In addition, we designed a time-weighted loss function to alleviate numerical instabilities, ensuring that the proposed method respects causality in evolution PDEs while accounting for the increasing training difficulty as time progresses. To evaluate $p(\cdot, t)$ for arbitrary initial conditions over bounded domains, we used an importance sampling strategy to enhance accuracy. A series of numerical experiments validated the effectiveness and robustness. The four-dimensional Beneš SDE test demonstrated that the proposed method has the potential for solving moderately high-dimensional FPEs. This research contributes to the numerical solution of complex FPEs by combining the approximation capability of the base model (the linearized SDE) with the expressive power of the conditional normalizing flow. In future work, we will further refine this framework by improving the accuracy of the base stochastic process and exploring its potential applications. Another promising direction is to generalize the proposed method to a broader class of evolutionary partial differential equations.

Appendix A. Proof of Proposition 3.1.

Proof. Notice that the solution to the linearized SDE at time t is a Gaussian process, with PDF $p_{\sigma_t}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \mathbb{E}_{\sigma_t}[\widehat{\boldsymbol{X}}_t|\widehat{\boldsymbol{X}}_0 = \boldsymbol{x}_0], \boldsymbol{\Sigma}_t[\widehat{\boldsymbol{X}}_t|\widehat{\boldsymbol{X}}_0 = \boldsymbol{x}_0])$ given the initial state $\boldsymbol{x}_0$. For notation simplicity, we omit the dependence on the initial state $\boldsymbol{x}_0$ and write $p_{\sigma_t}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \mathbb{E}_{\sigma_t}[\widehat{\boldsymbol{X}}_t], \boldsymbol{\Sigma}_t[\widehat{\boldsymbol{X}}_t])$. Moreover, the linearized SDE and the original SDE share the same constant diffusion matrix, and $\nabla \boldsymbol{f}_\nu$ is bounded by assumption. In particular, if $C_{\boldsymbol{f}} = 0$, then $\boldsymbol{f}_\sigma(\boldsymbol{x}) = \boldsymbol{f}_\nu(\boldsymbol{x})$ and consequently $\sigma_t = \nu_t$, which completes the proof. It remains to consider the case $C_{\boldsymbol{f}} > 0$. It is easy to check that $(1 + |\boldsymbol{x}|)^{-2} \left|G^{ij}\right|, (1 + |\boldsymbol{x}|)^{-1} |\boldsymbol{f}_\nu|, (1 + |\boldsymbol{x}|)^{-1} |\boldsymbol{f}_\nu - \boldsymbol{f}_\sigma| \in L^1\left(\mathbb{R}^d \times [0, t_f], \{\sigma_t\}_{0<t<t_f}\right)$ for some finite t_f where $\|f\|_{L^1\left(\mathbb{R}^d \times [0,t_f], \{\sigma_t\}_{0<t<t_f}\right)} = \int_0^{t_f} \int_{\mathbb{R}^d} |f(\boldsymbol{x})| \sigma_t(\mathrm{d}\boldsymbol{x})\mathrm{d}t$. Thus, according to Theorem 1.1 in [4], we have

$$H\left(\sigma_t \mid \nu_t\right) \leq \frac{1}{2} \int_0^t \int_{\mathbb{R}^d} \left| \boldsymbol{G}^{-1/2} (\boldsymbol{f}_\sigma(\boldsymbol{x}) - \boldsymbol{f}_\nu(\boldsymbol{x})) \right|^2 \mathrm{d}\sigma_s \mathrm{d}s. \tag{A.1}$$

Combining this result with the Pinsker–Csiszár–Kullback inequality [16] yields

$$\|\nu_t - \sigma_t\|_{\mathrm{TV}}^2 \leq \frac{1}{4} \int_0^t \int_{\mathbb{R}^d} \left| \boldsymbol{G}^{-1/2} \left(\boldsymbol{f}_\sigma(\boldsymbol{x}) - \boldsymbol{f}_\nu(\boldsymbol{x})\right) \right|^2 \mathrm{d}\sigma_s \mathrm{d}s. \tag{A.2}$$

Notice that $\boldsymbol{G}_\sigma = \boldsymbol{G}_\nu = \boldsymbol{G}$ is a constant, non-degenerate matrix,

$$\left| \boldsymbol{G}^{-1/2} (\boldsymbol{f}_\sigma - \boldsymbol{f}_\nu) \right|^2 = (\boldsymbol{f}_\sigma - \boldsymbol{f}_\nu)^T \boldsymbol{G}^{-1} (\boldsymbol{f}_\sigma - \boldsymbol{f}_\nu) \leq \frac{1}{\lambda_{\min}(\boldsymbol{G})} \left| \boldsymbol{f}_\sigma - \boldsymbol{f}_\nu \right|^2, \tag{A.3}$$

where $\lambda_{\min}(\boldsymbol{G})$ is the smallest eigenvalue of $\boldsymbol{G}$. Meanwhile, for any $s > 0$, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} |\boldsymbol{f}_\sigma(\boldsymbol{x}) - \boldsymbol{f}_\nu(\boldsymbol{x})|^2 \, \mathrm{d}\sigma_s \leq 2 \int_{\mathbb{R}^d} |\boldsymbol{f}_\sigma(\boldsymbol{x}) - \boldsymbol{f}_\sigma(\boldsymbol{x}_0)|^2 + |\boldsymbol{f}_\nu(\boldsymbol{x}_0) - \boldsymbol{f}_\nu(\boldsymbol{x})|^2 \, \mathrm{d}\sigma_s \\
&= 2 \int_{\mathbb{R}^d} |\nabla \boldsymbol{f}_\nu(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)|^2 + \left| \int_0^1 \nabla \boldsymbol{f}_\nu(\boldsymbol{x}_0 + \tau(\boldsymbol{x} - \boldsymbol{x}_0))(\boldsymbol{x} - \boldsymbol{x}_0) \mathrm{d}\tau \right|^2 \mathrm{d}\sigma_s \\
&\leq 4C_{\boldsymbol{f}}^2 \int_{\mathbb{R}^d} |\boldsymbol{x} - \boldsymbol{x}_0|^2 \, \mathrm{d}\sigma_s = 4C_{\boldsymbol{f}}^2 \mathrm{tr}\left(\boldsymbol{\Sigma}_s[\widehat{\boldsymbol{X}}_s]\right) + 4C_{\boldsymbol{f}}^2 \left| \mathbb{E}_{\sigma_s}[\widehat{\boldsymbol{X}}_s] - \boldsymbol{x}_0 \right|^2,
\end{aligned}$$

where the second inequality holds because $\|\nabla \boldsymbol{f}_\nu(\boldsymbol{x})\| \leq C_{\boldsymbol{f}}$, $\forall \boldsymbol{x} \in \mathbb{R}^d$ and the last equality holds because σ_s is a Gaussian measure with mean $\mathbb{E}_{\sigma_s}[\widehat{\boldsymbol{X}}_s]$ and covariance

$\boldsymbol{\Sigma}_s[\widehat{\boldsymbol{X}}_s]$. We give upper bounds for the two terms $\mathrm{tr}\left(\boldsymbol{\Sigma}_s[\widehat{\boldsymbol{X}}_s]\right)$ and $|\mathbb{E}_{\sigma_s}[\widehat{\boldsymbol{X}}_s]-\boldsymbol{x}_0|^2$.

$$\begin{aligned}
&\mathrm{tr}\left(\boldsymbol{\Sigma}_s[\widehat{\boldsymbol{X}}_s]\right)=\int_0^s \mathrm{tr}(e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)(s-\tau)}\boldsymbol{g}\boldsymbol{g}^\top e^{(\nabla \boldsymbol{f}(\boldsymbol{x}_0))^\top(s-\tau)})\mathrm{d}\tau\\
&\le 2\mathrm{tr}(\boldsymbol{G})\int_0^s \sigma_{\max}^2\left(e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)\tau}\right)\mathrm{d}\tau\\
&\le 2\mathrm{tr}(\boldsymbol{G})\int_0^s \left(e^{\sigma_{\max}(\nabla \boldsymbol{f}(\boldsymbol{x}_0)\tau)}\right)^2\mathrm{d}\tau \le 2\mathrm{tr}(\boldsymbol{G})\int_0^s e^{2C_{\boldsymbol{f}}\tau}\mathrm{d}\tau=\mathrm{tr}(\boldsymbol{G})\frac{e^{2C_{\boldsymbol{f}}s}-1}{C_{\boldsymbol{f}}},
\end{aligned}$$

where $\sigma_{\max}(\boldsymbol{A})$ is the spectral norm of the matrix $\boldsymbol{A}$. Meanwhile, equation (3.5) gives

$$|\mathbb{E}_{\sigma_s}[\widehat{\boldsymbol{X}}_s]-\boldsymbol{x}_0|^2=\left|\int_0^s e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)\tau}\boldsymbol{f}(\boldsymbol{x}_0)d\tau\right|^2\le|\boldsymbol{f}(\boldsymbol{x}_0)|^2\left(\frac{e^{C_{\boldsymbol{f}}s}-1}{C_{\boldsymbol{f}}}\right)^2.$$

Therefore, given $\boldsymbol{x}_0$, $\mathrm{tr}\left(\boldsymbol{\Sigma}_s[\widehat{\boldsymbol{X}}_s]\right)=\mathcal{O}(s)$, $\left|\mathbb{E}_{\sigma_s}[\widehat{\boldsymbol{X}}_s]-\boldsymbol{x}_0\right|^2=\mathcal{O}(s^2)$. Moreover,

$$\begin{aligned}
&H\left(\sigma_t\mid\nu_t\right)\le\frac{1}{2\lambda_{\min}(\boldsymbol{G})}\int_0^t\int_{\mathbb{R}^d}|\boldsymbol{f}_\sigma(\boldsymbol{x})-\boldsymbol{f}_\nu(\boldsymbol{x})|^2\,\mathrm{d}\sigma_s\mathrm{d}s,\\
&\le\frac{2C_{\boldsymbol{f}}^2}{\lambda_{\min}(\boldsymbol{G})}\int_0^t\left(\mathrm{tr}(\boldsymbol{G})\frac{e^{2C_{\boldsymbol{f}}s}-1}{C_{\boldsymbol{f}}}+|\boldsymbol{f}(\boldsymbol{x}_0)|^2\left(\frac{e^{C_{\boldsymbol{f}}s}-1}{C_{\boldsymbol{f}}}\right)^2\right)\mathrm{d}s\\
&=\frac{2C_{\boldsymbol{f}}^2}{\lambda_{\min}(\boldsymbol{G})}\left(\mathrm{tr}(\boldsymbol{G})\left(t^2+\mathcal{O}(t^3)\right)+|\boldsymbol{f}(\boldsymbol{x}_0)|^2\left(\frac{1}{3}t^3+\mathcal{O}(t^4)\right)\right)=\mathcal{O}(t^2)\text{ as }t\to0.
\end{aligned}$$

Hence, $\|(\nu_t-\sigma_t)\|_{\mathrm{TV}}^2\le\frac{1}{2}H\left(\sigma_t\mid\nu_t\right)=\mathcal{O}(t^2)$ as $t\to0$. Furthermore, if the Hessian of the drift $\boldsymbol{f}_\nu$ is uniformly bounded by $H_{\boldsymbol{f}}$, then

$$\begin{aligned}
&\int_{\mathbb{R}^d}|\boldsymbol{f}_\sigma(\boldsymbol{x})-\boldsymbol{f}_\nu(\boldsymbol{x})|^2\,\mathrm{d}\sigma_s\\
&=\int_{\mathbb{R}^d}\left|\int_0^1(1-h)(\boldsymbol{x}-\boldsymbol{x}_0)^\top\mathrm{Hess}(\boldsymbol{x}_0+h(\boldsymbol{x}-\boldsymbol{x}_0))(\boldsymbol{x}-\boldsymbol{x}_0)\mathrm{d}h\right|^2\mathrm{d}\sigma_s\le\frac{H_{\boldsymbol{f}}^2}{4}\int_{\mathbb{R}^d}|\boldsymbol{x}-\boldsymbol{x}_0|^4\mathrm{d}\sigma_s.
\end{aligned}$$

$$\begin{aligned}
&\int_{\mathbb{R}^d}|\boldsymbol{x}-\boldsymbol{x}_0|^4\mathrm{d}\sigma_s\le8\int_{\mathbb{R}^d}|\boldsymbol{x}-\mathbb{E}_{\sigma_s}[\widehat{\boldsymbol{X}}_s]|^4\mathrm{d}\sigma_s+8\int_{\mathbb{R}^d}|\mathbb{E}_{\sigma_s}[\widehat{\boldsymbol{X}}_s]-\boldsymbol{x}_0|^4\mathrm{d}\sigma_s\\
&\le24\mathrm{tr}^2\left(\boldsymbol{\Sigma}_s[\widehat{\boldsymbol{X}}_s]\right)+8\int_{\mathbb{R}^d}\left|\int_0^s e^{\nabla \boldsymbol{f}(\boldsymbol{x}_0)\tau}\boldsymbol{f}(\boldsymbol{x}_0)d\tau\right|^4\mathrm{d}\sigma_s\\
&\le24\mathrm{tr}^2(\boldsymbol{G})\frac{(e^{2C_{\boldsymbol{f}}s}-1)^2}{C_{\boldsymbol{f}}^2}+8|\boldsymbol{f}(\boldsymbol{x}_0)|^4\left(\frac{e^{C_{\boldsymbol{f}}s}-1}{C_{\boldsymbol{f}}}\right)^4=96\mathrm{tr}^2(\boldsymbol{G})s^2+\mathcal{O}(s^3),\quad s\to0,
\end{aligned}$$

having used that $\int_{\mathbb{R}^d}|\boldsymbol{x}-\mathbb{E}_{\sigma_s}[\widehat{\boldsymbol{X}}_s]|^4\mathrm{d}\sigma_s=\mathrm{tr}^2\left(\boldsymbol{\Sigma}_s[\widehat{\boldsymbol{X}}_s]\right)+2\mathrm{tr}\left(\boldsymbol{\Sigma}_s(\widehat{\boldsymbol{X}}_s)^2\right)$ and $\mathrm{tr}\left(\boldsymbol{\Sigma}_s^2[\widehat{X}_s]\right)\le\mathrm{tr}^2\left(\boldsymbol{\Sigma}_s[\widehat{\boldsymbol{X}}_s]\right)$. Hence, $\int_0^t\int_{\mathbb{R}^d}|\boldsymbol{f}_\sigma(\boldsymbol{x})-\boldsymbol{f}_\nu(\boldsymbol{x})|^2\,\mathrm{d}\sigma_s\mathrm{d}s\le\frac{H_{\boldsymbol{f}}^2}{4}\int_0^t\int_{\mathbb{R}^d}|\boldsymbol{x}-\boldsymbol{x}_0|^4\mathrm{d}\sigma_s\mathrm{d}s=8H_{\boldsymbol{f}}^2\mathrm{tr}^2(\boldsymbol{G})t^3+\mathcal{O}(t^4)$, as $t\to0$. Therefore, $H\left(\sigma_t\mid\nu_t\right)=\mathcal{O}(t^3)$, $\|\nu_t-\sigma_t\|_{TV}^2=\mathcal{O}(t^3)$ as $t\to0$. ∎

Appendix B. Proof of Proposition 3.2.

Proof. Let $\rho = \nu+\sigma$, then $\nu \ll \rho$, $\sigma \ll \rho$. Define $p_\nu \coloneqq \frac{d\nu}{d\rho}$, $p_\sigma \coloneqq \frac{d\sigma}{d\rho}$, thus $\nu = p_\nu \rho$, $\sigma = p_\sigma \rho$, and $p_\nu + p_\sigma = 1$. For $\boldsymbol{f} = (f^1, \ldots, f^m) \in L^2(\mathbb{R}^d, \nu; \mathbb{R}^m) \cap L^2(\mathbb{R}^d, \sigma; \mathbb{R}^m)$,

$$
\begin{aligned}
&\left| \int_{\mathbb{R}^d} \boldsymbol{f}(\boldsymbol{x}) \mathrm{d}\nu(\boldsymbol{x}) - \int_{\mathbb{R}^d} \boldsymbol{f}(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}) \right| = \left| \int_{\mathbb{R}^d} \boldsymbol{f}(\boldsymbol{x}) (p_\nu(\boldsymbol{x}) - p_\sigma(\boldsymbol{x})) \mathrm{d}\rho(\boldsymbol{x}) \right| \\
&\leq \int_{\mathbb{R}^d} \left| \boldsymbol{f}(\boldsymbol{x}) (p_\nu(\boldsymbol{x}) + p_\sigma(\boldsymbol{x}))^{\frac{1}{2}} \frac{p_\nu(\boldsymbol{x}) - p_\sigma(\boldsymbol{x})}{(p_\nu(\boldsymbol{x}) + p_\sigma(\boldsymbol{x}))^{\frac{1}{2}}} \right| \mathrm{d}\rho(\boldsymbol{x}) \\
&\leq \left(\int_{\mathbb{R}^d} |\boldsymbol{f}(\boldsymbol{x})|^2 (p_\nu(\boldsymbol{x}) + p_\sigma(\boldsymbol{x})) \mathrm{d}\rho(\boldsymbol{x}) \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} \frac{(p_\nu(\boldsymbol{x}) - p_\sigma(\boldsymbol{x}))^2}{p_\nu(\boldsymbol{x}) + p_\sigma(\boldsymbol{x})} \mathrm{d}\rho(\boldsymbol{x}) \right)^{\frac{1}{2}} \\
&\leq \left(\|\boldsymbol{f}\|_{L^2(\mathbb{R}^d, \nu; \mathbb{R}^m)} + \|\boldsymbol{f}\|_{L^2(\mathbb{R}^d, \sigma; \mathbb{R}^m)} \right) \sqrt{2 \|\nu - \sigma\|_{\mathrm{TV}}}.
\end{aligned}
$$

∎

Appendix C. Proof of Proposition 3.3.

Proof. Let $\Pi_k : \mathbb{R}^d \to \mathbb{R}^k$ denote the canonical projection onto the first k coordinates, that is, $\Pi_k(x_1, x_2, \cdots, x_d) = (x_1, x_2, \cdots, x_k)$. The KR map T_i admits $T_i(\boldsymbol{x}) = (T_i^1(x_1), T_i^2(x_1, x_2), \cdots, T_i^d(x_1, \ldots, x_d))$. For $1 \leq k \leq d$, denote its k-dimensional truncation by $T_i^{(k)}(\boldsymbol{x}^{(1:k)}) = \Pi_k \circ T_i(\boldsymbol{x})$ where $\boldsymbol{x}^{(1:k)} = (x_1, \ldots, x_k)$. We denote by $\mu^{(k)} \coloneqq \Pi_{k\#}\mu$ and $\nu_i^{(k)} \coloneqq \Pi_{k\#}\nu_i$ the k-dimensional marginal distribution of μ and ν_i, respectively. Their corresponding PDFs are denoted by $p_\mu^{(k)}$ and $p_{\nu_i}^{(k)}$. By the definition of total variation distance, we have $\|\nu_i^{(k)} - \mu^{(k)}\|_{\mathrm{TV}} \leq \|\nu_i - \mu\|_{\mathrm{TV}} \to 0$.

We now prove $T_i(\boldsymbol{x}) \to \boldsymbol{x}$ in μ-probability as $i \to \infty$ by induction.

We first show that $T_i^1(x_1) \to x_1$ in μ-probability as $i \to \infty$. Let F_i^1 and F^1 be the cumulative distribution functions of $\nu_i^{(1)}$ and $\mu^{(1)}$, respectively. Let $\delta_i^1 = \sup_{x \in \mathbb{R}} |F_i^1(x) - F^1(x)|$. Since $\|\nu_i^{(1)} - \mu^{(1)}\|_{\mathrm{TV}} \to 0$, we have $\lim_{i\to\infty} \delta_i^1 = 0$. Notice that $T_i^1(x_1) = (F_i^1)^{(-1)} \circ F^1(x_1)$, and $(F_i^1)^{(-1)}(s) = \inf\{y \in \mathbb{R} : F_i^1(y) \geq s\}$. For any $\epsilon > 0$, if $T_i^1(x_1) \geq x_1 + \epsilon$, then $F_i^1(x_1 + \epsilon) \leq F^1(x_1)$. Hence, $F^1(x_1 + \epsilon) - F^1(x_1) \leq F^1(x_1 + \epsilon) - F_i^1(x_1 + \epsilon) \leq \delta_i^1$. On the other hand, if $T_i^1(x_1) \leq x_1 - \epsilon$, then $F_i^1(x_1 - \epsilon) \geq F^1(x_1)$. Hence, $F^1(x_1) - F^1(x_1 - \epsilon) \leq F_i^1(x_1 - \epsilon) - F^1(x_1 - \epsilon) \leq \delta_i^1$. Therefore, we have $\{|T_i^1(x_1) - x_1| \geq \epsilon\} \subseteq \{F^1(x_1) - F^1(x_1 - \epsilon) \leq \delta_i^1\} \cup \{F^1(x_1 + \epsilon) - F^1(x_1) \leq \delta_i^1\}$. Let $\Omega^-(\delta) = \{x_1 \mid F(x_1) - F(x_1 - \epsilon) \leq \delta\}$, and $\Omega^+(\delta) = \{x_1 \mid F(x_1 + \epsilon) - F(x_1) \leq \delta\}$. Thus, $\mu^{(1)}\left(|T_i^1(x_1) - x_1| \geq \epsilon\right) \leq \mu^{(1)}\left(\Omega^-(\delta_i^1)\right) + \mu^{(1)}\left(\Omega^+(\delta_i^1)\right)$, where $\mu^{(1)}\left(\Omega^-(\delta_i^1)\right) = \int 1_{\Omega^-(\delta_i^1)}(x_1) \mathrm{d}\mu^{(1)}(x_1)$, and $\mu^{(1)}\left(\left(\Omega^+(\delta_i^1)\right)\right) = \int 1_{\Omega^+(\delta_i^1)}(x_1) \mathrm{d}\mu^{(1)}(x_1)$. According to the dominated convergence theorem, we have

$$
\lim_{i\to\infty} \mu^{(1)}\left(\Omega^-(\delta_i^1)\right) = \int 1_{\Omega^-(0)}(x_1) \mathrm{d}\mu^{(1)}(x_1) = \mu^{(1)}(\Omega^-(0)) = \sum_{j\in\mathbb{Z}} \mu^{(1)}(\Omega^-(0) \cap ((j-1)\epsilon, j\epsilon]).
$$

We note that $x_1 \in \Omega^-(0)$ implies $F(x_1) - F(x_1 - \epsilon) = 0$. For any $j \in \mathbb{Z}$, let $\overline{x}_j = \sup\{x_1 \in \Omega^-(0) \cap ((j-1)\epsilon, j\epsilon])\}$, then $\Omega^-(0) \cap ((j-1)\epsilon, j\epsilon] \subset ((j-1)\epsilon, \overline{x}_j]$ and

$$
\mu^{(1)}\left(\Omega^-(0)\right) \leq \sum_{j\in\mathbb{Z}} \mu^{(1)}\left(((j-1)\epsilon, \overline{x}_j]\right) \leq \sum_{j\in\mathbb{Z}} \mu^{(1)}((\overline{x}_j - \epsilon, \overline{x}_j]) = 0.
$$

Thus, $\lim_{i\to\infty} \mu^{(1)}\left(\Omega^-(\delta_i^1)\right) = 0$. Similarly, we have $\lim_{i\to\infty} \mu^{(1)}\left(\Omega^+(\delta_i^1)\right) = 0$. Hence, $\lim_{i\to\infty} \mu\left(\{|T_i^1(x_1) - x_1| \geq \epsilon\}\right) = 0$, and $T_i^1(x_1) \xrightarrow{i\to\infty} x_1$ in μ-probability.

Suppose that for any $1 \leq k \leq d-1$, $T_i^k(\boldsymbol{x}^{(1:k)}) \xrightarrow{i\to\infty} x_k$ in μ-probability. We next prove that $T_i^{k+1}(\boldsymbol{x}^{(1:k+1)}) \xrightarrow{i\to\infty} x_{k+1}$ in μ-probability. For notation simplicity,

we write $\zeta_{i,T_i^{(k)}}(\mathrm{d}x_{k+1}) \coloneqq \nu_i(dx_{k+1} \mid T_i^{(k)}(\boldsymbol{x}^{(1:k)}))$, $\zeta_i(\mathrm{d}x_{k+1}) \coloneqq \nu_i(dx_{k+1} \mid \boldsymbol{x}^{(1:k)})$, $\zeta_{T_i^{(k)}}(\mathrm{d}x_{k+1}) \coloneqq \mu(dx_{k+1} \mid T_i^{(k)}(\boldsymbol{x}^{(1:k)}))$, $\zeta(\mathrm{d}x_{k+1}) \coloneqq \mu(\mathrm{d}x_{k+1} \mid \boldsymbol{x}^{(1:k)})$. We show that

$$\int \|\zeta_{i,T_i^{(k)}} - \zeta\|_{\mathrm{TV}} \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{(1:k)}) \to 0, \text{ as } i \to \infty. \tag{C.1}$$

Notice that $\left\|\zeta_{i,T_i^{(k)}} - \zeta\right\|_{\mathrm{TV}} \leq \left\|\zeta_{i,T_i^{(k)}} - \zeta_{T_i^{(k)}}\right\|_{\mathrm{TV}} + \left\|\zeta_{T_i^{(k)}} - \zeta\right\|_{\mathrm{TV}}$, and

$$\begin{aligned}
&\int \|\zeta_{i,T_i^{(k)}} - \zeta_{T_i^{(k)}}\|_{\mathrm{TV}} \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{(1:k)}) = \int \|\zeta_i - \zeta\|_{\mathrm{TV}} \mathrm{d}\nu_i^{(k)}(\boldsymbol{x}^{(1:k)}) \\
=&\frac{1}{2}\int |p_{\nu_i}(x_{k+1}|\boldsymbol{x}^{(1:k)}) - p_\mu(x_{k+1}|\boldsymbol{x}^{(1:k)})| p_{\nu_i}^{(k)}(\boldsymbol{x}^{(1:k)}) \mathrm{d}x_{k+1}\mathrm{d}\boldsymbol{x}^{(1:k)} \\
\leq&\frac{1}{2}\int \left|p_{\nu_i}^{(k+1)} - p_\mu^{(k+1)}\right| \mathrm{d}\boldsymbol{x}^{(1:k+1)} + \frac{1}{2}\int \left|p_\mu^{(k)} - p_{\nu_i}^{(k)}\right| \mathrm{d}\boldsymbol{x}^{(1:k)} \\
=&\|\nu_i^{(k+1)} - \mu^{(k+1)}\|_{\mathrm{TV}} + \|\nu_i^{(k)} - \mu^{(k)}\|_{\mathrm{TV}} \leq 2\|\nu_i - \mu\|_{\mathrm{TV}} \to 0, \quad \text{as } i \to \infty.
\end{aligned}$$

Therefore, to prove (C.1), we only need to show $\int \|\zeta_{T_i^{(k)}} - \zeta\|_{\mathrm{TV}} \mathrm{d}\mu^{(k)}$ goes to 0 as $i \to \infty$. Since μ is absolutely continuous and $\int p_\mu(x_{k+1}|\boldsymbol{x}^{(1:k)})\mathrm{d}x_{k+1} = 1$ for $\mu^{(k)}$-a.e. $\boldsymbol{x}^{(1:k)}$, $p_\mu(x_{k+1}|\boldsymbol{x}^{(1:k)})$ can be regarded as a $L^1(\mathbb{R})$ function for $\mu^{(k)}$-a.e. $\boldsymbol{x}^{(1:k)}$. According to Lemma 1.2.31 in [20], for any $\epsilon > 0$, there exists a bounded continuous $L^1(\mathbb{R})$ function $\phi(x_{k+1}, \boldsymbol{x}^{(1:k)})$ with $\|\phi\|_\infty \coloneqq \sup_{\boldsymbol{z}\in\mathbb{R}^k} \|\phi(\cdot, \boldsymbol{z})\|_{L^1(\mathbb{R})}$ such that

$$H_\phi \coloneqq \int \|p_\mu(x_{k+1}|\boldsymbol{x}^{(1:k)}) - \phi(x_{k+1}, \boldsymbol{x}^{(1:k)})\|_{L^1(\mathbb{R})} \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{(1:k)}) < \epsilon.$$

Thus,

$$\begin{aligned}
&\int \|\zeta_{T_i^{(k)}} - \zeta\|_{\mathrm{TV}} \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{(1:k)}) \\
=&\frac{1}{2}\int \|p_\mu(x_{k+1}|T_i^{(k)}(\boldsymbol{x}^{(1:k)})) - p_\mu(x_{k+1}|\boldsymbol{x}^{(1:k)})\|_{L^1(\mathbb{R})} \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{(1:k)}) \\
\leq&\frac{1}{2}\int \|p_\mu(x_{k+1}|T_i^{(k)}(\boldsymbol{x}^{(1:k)})) - \phi(x_{k+1}, T_i^{(k)}(\boldsymbol{x}^{(1:k)}))\|_{L^1(\mathbb{R})} \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{(1:k)}) \\
&+ \frac{1}{2}\int \|\phi(x_{k+1}, T_i^{(k)}(\boldsymbol{x}^{(1:k)})) - \phi(x_{k+1}, \boldsymbol{x}^{(1:k)})\|_{L^1(\mathbb{R})} \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{(1:k)}) + \frac{1}{2}H_\phi.
\end{aligned} \tag{C.2}$$

Since ${T_i^{(k)}}_{\#}\mu^{(k)} = \nu_i^{(k)}$ and $\|\nu_i^{(k)} - \mu^{(k)}\|_{\mathrm{TV}} \to 0$, there exists $i_0 \in \mathbb{N}$ such that $\|\nu_i^{(k)} - \mu^{(k)}\|_{\mathrm{TV}} < \frac{\epsilon}{1+\|\phi\|_\infty}$ for any $i \geq i_0$. Then for $i \geq i_0$, we have

$$\begin{aligned}
&\int \|p_\mu(x_{k+1}|T_i^{(k)}(\boldsymbol{x}^{(1:k)})) - \phi(x_{k+1}, T_i^{(k)}(\boldsymbol{x}^{(1:k)}))\|_{L^1(\mathbb{R})} \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{(1:k)}) \\
=&\int \|p_\mu(x_{k+1}|\boldsymbol{x}^{(1:k)})) - \phi(x_{k+1}, \boldsymbol{x}^{(1:k)})\|_{L^1(\mathbb{R})} \mathrm{d}\nu_i^{(k)}(\boldsymbol{x}^{(1:k)}) \\
=&H_\phi + \int \|p_\mu(x_{k+1}|\boldsymbol{x}^{(1:k)})) - \phi(x_{k+1}, \boldsymbol{x}^{(1:k)})\|_{L^1(\mathbb{R})} \mathrm{d}(\nu_i^{(k)} - \mu^{(k)})(\boldsymbol{x}^{(1:k)}) \\
<&\epsilon + 2(1 + \|\phi\|_\infty)\|\nu_i^{(k)} - \mu^{(k)}\|_{\mathrm{TV}} < 3\epsilon.
\end{aligned} \tag{C.3}$$

Meanwhile, $T_i^{(k)}(\boldsymbol{x}^{(1:k)}) \to \boldsymbol{x}^{(1:k)}$ in μ-probability as $i \to \infty$ and $\phi(x_{k+1}, \boldsymbol{x}^{(1:k)})$ is a bounded continuous function. Hence, $\phi(x_{k+1}, T_i^{(k)}(\boldsymbol{x}^{(1:k)})) \to \phi(x_{k+1}, \boldsymbol{x}^{(1:k)})$ in $\mu^{(k)}$-probability with respect to the L^1 norm as $i \to \infty$. Since ϕ is bounded, hence integrable, according to the Vitali convergence theorem [14], we have $\int \|\phi(\cdot, T_i^{(k)}(\boldsymbol{x}^{1:k})) - \phi(\cdot, \boldsymbol{x}^{(1:k)})\|_{L^1(\mathbb{R})} \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{1:k}) \to 0$ as $i \to \infty$. Thus, there exists $i_1 \in \mathbb{N}$ such that

$$\int \|\phi(\cdot, T_i^{(k)}(\boldsymbol{x}^{1:k})) - \phi(\cdot, \boldsymbol{x}^{(1:k)})\|_{L^1(\mathbb{R})} \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{1:k}) < \epsilon, \text{ for any } i \geq i_1. \tag{C.4}$$

Substituting equations (C.3)-(C.4) into equation (C.2) gives $\int \|\zeta_{i,T_i^{(k)}} - \zeta\|_{\mathrm{TV}} \mathrm{d}\mu^{(k)} < \frac{5}{2}\epsilon$ as $i > \max\{i_0, i_1\}$. Hence, $\lim_{i\to\infty} \int \|\zeta_{i,T_i^{(k)}} - \zeta\|_{\mathrm{TV}} \mathrm{d}\mu^{(k)} = 0$ and equation (C.1) holds. For $\mu^{(k)}$-a.e. $\boldsymbol{x}^{(1:k)}$, $T_i^{k+1}(\boldsymbol{x}^{(1:k)}, x_{k+1})$ is a one-dimensional KR map from ζ to $\zeta_{i,T_i^{(k)}}$. Let F_i^{k+1} and F^{k+1} be the cumulative distribution functions of ζ and $\zeta_{i,T_i^{(k)}}$, respectively. Denote $\delta_i^{k+1}(\boldsymbol{x}^{(1:k)}) = \sup_{x_{k+1}} |F_i^{k+1}(x_{k+1}) - F^{k+1}(x_{k+1})|$. Then

$$\int \delta_i^{k+1}(\boldsymbol{x}^{(1:k)}) \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{(1:k)}) \leq 2 \int \left\| \zeta_{i,T_i^{(k)}} - \zeta \right\|_{\mathrm{TV}} \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{(1:k)}) \xrightarrow{i\to\infty} 0.$$

We want to show that $\mu^{(k+1)}\left(|T_i^{k+1}(\boldsymbol{x}^{(1:k+1)}) - x_{k+1}| \geq \epsilon\right) \xrightarrow{i\to\infty} 0$ or equivalently,

$$\forall \eta > 0, \exists i_\eta \in \mathbb{N} : \mu^{(k+1)}\left(|T_i^{k+1}(\boldsymbol{x}^{(1:k+1)}) - x_{k+1}| \geq \epsilon\right) \leq \eta, \forall i \geq i_\eta.$$

Let $R_\delta(\boldsymbol{x}^{(1:k)}) = \zeta\left(F^{k+1}(x_{k+1}) - F^{k+1}(x_{k+1} - \epsilon) \leq \delta\right) + \zeta\left(F^{k+1}(x_{k+1} + \epsilon) - F^{k+1}(x_{k+1}) \leq \delta\right)$. Following the same argument as for T_i^1 and since $0 \leq R_\delta \leq 2$, we have for any $\epsilon > 0$, $\int R_\delta \mathrm{d}\mu^{(k)} \to 0$ as $\delta \to 0$. Let here δ_η be s.t. $\int R_\delta \mathrm{d}\mu^{(k)} \leq \eta/2$. Since $\lim_{i\to\infty} \int \delta_i^{k+1}(\boldsymbol{x}^{(1:k)}) \mathrm{d}\mu^{(k)} = 0$, there exists i_η s.t. $\int \delta_i^{k+1}(\boldsymbol{x}^{(1:k)}) \mathrm{d}\mu^{(k)} \leq \delta_\eta \cdot \eta/2$, $\forall i \geq i_\eta$. Using that $\{|T_i^{k+1}(\boldsymbol{x}^{(1:k+1)}) - x_{k+1}| \geq \epsilon\} \cap \{\delta_i^{k+1} \leq \delta_\eta\} \subset \{F^{k+1}(x_{k+1}) - F^{k+1}(x_{k+1} - \epsilon) \leq \delta_\eta\} \cup \{F^{k+1}(x_{k+1} + \epsilon) - F^{k+1}(x_{k+1}) \leq \delta_\eta\}$, we have

$$\begin{aligned}
&\mu^{(k+1)}\left(|T_i^{k+1}(\boldsymbol{x}^{(1:k+1)}) - x_{k+1}| \geq \epsilon\right) \\
=&\mu^{(k+1)}\left(|T_i^{k+1}(\boldsymbol{x}^{(1:k+1)}) - x_{k+1}| \geq \epsilon, \delta_i^{k+1}(\boldsymbol{x}^{(1:k)}) \leq \delta_\eta\right) \\
&+ \mu^{(k+1)}\left(|T_i^{k+1}(\boldsymbol{x}^{(1:k+1)}) - x_{k+1}| \geq \epsilon, \delta_i^{k+1}(\boldsymbol{x}^{(1:k)}) > \delta_\eta\right) \\
\leq& \int \zeta(|T_i^{k+1}(\boldsymbol{x}^{(1:k+1)}) - x_{k+1}| \geq \epsilon) 1_{\delta_i^{k+1}(\boldsymbol{x}^{(1:k)}) \leq \delta_\eta} \mathrm{d}\mu^{(k)} + \mu^{(k+1)}(\delta_i^{k+1} > \delta_\eta) \\
\leq& \int R_{\delta_\eta}(\boldsymbol{x}^{(1:k)}) \mathrm{d}\mu^{(k)}(\boldsymbol{x}^{(1:k)}) + \mu^{k+1}(\delta_i^{k+1} > \delta_\eta) \leq \frac{\eta}{2} + \frac{1}{\delta_\eta} \int \delta_i^{k+1}(\boldsymbol{x}^{(1:k)}) \mathrm{d}\mu^{(k)} \leq \eta, \forall i \geq i_\eta,
\end{aligned}$$

which shows that $T_i^{k+1}(\boldsymbol{x}^{(1:k+1)}) \to x_{k+1}$ in $\mu^{(k+1)}$-probability.

We now prove that $\lim_{i\to\infty} \int_{\mathbb{R}^d} |T_i(\boldsymbol{x}) - \boldsymbol{x}|^2 \mathrm{d}\mu(\boldsymbol{x}) = 0$. Since $T_{i\#}\mu = \nu_i$,

$$\int_{\mathbb{R}^d} |T_i(\boldsymbol{x})|^2 \mathrm{d}\mu(\boldsymbol{x}) = \int_{\mathbb{R}^d} |\boldsymbol{x}|^2 \mathrm{d}\nu_i(\boldsymbol{x}) \to \int_{\mathbb{R}^d} |\boldsymbol{x}|^2 \mathrm{d}\mu(\boldsymbol{x}), \quad \text{as } i \to \infty.$$

That is to say, $\lim_{i\to\infty} \|T_i\|_{L^2(\mathbb{R}^d,\mu)} = \|id\|_{L^2(\mathbb{R}^d,\mu)}$. Meanwhile, $2|T_i(\boldsymbol{x}) \cdot \boldsymbol{x}| \leq |T_i(\boldsymbol{x})|^2 + |\boldsymbol{x}|^2$, and $\{T_i\}_{i\geq 1}$ are uniformly integrable in $L^1(\mu)$. Thus, according to the Vitali convergence theorem [14], we have $\lim_{i\to\infty} \int_{\mathbb{R}^d} T_i(\boldsymbol{x}) \cdot \boldsymbol{x} \mathrm{d}\mu(\boldsymbol{x}) = \int_{\mathbb{R}^d} |\boldsymbol{x}|^2 \mathrm{d}\mu(\boldsymbol{x})$. Hence,

$$\|T_i - id\|^2_{L^2(\mathbb{R}^d,\mu)} = \|T_i\|^2_{L^2(\mathbb{R}^d,\mu)} + \|id\|^2_{L^2(\mathbb{R}^d,\mu)} - 2\int_{\mathbb{R}^d} T_i(\boldsymbol{x}) \cdot \boldsymbol{x} \mathrm{d}\mu(\boldsymbol{x}) \xrightarrow{i\to\infty} 0.$$

Therefore, $\|T_i - id\|^2_{L^2(\mu)} \to 0$ as $i \to \infty$. ∎

Appendix D. Proof of Proposition 4.1.

Proof. For any $\tilde{p}(\boldsymbol{x}, t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}(t), \boldsymbol{\Sigma}(t))$, $\quad \boldsymbol{\Sigma}(t) \sim t\boldsymbol{G}$, we have,
$\frac{\partial \tilde{p}}{\partial t} = \tilde{p}\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\mathrm{d}\boldsymbol{\Sigma}}{\mathrm{d}t}\right) + \frac{\mathrm{d}\boldsymbol{m}^\top}{\mathrm{d}t}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^\top\boldsymbol{\Sigma}^{-1}\frac{\mathrm{d}\boldsymbol{\Sigma}}{\mathrm{d}t}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right)$,
$\nabla\tilde{p} = \tilde{p}(-\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m})), \quad \nabla\cdot\nabla\cdot(\boldsymbol{g}\boldsymbol{g}^\top\tilde{p}) = \tilde{p}(\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m}))^2 - \tilde{p}\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{g}$.

Hence, $r(\boldsymbol{x}, \boldsymbol{x}_0, t; \tilde{p}) = \tilde{p}\tilde{r}(\boldsymbol{x}, \boldsymbol{x}_0, t; \tilde{p})$,

$$\begin{aligned}\tilde{r}(\boldsymbol{x}, \boldsymbol{x}_0, t; \tilde{p}) =& -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\mathrm{d}\boldsymbol{\Sigma}}{\mathrm{d}t}\right) + \frac{\mathrm{d}\boldsymbol{m}^\top}{\mathrm{d}t}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^\top\boldsymbol{\Sigma}^{-1}\frac{\mathrm{d}\boldsymbol{\Sigma}}{\mathrm{d}t}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m})\\ &+\nabla\cdot\boldsymbol{f} - \boldsymbol{f}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m}) - \frac{1}{2}(\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m}))^2 + \frac{1}{2}\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{g}.\end{aligned}$$

Thus, $|\tilde{r}(\boldsymbol{x}, \boldsymbol{x}_0, t; \tilde{p})| \le \tilde{C}_0 + \tilde{C}_1 t^{-1} + \tilde{C}_2|\boldsymbol{x}-\boldsymbol{m}|t^{-1} + \tilde{C}_3|\boldsymbol{x}-\boldsymbol{m}|^2 t^{-1} + \tilde{C}_4|\boldsymbol{x}-\boldsymbol{m}|^2 t^{-2}$.

$$\begin{aligned}|\tilde{r}(\boldsymbol{x}, \boldsymbol{x}_0, t; \tilde{p})|^2 \le& (C_0 + C_1 t^{-1} + C_2 t^{-2}) + (C_3 t^{-1} + C_4 t^{-2})|\boldsymbol{x}-\boldsymbol{m}|\\ &+ (C_5 t^{-1} + C_6 t^{-2} + C_7 t^{-3})|\boldsymbol{x}-\boldsymbol{m}|^2 + (C_8 t^{-2} + C_9 t^{-3})|\boldsymbol{x}-\boldsymbol{m}|^3\\ &+ (C_{10} t^{-2} + C_{11} t^{-3} + C_{12} t^{-4})|\boldsymbol{x}-\boldsymbol{m}|^4.\end{aligned}$$

Notice that $\left(\mathbb{E}_{\boldsymbol{X}\sim\mathcal{N}(\boldsymbol{x};\boldsymbol{m},\boldsymbol{\Sigma})}\left[|\boldsymbol{X}-\boldsymbol{m}|\right]\right)^2 \le \mathbb{E}_{\boldsymbol{X}\sim\mathcal{N}(\boldsymbol{x};\boldsymbol{m},\boldsymbol{\Sigma})}\left[|\boldsymbol{X}-\boldsymbol{m}|^2\right]$.
$\mathbb{E}_{\boldsymbol{X}\sim\mathcal{N}(\boldsymbol{x};\boldsymbol{m},\boldsymbol{\Sigma})}\left[|\boldsymbol{X}-\boldsymbol{m}|^2\right] = \mathrm{tr}(\boldsymbol{\Sigma})$, $\mathbb{E}_{\boldsymbol{X}\sim\mathcal{N}(\boldsymbol{x};\boldsymbol{m},\boldsymbol{\Sigma})}\left[|\boldsymbol{X}-\boldsymbol{m}|^4\right] = \mathrm{tr}^2(\boldsymbol{\Sigma}) + 2\mathrm{tr}(\boldsymbol{\Sigma}^2)$,
$\mathbb{E}_{\boldsymbol{X}\sim\mathcal{N}(\boldsymbol{x};\boldsymbol{m},\boldsymbol{\Sigma})}\left[|\boldsymbol{X}-\boldsymbol{m}|^6\right] = \mathrm{tr}^3(\boldsymbol{\Sigma}) + 6\mathrm{tr}(\boldsymbol{\Sigma})\mathrm{tr}(\boldsymbol{\Sigma}^2) + 8\mathrm{tr}(\boldsymbol{\Sigma}^3)$. Thus,

$$\begin{aligned}&\int_{\mathbb{R}^d}|r(\boldsymbol{x}, \boldsymbol{x}_0, t; \tilde{p})|^2\mathrm{d}\boldsymbol{x} = (4\pi)^{-\frac{d}{2}}(\det(\boldsymbol{\Sigma}))^{-\frac{1}{2}}\mathbb{E}_{\boldsymbol{X}\sim\mathcal{N}\left(\boldsymbol{x};\boldsymbol{m},\frac{\boldsymbol{\Sigma}}{2}\right)}\left[|\tilde{r}(\boldsymbol{x}, \boldsymbol{x}_0, t; \tilde{p})|^2\right]\\ \le&(4\pi)^{-\frac{d}{2}}(\det(\boldsymbol{\Sigma}))^{-\frac{1}{2}}\left(C_0 + C_1 t^{-1} + C_2 t^{-2} + (C_3 t^{-1} + C_4 t^{-2})\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{2}}\right.\\ &+\frac{1}{2}(C_5 t^{-1} + C_6 t^{-2} + C_7 t^{-3})\mathrm{tr}(\boldsymbol{\Sigma}) + (C_8 t^{-2} + C_9 t^{-3})\sqrt{\frac{1}{8}\mathrm{tr}^3(\boldsymbol{\Sigma}) + \frac{3}{4}\mathrm{tr}(\boldsymbol{\Sigma})\mathrm{tr}(\boldsymbol{\Sigma}^2) + \mathrm{tr}(\boldsymbol{\Sigma}^3)}\\ &\left.+(C_{10} t^{-2} + C_{11} t^{-3} + C_{12} t^{-4})\left(\frac{1}{4}\mathrm{tr}^2(\boldsymbol{\Sigma}) + \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^2)\right)\right) = \mathcal{O}(t^{-d/2-2}) \text{ as } t \to 0.\end{aligned}$$

Similarly,

$$\begin{aligned}&\int_{\mathbb{R}^d}|r(\boldsymbol{x}, \boldsymbol{x}_0, t; \tilde{p})|^2\tilde{p}\mathrm{d}\boldsymbol{x} = (2\sqrt{3}\pi)^{-d}(\det(\boldsymbol{\Sigma}))^{-1}\mathbb{E}_{\boldsymbol{X}\sim\mathcal{N}\left(\boldsymbol{x};\boldsymbol{m},\frac{\boldsymbol{\Sigma}}{3}\right)}\left[|\tilde{r}(\boldsymbol{x}, \boldsymbol{x}_0, t; \tilde{p})|^2\right]\\ \le&(2\sqrt{3}\pi)^{-d}(\det(\boldsymbol{\Sigma}))^{-1}\left(C_0 + C_1 t^{-1} + C_2 t^{-2} + (C_3 t^{-1} + C_4 t^{-2})\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{3}}\right.\\ &+\frac{1}{3}(C_5 t^{-1} + C_6 t^{-2} + C_7 t^{-3})\mathrm{tr}(\boldsymbol{\Sigma}) + (C_8 t^{-2} + C_9 t^{-3})\sqrt{\frac{1}{27}\mathrm{tr}^3(\boldsymbol{\Sigma}) + \frac{2}{9}\mathrm{tr}(\boldsymbol{\Sigma})\mathrm{tr}(\boldsymbol{\Sigma}^2) + \frac{8}{27}\mathrm{tr}(\boldsymbol{\Sigma}^3)}\\ &\left.+(C_{10} t^{-2} + C_{11} t^{-3} + C_{12} t^{-4})\left(\frac{1}{9}\mathrm{tr}^2(\boldsymbol{\Sigma}) + \frac{2}{9}\mathrm{tr}(\boldsymbol{\Sigma}^2)\right)\right) = \mathcal{O}(t^{-(d+2)}) \text{ as } t \to 0.\end{aligned}$$ ∎

REFERENCES

[1] Z. Aldirany, R. Cottereau, M. Laforest, and S. Prudhomme, Operator approximation of the wave equation based on deep learning of Green's function, Computers & Mathematics with Applications, 159 (2024), pp. 21–30.

[2] R. Baptista, B. Hosseini, N. B. Kovachki, and Y. M. Marzouk, Conditional sampling with monotone GANs: From generative models to likelihood-free inference, SIAM/ASA Journal on Uncertainty Quantification, 12 (2024), pp. 868–900.
[3] N. M. Boffi and E. Vanden-Eijnden, Probability flow solution of the Fokker–Planck equation, Machine Learning: Science and Technology, 4 (2023), p. 035012.
[4] V. I. Bogachev, M. Röckner, and S. V. Shaposhnikov, Distances between transition probabilities of diffusions and applications to nonlinear Fokker–Planck–Kolmogorov equations, Journal of Functional Analysis, 271 (2016), pp. 1262–1300.
[5] N. Boullé, S. Kim, T. Shi, and A. Townsend, Learning Green's functions associated with time-dependent partial differential equations, Journal of Machine Learning Research, 23 (2022), pp. 1–34.
[6] J. Bruna, B. Peherstorfer, and E. Vanden-Eijnden, Neural Galerkin schemes with active learning for high-dimensional evolution equations, Journal of Computational Physics, 496 (2024), p. 112588.
[7] X. Chen, L. Yang, J. Duan, and G. E. Karniadakis, Solving inverse stochastic problems from discrete particle observations using the Fokker–Planck equation and physics-informed neural networks, SIAM Journal on Scientific Computing, 43 (2021), pp. B811–B830.
[8] A. Daw, J. Bu, S. Wang, P. Perdikaris, and A. Karpatne, Mitigating propagation failures in physics-informed neural networks using retain-resample-release (r3) sampling, Proceedings of Machine Learning Research, 202 (2023), pp. 7264–7302.
[9] W. Deng, Finite element method for the space and time fractional Fokker–Planck equation, SIAM journal on numerical analysis, 47 (2009), pp. 204–226.
[10] L. Dinh, J. Sohl-Dickstein, and S. Bengio, Density estimation using Real NVP, in International Conference on Learning Representations, 2017.
[11] W. E and Y. B, The deep Ritz method: A deep learning-based numerical algorithm for solving variational problems, Communications in Mathematics and Statistics, 6 (2018), pp. 1–12.
[12] N. El Bekri, L. Drumetz, and F. Vermet, FlowKac: An efficient neural fokker-planck solver using temporal normalizing flows and the feynman-kac formula, Transactions on Machine Learning Research (TMLR), (2025).
[13] X. Feng, L. Zeng, and T. Zhou, Solving time dependent Fokker-Planck equations via temporal normalizing flow, Communications in Computational Physics, 32 (2022), pp. 401–423.
[14] G. B. Folland, Real analysis: modern techniques and their applications, John Wiley & Sons, 1999.
[15] N. Gaby, X. Ye, and H. Zhou, Neural control of parametric solutions for high-dimensional evolution PDEs, SIAM Journal on Scientific Computing, 46 (2024), pp. C155–C185.
[16] G. L. Gilardoni, On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences, IEEE Transactions on Information Theory, 56 (2010), pp. 5377–5386.
[17] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, A kernel two-sample test, The journal of machine learning research, 13 (2012), pp. 723–773.
[18] J. He, Q. Liao, and X. Wan, Adaptive deep density approximation for stochastic dynamical systems, Journal of Scientific Computing, 102 (2025), p. 57.
[19] Z. Hu, Z. Zhang, G. E. Karniadakis, and K. Kawaguchi, Score-based physics-informed neural networks for high-dimensional Fokker–Planck equations, SIAM Journal on Scientific Computing, 47 (2025), pp. C680–C705.
[20] T. Hytonen, J. van Neerven, M. Veraar, and L. Weis, Analysis in Banach spaces, volume i: Martingales and littlewood-paley theory, Ergebnisse der Mathematik und ihrer Grenzgebiete, 3 (2016).
[21] R. Jordan, D. Kinderlehrer, and F. Otto, The variational formulation of the Fokker–Planck equation, SIAM journal on mathematical analysis, 29 (1998), pp. 1–17.
[22] P. Kumar and S. Narayanan, Solution of Fokker-Planck equation by finite element and finite difference methods for nonlinear systems, Sadhana, 31 (2006), pp. 445–461.
[23] W. Lee, L. Wang, and W. Li, Deep JKO: Time-implicit particle methods for general nonlinear gradient flows, Journal of Computational Physics, 514 (2024), p. 113187.
[24] M. Leutbecher and T. N. Palmer, Ensemble forecasting, Journal of computational physics, 227 (2008), pp. 3515–3539.
[25] Z. Li, N. B. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, Fourier neural operator for parametric partial differential equations, in International Conference on Learning Representations.
[26] S. Liu, W. Li, H. Zha, and H. Zhou, Neural parametric Fokker–Planck equation, SIAM Journal on Numerical Analysis, 60 (2022), pp. 1385–1449.
[27] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis, Learning nonlinear operators via

DeepONet based on the universal approximation theorem of operators, Nature machine intelligence, 3 (2021), pp. 218–229.
[28] Y. Lu and B. Huang, Structured output learning with conditional generative flows, in Proceedings of the AAAI conference on artificial intelligence, vol. 34, 2020, pp. 5005–5012.
[29] Y. Lu, R. Maulik, T. Gao, F. Dietrich, I. G. Kevrekidis, and J. Duan, Learning the temporal evolution of multivariate densities via normalizing flows, Chaos: An Interdisciplinary Journal of Nonlinear Science, 32 (2022).
[30] R. N. Miller, E. F. Carter Jr, and S. T. Blue, Data assimilation into nonlinear stochastic models, Tellus A, 51 (1999), pp. 167–194.
[31] H. Noh, J. Lee, and E. Yoon, FPL-net: A deep learning framework for solving the nonlinear Fokker–Planck–Landau collision operator for anisotropic temperature relaxation, Journal of Computational Physics, 523 (2025), p. 113665.
[32] L. Pichler, A. Masud, and L. A. Bergman, Numerical solution of the Fokker–Planck equation by finite difference and finite element methods—a comparative study, in Computational Methods in Stochastic Dynamics, Springer, 2013, pp. 69–85.
[33] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics, 378 (2019), pp. 686–707.
[34] F. Santambrogio, Optimal transport for applied mathematicians, Birkäuser, NY, 55 (2015), p. 94.
[35] R. Saporiti and F. Nobile, Neural Galerkin normalizing flow for transition probability density functions of diffusion models, arXiv preprint arXiv:2603.18907, (2026).
[36] R. Saporiti and F. Nobile, Neural Galerkin normalizing flows for bayesian inference of diffusions with inaccessible boundaries, arXiv preprint arXiv:2606.04324, (2026).
[37] J. Sirignano and K. Spiliopoulos, DGM: A deep learning algorithm for solving partial differential equations, Journal of Computational Physics, 375 (2018), pp. 1339–1364.
[38] K. Tang, X. Wan, and Q. Liao, Deep density estimation via invertible block-triangular mapping, Theoretical and Applied Mechanics Letters, 10 (2020), pp. 143–148.
[39] K. Tang, X. Wan, and Q. Liao, Adaptive deep density approximation for Fokker-Planck equations, Journal of Computational Physics, 457 (2022), p. 111080.
[40] Y. Teng, X. Zhang, Z. Wang, and L. Ju, Learning Green's functions of linear reaction-diffusion equations with application to fast numerical solver, in Mathematical and Scientific Machine Learning, PMLR, 2022, pp. 1–16.
[41] S. Wang, H. Wang, and P. Perdikaris, On the eigenvector bias of Fourier feature networks: From regression to solving multi-scale PDEs with physics-informed neural networks, Computer Methods in Applied Mechanics and Engineering, 384 (2021), p. 113938.
[42] T. Wang, Z. Hu, K. Kawaguchi, Z. Zhang, and G. E. Karniadakis, Tensor neural networks for high-dimensional Fokker–Planck equations, Neural Networks, 185 (2025), p. 107165.
[43] X. Wang, J. Feng, Q. Liu, C. Tan, Y. Liu, and Y. Xu, A deep learning framework for jointly solving transient fokker-planck equations with arbitrary parameters and initial distributions, arXiv preprint arXiv:2604.06001, (2026).
[44] Z. O. Wang, R. Baptista, Y. Marzouk, L. Ruthotto, and D. Verma, Efficient neural network approaches for conditional optimal transport with applications in Bayesian inference, SIAM Journal on Scientific Computing, 47 (2025), pp. C979–C1005.
[45] M. F. Wehner and W. Wolfer, Numerical evaluation of path-integral solutions to Fokker-Planck equations, Physical Review A, 27 (1983), p. 2663.
[46] Y. Wen, E. Vanden-Eijnden, and B. Peherstorfer, Coupling parameter and particle dynamics for adaptive sampling in neural Galerkin schemes, Physica D: Nonlinear Phenomena, 462 (2024), p. 134129.
[47] Y. Xu, H. Zhang, Y. Li, K. Zhou, Q. Liu, and J. Kurths, Solving Fokker-Planck equation using deep learning, Chaos: An Interdisciplinary Journal of Nonlinear Science, 30 (2020).
[48] M. Yang, P. Wang, D. del Castillo-Negrete, Y. Cao, and G. Zhang, A pseudoreversible normalizing flow for stochastic dynamical systems with various initial distributions, SIAM Journal on Scientific Computing, 46 (2024), pp. C508–C533.
[49] L. Zeng, X. Wan, and T. Zhou, Adaptive deep density approximation for fractional Fokker–Planck equations, Journal of Scientific Computing, 97 (2023), p. 68.
[50] L. Zeng, X. Wan, and T. Zhou, Bounded KRnet and its applications to density estimation and approximation, SIAM Journal on Scientific Computing, 47 (2025), pp. C1294–C1318.
[51] J. Zhai, M. Dobson, and Y. Li, A deep learning method for solving Fokker-Planck equations, in Mathematical and Scientific Machine Learning, PMLR, 2022, pp. 568–597.
[52] M. Zhou, S. Osher, and W. Li, Simulating Fokker-Planck equations via mean field control of score-based normalizing flows, arXiv preprint arXiv:2506.05723, (2025).