PROBABILITY FLOW SOLUTION OF THE FOKKER-PLANCK EQUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The method of choice for integrating the time-dependent Fokker-Planck equation in high-dimension is to generate samples from the solution via integration of the associated stochastic differential equation. Here, we introduce an alternative scheme based on integrating an *ordinary* differential equation that describes the flow of probability. Acting as a transport map, this equation deterministically pushes samples from the initial density onto samples from the solution at any later time. Unlike integration of the stochastic dynamics, the method has the advantage of giving direct access to quantities that are challenging to estimate from trajectories alone, such as the probability current, the density itself, and its entropy. The probability flow equation depends on the gradient of the logarithm of the solution (its "score"), and so is *a-priori* unknown. To resolve this dependence, we model the score with a deep neural network that is learned on-the-fly by propagating a set of samples according to the instantaneous probability current. We consider several high-dimensional examples from the physics of interacting particle systems to highlight the efficiency and precision of the approach; we find that the method accurately matches analytical solutions computed by hand and moments computed via Monte-Carlo.

1 INTRODUCTION

The time evolution of many dynamical processes occurring in the natural sciences, engineering, economics, and statistics are naturally described in the language of stochastic differential equations (SDE) (Gardiner, 2009; Oksendal, 2003; Evans, 2012). Typically, one is interested in the probability density function (PDF) of these processes, which describes the probability that the system will occupy a given state at a given time. The density can be obtained as the solution to a Fokker-Planck equation (FPE), which can generically be written as (Risken, 1996; Bass, 2011)

$$\partial_t \rho_t(x) = -\nabla \cdot (b_t(x)\rho_t(x) - D_t(x)\nabla \rho_t(x)) \quad x \in \Omega \subseteq \mathbb{R}^d \tag{FPE}$$

where $\rho_t(x) \in \mathbb{R}_{\geq 0}$ denotes the value of the density at time t, $b_t(x) \in \mathbb{R}^d$ is a vector field known as the drift, and $D_t(x) \in \mathbb{R}^{d\times d}$ is a positive-semidefinite tensor referred to as the diffusion matrix. In all but the simplest cases, the solution to this equation is not available analytically and can only be approximated via numerical integration.

High-dimensionality. For many systems of interest – such as interacting particle systems in statistical physics (Chandler, 1987; Spohn, 2012), stochastic control systems (Kushner et al., 2001), and models in mathematical finance (Oksendal, 2003) – the dimensionality d can be very large. This renders standard numerical methods for partial differential equations inapplicable, which become infeasible for d as small as five or six due to an exponential scaling of the computational complexity with d. The standard solution to this problem is a Monte-Carlo approach, whereby the SDE associated with (FPE)

$$dX_t = b_t(X_t)dt + \nabla \cdot D_t(X_t)dt + \sqrt{2}\sigma_t(X_t)dW_t, \tag{1}$$

is evolved via numerical integration to obtain a large number n of trajectories (Kloeden & Platen, 1992). In (1), $\sigma_t(x)$ satisfies $\sigma_t(x)\sigma_t^{\mathsf{T}}(x) = D_t(x)$ and W_t is a standard Brownian motion on $\mathbb{R}^d$. Assuming that the initial condition for (FPE) is a PDF ρ_0 from which we can draw samples $\{X_0^i\}_{i=1}^n$,

simulation of (1) enables the estimation of expectations via empirical averages

$$\int_\Omega \phi(x)\rho_t(x)dx \approx \frac{1}{n}\sum_{i=1}^n \phi(X_t^i), \tag{2}$$

where $\phi : \Omega \to \mathbb{R}$ is an observable of interest. While widely used, this method only provides samples from ρ_t, and hence other quantities of interest like the value of ρ_t itself or the time-dependent differential entropy of the system $S_t = -\int_\Omega \log \rho_t(x)\rho_t(x)dx$ require sophisticated interpolation methods that typically do not scale well to high-dimension.

A transport map approach. Another possibility, building on recent theoretical advances that connect transportation of measures to the Fokker-Planck equation (Jordan et al., 1998), is to recast (FPE) as the transport equation (Villani, 2009; Santambrogio, 2015)

$$\partial_t \rho_t(x) = -\nabla \cdot (v_t(x)\rho_t(x)) \tag{TE}$$

where we defined the velocity field

$$v_t(x) = b_t(x) - D_t(x)\nabla \log \rho_t(x). \tag{3}$$

This formulation reveals that ρ_t can be viewed as the pushforward of ρ_0 under the flow map $X_{\tau,t}(\cdot)$ of the ordinary differential equation

$$\frac{d}{dt}X_{\tau,t}(x) = v_t(X_{\tau,t}(x)), \qquad X_{\tau,\tau}(x) = x, \quad t,\tau \geq 0. \tag{4}$$

Equation (4) is known as the *probability flow equation*. Its solution has the remarkable property that if x is a sample from ρ_0, then $X_{0,t}(x)$ will be a sample from ρ_t. Viewing $X_{\tau,t}(\cdot)$ as a transport map, $\rho_t = X_{0,t}\sharp\rho_0$ can be evaluated at any position in Ω via the change of variables formula (Villani, 2009; Santambrogio, 2015)

$$\rho_t(x) = \rho_0(X_{t,0}(x)) \exp\left(-\int_0^t \nabla \cdot v_\tau(X_{t,\tau}(x))d\tau\right) \tag{5}$$

where $X_{t,0}(x)$ is obtained by solving (4) backward from some given x. Importantly, access to the PDF as provided by (5) immediately gives the ability to compute quantities such as the probability current or the entropy; by contrast, this capability is absent when directly simulating the SDE.

Learning the flow. The simplicity of the probability flow equation (4) is somewhat deceptive, because it depends explicitly on the solution ρ_t to the Fokker-Planck equation (FPE). Nevertheless, recent work in generative modeling via score-based diffusion (Song & Ermon, 2020a;b; Song & Kingma, 2021) has shown that it is possible to use deep neural networks to estimate v_t, or equivalently the so-called *score* $\nabla \log \rho_t$ of the solution density. Here, we introduce a variant of score-based diffusion modeling in which the score is learned on-the-fly over samples generated by the probability flow equation itself. The method is self-contained and allows us to bypass simulation of the SDE entirely.

1.1 Contributions

Our contributions are both theoretical and computational:

- We introduce several optimization problems that can be used to learn the velocity field (3) in the transport equation (TE) so that its solution coincides with that of the Fokker Planck equation (FPE). Due to its similarities with score-based diffusion (SBDM) approaches in generative modeling, we call the resulting method *score-based transport modeling* (SBTM). The primary difference is that SBTM does not requires input data from the SDE to learn the score and estimate the target ρ_t^*. Instead, SBTM uses samples obtained via the probability flow equation itself.
- We show that the Kullback-Leibler divergence of the estimate ρ_t from the target ρ_t^* is controlled by the SBTM objective, so that minimizing the objective systematically improves the accuracy of the approximate solution.
- We provide specific estimators for quantities that can be computed via SBTM but are not directly available from samples alone, like point-wise evaluation of ρ_t itself, the differential entropy, and the probability current.

- We implement and test SBTM on several examples involving interacting particles that pairwise repel but are kept close by common attraction to a moving trap. In such systems, high-dimensionality of the Fokker-Planck equation arises due to the large number of particles (which vary from 5 to 50 in the examples below). Problems of this type frequently appear in the molecular dynamics of externally driven soft matter systems (Frenkel & Smit, 2001; Spohn, 2012).

1.2 Notation and assumptions.

Throughout, we assume that the stochastic process (1) evolves over a domain $\Omega \subseteq \mathbb{R}^d$ in which it remains at all times $t \geq 0$. We assume that the drift vector $b_t : \Omega \to \mathbb{R}^d$ and the diffusion tensor $D_t : \Omega \to \mathbb{R}^{d\times d}$ are twice-differentiable and bounded in both x and t, so that the solution to the SDE (1) is well-defined at all times $t \geq 0$. The symmetric tensor $D_t(x) = D_t^\mathsf{T}(x)$ is assumed to be positive semi-definite for each (t, x), with Cholesky decomposition $D_t(x) = \sigma_t(x)\sigma_t^\mathsf{T}(x)$. We further assume that the initial PDF ρ_0 is three-times differentiable, positive everywhere on Ω, and such that $S_0 = -\int_\Omega \log \rho_0(x)\rho_0(x)dx < \infty$. This guarantees that ρ_t enjoys the same properties at all times $t > 0$. Finally, we assume that $\log \rho_t$ is K-smooth globally for $(t, x) \in [0, \infty) \times \Omega$, i.e.

$$\exists K > 0 \ : \quad \forall (t, x) \in [0, \infty) \times \Omega \quad |\nabla \log \rho_t(x) - \nabla \log \rho_t(y)| \leq K|x - y|. \tag{6}$$

This technical assumption is needed to guarantee global existence and uniqueness of the solution of the probability flow equation.

2 Related work

Score matching Our approach builds directly on the toolbox of score matching originally developed by Hyvärinen (Hyvärinen, 2005; Hyvarinen, 2007; Hyvärinen, 2007; 2008) and more recently extended in the context of diffusion-based generative modeling (Song & Ermon, 2020a;b; Song et al., 2021; De Bortoli et al., 2021; Dockhorn et al., 2022; Mittal et al., 2021). These approaches assume access to training samples from the target distribution (e.g., in the form of examples of natural images). Here, we bypass this need and use the probability flow equation itself to obtain the samples needed to learn an approximation of the score. Lu et al. (2022) recently showed that using the transport equation (TE) with a velocity field learned via SBDM can lead to inaccuracies in the likelihood unless higher-order score terms are well-approximated. Proposition 2 shows that the self-consistent approach used in SBTM solves these issues and ensures a systematic approximation of the target ρ_T^*.

Density estimation and Bayesian inference Our method shares commonalities with transport map-based approaches (Marzouk et al., 2016) for density estimation and variational inference (Zhang et al., 2019; Blei et al., 2017) such as normalizing flows (Tabak & Vanden-Eijnden, 2010; Tabak & Turner, 2013; Rezende & Mohamed, 2016; Huang et al., 2021; Papamakarios et al., 2021; Kobyzev et al., 2021). Moreover, because expectations are approximated over a set of samples according to (2), the method also inherits elements of classical "particle-based" approaches for density estimation such as Markov chain Monte Carlo (Robert & Casella, 2004) and sequential Monte Carlo (Dai et al., 2020; Del Moral et al., 2006).

This is reminiscent of a recent line of work in Bayesian inference that aims to combine the strengths of particle methods with those of variational approximations (Dai et al., 2016; Saeedi et al., 2017). In particular, the method we propose bears some similarity with Stein variational gradient descent (SVGD) (Liu, 2017; Liu & Wang, 2018; 2019) (see also (Lu et al., 2018; Li et al., 2020)), in that both methods approximate the target distribution via *deterministic* propagation of a set of samples. The key differences are that (i) our method learns the map used to propagate the samples, while the map in SVGD corresponds to optimization of the kernelized Stein discrepancy, and (ii) the methods have distinct goals, as we are interested in capturing the dynamical evolution of ρ_t rather than sampling at equilibrium.

Approaches for solving the FPE Most closely connected to our paper are the works by Maoutsa et al. (2020) and Shen et al. (2022), who similarly propose to bypass the SDE through use of the probability flow equation, building on earlier work by Degond & Mustieles (1990) and Russo (1990). The critical differences between Maoutsa et al. (2020) and our approach are that they perform

estimation over a linear space or a reproducing kernel Hilbert space rather than over the significantly richer class of neural networks, and that they train using the original score matching loss of Hyvärinen (2005). Because of this, Maoutsa et al. (2020) studies systems of dimension less than or equal to five; we show impressive accuracy on systems with dimensionality as high as 100.

Concurrently to our work, Shen et al. (2022) proposed a variational problem similar to SBTM. A key difference is that SBTM is not limited to Fokker-Planck equations that can be viewed as a gradient flow in the Wasserstein metric over some energy (i.e., the drift term in the SDE (1) need not be the gradient of a potential), and that it allows for spatially-dependent and rank-deficient diffusion matrices; moreover, our theoretical results are similar but avoid the use of costly Sobolev norms.

Neural-network solutions to PDEs Our approach can also be viewed as an alternative to recent neural network-based methods for the solution of partial differential equations (see e.g. E & Yu (2017); Raissi et al. (2019); Han et al. (2018); Sirignano & Spiliopoulos (2018); Bruna et al. (2022)). Unlike these existing approaches, our method is tailored to the solution of the Fokker-Planck equation and guarantees that the solution is a valid probability density. Our approach is also fundamentally Lagrangian in nature, which has the advantage that it only involves learning quantities locally at the positions of a set of evolving samples; this is naturally conducive to efficient scaling for high-dimensional systems.

3 METHODOLOGY

3.1 SCORE-BASED DIFFUSION MODELING

SBDM is based on the observation that the score $\nabla \log \rho_t(x)$ can be estimated by minimizing the following objective function over $s_t(x)$

$$\int_\Omega |s_t(x) - \nabla \log \rho_t(x)|^2 \rho_t(x) dx \\ = \int_\Omega \left(|s_t(x)|^2 + 2\nabla \cdot s_t(x) + |\nabla \log \rho_t(x)|^2 \right) \rho_t(x) dx. \tag{7}$$

Above, the equality is obtained by expanding the square and treating the cross term using Stein's identity $\int_\Omega s_t(x) \nabla \log \rho_t(x)\, \rho_t(x) dx = -\int_\Omega \nabla \cdot s_t(x)\, \rho_t(x) dx$. If $s_t(x) = \nabla \log \hat{\rho}_t(x)$ for some PDF $\hat{\rho}_t$, (7) is the Fisher divergence between ρ_t and $\hat{\rho}_t$.

Because the spatial integral in (7) is an expectation with respect to ρ_t, it can be evaluated empirically over samples. In SBDM, these samples are generated by passing available samples from the target distribution through an Ornstein-Uhlenbeck process whose explicit solution is available at all times. The term involving $|\nabla \log \rho_t|^2$ is unknown, but can be neglected since it is a constant when optimizing over s_t (Hyvärinen, 2005; Hyvarinen, 2007; Hyvärinen, 2007; 2008).

3.2 SCORE-BASED TRANSPORT MODELING

The strategy employed by SBDM could in principle be applied to the problem considered here, whereby the expectation with respect to ρ_t in the loss (7) could be approximated over an empirical measure via integration of the SDE (1). However, because our goal is to bypass this integration, an alternative is to define ρ_t as the solution of the transport equation (TE) with $v_t(x) = b_t(x) - D_t(x)s_t(x)$. The expectation may then be approximated via a set of samples self-consistently propagated according to the probability flow equation (4). This implies that ρ_t must be considered a functional of s_t, since the velocity v_t depends on s_t. The following result demonstrates that if ρ_t satisfies (TE), we can solve for both ρ_t and $\nabla \log \rho_t$ by the method of characteristics.

Proposition 1 (SBTM in the Lagrangian frame)**.** *Assume that the conditions listed in Sec. 1.2 hold. Fix $T \in (0, \infty]$ and let $\lambda : [0, T) \to \mathbb{R}_{>0}$ be a positive function. Define $v_t(x) = b_t(x) - D_t(x)s_t(x)$*

and consider the optimization problem

$$\min_{s_t\,:\,t\in[0,T)} \int_0^T \lambda(t) \int_\Omega |s_t(X_t(x)) - G_t(x)|^2 \rho_0(x)dxdt,$$
$$\text{subject to:} \quad \frac{d}{dt}X_t(x) = v_t(X_t(x)), \tag{SBTM}$$
$$\frac{d}{dt}G_t(x) + [\nabla v_t(X_t(x))]^\mathsf{T} G_t(x) = -\nabla\nabla\cdot v_t(X_t(x)),$$

with initial conditions $X_0(x) = x$ and $G_0(x) = \nabla \log \rho_0(x) = s_0(x)$. Then, the unique minimizer of (SBTM) is $s_t^*(x) = -\nabla \log \rho_t^*(x)$ where $\rho_t^* : \Omega \to \mathbb{R}_{>0}$ solves the Fokker-Planck equation (FPE). Moreover, the map X_t^* associated to this minimizer is a transport map from ρ_0 to ρ_t^*:

$$x \sim \rho_0 \qquad \textit{implies that} \qquad X_t^*(x) \sim \rho_t^*, \qquad t \in [0,T]. \tag{8}$$

Proposition 1 is proven in Appendix B.2, where it is shown that $G_t(x) = \nabla \log \rho_t(X_t(x))$ if $\rho_t = X_t\sharp\rho_0$. The objective function in (SBTM) is similar to Eq. (15) of Shen et al. (2022); the equation for G_t also appears in their Proposition 2 and in Theorem D.1 of Lu et al. (2022).

The objective in (SBTM) has the important advantage that its minimum value is zero, which gives the ability to assess the quality of the learned score. This is in contrast to the loss used in SBDM, which attains an unknown negative value at the minimum due to the neglected integral $\int |\nabla \log \rho_t(x)|^2 \rho_t(x)dx$. Moreover, as highlighted by the following proposition, the objective in (SBTM) with uniform weight $\lambda(t) = 1$ controls the KL divergence of the approximate ρ_t from the true ρ_t^*. This implies that solution of the optimization problem (SBTM) to a tolerance ϵ gives a bound $\delta_t(\epsilon)$ on the distance between ρ_t and ρ_t^*:

Proposition 2 (Control of the KL divergence)**.** *Let ρ_t denote the solution to the transport equation (TE) with $v_t(x) = b_t(x) - D_t(x)s_t(x)$ and let ρ_t^* denote the solution to the Fokker-Planck equation (FPE). Assume that $\rho_{t=0}(x) = \rho_{t=0}^*(x)$ for all $x \in \Omega$. Then for any $T \in [0,\infty)$*

$$D_{KL}(\rho_T|\rho_T^*) \le \frac{D_m}{2} \int_0^T \int_\Omega |s_t(X_t(x)) - G_t(x)|^2 \rho_0(x)dxdt \tag{9}$$

where $X_t(\cdot)$ and $G_t(\cdot)$ obey the dynamics in (SBTM) and $D_m = \sup_{t\ge 0}\sup_{x\in\Omega}\|D_t(x)\| < \infty$ with $\|D_t(x)\| = \sup_{z:\|z\|=1} z^\mathsf{T} D_t(x) z$.

The proof can be found in Appendix B.3. In practice, the objective in (SBTM) can be estimated empirically by generating samples from ρ_0 and solving the equations for $X_t(x)$ and $G_t(x)$ with $x \sim \rho_0$ to form an approximation of ρ_t. The procedure only requires knowledge of the functional form of ρ_0 up to a normalization constant, and the constrained minimization problem (SBTM) can be solved using the adjoint method. Interestingly, this optimization can be performed using online learning, as fresh samples can always be drawn from ρ_0. In the special case when the SDE is an Ornstein-Uhlenbeck process, the score and the equations for both X_t and G_t can be written explicitly; they are studied in Appendix C.

3.3 Sequential SBTM

If $\lambda(t)$ is scaled so that all of its mass lies at a single time $t > 0$, the objective in (SBTM) becomes

$$\int_\Omega |s_t(X_t(x)) - G_t(x)|^2 \rho_0(x)dx$$
$$= \int_\Omega \left(|s_t(X_t(x))|^2 - 2s_t(X_t(x))\cdot G_t(x) + |G_t(x)|^2\right)\rho_0(x)dx,$$

where the term involving $|G_t(x)|^2$ can be neglected, because it only depends on s_τ at times $\tau < t$. The term $s_t(X_t(x))\cdot G_t(x)$ can be simplified using Stein's identity

$$\int_\Omega s_t(X_t(x))\cdot G_t(x)\rho_0(x)dx = -\int_\Omega \nabla\cdot s_t(X_t(x))\rho_0(x)dx \tag{10}$$

which holds as long as $G_t(x) = -\nabla \log \rho_t(X_t(x))$ and X_t is a transport map from ρ_0 to ρ_t, i.e. $\rho_t = X_t\sharp\rho_0$. This offers a variational way to estimate the score at a given time, as summarized in:

Algorithm 1 Sequential score-based transport modeling.

1: **Input**: An initial time $t_0 \in \mathbb{R}_{\geq 0}$. A set of n samples $\{x_i\}_{i=1}^n$ from ρ_{t_0}. A set of N_T timesteps $\{\Delta t_k\}_{k=0}^{N_t-1}$.
2: Initialize sample locations $X_{t_0}(x_i) = x_i$ for $i = 1, \dots, n$.
3: **for** $k = 0, \dots, N_t - 1$ **do**
4: Optimize: $s_{t_k} = \arg\min_s \frac{1}{n} \sum_{i=1}^n \left[|s\left(X_{t_k}(x_i)\right)|^2 + 2\nabla \cdot s\left(X_{t_k}(x_i)\right) \right]$.
5: Propagate samples:

$$X_{t_{k+1}}(x_i) = X_{t_k}(x_i) + \Delta t_k \left(b_{t_k}(X_{t_k}(x_i)) + D_{t_k}(X_{t_k}(x_i)) s_{t_k}(X_{t_k}(x_i)) \right).$$

6: Set $t_{k+1} = t_k + \Delta t_k$.
7: **Output**: A set of n samples $\{X_{t_k}(x_i)\}_{i=1}^n$ from ρ_{t_k} for all $\{t_k\}_{k=0}^{N_T}$.

Proposition 3 (Sequential SBTM)**.** *In the same setting as Proposition 1, let X_t be a transport map from ρ_0 to ρ_t such that $X_t \sharp \rho_0 = \rho_t$. Fix $t \geq 0$ and consider the optimization problem*

$$\min_{s_t} \int_\Omega \left(|s_t(X_t(x))|^2 + 2\nabla \cdot s_t(X_t(x)) \right) \rho_0(x) dx. \tag{seqSBTM}$$

Then the minimizer s_t^* of (seqSBTM) is unique and is given by $s_t^* = \nabla \log \rho_t$.

Proposition 3 is proven in Appendix B.4. We can use this result to self-consistently propagate samples from ρ_t using the probability flow equation (4) rather than the SDE (1), alternating between score estimation and sample propagation. The resulting procedure is presented in Algorithm 1.

To avoid computation of the divergence $\nabla \cdot s_t$ (which is often costly for neural networks), we can use the denoising score matching loss function introduced by Vincent (2011), which we discuss in Appendix B.5. Moreover, given knowledge of ρ_0 up to a normalization factor, the $\mathcal{L}_2$ loss $\int_\Omega |s_0(x) - \nabla \log \rho_0(x)|^2 \rho_0(x) dx$ may be used to fit the initial condition.

As stated previously, the score-based transport method outlined above gives access to quantities that cannot be obtained directly by simulating the SDE (1). For example, it allows for pointwise evaluation of ρ_t and the probability current $v_t(x)\rho_t(x)$, calculation of the entropy $S_t = -\int_\Omega \log \rho_t(x) \rho_t(x) dx$, and resampling of the density $\rho_t(x)$ at any time t. Computation of these quantities is explained in Appendix A and explored in the numerical examples.

4 Numerical Experiments

4.1 Interacting particle systems

In the following, we study two high dimensional examples from the physics of interacting particle systems, where the spatial variable of the Fokker-Planck equation (FPE) can be written as $x = \left(x^{(1)}, x^{(2)}, \dots, x^{(N)}\right)^\mathsf{T}$ with each $x^{(i)} \in \mathbb{R}^{\bar{d}}$. Here, $\bar{d}$ describes a lower-dimensional ambient space, e.g. $\bar{d} = 2$, so that the dimensionality of the Fokker-Planck equation $d = N\bar{d}$ will be high if the number of particles N is even moderate.

4.1.1 Harmonically interacting particles in a harmonic trap

Setup. Here we study a problem that admits a tractable analytical solution for direct comparison. We consider N two-dimensional particles ($\bar{d} = 2$) that repel according to a harmonic interaction but experience harmonic abut experience ttraction towards a moving trap $\beta_t \in \mathbb{R}^2$. The motion of the particles is governed by the stochastic dynamics

$$dX_t^{(i)} = (\beta_t - X_t^{(i)})dt - \alpha\Big(X_t^{(i)} - \frac{1}{N}\sum_{j=1}^N X_t^{(j)}\Big)dt + \sqrt{2D}\, dW_t, \quad i = 1, \dots, N \tag{11}$$

where $\alpha \in (0, 1)$ is a fixed coefficient that sets the magnitude of the repulsion. The dynamics (11) is an Ornstein-Uhlenbeck process in the extended variable $x \in \mathbb{R}^{\bar{d}N}$ with block components $x^{(i)}$. Assuming a Gaussian initial condition, the solution to the Fokker-Planck equation associated with (11)

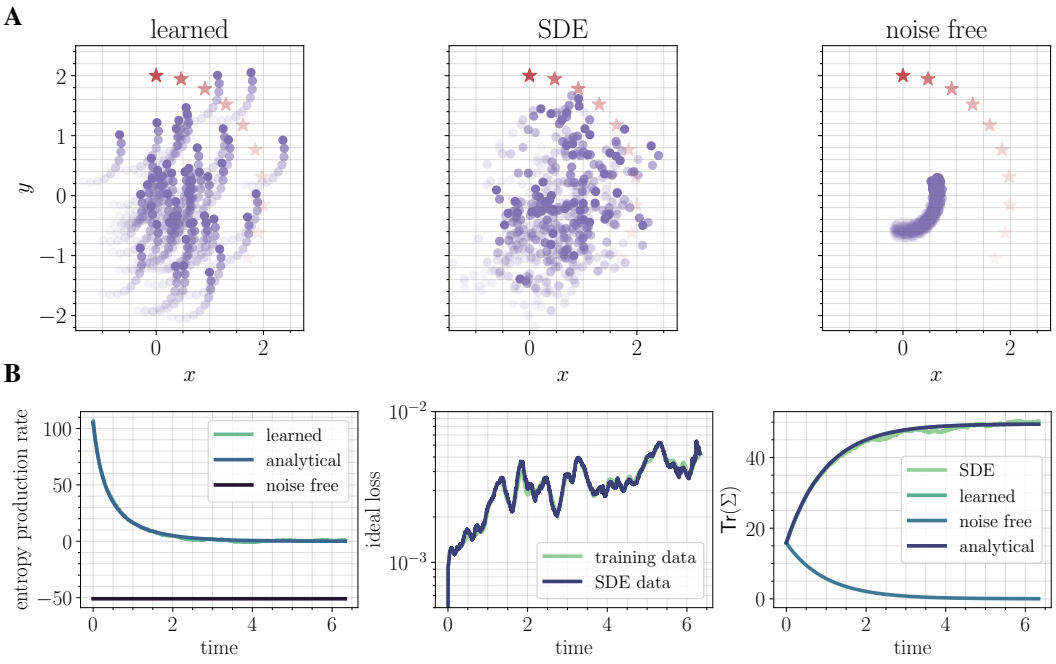

Figure 1: *A system of $N = 50$ particles in a harmonic trap with a harmonic interaction:* (A) A single sample trajectory. The mean of the trap β_t is shown with a red star, while past positions of the particles are indicated by a fading trajectory. The noise-free system (right) is too concentrated, and fails to capture the variance of the stochastic dynamics (center). The learned system (left) accurately captures the variance, and in addition generates physically interpretable trajectories for the particles. (B) Quantitative comparison to the analytical solution. The learned solution matches the entropy production rate, score, and covariance well.

is a Gaussian for all time and hence can be characterized entirely by its mean m_t and covariance C_t. These can be obtained analytically (Appendices C and D), which facilitates a quantitative comparison to the learned model. The differential entropy S_t is given by (see Appendix D)

$$S_t = \tfrac{1}{2}\bar{d}N\left(\log\left(2\pi\right) + 1\right) + \tfrac{1}{2}\log\det C_t \tag{12}$$

In the experiments, we take $\beta_t = a(\cos \pi\omega t, \sin \pi\omega t)^\mathsf{T}$ with $a = 2$, $\omega = 1$, $D = 0.25$, $\alpha = 0.5$, and $N = 50$, giving rise to a 100-dimensional Fokker-Planck equation. The particles are initialized from an isotropic Gaussian with mean β_0 (the initial trap position) and variance $\sigma_0^2 = 0.25$.

Network architecture. We take $s_t(x) = -\nabla U_{\theta_t}(x)$, where the potential $U_{\theta_t}(\cdot)$ is given as a sum of one- and two-particle terms

$$U_{\theta_t}\big(x^{(1)}, \ldots, x^{(N)}\big) = \sum_{i=1}^{N} U_{\theta_t,1}\big(x^{(i)}\big) + \frac{1}{N}\sum_{\substack{i,j=1\\ i\neq j}}^{N} U_{\theta_t,2}\big(x^{(i)}, x^{(j)}\big). \tag{13}$$

To ensure permutation symmetry amongst the particles, we require that $U_{\theta_t,2}\left(x, y\right) = U_{\theta_t,2}\left(y, x\right)$ for each $x, y \in \mathbb{R}^{\bar{d}}$. Modeling at the level of the potential introduces an additional gradient into the loss function, but makes it simple to enforce permutation symmetry; moreover, by writing the potential as a sum of one- and two-particle terms, the dimensionality of the function estimation problem is reduced. As motivation for this choice of architecture, we show in Appendix D.1 that the class of scores representable by (13) contains the analytical score for the harmonic problem considered in this section. To obtain the parameters $\theta_{t_k+\Delta t_k}$, we perform a warm start and initialize from θ_{t_k}, which reduces the number of optimization steps that need to be performed at each iteration. All networks are taken to be multi-layer perceptrons with the `swish` activation function (Ramachandran et al., 2017); further details on the architectures used can be found in Appendix D.

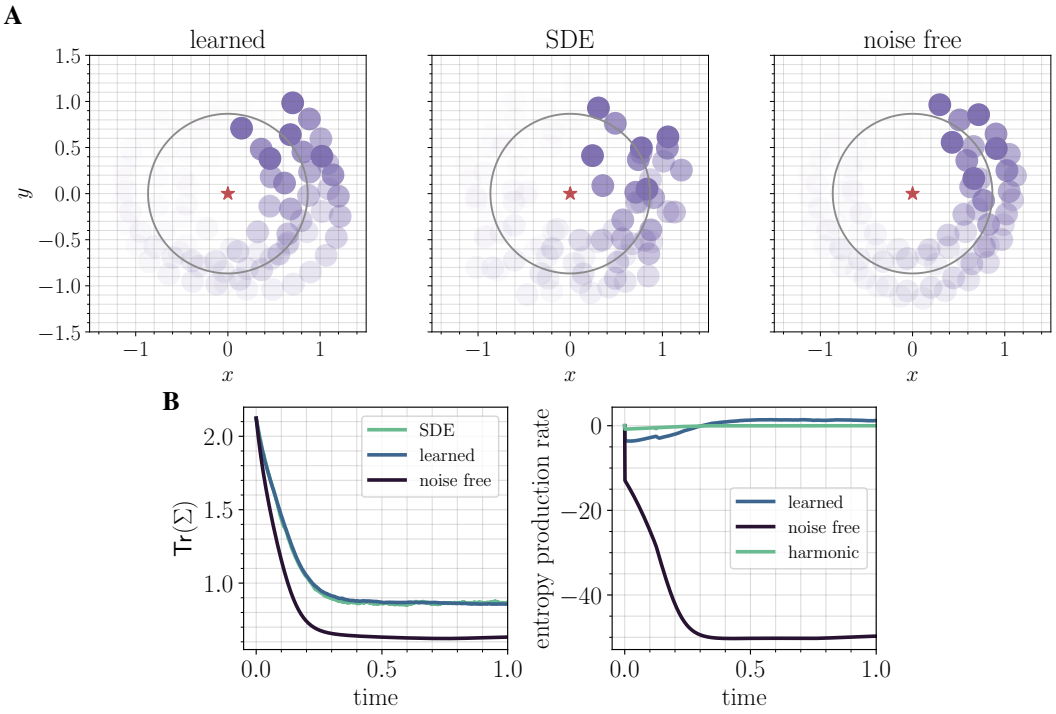

Figure 2: *A system of $N = 5$ soft-spheres in an anharmonic trap:* (A) Example particle trajectories, shown relative to the moving mean of the trap (here shown fixed at the origin in red). (B) Trace of the covariance and entropy production rate predictions for the SDE, noise-free, and learned systems. The learned system accurately predicts the moments and captures the temporal evolution of the entropy rate. The noise-free system is inaccurate in both cases, and a harmonic approximation used to compute the entropy from SDE trajectories misses the temporal evolution of the entropy.

Quantitative comparison. For a quantitative comparison between the learned model and the exact solution, we study the empirical covariance Σ over the samples and the entropy production rate $\frac{dS_t}{dt}$. Because an analytical solution is available for this system, we may also compute the target $\nabla \log \rho_t(x) = -C_t^{-1}(x - m_t)$ and measure the goodness of fit via the relative Fisher divergence

$$\frac{\int_\Omega |s_t(x) - \nabla \log \rho_t(x)|^2 \bar{\rho}(x) dx}{\int_\Omega |\nabla \log \rho_t(x)|^2 \bar{\rho}(x) dx}. \tag{14}$$

In Equation (14), $\bar{\rho}$ can be taken to be equal to the current particle estimate of ρ_t (the training data), or estimated using samples from the stochastic differential equation (the SDE data).

Results. The representation of the dynamics (11) in terms of the flow of probability leads to an intuitive deterministic motion that accurately captures the statistics of the underlying stochastic process. Snapshots of particle trajectories from the learned probability flow (4), the SDE (11), and the noise-free equation obtained by setting $D = 0$ in (11) are shown in Figure 1A.

Results for this quantitative comparison are shown in Figure 1B. The learned model accurately predicts the entropy production rate of the system and minimizes the relative metric (14) to the order of 10^{-2}. The noise-free system incorrectly predicts a constant and negative entropy production rate, while the SDE cannot make a prediction for the entropy production rate. In addition, the learned model accurately predicts the high-dimensional covariance Σ of the system (curves lie directly on top of the analytical result, trace shown for simplicity). The SDE also captures the covariance, but exhibits more fluctuations in the estimate; the noise-free system incorrectly estimates all covariance components as converging to zero.

4.1.2 SOFT SPHERES IN AN ANHARMONIC TRAP

Setup. Here, we consider a system of $N = 5$ particles in an *anharmonic* trap in dimension $\bar{d} = 2$ that exhibit soft-sphere repulsion. This system gives rise to a 10-dimensional Fokker-Planck equation, a dimensionality that is already significantly too high for standard PDE solvers. The stochastic dynamics is given by

$$dX_t^{(i)} = 4B\big(\beta_t - X_t^{(i)}\big)|X_t^{(i)} - \beta_t|^2 dt + \frac{A}{Nr^2}\sum_{j=1}^{N}\big(X_t^{(i)} - X_t^{(j)}\big)e^{-|X_t^{(i)}-X_t^{(j)}|^2/2r^2}dt + \sqrt{2D}\,dW_t, \quad i = 1, \dots, N,$$

where β_t again represents a moving trap, $A > 0$ sets the strength of the repulsion between the spheres, r sets their size, and $B > 0$ sets the strength of the trap. We set $\beta(t) = a\cos(\pi\omega t, \sin \pi\omega t)^\intercal$ with $a = 2, \omega = 1, D = 0.25, A = 30$, and $r = 0.15$. We fix $B = D/R^2$ with $R = \sqrt{\gamma N}r$ and $\gamma = 1.25$. This ensures that the trap scales with the number of particles and that they have sufficient room in the trap to generate a complex dynamics. We make use of the same network architecture as in Sec. 4.1.1.

Results. Similar to Section 4.1.1, an example trajectory from the learned system, the SDE (15) and the noise-free system obtained by setting $D = 0$ in (15) are shown in Figure 2A. Here, the particles are visualized in a Lagrangian frame that moves with the mean of the trap, so that the trap (red star) appears fixed at zero. A single level set of the trap at energy $4 \times D$ is shown in gray. The learned particle trajectories exhibit an intuitive "rolling" motion with structural rearrangements that accurately capture the statistics of the stochastic dynamics. The noise-free trajectory is too confined to the level set, and does not accurately capture the structural rearrangements (links to movies highlighting these observations can be found in Appendix D.2). Numerical estimates of the moments (trace of the covariance plotted for visual ease) and entropy production rate are shown in Figure 2B. The learned and SDE systems accurately capture the covariance, while $\mathrm{Tr}(\Sigma)$ is under-predicted by the noise-free system. The noise-free system predicts a large and negative entropy production rate, while the learned system predicts a small positive entropy production rate. Here, no analytical solution is available for comparison, and the SDE cannot be systematically used to make a prediction for the entropy. Instead, we apply a common technique from the molecular dynamics community, and compute the entropy production rate of a Gaussian with covariance computed from the stochastic trajectories (Karplus & Kushick, 1981). This "harmonic approximation" misses the temporal evolution of the entropy production rate captured by the learned system.

5 OUTLOOK AND CONCLUSIONS

Building on the toolbox of score-based diffusion recently developed for generative modeling, we introduced a related approach – score-based transport modeling (SBTM) – that gives an alternative to simulating the corresponding SDE to solve the Fokker-Planck equation. While SBTM is more costly than integration of the SDE because it involves a learning component, it gives access to quantities that are not directly accessible from the samples given by integrating the SDE, such as pointwise evaluation of the PDF, the probability current, or the entropy. Our numerical examples indicate that SBTM is scalable to systems in high dimension where standard numerical techniques for partial differential equations are inapplicable. The method can be viewed as a deterministic Lagrangian integration method for the Fokker-Planck equation, and our results show that its trajectories are more easily interpretable than the corresponding trajectories of the SDE. Future work will investigate applications of the method to other systems with simulation of the SDE is known to be difficult, as well as to other partial differential equations where SBTM may provide an appealing integration scheme even in low dimension.

REFERENCES

Richard F Bass. *Stochastic processes*, volume 33. Cambridge University Press, 2011.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1285773.

Joan Bruna, Benjamin Peherstorfer, and Eric Vanden-Eijnden. Neural galerkin scheme with active learning for high-dimensional evolution equations. *arXiv:2203.01360*, 2022.

David Chandler. Introduction to modern statistical. *Mechanics. Oxford University Press, Oxford, UK*, 5, 1987.

Bo Dai, Niao He, Hanjun Dai, and Le Song. Provable Bayesian Inference via Particle Mirror Descent. *arXiv:1506.03101*, May 2016.

Chenguang Dai, Jeremy Heng, Pierre E. Jacob, and Nick Whiteley. An invitation to sequential Monte Carlo samplers. *arXiv:2007.11936*, 2020.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. *arXiv:2106.01357*, 2021.

Pierre Degond and Francisco-José Mustieles. A deterministic approximation of diffusion equations using particles. *SIAM Journal on Scientific and Statistical Computing*, 11(2):293–310, 1990. doi: 10.1137/0911018.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-Based Generative Modeling with Critically-Damped Langevin Diffusion. *arXiv:2112.07068*, 2022.

Weinan E and Bing Yu. The deep ritz method: A deep learning-based numerical algorithm for solving variational problems, 2017. URL https://arxiv.org/abs/1710.00211.

Lawrence C Evans. *An introduction to stochastic differential equations*, volume 82. American Mathematical Soc., 2012.

Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Elsevier, 2001.

Crispin Gardiner. *Stochastic Methods*. Springer-Verlag Berlin Heidelberg, 4th edition, 2009. ISBN 978-3-540-70712-7.

Jiequn Han, Arnulf Jentzen, and Weinan E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018. doi: 10.1073/pnas.1718942115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1718942115.

Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, and Aaron Courville. Convex Potential Flows: Universal Probability Distributions with Optimal Transport and Convex Optimization. *arXiv:2012.05942*, February 2021.

Aapo Hyvarinen. Connections Between Score Matching, Contrastive Divergence, and Pseudolikelihood for Continuous-Valued Variables. *IEEE Transactions on Neural Networks*, 18(5), 2007. doi: 10.1109/TNN.2007.895819.

Aapo Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24), 2005. ISSN 1533-7928.

Aapo Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51 (5):2499–2512, 2007. doi: 10.1016/j.csda.2006.09.003.

Aapo Hyvärinen. Optimal Approximation of Signal Priors. *Neural Computation*, 20(12), 2008. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.2008.10-06-384.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167*, 2015.

Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. ISSN 0036-1410, 1095-7154. doi: 10.1137/S0036141096303359.

Martin Karplus and Joseph N Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, 1981.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*, 2017. arXiv: 1412.6980.

Peter E Kloeden and Eckhard Platen. Stochastic differential equations. In *Numerical Solution of Stochastic Differential Equations*, pp. 103–160. Springer, 1992.

Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (11):3964–3979, 2021. ISSN 1939-3539. doi: 10.1109/TPAMI.2020.2992934.

Harold Joseph Kushner Kushner, Harold J Kushner, Paul G Dupuis, and Paul Dupuis. *Numerical methods for stochastic control problems in continuous time*, volume 24. Springer Science & Business Media, 2001.

Lei Li, Yingzhou Li, Jian-Guo Liu, Zibu Liu, and Jianfeng Lu. A stochastic version of Stein variational gradient descent for efficient sampling. *Communications in Applied Mathematics and Computational Science*, 15(1):37–63, 2020. ISSN 1559-3940, 2157-5452.

Qiang Liu. Stein Variational Gradient Descent as Gradient Flow. *arXiv:1704.07520*, 2017.

Qiang Liu and Dilin Wang. Stein Variational Gradient Descent as Moment Matching. *arXiv:1810.11693*, 2018.

Qiang Liu and Dilin Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. *arXiv:1608.04471*, 2019.

Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion odes by high-order denoising score matching, 2022. URL https://arxiv.org/abs/2206.08265.

Jianfeng Lu, Yulong Lu, and James Nolen. Scaling limit of the Stein variational gradient descent: the mean field regime. *arXiv:1805.04035*, November 2018.

Dimitra Maoutsa, Sebastian Reich, and Manfred Opper. Interacting particle solutions of Fokker-Planck equations through gradient-log-density estimation. *Entropy*, 22, 2020. ISSN 1099-4300.

Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. An introduction to sampling via measure transport. In *Handbook of Uncertainty Quantification; R. Ghanem, D. Higdon, and H. Owhadi, editors*, pp. 1–41. Springer, 2016. doi: 10.1007/978-3-319-11259-6_23-1.

Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic Music Generation with Diffusion Models. *arXiv:2103.16091*, 2021. arXiv: 2103.16091.

Bernt Oksendal. *Stochastic Differential Equations*. Springer-Verlag Berlin Heidelberg, 6 edition, 2003. ISBN 978-3-642-14394-6.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. ISSN 1533-7928.

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*, 378:686–707, 2019. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for Activation Functions. *arXiv:1710.05941*, 2017.

Danilo Jimenez Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. *arXiv:1505.05770*, 2016.

Hannes Risken. Fokker-planck equation. In *The Fokker-Planck Equation*, pp. 63–95. Springer, 1996.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

Giovanni Russo. Deterministic diffusion of particles. *Communications on Pure and Applied Mathematics*, 43(6):697–733, 1990. ISSN 1097-0312.

Ardavan Saeedi, Tejas D. Kulkarni, Vikash K. Mansinghka, and Samuel J. Gershman. Variational Particle Approximations. *Journal of Machine Learning Research*, 18(69):1–29, 2017. ISSN 1533-7928.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Zebang Shen, Zhenfu Wang, Satyen Kale, Alejandro Ribeiro, Aim Karbasi, and Hamed Hassani. Self-consistency of the fokker-planck equation. *arXiv:2206.00860*, 2022.

Justin Sirignano and Konstantinos Spiliopoulos. Dgm: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339–1364, 2018. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2018.08.029. URL `https://www.sciencedirect.com/science/article/pii/S0021999118305527`.

Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. *arXiv:1907.05600*, 2020a.

Yang Song and Stefano Ermon. Improved Techniques for Training Score-Based Generative Models. *arXiv:2006.09011*, 2020b.

Yang Song and Diederik P. Kingma. How to Train Your Energy-Based Models. *arXiv:2101.03288*, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. *arXiv:2011.13456*, 2021.

Herbert Spohn. *Large scale dynamics of interacting particles*. Springer Science & Business Media, 2012.

E. G. Tabak and Cristina V. Turner. A Family of Nonparametric Density Estimation Algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013. ISSN 1097-0312. doi: 10.1002/cpa.21423.

Esteban G. Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010. ISSN 15396746, 19450796. doi: 10.4310/CMS.2010.v8.n1.a11.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. ISSN 0899-7667, 1530-888X. doi: 10.1162/NECO_a_00142.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in Variational Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, August 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2889774.