PROBABILITY FLOW SOLUTION OF THE FOKKER-PLANCK EQUATION

NICHOLAS M. BOFFI AND ERIC VANDEN EIJNDEN

Courant Institute of Mathematical Sciences
New York University, New York, NY 10012

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. Unlike the stochastic dynamics, this equation deterministically pushes samples from the initial density onto samples from the solution at any later time. The method has the advantage of giving direct access to quantities that are challenging to estimate only given samples from the solution, 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 particles according to the instantaneous probability current. Our approach is based on recent advances in score-based diffusion for generative modeling, with the important difference that the training procedure is self-contained and does not require samples from the target density to be available beforehand. To demonstrate the validity of the approach, we consider several examples from the physics of interacting particle systems; we find that the method scales well to high-dimensional systems, and accurately matches available analytical solutions 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) [15, 36, 13]. 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 [41, 1]

$$\partial_t \rho_t(x) = -\nabla \cdot (b_t(x)\rho_t(x) - D_t(x)\nabla \rho_t(x)) \quad x \in \Omega \subseteq \mathbb{R}^d \tag{1}$$

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.

E-mail address: boffi@cims.nyu.edu and eve2@cims.nyu.edu.

1.1. **High-dimensionality.** For many systems of interest – such as interacting particle systems in statistical physics [4, 51], stochastic control systems [27], and models in mathematical finance [36] – the dimensionality d can be very large. This renders standard numerical methods for partial differential equations inapplicable for solving (1), 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 (1)

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

is discretized to obtain a large number n of trajectories. In (2), $\sigma_t(x)$ 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 (1) is a PDF ρ_0 from which we can draw samples $\{X_0^i\}_{i=1}^n$, simulation of (2) via stochastic integration techniques such as the Euler-Maruyama method [25] enables the estimation of expectations via empirical averages

$$\int_\Omega \phi(x)\rho_t(x)dx \approx \frac{1}{n}\sum_{i=1}^n \phi(X_t^i), \tag{3}$$

where $\phi : \Omega \to \mathbb{R}$ is an observable of interest. While widely used, this method only provides samples from ρ_t, and other quantities of interest like the value of ρ_t itself or the time-dependent 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.

1.2. **A transport map approach.** Another possibility, building on recent theoretical advances connecting transportation of measures to the Fokker-Planck equation [23], is to recast (1) as the transport equation [54, 44]

$$\partial_t \rho_t(x) = -\nabla \cdot (v_t(x)\rho_t(x)) \tag{4}$$

where we defined the velocity field

$$v_t(x) = b_t(x) - D_t(x)\nabla \log \rho_t(x). \tag{5}$$

This formulation allows us to view ρ_t as the pushforward of ρ_0 under the flow map $X_{\tau,t}(\cdot)$ of the ordinary differential equation

$$\frac{d}{dt}X_{\tau,t}(x) = v_t(X_{\tau,t}(x)), \qquad X_{\tau,\tau}(x) = x, \quad t,\tau \geq 0. \tag{6}$$

(6) 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 [54, 44]

$$\rho_t(x) = \rho_0(X_{t,0}(x)) \exp\left(-\int_0^t \nabla \cdot v_\tau(X_{t,\tau}(x))d\tau\right) \tag{7}$$

where $X_{t,0}(x)$ is obtained by solving (6) backward from some given x. Importantly, access to the PDF as provided by (7) 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.

1.3. **Learning the flow.** The simplicity of the probability flow equation (6) is somewhat deceptive, because it depends explicitly on the solution ρ_t to the Fokker-Planck equation (1). Nevertheless, recent work in generative modeling via score-based diffusion [47, 48, 49] 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.

1.4. **Main contributions.** Our results are both theoretical and computational:

- We introduce several optimization problems that can be used to learn the velocity field (5) in the transport equation (4) so that its solution coincides with that of the Fokker Planck equation (1). 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 is self-contained, in that it does not requires input data from the SDE to learn the score and estimate ρ_t. Instead, SBTM uses samples obtained via the probability flow ODE itself.
- We show that SBTM can be used to assess the quality of the score regardless of how it was learned. This is because SBTM allows for evaluation of the full Fisher divergence, which has a minimum at zero attained only when the score function is perfect.
- We discuss and give specific estimators for quantities that can be computed via SBTM but that are not directly available from samples alone, like point-wise evaluation of ρ_t itself, the entropy, and the probability current.
- We implement and test SBTM on several examples of high and moderate dimensionality. We primarily focus on systems of interacting particles in a low-dimensional ambient space 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. Problems of this type frequently appear in the molecular dynamics of externally driven soft matter systems [14, 51].

1.5. **Related work.** Our approach builds directly on the toolbox of score matching originally developed by Hyvärinen [19, 18, 20, 21] and more recently extended in the context of diffusion-based generative modeling [47, 48, 50, 8, 11, 35]. As stated previously, the key difference between the method introduced here and these past approaches is that the probability flow equation itself is used to obtain samples needed to learn the approximation of the score. By contrast, it is common in the prior literature to assume access to samples from the target distribution (e.g., in the form of examples of natural images).

Our method shares commonalities with transport map-based approaches [34] for density estimation and variational inference [56, 2] such as normalizing flows [53, 52, 40, 17, 37, 26]. Moreover, because expectations are approximated over a set of particles according to (3), the method also inherits elements of classical particle-based approaches for density estimation such as Markov chain Monte Carlo [42] and sequential Monte Carlo [7, 10]. 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 [6, 43]. In particular, the method we propose

bears some similarity with Stein variational gradient descent (SVGD) [29, 30, 31] (see also [32, 28]), in that both methods approximate the target distribution via *deterministic* propagation of a set of particles. The key differences are that (i) our method learns the map used to propagate the particles, 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, which may not have a gradient flow structure, rather than sampling at equilibrium. In fact, for the examples considered here, an equilibrium density does not exist due to the continual presence of an external drive.

Most closely connected to our paper are the works by Maoutsa *et al.* [33] and Shen *et al.* [45], who similarly propose to bypass the SDE through use of the probability flow equation, building on earlier work by Degond and Mustieles [9]. The critical difference between Ref. [33] and our work is that the learning component in [33] is minimal, in that the authors perform estimation over a linear space or a reproducing kernel Hilbert space rather than over the significantly richer class of neural networks. Moreover, we present a variational problem that allows us to learn the score over a finite time interval, and thereby estimate its quality through the Fisher divergence. This variational problem is similar to the one proposed by Ref. [45] concurrently to our work, with key differences that ours 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 (2) need not be the gradient of a potential), and it allows for spatially-dependent and rank-deficient diffusion matrices. In addition, we derive variational problems for the solution of the Fokker-Planck equation that exploit locality in time to avoid the complexity of the NeuralODE adjoint equations and instead can be solved sequentially.

Our approach can also be viewed as an alternative to other recent neural network-based methods for the solution of partial differential equations (see e.g. [12, 38, 16, 46, 3]). 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 (i.e. that it satisfies positivity and remains normalized). 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 particles; this is naturally conducive to efficient scaling for high-dimensional systems.

1.6. **Notation and assumptions.** Throughout, we assume that the stochastic process (2) 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 (2) 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 in $(t, x) \in [0, \infty) \times \Omega$, i.e.

$$\exists K > 0 \ : \quad \forall (t, x) \in [0, \infty) \times \Omega \quad |\nabla \log \rho_t(x) - \nabla \log \rho_t(y)| \leq K|x - y|. \tag{8}$$

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

2. Methodology

2.1. **Score-based diffusion modeling.** SBDM is based on the observation that the score $\nabla \log \rho_t(x)$ can be estimated for all $(t,x) \in [0,T) \times \Omega$ with $T \in (0,\infty]$ by minimizing the following objective function over $s_t(x)$

$$\int_0^T \lambda(t) \int_\Omega |s_t(x) - \nabla \log \rho_t(x)|^2 \rho_t(x) dx dt, \tag{9}$$

where $\lambda : [0,T) \to \mathbb{R}_{\geq 0}$ is a positive function such that $\int_0^T \lambda(t) dt < \infty$. If $s_t(x) = \nabla \log \hat{\rho}_t(x)$ for some PDF $\hat{\rho}_t$, (9) is the (weighted) time integral of the Fisher divergence between ρ_t and $\hat{\rho}_t$. Expanding the square and treating the cross term using Stein's identity $\int_\Omega s_t(x) \nabla \log \rho_t(x)\, \rho_t(x) dx = -\int_\Omega \nabla \cdot s_t(x)\, \rho_t(x) dx$, we can equivalently minimize

$$\int_0^T \lambda(t) \int_\Omega \left(|s_t(x)|^2 + 2\nabla \cdot s_t(x) + |\nabla \log \rho_t(x)|^2 \right) \rho_t(x) dx dt. \tag{10}$$

Because the spatial integral in (10) is an expectation with respect to ρ_t, it can be evaluated empirically over samples. In SBDM, these samples are generated by passing the available samples from the target distribution (such as example natural images, here playing the role of ρ_0) 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 [19, 18, 20, 21].

A similar strategy could in principle be applied to the problem considered in this work, whereby the expectation with respect to ρ_t in the loss (10) could be approximated over an empirical measure via integration of the SDE (2). However, because our goal is to bypass this integration, an alternative is to define ρ_t as the solution of the transport equation (4). The expectation may then be approximated via a set of particles propagated according to the probability flow equation (6).

2.2. **Score-based transport modeling.** In SBDM, ρ_t in (9) or (10) is assumed to be the solution to the Fokker-Planck equation (1); by contrast, we take ρ_t to be the solution of the transport equation (4) in SBTM. This implies that ρ_t must be considered as a function of s_t, since the velocity v_t used in (4) depends on s_t. The following result, which studies ρ_t in a fixed (Eulerian) reference frame, demonstrates that this is not a problem.

Proposition 1 (SBTM in the Eulerian frame)**.** *Assume that the conditions listed in Sec. 1.6 hold. Fix $T \in (0,\infty]$, let $\lambda : [0,T) \to \mathbb{R}_{>0}$ be a positive function, and consider the optimization problem*

$$\min_{\{s_t : t \in [0,T)\}} \int_0^T \lambda(t) \int_\Omega |s_t(x) - \nabla \log \rho_t(x)|^2 \rho_t(x) dx dt \tag{SBTM1}$$
$$\text{subject to:} \quad \partial_t \rho_t(x) = -\nabla \cdot (v_t(x) \rho_t(x)), \quad x \in \Omega$$

with $v_t(x) = b_t(x) - D_t(x) s_t(x)$. Then the minimizer of (SBTM1) is unique and given by $s_t^*(x) = \nabla \log \rho_t^*(x)$ where $\rho_t^* : \Omega \to \mathbb{R}_{>0}$ solves

$$\partial_t \rho_t^*(x) = -\nabla \cdot (b_t(x) \rho_t^*(x) - D_t(x) \nabla \rho_t^*(x)), \quad x \in \Omega. \tag{FPE}$$

Proposition 1 is proven in Appendix B.1. In words, it states that solving the constrained optimization problem (SBTM1) is equivalent to solving the Fokker-Planck equation (FPE).

Remark 2.1. *A similar result holds if we replace* (SBTM1) *by the diffusion-weighted loss*

$$\text{(SBTM1')} \qquad \min_{\{s_t:t\in[0,T]\}} \int_0^T \int_\Omega |s_t(x) - \nabla \log \rho_t(x)|^2_{D_t(x)}\, \rho_t(x) dx dt,$$

subject to the same constraints, with $|\cdot|^2_{D_t(x)} = \langle \cdot, D_t(x)\cdot\rangle$. *In this case, the minimizer need not be unique if* $D_t(x)$ *is not invertible. Nevertheless, all minimizers agree in the range of* $D_t(x)$, *in the sense that they satisfy* $D_t(x)s_t^*(x) = D_t(x)\nabla \log \rho_t^*(x)$, *where* ρ_t^* *is the solution to* (FPE). *Since* $D_t(x)s_t^*(x)$ *is the quantity that enters the transport equation* (4) *and the probability flow ODE* (6), *agreement in the range is all that matters.*

As stated, Proposition 1 is not practical, because it is phrased in terms of the density ρ_t. The following result demonstrates that the transport map identity (7) can be used to re-express Proposition 1 entirely in terms of known quantities. The key to doing so is to change to a Lagrangian reference frame that moves with the flow of probability:

Proposition 2 (SBTM in the Lagrangian frame)**.** *In the same setting as Proposition 9, let* $v_t(x) = b_t(x) - D_t(x)s_t(x)$ *and consider the optimization problem*

$$\text{(SBTM2)} \qquad \begin{aligned} &\min_{\{s_t:t\in[0,T)\}} \int_0^T \lambda(t) \int_\Omega |s_t(X_t(x)) - G_t(x)|^2\, \rho_0(x) dx dt, \\ &\textit{subject to:} \quad \frac{d}{dt} X_t(x) = v_t(X_t(x)), \\ &\qquad\qquad \frac{d}{dt} G_t(x) + [\nabla v_t(X_t(x))]^\mathsf{T} G_t(x) = -\nabla\nabla\cdot v_t(X_t(x)). \end{aligned}$$

with initial conditions $X_0(x) = x$ *and* $G_0(x) = \nabla \log \rho_0(x)$. *Then, the minimizer* s_t^* *of* (SBTM2) *is unique and is identical to the minimizer of* (SBTM1). *Moreover, the map* X_t^* *associated to this minimizer is a transport map from* ρ_0 *to* ρ_t^*, *the solution of* (FPE), *in the sense that*

$$x \sim \rho_0 \qquad \textit{implies that} \qquad X_t^*(x) \sim \rho_t^*, \qquad t \in [0,T]. \tag{11}$$

Proposition 2 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 (SBTM2) is similar the one in Eq. (15) of Ref. [45], and the equation for G_t is given in their Proposition 2. 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. Several comments on the result in Proposition 2 are in order.

Remark 2.2. *Following Remark 2.1, a similar result holds if we replace* (SBTM2) *by the diffusion-weighted loss*

$$\text{(SBTM2')} \qquad \min_{\{s_t:t\in[0,T]\}} \int_0^T \lambda(t) \int_\Omega |s_t(X_t(x)) - G_t(x)|^2_{D_t(X_t(x))}\, \rho_0(x) dx dt$$

subject to the same constraints.

Remark 2.3. *The objectives in* (SBTM2) *and* (SBTM2') *can be estimated empirically by generating samples from* ρ_0, *evaluating* $\nabla \log \rho_0$, *and solving the equation for* $X_t(x)$ *and* $G_t(x)$ *with* $x \sim \rho_0$. *This only requires knowledge of the functional form of* ρ_0 *up to a normalization constant. In practice, the constrained minimization*

Algorithm 1 Sequential score-based transport modeling.

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 $\{\Delta t_k\}_{k=0}^{N_t-1}$.
Initialize particle locations $X_{t_0}(x_i) = x_i$ for $i = 1, \dots, n$.
for $k = 0, \dots, N_t - 1$ **do**
  Optimize: $s_{t_k} = \operatorname{argmin}_s \frac{1}{n} \sum_{i=1}^n \left[|s\left(X_{t_k}(x_i)\right)|^2 + 2\nabla \cdot s\left(X_{t_k}(x_i)\right) \right]$.
  Propagate particles:

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

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

problem (SBTM2) *can be performed using the NeuralODE methodology [5], and the corresponding adjoint equations required to impose the constraints are written in Appendix B.2. Interestingly, this optimization can be performed using online learning, as fresh samples can always be drawn from* ρ_0.

Remark 2.4. *The objectives in* (SBTM2) *and* (SBTM2') *have the important advantage that their minimum value is zero, which gives the ability to assess the quality of the learned score. This is in contrast to the loss* (10)*, which attains an unknown negative value at the minimum due to the neglected integral* $\int_0^T \lambda(t) \int |\nabla \log \rho_t(x)|^2 \rho_t(x) dx dt$.

2.3. **Sequential SBTM.** If $\lambda(t)$ is scaled so that all of its mass lies at a single time $t \geq 0$, the objective (10) becomes

$$\int_\Omega \left(|s_t(x)|^2 + 2\nabla \cdot s_t(x) + |\nabla \log \rho_t(x)|^2 \right) \rho_t(x) dx. \tag{12}$$

Given samples from ρ_t (regardless of their origin), this objective can be optimized to determine s_t by neglecting the $|\nabla \log \rho_t|^2$ term, as it only depends on s_τ for $\tau < t$ and is independent of s_t. This is summarized in the following proposition.

Proposition 3 (Sequential SBTM)**.** *In the same setting as Proposition 9, 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*

$$\min_{s_t} \int_\Omega \left(|s_t(X_t(x))|^2 + 2\nabla \cdot s_t(X_t(x)) \right) \rho_0(x) dx. \tag{SBTM3}$$

Then the minimizer s_t^* *of* (SBTM3) *is unique and is given by* $s_t^* = -\nabla \log \rho_t$.

Proposition 3 is proven in Appendix B.3. We can use this result to self-consistently propagate samples from ρ_t using the probability flow equation (6) rather than the SDE (2), alternating between score estimation and sample propagation. The resulting procedure is presented in Algorithm 1. A few comments about this algorithm are in order:

Sources of error. The method in Algorithm 1 is only approximate due to the finite particle approximation of the densities $\{\rho_{t_k}\}_{k=0}^{N_T}$ and the timestepping method used to propagate the particles. Both of these sources of error can be reduced systematically: the number of particles n and number of timesteps N_t can be increased, and the first-order accurate forward-Euler discretization can be replaced

with any higher-order integrator, such as a Runge-Kutta method. In practice, the minimization over s_{t_k} is performed by optimizing the parameters of a neural network for a fixed number of steps of a gradient-based optimization algorithm, which introduces an additional source of estimation error.

Loss function. To avoid computation of the divergence $\nabla \cdot s_t$ (which is often costly for neural networks), we can use a variant of the denoising score matching loss function introduced by [55], which we discuss in Appendix B.4. 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 obtain an estimate of the initial $s_0(x)$.

Quality measure. To measure the quality of the estimated score $\{s_{t_k}\}_{k=0}^{N_T}$ *a-posteriori*, we can integrate the equation for G_t in (SBTM2) along with the equation for X_t and use the result in the expression for the Fisher divergence

$$\frac{1}{n} \sum_{i=1}^{n} |s_{t_k}(X_{t_k}(x_i)) - G_{t_k}(x_i)|^2 . \tag{13}$$

For example, a forward Euler discretization gives

$$G_{t_{k+1}}(x_i) = G_{t_k}(x_i) - \Delta t_k \left([\nabla v_{t_k}(X_{t_k}(x_i))]^{\mathsf{T}} G_{t_k}(x_i) + \nabla \nabla \cdot v_{t_k}(X_{t_k}(x_i)) \right) . \tag{14}$$

Because the quantity in (13) will be identically zero if the score is perfect, it provides a quantitative measure of the validity of the score approximation.

2.4. **Outputs of the method.** The score-based transport method outlined above allows us to compute quantities that cannot be obtained directly by simulating the stochastic differential equation (2). Here we discuss several interesting options, some of which are explored in the numerical examples.

Pointwise evaluation of ρ_t and the probability current. As stated in Section 1.2, if ρ_0 can be evaluated pointwise in Ω, then ρ_t can be estimated at any point $x \in \Omega$ by running the probability flow equation (6) backward in time and applying the formula (7)[1]. Similarly, the probability current at any $x \in \Omega$ is given by

$$v_t(x)\rho_t(x) = v_t(x)\rho_0(X_{t,0}(x)) \exp\left(- \int_0^t \nabla \cdot v_\tau(X_{\tau,t}(x)) d\tau \right) . \tag{15}$$

These expressions are exact in principle, but their accuracy relies on the quality of the approximation of v_t, which will deteriorate away from the samples on which it was learned; as a result, these estimates will have highest accuracy locally around samples.

Calculation of the entropy $S_t = \int_\Omega \log \rho_t(x) \rho_t(x) dx$. Assuming that the initial value of the entropy $S_0 = \int_\Omega \log \rho_0(x) \rho_0(x) dx$ is known, the entropy can be estimated at any later time. Using that the flow map $X_{0,t}(\cdot)$ for the probability flow equation (6) is a transport map from ρ_0 to ρ_t and applying the expression for ρ_t (7),

$$S_t = \int_\Omega \log \rho_t(X_{0,t}(x)) \, \rho_0(x) dx = S_0 - \int_0^t \int_\Omega \nabla \cdot v_\tau(X_{0,\tau}(x)) \rho_0(x) dx d\tau. \tag{16}$$

[1]This formula is derived in Appendix A for completeness, but may also be found in the appendix of [5].

Applying Stein's identity and assuming a perfectly learned score, equation (16) is equivalent to the expression

$$S_t = S_0 + \int_0^t \int_\Omega s_\tau(X_{0,\tau}(x)) \cdot v_\tau(X_{0,\tau}(x))\rho_0(x)dxd\tau. \tag{17}$$

Equations (16) and (17) are derived in Appendix A. Importantly, they can both be evaluated empirically using only the samples propagated by the probability flow equation (6). If the learned s_t is not exact, these two expressions for S_t give different results; evaluating S_t via both formulas gives an indirect method to assess the accuracy of the score.

Resampling of ρ_t at any time t. If the score $s_t \approx \nabla \log \rho_t$ is known to sufficient accuracy, ρ_t can be resampled at any time t using the dynamics

$$dX_\tau = s_t(X_\tau)d\tau + dW_\tau. \tag{18}$$

In (18), τ is an artificial time used for sampling that is distinct from the physical time in (2). For $s_t = \nabla \log \rho_t$, the equilibrium distribution of (18) is exactly ρ_t. In practice, s_t will be imperfect and will have an error that increases away from the samples used to learn it; as a result, (18) should be used near samples for a fixed amount of time to avoid the introduction of additional errors.

3. Numerical Experiments

3.1. **Interacting particles systems.** In the following, we study two examples from the physics of interacting particle systems, where the spatial variable of the Fokker-Planck equation (1) can be written as $x = \left(x^{(1)}, x^{(2)}, \ldots, x^{(N)}\right)^\mathsf{T}$ 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 overall system $d = N\bar{d}$ can be high if the number of particles N is large. All numerical experiments were performed in `jax` using the `dm-haiku` package to implement the networks and the `optax` package for optimization.

Network architecture. In each case 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

$$U_{\theta_t}\left(x^{(1)}, \ldots, x^{(N)}\right) = \sum_{i=1}^N U_{\theta_t,1}\left(x^{(i)}\right) + \frac{1}{N}\sum_{\substack{i,j=1 \\ i\neq j}}^N U_{\theta_t,2}\left(x^{(i)}, x^{(j)}\right). \tag{19}$$

To ensure permutation symmetry amongst the particles, we require that $U_{\theta_t,2}(x,y) = U_{\theta_t,2}(y,x)$ 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. By contrast, modeling at the level of the score would lead to a requirement that the neural network satisfy permutation equivariance rather than permutation invariance. As motivation for the choice of architecture in (19), we show in Appendix D.1 that the class of scores representable by (19) contains the analytical score for a high-dimensional problem considered in the next 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 simple multi-layer perceptrons that use the `swish` activation function [39]; further details on the architectures used can be found in Appendix D.

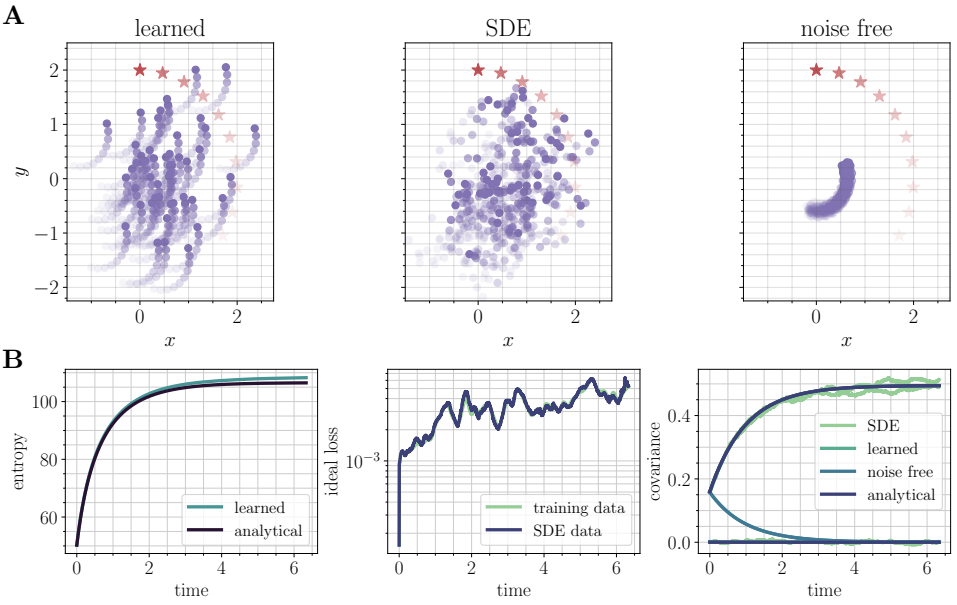

FIGURE 1. *A system of $N = 50$ particles in a harmonic trap with a harmonic interaction:* (A) A single sample trajectory. The mean of the trap β_t is shown with a red star, while past positions of the particles are indicated by a fading trajectory. The noise-free system (right) is too concentrated, 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

3.1.1. *Harmonically interacting particles in a harmonic trap.*
Setup. Here we consider a model problem that admits a tractable analytical solution for direct comparison. We consider N two-dimensional particles ($\bar{d} = 2$) that repel each other according to a harmonic interaction while experiencing harmonic attraction towards a moving trap $\beta_t \in \mathbb{R}^2$. The motion of the particles is governed by the stochastic dynamics

$$dX_t^{(i)} = (\beta_t - X_t^{(i)})dt - \alpha\Big(X_t^{(i)} - \frac{1}{N}\sum_{j=1}^{N} X_t^{(j)}\Big)dt + \sqrt{2D}\, dW_t, \quad i = 1, \dots, N \tag{20}$$

where $\alpha \in (0, 1)$ is a fixed coefficient that sets the magnitude of the repulsion. The dynamics (20) 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 (20) 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 enables a quantitative comparison to the learned model. The entropy S_t can be obtained by direct calculation (see

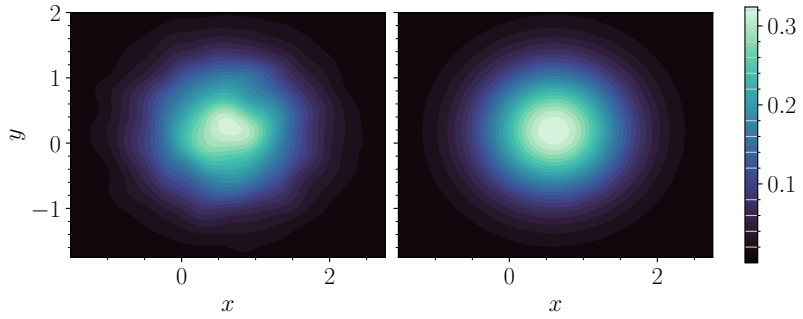

FIGURE 2. *A system of $N = 50$ particles in a harmonic trap with a harmonic interaction:* Kernel density estimate for the one-particle PDF (left) compared to the ideal one-particle PDF (right).

Appendix D) as

$$S_t = N\left(\log\left(2\pi\right) + 1\right) + \tfrac{1}{2}\log\det C_t \tag{21}$$

In the experiments, we take $\beta_t = a(\cos\pi\omega t, \sin\pi\omega t)^\mathsf{T}$ with $a = 2$ and $\omega = 1$, $D = 0.25$, $\alpha = 0.5$, and $N = 50$, corresponding 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 $\sigma_0^2 = 0.25$.

Results. The representation of the dynamics (20) 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 (6), the stochastic differential equation (20), and the noise-free equation obtained by setting $D = 0$ in (20) are shown in Figure 1A.

For a quantitative comparison between the learned and the exact models, we compare the mean m_t, the covariance C_t, and the entropy S_t. 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 score discrepancy

$$\frac{\int_\Omega |s_t(x) - \nabla \log \rho_t(x)|^2 \bar{\rho}(x) dx}{\int_\Omega |\nabla \log \rho_t(x)|^2 \bar{\rho}(x) dx}. \tag{22}$$

In Equation (22), $\bar{\rho}$ 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 for this quantitative comparison are shown in Figure 1B. The learned model accurately predicts the entropy of the system and minimizes the relative metric (22) to the order of 10^{-2}. In addition, the learned model accurately predicts the covariance of the one-particle solution obtained by marginalizing over all particles except one (curves lie directly on top of the analytical result). The SDE also captures the covariance, but exhibits more fluctuations in the estimate; by contrast, the noise-free system incorrectly estimates all covariance components as converging to zero. A kernel density estimate of the one-particle PDF also agrees well with the analytical PDF (Figure 2).

3.1.2. *Soft spheres in an anharmonic trap.*

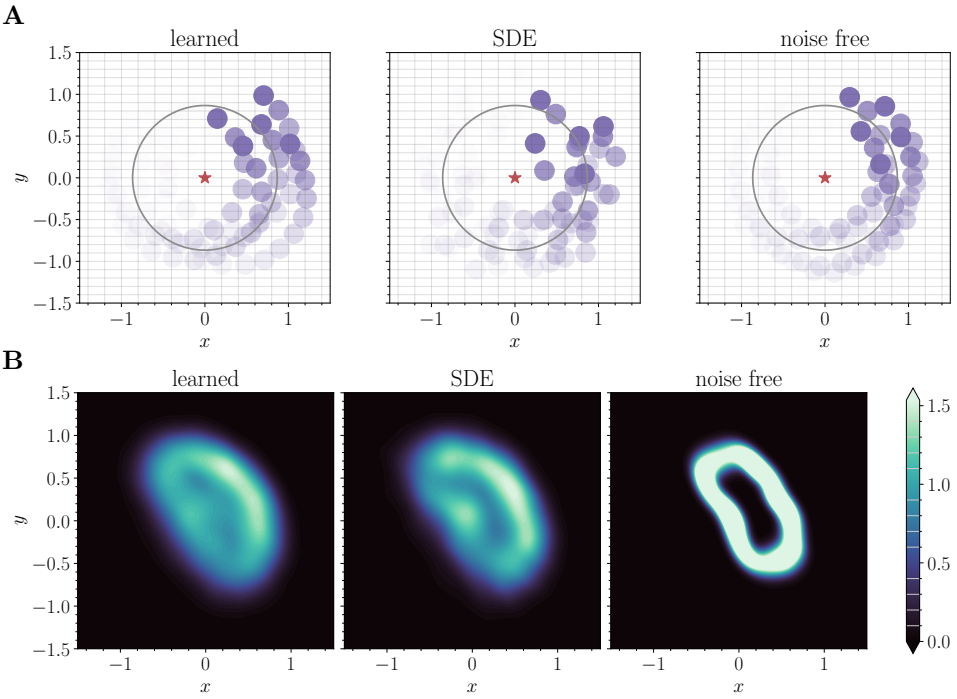

FIGURE 3. *A system of $N = 5$ soft-spheres in an anharmonic trap:* (A) Example particle trajectories, shown relative to the moving mean of the trap (here shown fixed at the origin in red). (B) Kernel density estimates for the one-particle PDF

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; such moderate dimensionality 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 $\beta(t) = a\cos(\pi\omega t, \sin\pi\omega t)^\mathsf{T}$ 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.

Results. Similar to Section 3.1.1, an example trajectory from the learned system, the SDE (23) and the noise-free system obtained by setting $D = 0$ in (23) are shown in Figure 3A. 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

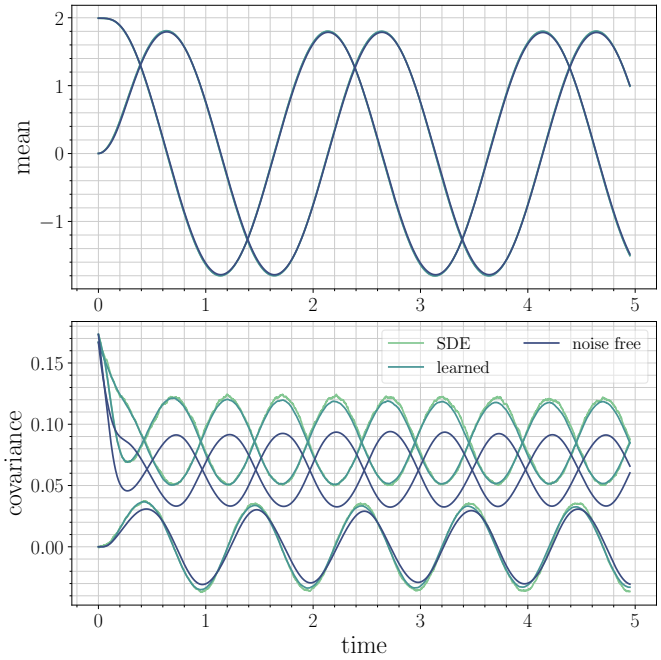

FIGURE 4. *System of $N = 5$ soft-spheres in an anharmonic trap:* Numerical estimates of the one-particle mean and covariance ($\bar{d}$ and $\bar{d}^2$ dimensional, respectively).

trajectories exhibit an intuitive "rolling" motion with structural rearrangements that accurately capture the statistics of the stochastic dynamics. By contrast, 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). Kernel density estimates of the single-particle PDF are shown in Figure 3B and numerical estimates of the moments are shown in Figure 4, Both highlight that the learned dynamics accurately captures the statistics of the process while the noise-free dynamics only captures the mean.

3.2. **Active swimmer.**

Setup. Here, we study an "active swimmer" model in two dimensions given by the stochastic differential equation

$$\dot{x} = -x^3 + v \tag{23}$$

$$dv = -\gamma v dt + \sqrt{2\gamma D} dW_t \tag{24}$$

with $\gamma = 0.1$ and $D = 1.0$. This example describes the motion of a particle in one dimension in a fixed anharmonic trap with a tendency to continue traveling in a noisy direction. While the example is low-dimensional, it exhibits convergence to a nonequilibrium statistical steady state in which the probability current is nonzero. Because the noise only enters the dynamics through the velocity variable v in (23) & (24), the learned score only enters the dynamics for v in the probability

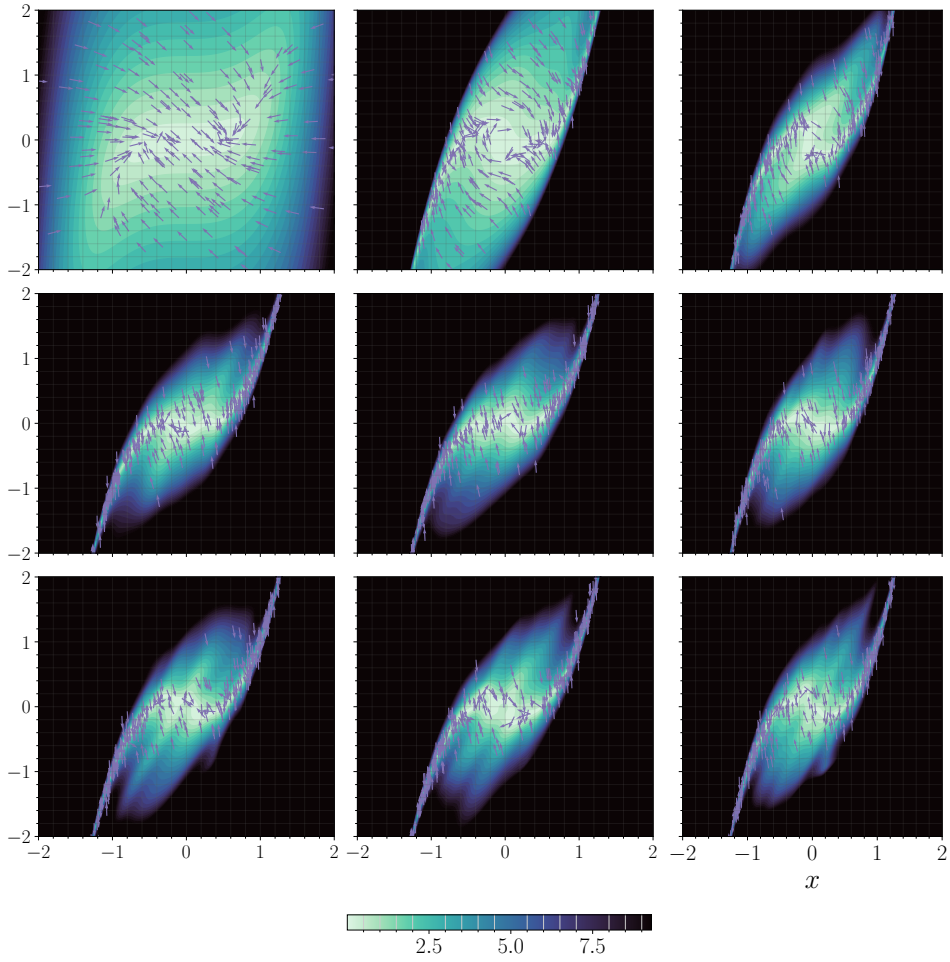

FIGURE 5. *An active swimmer: learned velocity.* The learned velocity field (right-hand side of (6)) for the active swimmer example. Color indicates the magnitude of the velocity field computed on a grid, while arrows indicate the direction of the velocity field on samples. Time corresponds to progressing in the grid along columns from the top-left to the bottom-right image ($t = k \times .75$ with k the image number, zero-indexed). The learned velocity field converges to closed streamlines that enforce a nonzero steady-state current.

equation (6), and is hence one-dimensional. To exploit this structure, we parameterize the score directly $s_t : \mathbb{R}^2 \to \mathbb{R}$ (rather than as the gradient of a potential) using a three-hidden layer neural network with `n_hidden` $= 32$ neurons per hidden layer.

Results. In Figure 5, we show that the nonequilibrium steady-state current can be captured by the learned velocity field. Further results, including example particle trajectories and estimated densities computed via kernel density estimation can be found in Appendix D.3. Similar to the previous examples, these results highlight the

inability of the noise-free system to capture the statistics of the underlying stochastic process; furthermore, the noise-free systems fails to capture the steady-state current.

4. Outlook and Conclusions

Building on the toolbox of score-based diffusion recently developed in the context of generative modeling, we introduced a related approach – score-based transport modeling (SBTM) – that gives an alternative to simulating the corresponding stochastic differential equation to solve the Fokker-Planck equation. While SBTM is more costly than integration of the SDE because it involves a learning component, it also gives access to quantities that are not directly accessible from the samples given by integrating the SDE, like the pointwise value of the PDF, the probability current, or the entropy. Our numerical examples on interacting particle systems indicate that SBTM is scalable to systems in high or moderate 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 numerical results show that its trajectories are more easily interpretable than the corresponding trajectories of the SDE. This suggests that the method may be useful in other, more challenging contexts, where the calculation of quantities like the entropy is difficult but desirable. Future work will investigate applications of the method to such systems, as well as to other partial differential equations where SBTM may provide an appealing Lagrangian integration scheme even in low dimension. The loss used in SBTM may also have application in the context of SBDM, where it could be used to estimate the Fisher divergence and thereby the quality of the score learned.

References

1. Richard F Bass, *Stochastic processes*, vol. 33, Cambridge University Press, 2011.
2. David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe, *Variational Inference: A Review for Statisticians*, Journal of the American Statistical Association **112** (2017), no. 518, 859–877 (en).
3. Joan Bruna, Benjamin Peherstorfer, and Eric Vanden-Eijnden, *Neural galerkin scheme with active learning for high-dimensional evolution equations*, 2022.
4. David Chandler, *Introduction to modern statistical*, Mechanics. Oxford University Press, Oxford, UK **5** (1987).
5. Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud, *Neural Ordinary Differential Equations*, arXiv:1806.07366 (2019).
6. Bo Dai, Niao He, Hanjun Dai, and Le Song, *Provable Bayesian Inference via Particle Mirror Descent*, arXiv:1506.03101 (2016).
7. Chenguang Dai, Jeremy Heng, Pierre E. Jacob, and Nick Whiteley, *An invitation to sequential Monte Carlo samplers*, (2020).
8. Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet, *Diffusion Schr\"odinger Bridge with Applications to Score-Based Generative Modeling*, arXiv:2106.01357 (2021).
9. Pierre Degond and Francisco-José Mustieles, *A deterministic approximation of diffusion equations using particles*, SIAM Journal on Scientific and Statistical Computing **11** (1990), no. 2, 293–310.

10. Pierre Del Moral, Arnaud Doucet, and Ajay Jasra, *Sequential Monte Carlo samplers*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68** (2006), no. 3, 411–436 (en).
11. Tim Dockhorn, Arash Vahdat, and Karsten Kreis, *Score-Based Generative Modeling with Critically-Damped Langevin Diffusion*, arXiv:2112.07068 (2022).
12. Weinan E and Bing Yu, *The deep ritz method: A deep learning-based numerical algorithm for solving variational problems*, 2017.
13. Lawrence C Evans, *An introduction to stochastic differential equations*, vol. 82, American Mathematical Soc., 2012.
14. Daan Frenkel and Berend Smit, *Understanding molecular simulation: from algorithms to applications*, vol. 1, Elsevier, 2001.
15. Crispin Gardiner, *Stochastic methods*, 4th ed., Springer-Verlag Berlin Heidelberg, 2009.
16. Jiequn Han, Arnulf Jentzen, and Weinan E, *Solving high-dimensional partial differential equations using deep learning*, Proceedings of the National Academy of Sciences **115** (2018), no. 34, 8505–8510.
17. 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 [cs, math] (2021).
18. Aapo Hyvarinen, *Connections Between Score Matching, Contrastive Divergence, and Pseudolikelihood for Continuous-Valued Variables*, IEEE Transactions on Neural Networks **18** (2007), no. 5.
19. Aapo Hyvärinen, *Estimation of Non-Normalized Statistical Models by Score Matching*, Journal of Machine Learning Research **6** (2005), no. 24.
20. ———, *Some extensions of score matching*, Computational Statistics & Data Analysis **51** (2007), no. 5, 2499–2512.
21. ———, *Optimal Approximation of Signal Priors*, Neural Computation **20** (2008), no. 12.
22. Sergey Ioffe and Christian Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, arXiv:1502.03167 (2015).
23. Richard Jordan, David Kinderlehrer, and Felix Otto, *The Variational Formulation of the Fokker–Planck Equation*, SIAM Journal on Mathematical Analysis **29** (1998), no. 1, 1–17.
24. Diederik P. Kingma and Jimmy Ba, *Adam: A Method for Stochastic Optimization*, arXiv:1412.6980 (2017), arXiv: 1412.6980.
25. Peter E Kloeden and Eckhard Platen, *Stochastic differential equations*, Numerical Solution of Stochastic Differential Equations, Springer, 1992, pp. 103–160.
26. 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** (2021), no. 11, 3964–3979.
27. Harold Joseph Kushner Kushner, Harold J Kushner, Paul G Dupuis, and Paul Dupuis, *Numerical methods for stochastic control problems in continuous time*, vol. 24, Springer Science & Business Media, 2001.
28. 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** (2020), no. 1, 37–63.

29. Qiang Liu, *Stein Variational Gradient Descent as Gradient Flow*, arXiv:1704.07520 (2017).
30. Qiang Liu and Dilin Wang, *Stein Variational Gradient Descent as Moment Matching*, arXiv:1810.11693 (2018).
31. ———, *Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm*, arXiv:1608.04471 (2019).
32. Jianfeng Lu, Yulong Lu, and James Nolen, *Scaling limit of the Stein variational gradient descent: the mean field regime*, arXiv:1805.04035 (2018).
33. Dimitra Maoutsa, Sebastian Reich, and Manfred Opper, *Interacting particle solutions of Fokker-Planck equations through gradient-log-density estimation*, Entropy **22** (2020).
34. Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini, *An introduction to sampling via measure transport*, Handbook of Uncertainty Quantification; R. Ghanem, D. Higdon, and H. Owhadi, editors, Springer, 2016, pp. 1–41.
35. Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon, *Symbolic Music Generation with Diffusion Models*, arXiv:2103.16091 (2021), arXiv: 2103.16091.
36. Bernt Oksendal, *Stochastic differential equations*, 6 ed., Springer-Verlag Berlin Heidelberg, 2003.
37. 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** (2021), no. 57, 1–64.
38. 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), 686–707.
39. Prajit Ramachandran, Barret Zoph, and Quoc V. Le, *Searching for Activation Functions*, arXiv:1710.05941 (2017).
40. Danilo Jimenez Rezende and Shakir Mohamed, *Variational Inference with Normalizing Flows*.
41. Hannes Risken, *Fokker-planck equation*, The Fokker-Planck Equation, Springer, 1996, pp. 63–95.
42. Christian P. Robert and George Casella, *Monte carlo statistical methods*, Springer, 2004.
43. Ardavan Saeedi, Tejas D. Kulkarni, Vikash K. Mansinghka, and Samuel J. Gershman, *Variational Particle Approximations*, Journal of Machine Learning Research **18** (2017), no. 69, 1–29.
44. Filippo Santambrogio, *Optimal transport for applied mathematicians*, Birkäuser, NY **55** (2015), no. 58-63, 94.
45. Zebang Shen, Zhenfu Wang, Satyen Kale, Alejandro Ribeiro, Aim Karbasi, and Hamed Hassani, *Self-consistency of the fokker-planck equation*, arXiv:2206.00860 (2022).
46. Justin Sirignano and Konstantinos Spiliopoulos, *Dgm: A deep learning algorithm for solving partial differential equations*, Journal of Computational Physics **375** (2018), 1339–1364.
47. Yang Song and Stefano Ermon, *Generative Modeling by Estimating Gradients of the Data Distribution*, arXiv:1907.05600 (2020).

48. ———, *Improved Techniques for Training Score-Based Generative Models*, arXiv:2006.09011 (2020).
49. Yang Song and Diederik P. Kingma, *How to Train Your Energy-Based Models*, arXiv:2101.03288 (2021).
50. 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).
51. Herbert Spohn, *Large scale dynamics of interacting particles*, Springer Science & Business Media, 2012.
52. E. G. Tabak and Cristina V. Turner, *A Family of Nonparametric Density Estimation Algorithms*, Communications on Pure and Applied Mathematics **66** (2013), no. 2, 145–164.
53. Esteban G. Tabak and Eric Vanden-Eijnden, *Density estimation by dual ascent of the log-likelihood*, Communications in Mathematical Sciences **8** (2010), no. 1, 217–233 (en).
54. Cédric Villani, *Optimal transport: old and new*, vol. 338, Springer, 2009.
55. Pascal Vincent, *A Connection Between Score Matching and Denoising Autoencoders*, Neural Computation **23** (2011), no. 7, 1661–1674.
56. Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt, *Advances in Variational Inference*, IEEE Transactions on Pattern Analysis and Machine Intelligence **41** (2019), no. 8, 2008–2026.

Appendix A. Some basic formulas

Here, we derive some results linking the solution of the transport equation (4) with that of the probability flow equation (6). First we have:

Lemma 1. *Let $\rho_t : \Omega \to \mathbb{R}_{\geq 0}$ satisfy the transport equation*

$$\partial_t \rho_t(x) = -\nabla \cdot (v_t(x)\rho_t(x)) \,. \tag{A.1}$$

Assume that $v_t(x)$ is C^2 in both t and x for $t \geq 0$ and globally Lipschitz in x. Then, given any $t, t' \geq 0$, the solution of (A.1) satisfies

$$\rho_t(x) = \rho_{t'}(X_{t,t'}(x)) \exp\left(- \int_{t'}^{t} \nabla \cdot v_\tau(X_{t,\tau}(x)) d\tau \right) \tag{A.2}$$

where $X_{\tau,t}$ is the probability flow solution to (6). In addition, given any test function $\phi : \Omega \to \mathbb{R}$, we have

$$\int_\Omega \phi(x)\rho_t(x)dx = \int_\Omega \phi(X_{t',t}(x))\rho_{t'}(x)dx. \tag{A.3}$$

In words, Lemma 1 states that an evaluation of the PDF ρ_t at a given point x may be obtained by evolving the probability flow equation (6) backwards to some earlier time t' to find the point x' that evolves to x at time t, assuming that $\rho_{t'}(x')$ is available. In particular, for $t' = 0$, we obtain

$$\rho_t(x) = \rho_0(X_{t,0}(x)) \exp\left(- \int_{0}^{t} \nabla \cdot v_\tau(X_{t,\tau}(x)) d\tau \right), \tag{A.4}$$

and

$$\int_\Omega \phi(x)\rho_t(x)dx = \int_\Omega \phi(X_{0,t}(x))\rho_0(x)dx. \tag{A.5}$$

Since the probability current is by definition $v_t(x)\rho_t(x)$, using (A.4) to express $\rho_t(x)$ also gives equation (15) for the current.

Proof. The assumed C^2 and globally Lipschitz conditions on v_t guarantee global existence (on $t \geq 0$) and uniqueness of the solution to (6). Differentiating $\rho_t(X_{t',t}(x))$ with respect to t and using (6) and (A.1) we deduce

$$\begin{aligned} \frac{d}{dt}\rho_t(X_{t',t}(x)) &= \partial_t \rho_t(X_{t',t}(x)) + \frac{d}{dt} X_{t',t}(x) \cdot \nabla \rho_t(X_{t',t}(x)) \\ &= \partial_t \rho_t(X_{t',t}(x)) + v_t(X_{t',t}(x)) \cdot \nabla \rho_t(X_{t',t}(x)) \\ &= -\nabla \cdot v_t(X_{t',t}(x))\, \rho_t(X_{t',t}(x)) \end{aligned} \tag{A.6}$$

Integrating this equation in t from $t = t'$ to $t = t$ gives

$$\rho_t(X_{t',t}(x)) = \rho_{t'}(x) \exp\left(- \int_{t'}^{t} \nabla \cdot v_\tau(X_{t',\tau}(x)) d\tau \right) \tag{A.7}$$

Evaluating this expression at $x = X_{t,t'}(x)$ and using the group properties (i) $X_{t',t}(X_{t,t'}(x)) = x$ and (ii) $X_{t',\tau}(X_{t,t'}(x)) = X_{t,\tau}(x)$ gives (A.2). Equation (A.3) can be derived by using (A.2) to express $\rho_t(x)$ in the integral at the left hand-side, changing integration variable $x \to X_{t',t}(x)$ and noting that the factor $\exp\left(- \int_{t'}^{t} \nabla \cdot v_\tau(X_{t,\tau}(x)) \right)$ is precisely the Jacobian of this change of variable. The result is the integral at the right hand-side of (A.3). □

Lemma 1 also holds locally in time for any $v_t(x)$ that is C^2 in both t and x. In particular, it holds locally if we set $s_t(x) = \nabla \log \rho_t(x)$ and if we assume that $\rho_0(x)$ is (i) positive everywhere on Ω and (ii) C^3 in x. In this case, (A.1) is the Fokker-Planck equation (1) and (A.2) holds for the solution to that equation.

We now consider computation of the entropy, and state a similar result.

Lemma 2. *Assume that $\rho_0 : \Omega \to \mathbb{R}_{\geq 0}$ is positive everywhere on Ω and C^3 in its argument. Let $\rho_t : \Omega \to \mathbb{R}_{\geq 0}$ denote the solution to the FPE (1) (or equivalently, to the transport equation (A.1) with $s_t(x) = \nabla \log \rho_t(x)$ in the definition of $v_t(x)$). Then the entropy $S_t = \int_\Omega \log \rho_t(x)\, \rho_t(x) dx$ can expressed as*

$$S_t = \int_\Omega \log \rho_t(X_{0,t}(x))\, \rho_0(x) dx = S_0 - \int_0^t \int_\Omega \nabla \cdot v_\tau(X_{0,\tau}(x)) \rho_0(x) dx d\tau \tag{A.8}$$

or

$$S_t = S_0 + \int_0^t \int_\Omega s_\tau(X_{0,\tau}(x)) \cdot v_\tau(X_{0,\tau}(x)) \rho_0(x) dx d\tau \tag{A.9}$$

Equations (A.8) and (A.9) are repeats of (16) and (17), respectively.

Proof. We first derive (A.8). Observe that applying (A.5) with $\phi = \log \rho_t$ leads to the first equality. The second can then be deduced from (A.4). To derive (A.9), notice that from (A.1),

$$\begin{aligned} \frac{d}{dt} S_t &= -\int_\Omega \log \rho_t(x) \nabla \cdot (v_t(x) \rho_t(x))\, dx, \\ &= \int_\Omega \nabla \log \rho_t(x) \cdot v_t(x) \rho_t(x) dx, \\ &= \int_\Omega s_t(x) \cdot v_t(x) \rho_t(x) dx. \end{aligned} \tag{A.10}$$

Above, we used integration by parts to obtain the second equality and $s_t = \nabla \log \rho_t$ to get the third. Now, using (A.5) with $\phi = s_t \cdot v_t$ integrating the result gives (A.9). □

Appendix B. Further details on Score-Based Transport Modeling

Like SBDM, SBTM is based on the observation that we can learn the score $s_t = \nabla \log \rho_t$ globally on some interval $t \in [0, T]$ via the minimization problem

$$\min_{\{s_t : t \in [0,T)\}} \int_0^T \lambda(t) \int_\Omega |s_t(x) - \nabla \log \rho_t(x)|^2 \rho_t(x) dx dt \tag{B.1}$$

where $\lambda(t) > 0$ is some pre-defined function that weights the data over the time interval (e.g. $\lambda(t) = 1$ or $\lambda(t) = e^{-t}$). The primary difference between SBDM and SBTM is the definition of ρ_t: in SBDM ρ_t is an external input given by the solution to the Fokker-Planck equation (1). By contrast, in SBTM ρ_t is the solution to the transport equation (A.1), which itself depends on s_t. As a result, unlike in SBDM, ρ_t must be treated as a functional of s_t. We now study what this entails, first working with the transport equation (A.1) in App. B.1 and then with the probability flow equation (6) in App. B.2. While the second approach is more amenable to a practical implementation, the first is conceptually simpler. For simplicity of exposition, we restrict ourselves to the case $\lambda(t) = 1$.

B.1. **Proof of Proposition 1.**

Proposition 1 (SBTM in the Eulerian frame)**.** *Assume that the conditions listed in Sec. 1.6 hold. Fix $T \in (0,\infty]$, let $\lambda : [0,T) \to \mathbb{R}_{>0}$ be a positive function, and consider the optimization problem*

$$\text{(SBTM1)} \qquad \min_{\{s_t : t\in[0,T)\}} \int_0^T \lambda(t) \int_\Omega \left|s_t(x) - \nabla \log \rho_t(x)\right|^2 \rho_t(x) dx dt$$
$$\text{subject to:} \quad \partial_t \rho_t(x) = -\nabla \cdot (v_t(x)\rho_t(x)), \quad x \in \Omega$$

with $v_t(x) = b_t(x) - D_t(x)s_t(x)$. Then the minimizer of (SBTM1) is unique and given by $s_t^*(x) = \nabla \log \rho_t^*(x)$ where $\rho_t^* : \Omega \to \mathbb{R}_{>0}$ solves

$$\text{(FPE)} \qquad \partial_t \rho_t^*(x) = -\nabla \cdot (b_t(x)\rho_t^*(x) - D_t(x)\nabla \rho_t^*(x)), \quad x \in \Omega.$$

Proof. The constrained minimization problem (SBTM1') can be handled by considering the extended objective

$$\text{(B.2)} \qquad \int_0^T \int_\Omega \left(\left|s_t - \nabla \log \rho_t\right|^2 \rho_t + \mu_t \left(\partial_t \rho_t + \nabla \cdot (v_t \rho_t)\right)\right) dx dt$$

where $v_t = b_t - D_t s_t$ and $\mu_t : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is a Lagrange multiplier. The Euler-Lagrange equations associated with (B.2) read

$$\text{(B.3)} \qquad \begin{aligned} &\partial_t \rho_t = -\nabla \cdot ((b_t - D_t s_t)\rho_t) \\ &\partial_t \mu_t = (b_t - D_t s_t) \cdot \nabla \mu_t + |s_t|^2 - |\nabla \log \rho_t|^2 + 2\nabla \cdot (s_t - \nabla \log \rho_t), \\ &0 = \mu_T(x), \\ &0 = s_t - \nabla \log \rho_t - D_t \nabla \mu_t \end{aligned}$$

Clearly, these equations are satisfied if $s_t^*(x) = \nabla \log \rho_t^*(x)$ for all $x \in \Omega$, $\mu_t^*(x) = 0$ for all x, and ρ_t^* solves (FPE). This solution is also a global minimizer, because it zeroes the value of the objective. Moreover, all global minimizers must satisfy $s_t^*(x) = \nabla \log \rho_t^*(x)$ (ρ_t-almost everywhere), as this is the *only* way to zero the objective. It is also easy to see that there are no other local minimizers. To check this, we can eliminate s_t from (B.3) using the fourth equation. This reduces the first three to

$$\text{(B.4)} \qquad \begin{aligned} &\partial_t \rho_t = -\nabla \cdot \left(b_t \rho_t - D_t \nabla \rho_t - \rho_t D_t^2 \nabla \mu_t\right) \\ &\partial_t \mu_t = b_t \cdot \nabla \mu_t + D_t \nabla \log \rho_t \cdot \nabla \mu_t + 2\nabla \cdot (D_t \nabla \mu_t), \quad \mu_T(x) = 0, \end{aligned}$$

Since the equation for μ_t is homogeneous in μ_t and $\mu_T = 0$, we must have $\mu_t = 0$ for all $t \in [0,T)$, and the equation for ρ_t reduces to (FPE). □

Note that (SBTM1) is nontrivial because ρ_t is a functional of s_t. In particular, we can expand the integrand in the objective function of (SBTM1) and use integration by parts to rewrite it as

$$\begin{aligned} &\int_0^T \int_\Omega \left|s_t - \nabla \log \rho_t\right|^2 \rho_t dx dt \\ &= \int_0^T \int_\Omega \left(|s_t|^2 + 2\nabla \cdot s_t + |\nabla \log \rho_t|^2\right) \rho_t dx dt, \end{aligned}$$

However the last term cannot be neglected because it is not a constant in s_t, in contrast to SBDM.

B.2. **Proof of Proposition 2.**

Proposition 2 (SBTM in the Lagrangian frame)**.** *In the same setting as Proposition 9, let* $v_t(x) = b_t(x) - D_t(x)s_t(x)$ *and consider the optimization problem*

$$\text{(SBTM2)} \qquad \begin{aligned} &\min_{\{s_t : t\in[0,T)\}} \int_0^T \lambda(t) \int_\Omega \left|s_t(X_t(x)) - G_t(x)\right|^2 \rho_0(x) dx dt, \\ &\textit{subject to:} \quad \frac{d}{dt} X_t(x) = v_t(X_t(x)), \\ &\qquad\qquad \frac{d}{dt} G_t(x) + [\nabla v_t(X_t(x))]^\mathsf{T} G_t(x) = -\nabla\nabla\cdot v_t(X_t(x)). \end{aligned}$$

with initial conditions $X_0(x) = x$ *and* $G_0(x) = \nabla \log \rho_0(x)$*. Then, the minimizer* s_t^* *of* (SBTM2) *is unique and is identical to the minimizer of* (SBTM1)*. Moreover, the map* X_t^* *associated to this minimizer is a transport map from* ρ_0 *to* ρ_t^**, the solution of* (FPE)*, in the sense that*

$$x \sim \rho_0 \qquad \textit{implies that} \qquad X_t^*(x) \sim \rho_t^*, \qquad t \in [0,T]. \tag{11}$$

Proof. Let us first show that $G_t(x) = \nabla \log \rho_t(X_t(x))$ satisfies (SBTM2) if $\rho_t = X_t \sharp \rho_0$, i.e. ρ_t satisfies the transport equation (4). Since (4) implies that

$$\partial_t \log \rho_t(x) + v_t(x)\cdot \nabla \log \rho_t(x) = -\nabla \cdot v_t(x), \tag{B.5}$$

taking the gradient of this last equation gives

$$\partial_t \nabla \log \rho_t(x) + [\nabla v_t(x)]^\mathsf{T} \nabla \log \rho_t(x) + v_t(x)\cdot \nabla\nabla \log \rho_t(x) = -\nabla\nabla\cdot v_t(x) \tag{B.6}$$

Therefore $G_t(x) = \nabla \log \rho_t(X_t(x))$ solves

$$\begin{aligned} \frac{d}{dt} G_t(x) &= \partial_t \nabla \log \rho_t(X_t(x)) + \frac{d}{dt} X_t(x) \cdot \nabla\nabla \log \rho_t(X_t(x)) \\ &= \partial_t \nabla \log \rho_t(X_t(x)) + v_t(x)\cdot \nabla\nabla \log \rho_t(X_t(x)) \\ &= -\nabla\nabla\cdot v_t(X_t(x)) - [\nabla v_t(X_t(x))]^\mathsf{T} \nabla \log \rho_t(X_t(x)) \end{aligned} \tag{B.7}$$

and we recover the equation for $G_t(x)$ in (SBTM2). Hence, the objective in (SBTM2) can also be written as

$$\begin{aligned} &\int_0^T \int_\Omega \left|s_t(X_t(x)) - \nabla \log \rho_t(X_t(x))\right|^2 \rho_0(x) dx dt \\ &= \int_0^T \int_\Omega \left|s_t(x) - \nabla \log \rho_t(x)\right|^2 \rho_t(x) dx dt \end{aligned} \tag{B.8}$$

where the second equality follows from (A.5) if $\rho_t(x)$ satisfies (A.1). Therefore, (SBTM2) is equivalent to (SBTM1). □

In terms of a practical implementation, the objective in (SBTM2) can be evaluated by generating samples $\{x_i\}_{i=1}^n$ from ρ_0 and solving the equations for X_t and G_t using the initial conditions $X_0(x_i) = x_i$ and $G_0(x_i) = \nabla \log \rho_0(x_i)$. Note that evaluating this second initial condition only requires one to know ρ_0 up to a normalization factor. To evaluate the gradient of the objective, we can introduce equations adjoint

to those for X_t and G_t. They read, respectively

$$\begin{aligned}
\frac{d}{dt}\theta_t(x) + [\nabla v_t(X_t(x))]^\mathsf{T}\theta_t(x) &= \eta_t(x)\cdot\nabla\nabla v_t(X_t(x))G_t(x) \\
&\quad + \eta_t(x)\cdot\nabla\nabla\nabla v_t(X_t(x))G_t(x) \\
&\quad + 2\nabla s_t(X_t(x))(s_t(X_t(x)) - G_t(x)), \\
\theta_T(x) &= 0 \\
\frac{d}{dt}\eta_t(x) - \nabla v_t(X_t(x))\eta_t(x) &= 2(G_t(x) - s_t(X_t(x))), \\
\eta_T(x) &= 0.
\end{aligned} \tag{B.9}$$

In terms of these functions, the gradient of the objective is the gradient with respect to $s_t(x)$ (or the parameters in this function when it is modeled by a neural network) of the extended objective:

$$\begin{aligned}
L[s_t] = &\int_0^T \int_\Omega |s_t(X_t(x)) - G_t(x)|^2 \rho_0(x)dxdt \\
&+ \int_0^T \int_\Omega \theta_t(x)\cdot\left(\dot{X}_t(x) - v_t(X_t(x))\right)\rho_0(x)dxdt \\
&+ \int_0^T \int_\Omega \eta_t(x)\cdot\Big(\dot{G}_t(x) + [\nabla v_t(X_t(x))]^\mathsf{T} G_t(x) \\
&\qquad + \nabla\nabla\cdot v_t(X_t(x))\Big)\rho_0(x)dxdt,
\end{aligned} \tag{B.10}$$

where $v_t(x) = b_t(x) + D_t(x)s_t(x)$.

B.3. **Proof of Proposition 3.**

Proposition 3 (Sequential SBTM)**.** *In the same setting as Proposition 9, 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*

$$\min_{s_t} \int_\Omega \left(|s_t(X_t(x))|^2 + 2\nabla\cdot s_t(X_t(x))\right)\rho_0(x)dx. \tag{SBTM3}$$

Then the minimizer s_t^* of (SBTM3) is unique and is given by $s_t^* = -\nabla\log\rho_t$.

Proof. If $X_t\sharp\rho_0 = \rho_t$, then by definition we have the identity

$$\begin{aligned}
&\int_\Omega \left(|s_t(X_t(x))|^2 + 2\nabla\cdot s_t(X_t(x))\right)\rho_0(x)dx \\
&\qquad = \int_\Omega \left(|s_t(x)|^2 + 2\nabla\cdot s_t(x)\right)\rho_t(x)dx.
\end{aligned} \tag{B.11}$$

This means that the optimization problem in (SBTM3) is equivalent to

$$\min_{s_t} \int_\Omega \left(|s_t(x)|^2 + 2\nabla\cdot s_t(x)\right)\rho_t(x)dx. \tag{SBTM3}$$

The minimizer of this problem is unique and given by $s_t^*(x) = \nabla\log\rho_t(x)$. □

B.4. **Denoising Loss.** The following standard trick [55] can be used to avoid computing the divergence of $s_t(x)$:

Lemma 1. *Given $\xi = N(0, I)$, we have*

$$\begin{aligned} &\lim_{\alpha\downarrow 0} \alpha^{-1}\mathbb{E}\big(s_t(x+\alpha\xi)\cdot\xi\big) = \nabla\cdot s_t(x), \\ &\lim_{\alpha\downarrow 0} \alpha^{-1}\mathbb{E}\big(s_t(x+\alpha\sigma_t(x)\xi)\cdot\sigma_t(x)\xi\big) = \mathrm{tr}\,(D_t(x)\nabla s_t(x)) \end{aligned} \tag{B.12}$$

Proof. We have

$$\alpha^{-1}s_t(x+\alpha\xi)\cdot\xi = \alpha^{-1}s_t(x)\cdot\xi + (\nabla s_t(x)\sigma_t(x)\xi)\cdot\xi + o(\alpha) \tag{B.13}$$

The expectation of the first term on the right hand-side of this equation is zero; the expectation of the second gives the result in (B.12). Hence, taking the expectation of (B.13) and evaluating the result in the limit as $\alpha \downarrow 0$ gives the first equation in (B.12). The second equation in (B.12) can be proven similarly using $\sigma_t(x)\sigma_t(x)^{\mathsf{T}} = D_t(x)$. □

In practice, we observe that the smoothing effect of the noise in (B.14) allows the objective to probe regions nearby the samples, and hence improves exploration. We can improve the accuracy of the approximation with a "doubling trick" that uses two moves with noise of opposite sign to reduce the variance. This amounts to replacing the expectations in (B.12) with

$$\begin{aligned} &\tfrac{1}{2}\alpha^{-1}\mathbb{E}\left(s_t(x+\alpha\xi)\cdot\xi - \mathbb{E}\big(s_t(x-\alpha\xi)\cdot\xi\right), \\ &\tfrac{1}{2}\alpha^{-1}\mathbb{E}\left(s_t(x+\alpha\sigma_t(x)\xi)\cdot\sigma_t(x)\xi - s_t(x-\alpha\sigma_t(x)\xi)\cdot\sigma_t(x)\xi\right), \end{aligned} \tag{B.14}$$

whose limits as $\alpha \to 0$ are $\nabla\cdot s_t(x)$ and $\mathrm{tr}\,(D_t(x)\nabla s_t(x))$, respectively. In practice, we observe that this approach always helps.

Appendix C. Gaussian case

Here, we consider the case of an Ornstein-Uhlenbeck (OU) process where the score can be written analytically, thereby providing a benchmark for our approach. The example treated in Section 3.1.1 with details in Appendix D.1 is a special case of such an OU process with additional symmetry arising from permutations of the particles.

The SDE reads

$$dX_t = -\Gamma_t(X_t - b_t)dt + \sqrt{2}\sigma_t dW_t \tag{C.1}$$

where $X_t \in \mathbb{R}^d$, $\Gamma_t \in \mathbb{R}^{d\times d}$ is a time-dependent positive-definite tensor (not necessarily symmetric), $b_t \in \mathbb{R}^d$ is a time-dependent vector, and $\sigma_t \in \mathbb{R}^{d\times d}$ is a time-dependent tensor. The Fokker-Planck equation associated with (C.1) is

$$\partial_t\rho_t(x) = -\nabla\cdot((\Gamma_t x - b_t)\rho_t(x) - D_t\nabla\rho_t(x)) \tag{C.2}$$

where $D_t = \sigma_t\sigma_t^{\mathsf{T}}$. Assuming that the initial condition is Gaussian, $\rho_0 = \mathsf{N}(m_0, C_0)$ with $C_0 = C_0^{\mathsf{T}} \in \mathbb{R}^{d\times d}$ positive-definite, the solution is Gaussian at all times $t \geq 0$, $\rho_t = \mathsf{N}(m_t, C_t)$ with m_t and $C_t = C_t^{\mathsf{T}}$ solutions to

$$\begin{aligned} \dot{m}_t &= -\Gamma_t(m_t - b_t) \\ \dot{C}_t &= -\Gamma_t C_t - C_t\Gamma_t^{\mathsf{T}} + 2D_t \end{aligned} \tag{C.3}$$

This implies in particular that

$$s_t(x) = \nabla\log\rho_t(x) = -C_t^{-1}(x - m_t). \tag{C.4}$$

so that the probability flow equation for X_t and the equation for G_t written in (SBTM2) read

$$\begin{aligned}\dot{X}_t(x) &= (D_t C_t^{-1} - \Gamma_t)X_t(x) + \Gamma_t b_t - D_t C_t^{-1} m_t,\\ \dot{G}_t(x) &= (\Gamma_t^\mathsf{T} - C_t^{-1} D_t)G_t(x),\end{aligned} \tag{C.5}$$

with initial condition $X_0(x) = x$ and $G_0(x) = \nabla \log \rho_0(x) = -C_0^{-1}(x - m_0)$. It is easy to see that with $x \sim \rho_0 = \mathsf{N}(m_0, C_0)$ we have $X_t(x) \sim \rho_t = \mathsf{N}(m_t, C_t)$ since, from the first equation in (C.5), the mean and variance of X_t satisfy (C.3). Similarly, when $x \sim \rho_0 = \mathsf{N}(m_0, C_0)$, $G_0(x) \sim N(0, C_0^{-1})$, so that $G_t(x) \sim \mathsf{N}(0, C_t^{-1})$ because the second equation in (C.5) is linear and hence preserves Gaussianity. Moreover, $\mathbb{E}_0 G_t(x) = 0$ and $B_t = B_t^\mathsf{T} = \mathbb{E}_0[G_t(x)G_t^\mathsf{T}(x)]$ satisfies

$$\frac{d}{dt}B_t = (\Gamma_t^\mathsf{T} - C_t^{-1}D_t)B_t + B_t(\Gamma_t - D_t C_t^{-1}) \tag{C.6}$$

The solution to this equation is $B_t = C_t^{-1}$ since substituting this ansatz into (C.6) gives the equation for C_t^{-1} that we can deduce from (C.3)

$$\frac{d}{dt}C_t^{-1} = C_t^{-1}\dot{C}_t C_t^{-1} = -C_t^{-1}\Gamma_t - \Gamma_t^\mathsf{T} C_t^{-1} + 2C_t^{-1} D_t C_t^{-1}. \tag{C.7}$$

Note that if $\Gamma_t = \Gamma$, $b_t = b$, and $D_t = D$ are all time-independent, then $\lim_{t\to\infty} \rho_t = N(m_\infty, C_\infty)$ with $m_\infty = b$ and C_∞ the solution to the Lyapunov matrix equation

$$\Gamma C_\infty + C_\infty \Gamma^\mathsf{T} = 2D. \tag{C.8}$$

This means that at long times the coefficients at the right-hand sides of (C.5) also settle on constant values. However, X_t and G_t do not necessarily stop evolving; one situation where they too tend to fix values is when the OU process is in detailed balance, i.e. when $\Gamma = DA$ for some $A = A^\mathsf{T} \in \mathbb{R}^{d\times d}$ positive-definite. In that case, the solution to (C.8) is $C_\infty = A^{-1}$ and it is easy to see that at long times the right hand sides of (C.5) tend to zero.

Remark C.1. *This last conclusion is actually more generic than for a simple OU process. For any SDE in detailed balance, i.e. that can be written as*

$$dX_t = -D(X_t)\nabla U(X_t)dt + \nabla \cdot D(X_t)dt + \sqrt{2}\sigma_t(X_t)dW_t \tag{C.9}$$

where $U : \mathbb{R}^d \to \mathbb{R}_{>0}$ is a C^2-potential such that $Z = \int_{\mathbb{R}^d} e^{-U(x)}dx < \infty$, we have that $\lim_{t\to\infty} \rho_t(x) = Z^{-1}e^{-U(x)}$, and the corresponding flows X_t and G_t eventually stop as $t \to \infty$. In this case, ρ_t follows gradient descent in W_2 over the energy

$$E[\rho] = \int_{\mathbb{R}^d} (U(x) + \log \rho(x))\rho(x)dx \tag{C.10}$$

The unique PDF minimizing this energy is $Z^{-1}e^{-U(x)}$, and as $t \to \infty$ X_t converges towards a transport map between the initial ρ_0 and $Z^{-1}e^{-U(x)}$.

Appendix D. Experimental details and additional examples

D.1. **Harmonically interacting particles in a harmonic trap.**

D.1.1. *Network architecture.* Both the single-particle energy $U_{\theta_t,1} : \mathbb{R}^d \to \mathbb{R}$ and two-particle interaction energy $U_{\theta_t,2} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ are parameterized as single hidden-layer neural networks with the `swish` activation function [39] and `n_hidden` $= 100$ hidden neurons. The hidden layer biases are initialized to zero while the hidden layer weights are initialized from a truncated normal distribution with variance $1/$`fan_in`, following the guidelines recommended in [22].

D.1.2. *Optimization.* The Adam [24] optimizer is used with an initial learning rate of $\eta = 10^{-4}$ and otherwise default settings. At time $t = 0$, the analytical relative loss

$$L[s_0] = \frac{\int |s_0(x) - \nabla \log \rho_0(x)|^2 \rho_0(x) dx}{\int |\nabla \log \rho_0(x)|^2 \rho_0(x) dx} \tag{D.1}$$

is minimized to a value less than 10^{-4} using knowledge of the initial condition $\rho_0 = \mathsf{N}\left(\beta_0, \sigma_0^2 I\right)$ with $\sigma_0 = 0.25$. In (D.1), the expectation with respect to ρ_0 is approximated by an initial set of samples $x_j = \left(x_j^{(1)}, x_j^{(2)}, \dots, x_j^{(N)}\right)^{\mathsf{T}}$ with $j = 1, \dots, n$ drawn from ρ_0. In the experiments, we set $n = 100$. We set the physical timestep $\Delta t = 10^{-3}$ and take `n_opt_steps` $= 25$ steps of Adam until the norm of the gradient is below `gtol` $= 0.1$.

D.1.3. *Analytical moments.* First define the mean, second moment, and covariance according to

$$\begin{aligned} m_t^{(i)} &= \mathbb{E}\big[X_t^{(i)}\big], \\ M_t^{(ij)} &= \mathbb{E}\big[X_t^{(i)} \big(X_t^{(j)}\big)^{\mathsf{T}}\big], \\ C_t^{(ij)} &= M^{(ij)} - m^{(i)} \big(m^{(j)}\big)^{\mathsf{T}}. \end{aligned}$$

It is straightforward to show that the mean and covariance obey the dynamics

$$\dot{m}_t^{(i)} = -(m_t^{(i)} - \beta_t) + \frac{\alpha}{N} \sum_{k=1}^{N} \left(m_t^{(i)} - m_t^{(k)}\right), \tag{D.2}$$

$$\dot{C}_t^{(ij)} = -2(1-\alpha) C_t^{(ij)} + 2DI\delta_{ij} - \frac{\alpha}{N} \sum_{k=1}^{N} \left(C_t^{(kj)} + C_t^{(ik)}\right) \tag{D.3}$$

Because the particles are indistinguishable so long as they are initialized from a distribution that is symmetric with respect to permutations of their labeling, the moments will satisfy the ansatz

$$m_t^{(i)} = \bar{m}(t), \quad i = 1, \dots, N \tag{D.4}$$

$$C_t^{(ij)} = C_d(t)\delta_{ij} + C_o(t)(1 - \delta_{ij}), \quad i, j = 1, \dots, N. \tag{D.5}$$

The dynamics for the vector $\bar{m} : \mathbb{R}_{\geq 0} \to \mathbb{R}^{\bar{d}}$, as well as the matrices $C_d : \mathbb{R}_{\geq 0} \to \mathbb{R}^{\bar{d} \times \bar{d}}$ and $C_o : \mathbb{R}_{\geq 0} \to \mathbb{R}^{\bar{d} \times \bar{d}}$ can then be obtained from (D.2) and (D.3) as

$$\begin{aligned} \dot{\bar{m}} &= \beta_t - \bar{m}, \\ \dot{C}_d &= 2(\alpha - 1) C_d - 2\frac{\alpha}{N} \left(C_d + (n-1) C_o\right) + 2DI, \\ \dot{C}_o &= 2(\alpha - 1) C_o - 2\frac{\alpha}{N} \left(C_d + (n-1) C_o\right). \end{aligned}$$

For a given $\beta : \mathbb{R} \to \mathbb{R}^{\bar{d}}$, these equations can be solved analytically in Mathematica as a function of time, giving the mean $m_t = \bar{m}(t) \otimes 1_N \in \mathbb{R}^{N\bar{d}}$ and covariance $C_t = (C_d(t) - C_o(t)) \otimes I_{N\times N} + C_o(t) \otimes \left(1_N 1_N^\mathsf{T}\right) \in \mathbb{R}^{N\bar{d}\times N\bar{d}}$. Because the solution is Gaussian for all t, we then obtain the analytical solution to the Fokker-Planck equation $\rho_t^* = \mathsf{N}(m_t, C_t)$ and the corresponding analytical score $-\nabla \log \rho_t^*(x) = C_t^{-1}(x - m_t)$.

D.1.4. *Potential structure.* Here, we show that the potential for this example lies in the class of potentials described by (19). From Equation D.5, we have a characterization of the structure of the covariance matrix C_t for the analytical potential $U_t(x) = \frac{1}{2}(x-m_t)^\mathsf{T} C_t^{-1}(x-m_t)$. In particular, C_t is block circulant, and hence is block diagonalized by the roots of unity (the block discrete Fourier transform). That is, we may take a "block eigenvector" of the form $\omega_k = \left(I_{\bar{d}\times\bar{d}}\rho^k, I_{\bar{d}\times\bar{d}}\rho^{2k}, \ldots, I_{\bar{d}\times\bar{d}}\rho^{(N-1)k}\right)^\mathsf{T}$ with $\rho = \exp(-2\pi i/N)$ for $k = 0, \ldots N-1$. By direct calculation, this block diagonalization leads to two distinct block eigenmatrices,

$$C_t = V \begin{pmatrix} C_d(t) + (N-1)C_o(t) & 0 & 0 & \ldots & 0 \\ 0 & C_d(t) - C_o(t) & 0 & \ldots & 0 \\ & & \ddots & & \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & C_d(t) - C_o(t) \end{pmatrix} V^{-1}$$

where $V \in \mathbb{R}^{N\bar{d}\times N\bar{d}}$ denotes the matrix with block columns ω_k. The inverse matrix C_t^{-1} then must similarly have only two distinct block eigenmatrices given by $(C_d(t) + (N-1)C_o(t))^{-1}$ and $(C_d(t) - C_o(t))^{-1}$. By inversion of the block Fourier transform, we then find that

$$\left(C_t^{-1}\right)^{(ij)} = \bar{C}_d \delta_{ij} + \bar{C}_o(1 - \delta_{ij})$$

for some matrices $\bar{C}_d, \bar{C}_o$. Hence, by direct calculation

$$\begin{aligned}
(x - m_t)^\mathsf{T} C_t^{-1} (x - m_t) &= \sum_{i,j}^{N} \left(x^{(i)} - m_t^{(i)}\right)^\mathsf{T} \left(C_t^{-1}\right)^{(ij)} \left(x^{(j)} - m_t^{(j)}\right) \\
&= \sum_{i,j}^{N} \left(x^{(i)} - \bar{m}(t)\right)^\mathsf{T} \left(\bar{C}_d \delta_{ij} + \bar{C}_o(1 - \delta_{ij})\right) \left(x^{(j)} - \bar{m}(t)\right) \\
&= \sum_{i}^{N} \left(x^{(i)} - \bar{m}(t)\right)^\mathsf{T} \bar{C}_d \left(x^{(i)} - \bar{m}(t)\right)^\mathsf{T} \\
&\quad + \sum_{i\neq j}^{N} \left(x^{(i)} - \bar{m}(t)\right)^\mathsf{T} \bar{C}_o \left(x^{(j)} - \bar{m}(t)\right)
\end{aligned}$$

Above, we may identify the first term in the last line as $\sum_{i=1}^{N} U_1(x^{(i)})$ and the second term in the last line as $\frac{1}{N}\sum_{i\neq j}^{N} U_2(x^{(i)}, x^{(j)})$. Moreover, $U_2(\cdot,\cdot)$ is symmetric with respect to its arguments.

D.1.5. *Analytical Entropy.* For this example, the entropy can be computed analytically and compared directly to the learned numerical estimate. By definition,

$$\begin{aligned} S_t &= -\int_{\mathbb{R}^{Nd}} \log \rho_t(x)\rho_t(x)dx, \\ &= -\int_{\mathbb{R}^{Nd}} \left(-\frac{Nd}{2}\log(2\pi) - \frac{1}{2}\log\det C_t - \frac{1}{2}(x-m_t)^{\mathsf{T}} C_t^{-1}(x-m_t) \right) \rho_t(x)dx, \\ &= \frac{Nd}{2}\left(\log(2\pi)+1\right) + \frac{1}{2}\log\det C_t. \end{aligned}$$

D.1.6. *Additional figures.* Images of the learned velocity field and potential in comparison to the corresponding analytical solutions can be found in Figures D.1 and D.2, respectively. Further detail can be found in the corresponding captions. We stress that the two-dimensional images represent single-particle slices of the high-dimensional functions. A movie of the particle motion can be found at the link `https://drive.google.com/file/d/1G6-c0NNFtXW3UxFM0RwqPSDsVD6mDGkq/view?usp=sharing`. The movie highlights the similarity between the learned and SDE trajectories, while the noise free system collapses to a point.

D.2. **Soft spheres in an anharmonic trap.**

D.2.1. *Network architecture.* Both potential terms $U_{\theta_t,1}$ and $U_{\theta_t,2}$ are modeled as four hidden-layer deep fully connected networks with `n_hidden` $= 32$ neurons in each layer. The initialization is identical to Appendix D.2.

D.2.2. *Optimization and initialization.* The Adam optimizer is used with an initial learning rate of $\eta = 5 \times 10^{-3}$ and otherwise default settings. At time $t = 0$, the loss (D.1) is minimized to a value less than 10^{-4} over n samples $x_{0,j} \sim \mathsf{N}(\beta_0, \sigma_0^2 I)$ with $\sigma_0 = 0.5$ and $n = 1000$, similar to Appendix D.2. After this initial optimization, 100 steps of the SDE (23) are taken in artificial time τ with fixed physical $t = 0$ to ensure that no spheres are overlapping at initialization. Past this initial stage, the denoising loss is used with a noise scale $\sigma = 0.025$. The loss is minimized by taking `n_opt_steps` $= 25$ steps of Adam until the norm of the gradient is below `gtol` $= 0.5$. The physical timestep is set to $\Delta t = 10^{-3}$.

D.2.3. *Additional figures.* A depiction of the one-particle potential, estimated as the negative logarithm of the one-particle PDF obtained via kernel density estimation, can be found in Figure D.3 (for further details, see the caption). Movies of the particle motion with respect to the moving trap can be found at `https://drive.google.com/file/d/111HPnZD37pjgO2tDgXQRbabvlELXwTC3/view?usp=sharing` and `https://drive.google.com/file/d/1j6T7vJVulF46aN_ByWxXOxtTULv17Emv/view?usp=sharing`, while movies in a fixed reference frame can be found at `https://drive.google.com/file/d/18PWSW1YOfCsJt5v7szyCf4IDXzJtggIt/view?usp=sharing` and `https://drive.google.com/file/d/1SbLtFaAB-tAteUfJWwlTUcdoSapYPHsY/view?usp=sharing`. The movies highlight configurational re-arrangements and a "rolling motion" that preserves the statistics of the SDE not seen in the noise free system.

D.3. **Active swimmer.**

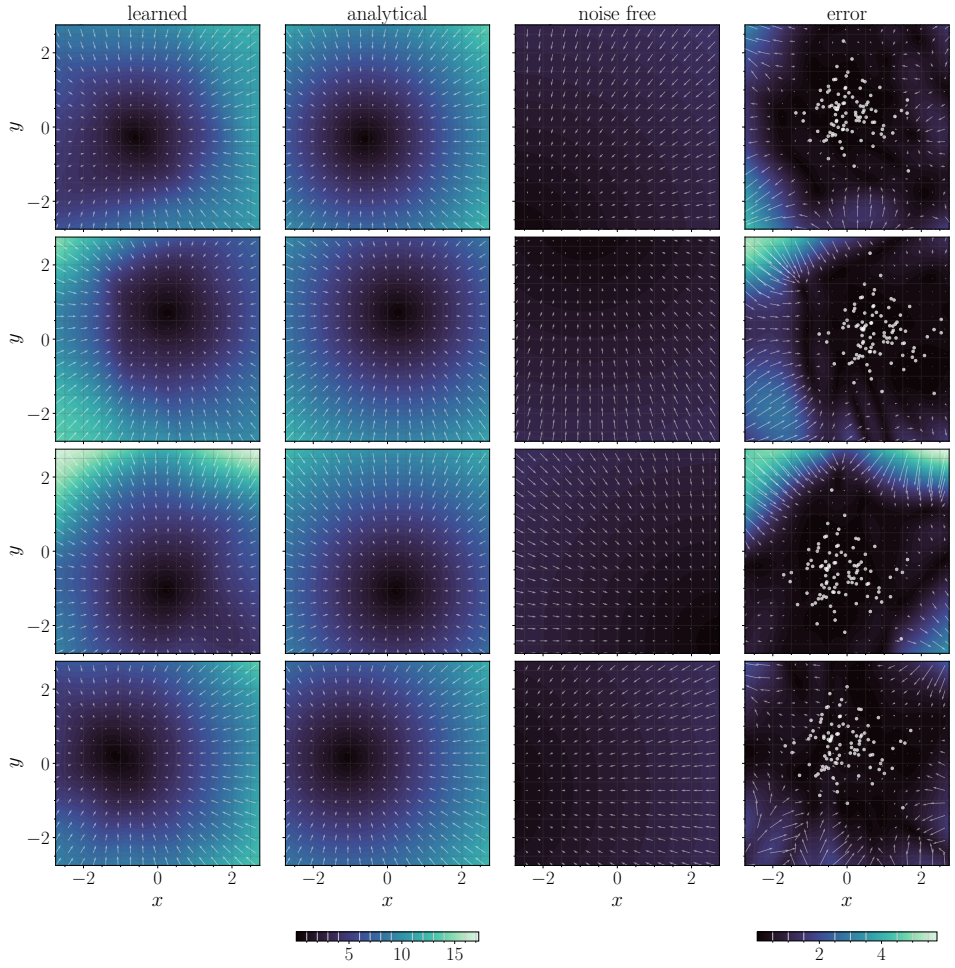

FIGURE D.1. *A system of $N = 50$ harmonically interacting particles in a harmonic trap: slices of the high-dimensional velocity field.* Cross sections of the velocity field for $N = 50$ harmonically interacting particles in a moving harmonic trap. Columns depict the learned, analytical, noise-free, and error between the learned and analytical velocity fields, respectively. Rows indicate different time points, corresponding to $t = 1.25, 2.5, 3.75$, and 5.0, respectively. Each velocity field is plotted as a function of a single particle's coordinate (denoted as x and y); all other particle coordinates are fixed to be at the location of a sample. Color depicts the magnitude of the velocity field while arrows indicate the direction. Learned, analytical, and noise-free share a colorbar for direct comparison; the error occurs on a different scale and is plotted with its own colorbar. White circles in the error plot indicate samples projected onto the xy plane; locations of low error correlate well with the presence of samples.

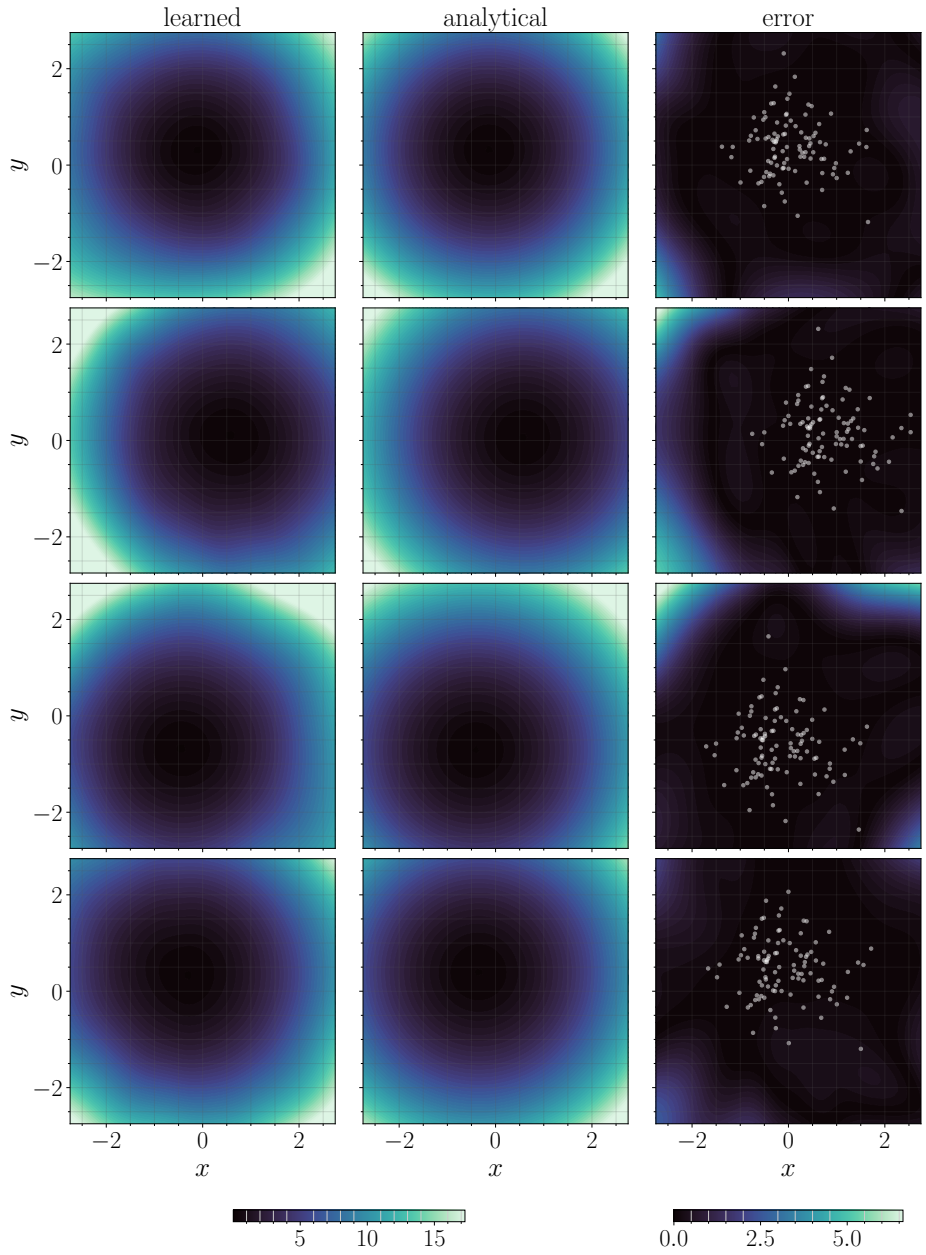

FIGURE D.2. *A system of $N = 50$ harmonically interacting particles in a harmonic trap: slices of the high-dimensional potential.* Cross sections of the potential field $U_{\theta_t}(x)$*computedvia* (19).. Columns depict the learned, analytical, and error between the learned and analytical, respectively. Rows indicate distinct time points, corresponding to $t = 1.25, 2.5, 3.75$, and 5.0, respectively. As in Figure D.1, each potential field is plotted as a function of a single particle's coordinate (denoted as x and y) with other particle coordinates fixed on a sample. All potentials are normalized via an overall shift so that the minimum value is zero. White circles in the error plot indicate samples from the learned system projected onto the xy plane.

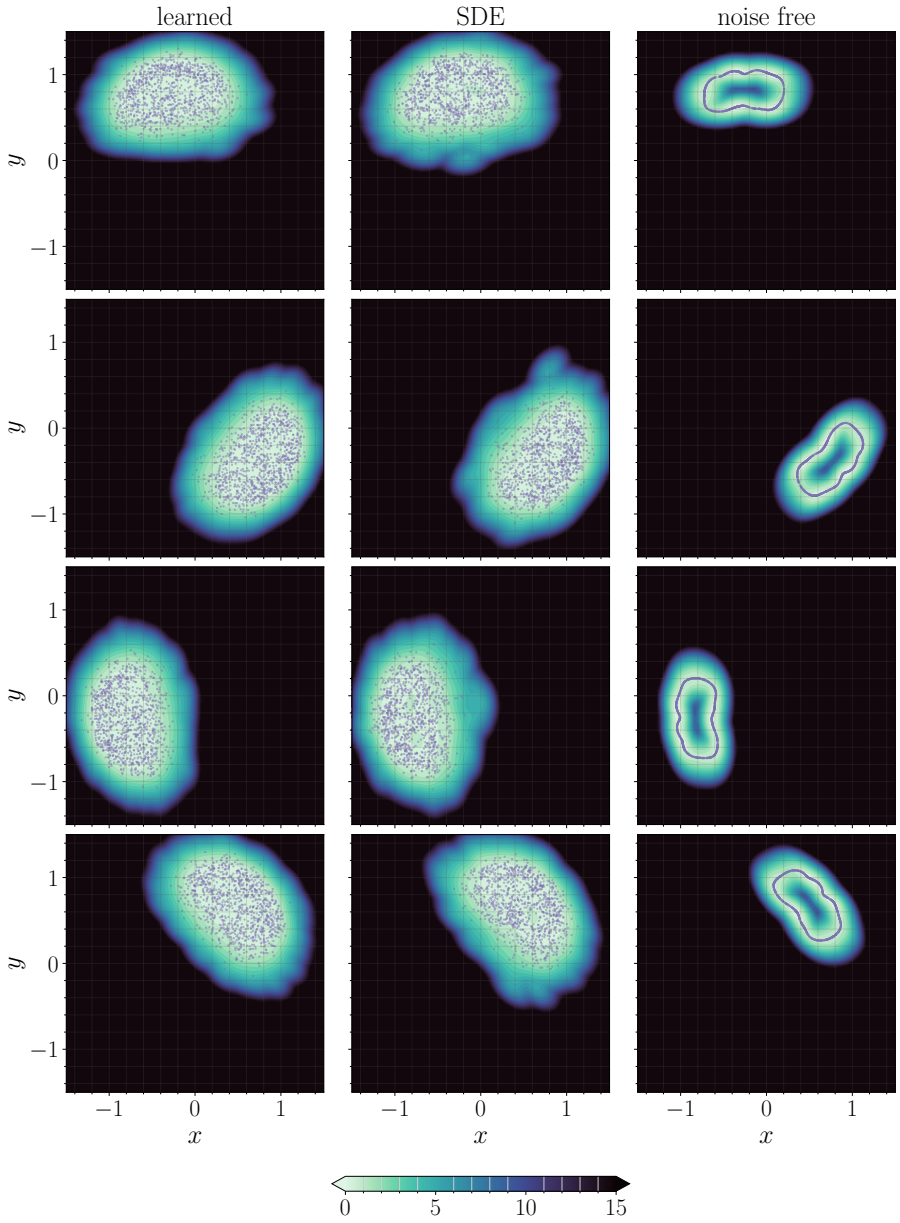

FIGURE D.3. *A system of $N = 5$ soft-sphere particles in an anharmonic trap: one-particle potential.* Cross sections of the one-particle potential field $U(x) = -\log \rho_{\mathrm{KDE}}(x)$ where ρ_{KDE} denotes a kernel density estimate of the one-particle density obtained by pooling all particle samples, shown relative to the moving mean. Columns depict the learned, SDE, and noise free systems, respectively. Purple dots indicate samples from the corresponding system. Rows indicate distinct time points, corresponding to $t = 1.25, 2.5, 3.75$, and 4.95, respectively. All potentials are normalized via an overall shift so that the minimum value is zero, and are clipped to a maximum value of 15. The learned and SDE potentials match well, while the noise free KDE becomes too peaked and develops a spurious maximum that causes the particles to align in a ring.

D.3.1. *Optimization and initialization.* The network initialization is identical to the previous two experiments. The physical timestep is set to $\Delta t = 10^{-3}$. The Adam optimizer is used with an initial learning rate of $\eta = 10^{-4}$. At time $t = 0$ the loss (D.1) is minimized to a tolerance of 10^{-4} over $n = 5000$ samples drawn from an initial distribution $\mathsf{N}(0, \sigma_0^2 I)$ with $\sigma_0 = 1$. The denoising loss is used with a noise scale $\sigma = 0.05$, using `n_opt_steps` $= 25$ steps of Adam until the norm of the gradient is below `gtol` $= 0.5$.

D.3.2. *Figures.* Depictions of the particle trajectories, the PDF obtained via kernel density estimation, and the noise-free velocity fields can be seen in Figures D.4- D.6, respectively. A movie of the particle motion can be seen at `https://drive.google.com/file/d/1YqMEF7HO1z47CRwC8JJUTD1ehIQ_fbjj/view?usp=sharing`. The movie highlights convergence of the learned solution to a nonzero steady-state probability current matching the SDE, while the noise free system becomes increasingly concentrated.

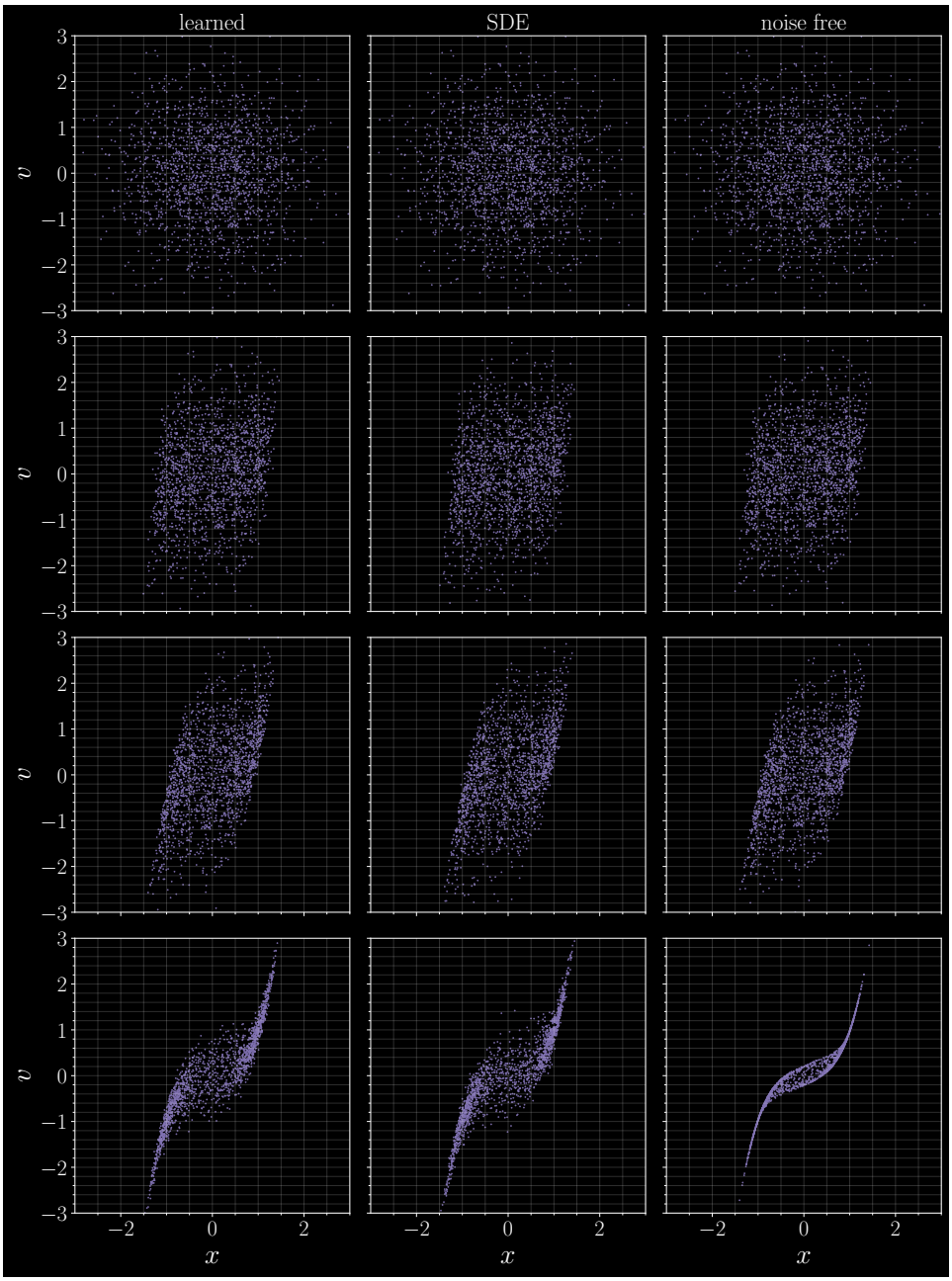

FIGURE D.4. *An active swimmer: samples.* Samples in the xv plane for the active swimmer example, with columns denoting solution type and rows indicating snapshots in time ($t = 0, 0.25, 0.5, 3.0$, respectively). The presence of the particle velocity causes the learned and SDE systems to develop bimodality despite corresponding to an anharmonic trap centered at the origin. The noise free system collapses with time, and does not correctly capture the variance.

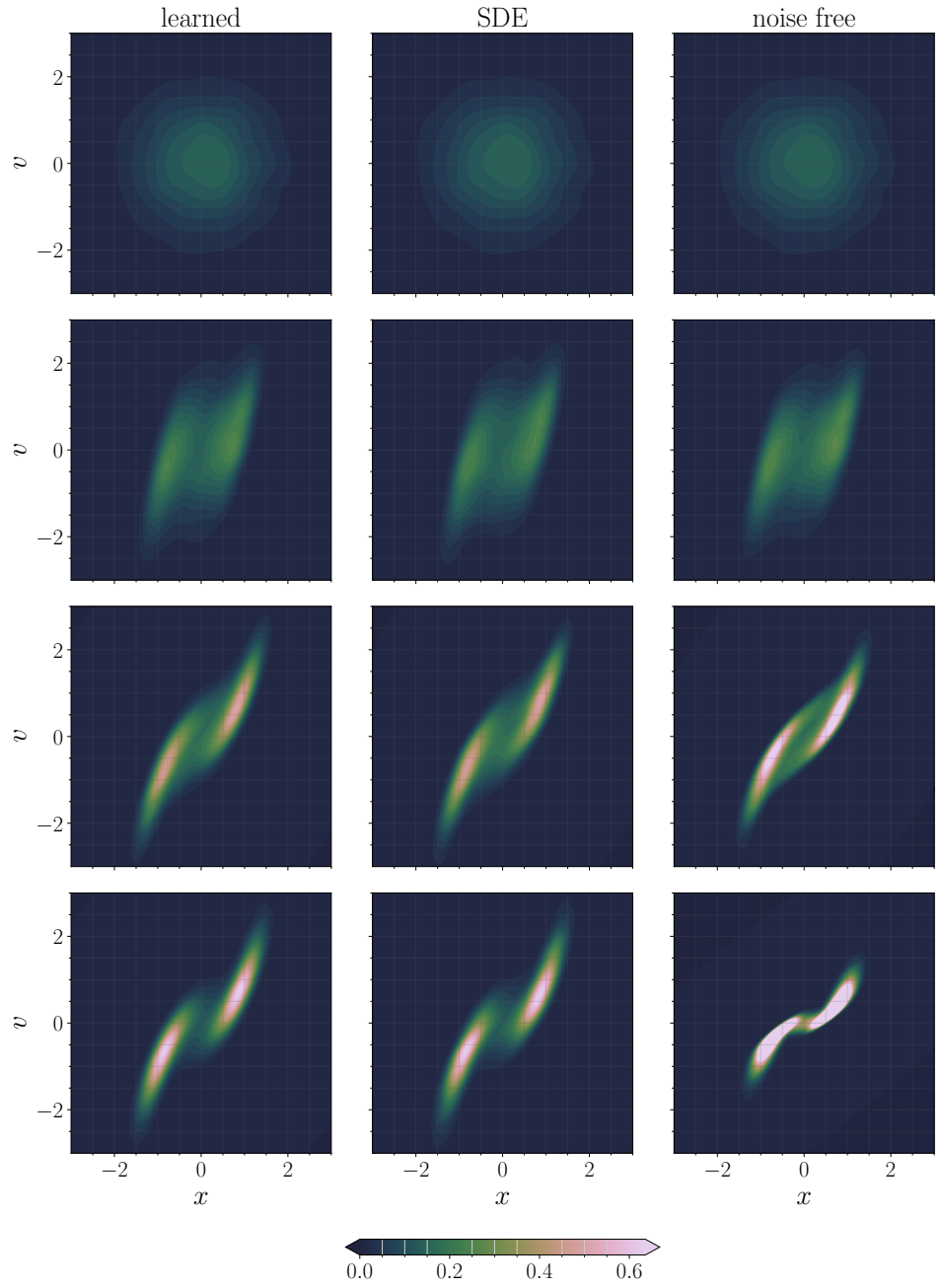

FIGURE D.5. *An active swimmer: density.* PDFs computed via kernel density estimation in the xv plane. Columns denote solution type and rows denote snapshots in time ($t = 0, 0.5, 1.5, 6.0$, respectively). Similar to the samples presented in Figure D.4, the KDE reveals bimodality in the probability density due to the presence of the particle velocity field. The noise free system becomes too concentrated and does not accurately capture the shape of the SDE and learned solutions, while the SDE and learned solutions are nearly identical.

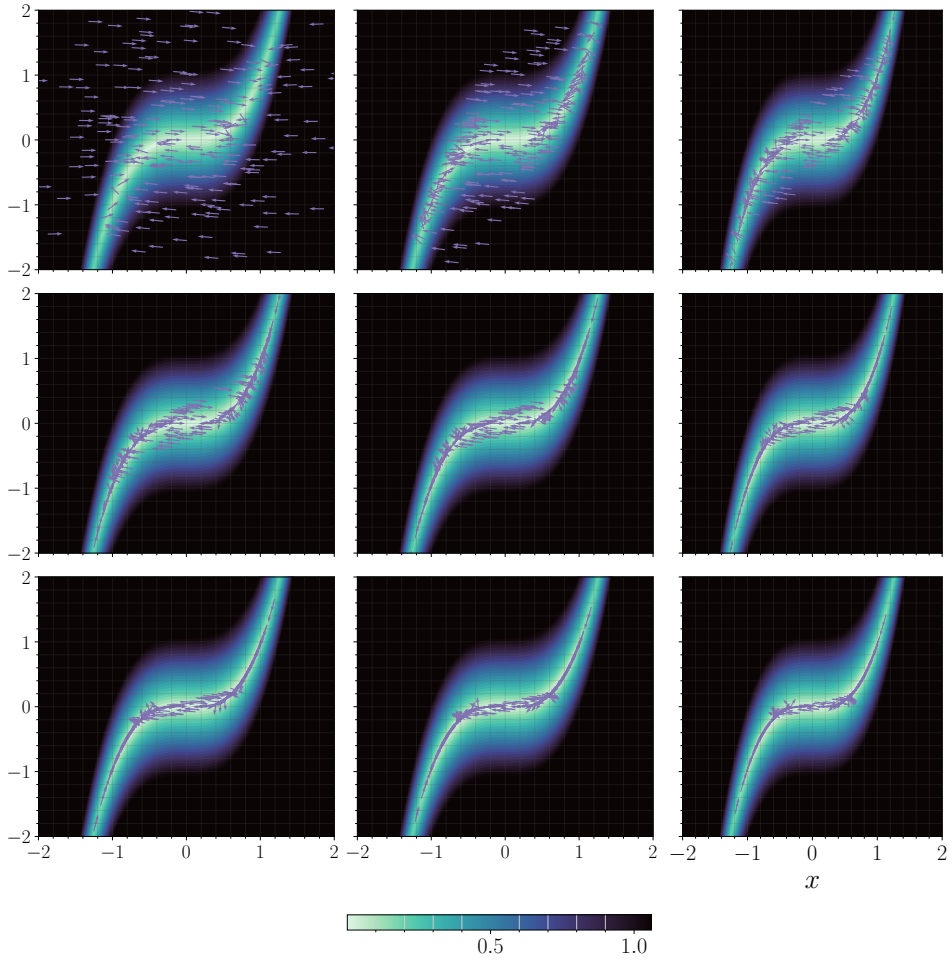

FIGURE D.6. *An active swimmer: noise free velocity.* Noise free velocity field. As in Figure 5, color indicates the magnitude of the velocity field while arrows indicate the direction, and time corresponds to progressing in the grid along columns from the top-left to the bottom-right image ($t = k \times .75$ with k the image number, zero-indexed). The velocity field in the noise-free case incorrectly pushes the swimmers to lie along a thin band.