A SOME BASIC FORMULAS

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

A.1 PROBABILITY DENSITY AND PROBABILITY CURRENT

We begin with a lemma.

Lemma A.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 (4). 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 A.1 states that an evaluation of the PDF ρ_t at a given point x may be obtained by evolving the probability flow equation (4) 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 the follwing equation for the current:

$$v_t(x)\rho_t(x) = v_t(x)\rho_0(X_{t,0}(x)) \exp\left(-\int_0^t \nabla \cdot v_\tau(X_{\tau,t}(x))d\tau\right). \tag{A.6}$$

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 (4). Differentiating $\rho_t(X_{t',t}(x))$ with respect to t and using (4) and (A.1) we deduce

$$\begin{aligned}
\frac{d}{dt}\rho_t(X_{t',t}(x)) &= \partial_t \rho_t(X_{t',t}(x)) + \frac{d}{dt}X_{t',t}(x) \cdot \nabla \rho_t(X_{t',t}(x)) \\
&= \partial_t \rho_t(X_{t',t}(x)) + v_t(X_{t',t}(x)) \cdot \nabla \rho_t(X_{t',t}(x)) \\
&= -\nabla \cdot v_t(X_{t',t}(x))\, \rho_t(X_{t',t}(x))
\end{aligned} \tag{A.7}$$

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

$$\rho_t(X_{t',t}(x)) = \rho_{t'}(x) \exp\left(-\int_{t'}^t \nabla \cdot v_\tau(X_{t',\tau}(x))d\tau\right) \tag{A.8}$$

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 A.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 (FPE) and (A.2) holds for the solution to that equation.

A.2 CALCULATION OF THE DIFFERENTIAL ENTROPY

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

Lemma A.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 Fokker Planck equation (FPE) (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 differential entropy $S_t = -\int_\Omega \log \rho_t(x)\, \rho_t(x) dx$ can expressed as*

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

or

$$S_t = S_0 - \int_0^t \int_\Omega s_\tau(X_{0,\tau}(x)) \cdot v_\tau(X_{0,\tau}(x)) \rho_0(x) dx d\tau \tag{A.10}$$

Proof. We first derive (A.9). 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.10), notice that from (A.1),

$$\begin{aligned} \frac{d}{dt} S_t &= \int_\Omega \log \rho_t(x) \nabla \cdot (v_t(x) \rho_t(x))\, dx, \\ &= -\int_\Omega \nabla \log \rho_t(x) \cdot v_t(x) \rho_t(x) dx, \\ &= -\int_\Omega s_t(x) \cdot v_t(x) \rho_t(x) dx \end{aligned} \tag{A.11}$$

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.10). □

A.3 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

$$dX_\tau = s_t(X_\tau) d\tau + dW_\tau. \tag{A.12}$$

In (A.12), τ is an artificial time used for sampling that is distinct from the physical time in (1). For $s_t = \nabla \log \rho_t$, the equilibrium distribution of (A.12) 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, (A.12) should be used near samples for a fixed amount of time to avoid the introduction of additional errors.

B FURTHER DETAILS ON SCORE-BASED TRANSPORT MODELING

Like SBDM, SBTM is based on the observation that we can learn the score $\nabla \log \rho_t$ of a target distribution ρ_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 a pre-defined function that weights the data over the time interval (e.g. $\lambda(t) = 1$ or $\lambda(t) = e^{-t}$). As stated in the main text, 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 (FPE). 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) directly in App. B.1 and then with the probability flow equation (4) in App. B.2. While the second approach is the one that is amenable to a practical implementation, the first is conceptually simpler.

B.1 SBTM IN THE EULERIAN FRAME

The Eulerian equivalent of Proposition 1 can be stated as:

Proposition B.1 (SBTM in the Eulerian frame)**.** *Assume that the conditions listed in Sec. 1.2 hold. Fix $T \in (0, \infty]$, let $\lambda : [0, T) \to \mathbb{R}_{>0}$ be a positive function, and consider the optimization problem*

$$\min_{\{s_t : t\in[0,T)\}} \int_0^T \lambda(t) \int_\Omega \left| s_t(x) - \nabla \log \rho_t(x) \right|^2 \rho_t(x) dx dt \tag{SBTM2}$$

$$\textit{subject to:} \quad \partial_t \rho_t(x) = -\nabla \cdot \left(v_t(x) \rho_t(x) \right), \quad x \in \Omega$$

with $v_t(x) = b_t(x) - D_t(x)s_t(x)$. Then the minimizer of (SBTM2) is unique and given by $s_t^*(x) = \nabla \log \rho_t^*(x)$ where $\rho_t^* : \Omega \to \mathbb{R}_{>0}$ solves

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

In words, this proposition states that solving the constrained optimization problem (SBTM2) is equivalent to solving the Fokker-Planck equation (FPE).

Remark B.2. A similar result holds if we replace (SBTM2) by the diffusion-weighted loss

$$\min_{\{s_t : t\in[0,T]\}} \int_0^T \int_\Omega |s_t(x) - \nabla \log \rho_t(x)|^2_{D_t(x)} \rho_t(x) dx dt, \tag{SBTM2'}$$

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 (TE) and the probability flow ODE (4), agreement in the range is all that matters.

Proof. The constrained minimization problem (SBTM2') can be handled by considering the extended objective

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

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

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

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

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

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). □

Remark B.3. We would like to stress that (SBTM2) is nontrivial because ρ_t is a functional of s_t. In particular, we can expand the integrand in the objective function of (SBTM2) and use integration by parts to rewrite it as

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

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

B.2 SBTM IN THE LAGRANGIAN FRAME

As stated, Proposition B.1 is not practical, because it is phrased in terms of the density ρ_t. The following result demonstrates that the transport map identity (5) can be used to re-express Proposition B.1 entirely in terms of known quantities.

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*

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

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^*:

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

Remark B.4. Following Remark B.2, a similar result holds if we replace (SBTM) by the diffusion-weighted loss

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

subject to the same constraints.

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

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

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

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

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

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

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

In terms of a practical implementation, the objective in (SBTM) 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:

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

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

B.3 Bounding the KL divergence

Let us restate Proposition 2 for convenience:

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 $D_m = \sup_{t\ge0}\sup_{x\in\Omega}\|D_t(x)\| < \infty$ with $\|D_t(x)\| = \sup_{z:\|z\|=1} z^\mathsf{T}D_t(x)z$.

Proof. By assumption, ρ_t solves (TE) and ρ_t^* solves (FPE). Denote by $v_t(x) = b_t(x) - D_t(x)s_t(x)$ and $v_t^*(x) = b_t(x) - D_t(x)s_t^*(x)$ with $s_t^*(x) = \nabla\log\rho_t^*(x)$. Then, we have

$$\begin{aligned}
\frac{d}{dt}D_{\mathrm{KL}}(\rho_t|\rho_t^*) &= \frac{d}{dt}\int_\Omega \log\left(\frac{\rho_t(x)}{\rho_t^*(x)}\right)\rho_t(x)dx, \\
&= -\int_\Omega \frac{\rho_t(x)}{\rho_t^*(x)}\partial_t\rho_t^*(x)dx + \int_\Omega \log\left(\frac{\rho_t(x)}{\rho_t^*(x)}\right)\partial_t\rho_t(x)dx, \\
&= -\int_\Omega v_t^*(x)\cdot\nabla\left(\frac{\rho_t(x)}{\rho_t^*(x)}\right)\rho_t^*(x)dx + \int_\Omega v_t(x)\cdot\nabla\log\left(\frac{\rho_t(x)}{\rho_t^*(x)}\right)\rho_t(x)dx, \\
&= -\int_\Omega (v_t^*(x) - v_t(x))\cdot(\nabla\log\rho_t(x) - \nabla\log\rho_t^*(x))\,\rho_t(x)dx, \\
&= \int_\Omega (s_t^*(x) - s_t(x))\cdot D_t(x)\,(\nabla\log\rho_t(x) - s_t^*(x))\,\rho_t(x)dx.
\end{aligned}$$

Above, we used integration by parts to get the third equality. Using the identity (A.5), this can be expressed in terms of the transport map X_t as

$$\begin{aligned}
&\frac{d}{dt} D_{\mathrm{KL}}(\rho_t|\rho_t^*) \\
&= \int_\Omega \left(s_t^*(X_t(x)) - s_t(X_t(x))\right) \cdot D_t(X_t(x)) \left(\nabla \log \rho_t(X_t(x)) - s_t^*(X_t(x))\right) \rho_0(x) dx \\
&= \int_\Omega \left(s_t^*(X_t(x)) - s_t(X_t(x))\right) \cdot D_t(X_t(x)) \left(G_t(x) - s_t^*(X_t(x))\right) \rho_0(x) dx
\end{aligned}$$

where we used the definition of $G_t(x)$ to get the second equality. Since (dropping the argument for simplicity of notation and denoting $|a|^2_{D_t} = a \cdot D_t a$)

$$\begin{aligned}
|G_t - s_t|^2_{D_t} &= |G_t - s_t^* + s_t^* - s_t|^2_{D_t} \\
&= |G_t - s_t^*|^2_{D_t} + |s_t^* - s_t|^2_{D_t} + 2(G_t - s_t^*) \cdot D_t(s_t^* - s_t) \\
&\geq 2(G_t - s_t^*) \cdot D_t(s_t^* - s_t)
\end{aligned} \tag{B.11}$$

we deduce that

$$\frac{d}{dt} D_{\mathrm{KL}}(\rho_t|\rho_t^*) \leq \frac{1}{2} \int_\Omega |s_t(X_t(x)) - G_t(x)|^2_{D_t(x)} \rho_0(x) dx \tag{B.12}$$

Integrating this equation on $t \in [0, T]$ using $D_{\mathrm{KL}}(\rho_0|\rho_0^*) = 0$ implies (9) by definition of D_m. □

Remark B.5. Note that (B.12) gives a better bound than (9), thereby justifying use of the diffusion-weighted objective in (SBTM'). Note also that if $\rho_0 \neq \rho_0^*$ we simply get

$$D_{\mathrm{KL}}(\rho_T|\rho_T^*) \leq D_{\mathrm{KL}}(\rho_0|\rho_0^*) + \frac{1}{2} \int_0^T \int_\Omega |s_t(X_t(x)) - G_t(x)|^2_{D_t(x)} \rho_0(x) dx dt \tag{B.13}$$

B.4 Sequential SBTM

Let us restate Proposition 3 for convenience:

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 $s_t^* = \nabla \log \rho_t$.

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

$$\begin{aligned}
&\int_\Omega \left(|s_t(X_t(x))|^2 + 2\nabla \cdot s_t(X_t(x))\right) \rho_0(x) dx \\
&\qquad = \int_\Omega \left(|s_t(x)|^2 + 2\nabla \cdot s_t(x)\right) \rho_t(x) dx.
\end{aligned} \tag{B.14}$$

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

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

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

B.5 Denoising Loss

The following standard trick can be used to avoid computing the divergence of $s_t(x)$:

Lemma B.6. *Given $\xi = N(0, I)$, we have*

$$\begin{aligned}
&\lim_{\alpha \downarrow 0} \alpha^{-1} \mathbb{E}\big(s_t(x + \alpha\xi) \cdot \xi\big) = \nabla \cdot s_t(x), \\
&\lim_{\alpha \downarrow 0} \alpha^{-1} \mathbb{E}\big(s_t(x + \alpha\sigma_t(x)\xi) \cdot \sigma_t(x)\xi\big) = \mathrm{tr}\left(D_t(x) \nabla s_t(x)\right)
\end{aligned} \tag{B.15}$$

Proof. We have

$$\alpha^{-1} s_t(x + \alpha\xi) \cdot \xi = \alpha^{-1} s_t(x) \cdot \xi + (\nabla s_t(x)\xi) \cdot \xi + o(\alpha) \tag{B.16}$$

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.15). Hence, taking the expectation of (B.16) and evaluating the result in the limit as $\alpha \downarrow 0$ gives the first equation in (B.15). The second equation in (B.15) can be proven similarly using $\sigma_t(x)\sigma_t(x)^\mathsf{T} = D_t(x)$. □

Replacing $\nabla \cdot s_t(x)$ in (seqSBTM) with the first expression in (B.16) for a fixed $\alpha > 0$ gives the loss

$$\mathcal{L}[s_t] = \mathbb{E}_\xi \left[\int_\Omega \left(|s_t(X_t(x))|^2 + 2 s_t(X_t(x) + \alpha\xi) \cdot \xi \right) \rho_0(x) dx \right]. \tag{B.17}$$

Evaluating the square term at a perturbed data point and re-weighting by α^2 recovers the denoising loss of Vincent (2011)

$$\mathcal{L}[s_t] = \mathbb{E}_\xi \left[\int_\Omega \left| s_t(X_t(x) + \alpha\xi) + \frac{\xi}{\alpha^2} \right|^2 \rho_0(x) dx \right]. \tag{B.18}$$

In practice, we observe that the smoothing effect of the noise in (B.17) & (B.18) allows the objective to probe regions nearby the samples, and hence improves exploration.

We can also improve the accuracy of the approximation with a "doubling trick" that applies two draws of the noise of opposite sign to reduce the variance. This amounts to replacing the expectations in (B.15) with

$$\begin{aligned} &\tfrac{1}{2}\alpha^{-1}\mathbb{E}\big[s_t(x + \alpha\xi) \cdot \xi - s_t(x - \alpha\xi) \cdot \xi\big], \\ &\tfrac{1}{2}\alpha^{-1}\mathbb{E}\big[s_t(x + \alpha\sigma_t(x)\xi) \cdot \sigma_t(x)\xi - s_t(x - \alpha\sigma_t(x)\xi) \cdot \sigma_t(x)\xi\big], \end{aligned} \tag{B.19}$$

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. Moreover, we observe that use of the denoising loss stabilizes training, so that it is preferable to full computation of $\nabla \cdot s_t(x)$ even when the latter is feasible.

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 4.1.1 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 (SBTM) 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 $Z^{-1}e^{-U(x)}$.

D Experimental details and additional examples

All numerical experiments were performed in `jax` using the `dm-haiku` package to implement the networks and the `optax` package for optimization.

D.1 Harmonically interacting particles in a harmonic trap

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 (Ramachandran et al., 2017) 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 (Ioffe & Szegedy, 2015).

Optimization The Adam (Kingma & Ba, 2017) 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 $x_j = \left(x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(N)}\right)^\mathsf{T}$ with $j = 1, \ldots, n$ 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` $= 0.1$.

Analytical moments First define the mean, second moment, and covariance according to

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

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

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

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

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 $\rho_t^* = \mathsf{N}\left(m_t, C_t\right)$ and the corresponding analytical score $-\nabla \log \rho_t^*(x) = C_t^{-1}(x - m_t)$.

Potential structure Here, we show that the potential for this example lies in the class of potentials described by (13). 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 $\left(C_d(t) + (N-1)C_o(t)\right)^{-1}$ and $\left(C_d(t) - C_o(t)\right)^{-1}$. 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.

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

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

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

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.

Optimization and initialization The Adam optimizer is used with an initial learning rate of $\eta = 5\times10^{-3}$ 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 (15) 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 $\Delta t = 10^{-3}$.

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.`

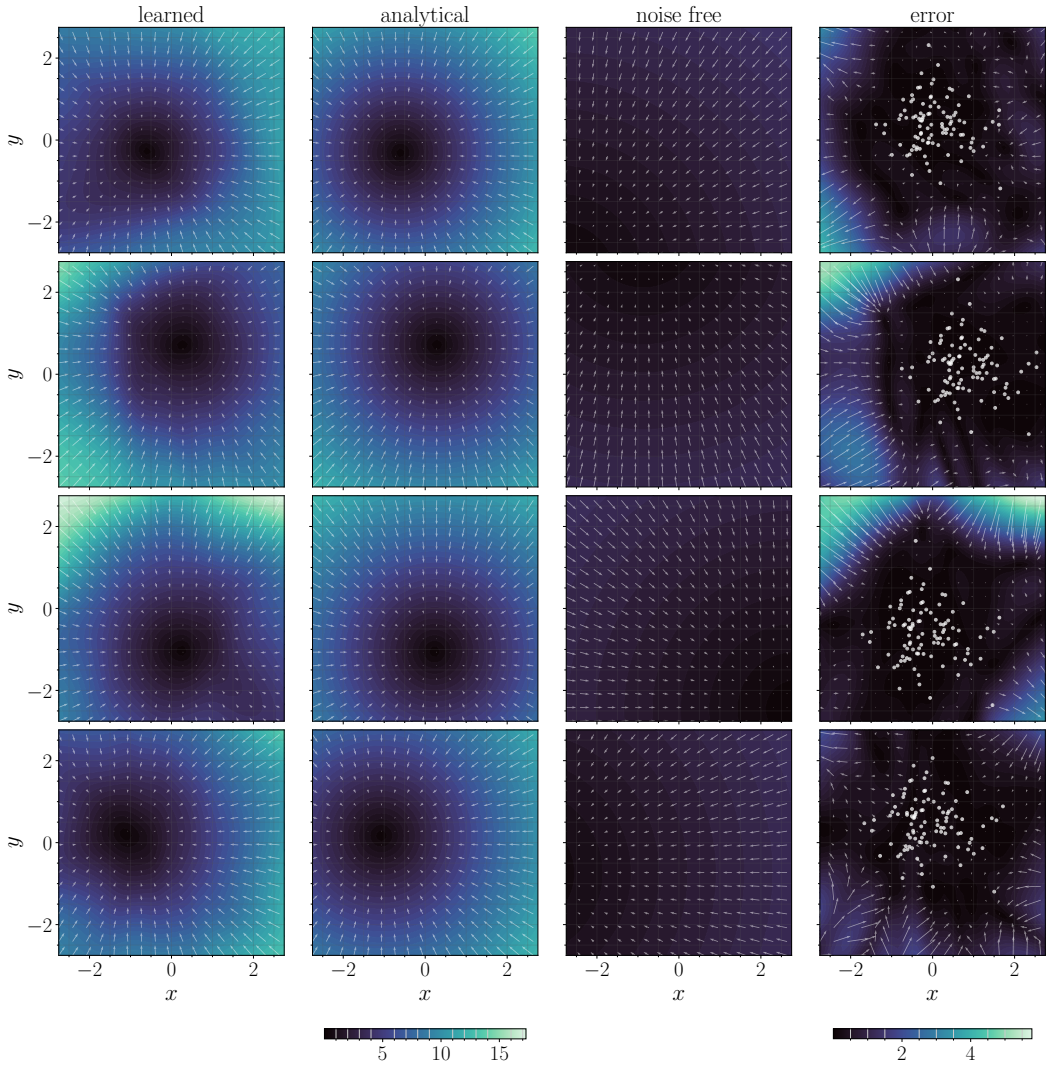

Figure D.1: *A system of $N = 50$ harmonically interacting particles in a harmonic trap: slices of the high-dimensional velocity field.* Cross sections of the velocity field for $N = 50$ harmonically interacting particles in a moving harmonic trap. Columns depict the learned, analytical, noise-free, 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.

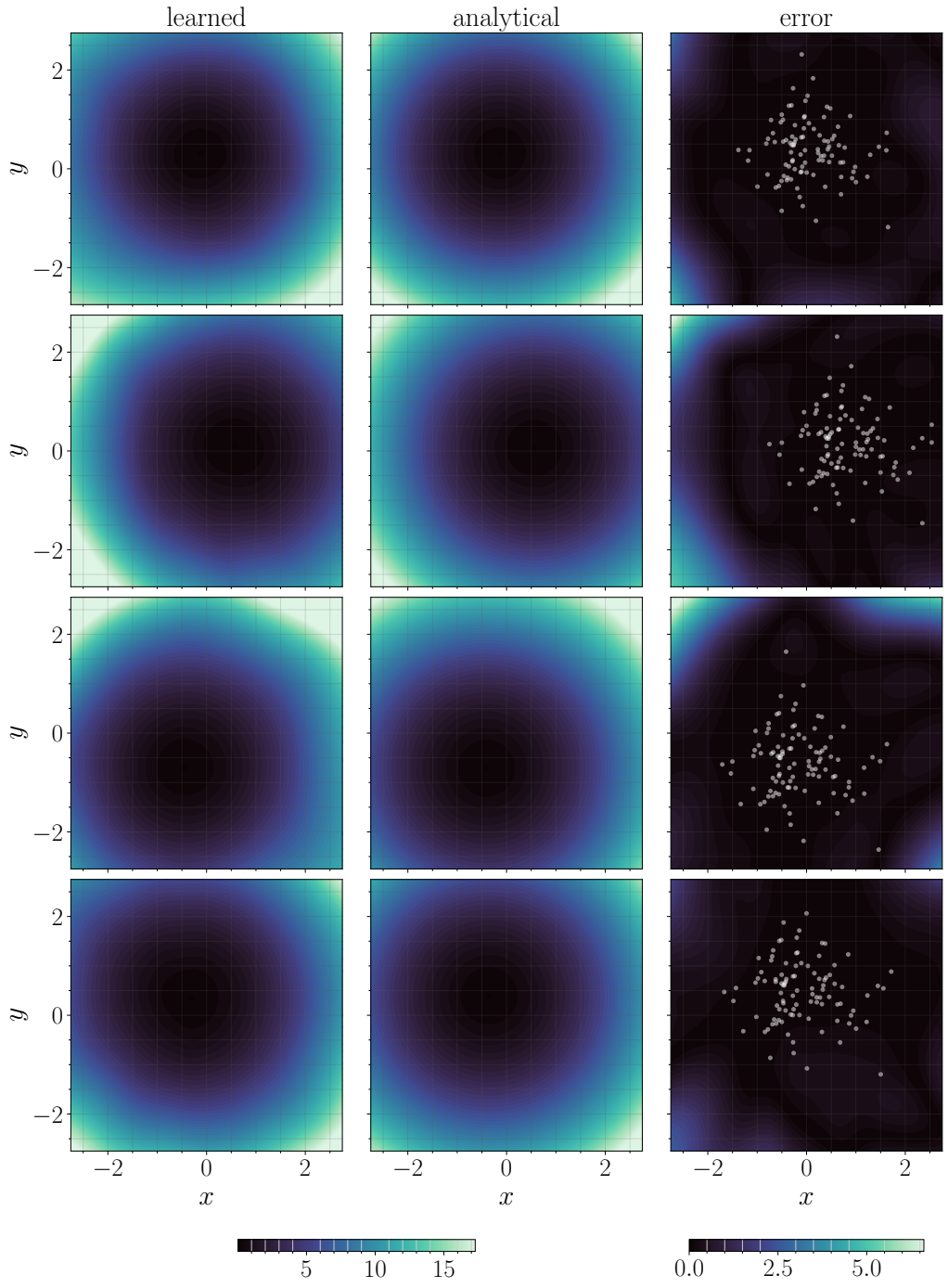

Figure D.2: *A system of $N = 50$ harmonically interacting particles in a harmonic trap: slices of the high-dimensional potential.* Cross sections of the potential field $U_{\theta_t}(x)$ computed via (13). Columns depict the learned, analytical, and error between the learned and analytical, respectively. 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.

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 AN ACTIVE SWIMMER

Here, we study an "active swimmer" model that describes the motion of a particle in an anharmonic trap with a preference to travel in a noisy direction. The system is two-dimensional, and is given by the stochastic differential equation for the position x and velocity v

$$\begin{aligned} dx &= \left(-x^3 + v\right) dt, \\ dv &= -\gamma v dt + \sqrt{2\gamma D} dW_t. \end{aligned} \tag{D.6}$$

Despite its low-dimensionality, (D.6) exhibits convergence to a non-equilibrium statistical steady state in which the probability current $j_t(x) = v_t(x)\rho_t(x)$ is non-zero.

Setup We set $\gamma = 0.1$ and $D = 1.0$. Because noise only enters the system through the velocity variable v in (D.6), the score can be taken to be one-dimensional. This is equivalent to learning the score only in the range of the rank-deficient diffusion matrix. We parameterize the score directly $s_t : \mathbb{R}^2 \to \mathbb{R}$ using a three-hidden layer neural network with $\texttt{n\_hidden} = 32$ neurons per hidden layer.

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 $\texttt{n\_opt\_steps} = 25$ steps of Adam until the norm of the gradient is below $\texttt{gtol} = 0.5$.

Results Depictions of the sample trajectories $\{x_i(t), v_i(t)\}_{i=1}^n$ in phase space are shown in Figure D.4. The trajectories demonstrate that the distribution of the learned samples qualitatively matches the distribution of the SDE samples. The noise-free system grows increasingly and overly compressed with time. The learned velocity field effectively captures a non-zero rotational steady-state current that qualitatively matches the current of the SDE but enjoys more interpretable sample trajectories.

A movie of the motion of the samples (x_i, v_i) in phase space 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 non-zero steady-state probability current that qualitatively matches that of the SDE. By contrast, the noise-free system becomes increasingly concentrated with time, failing to accurately capture the current. Figure D.5 depicts the learned velocity field $v_t(x) = b_t(x) - Ds_t(x)$. The figure highlights the structure of the steady-state current, which contains an elliptical region with closed orbits. The elliptical region remains roughly fixed in size as time proceeds, while the orbits of the noise-free system in Figure D.6 become increasing compressed. Kernel density estimation demonstrates that an estimated PDF for the samples of learned solution qualitatively matches that of the SDE (Figure D.7).

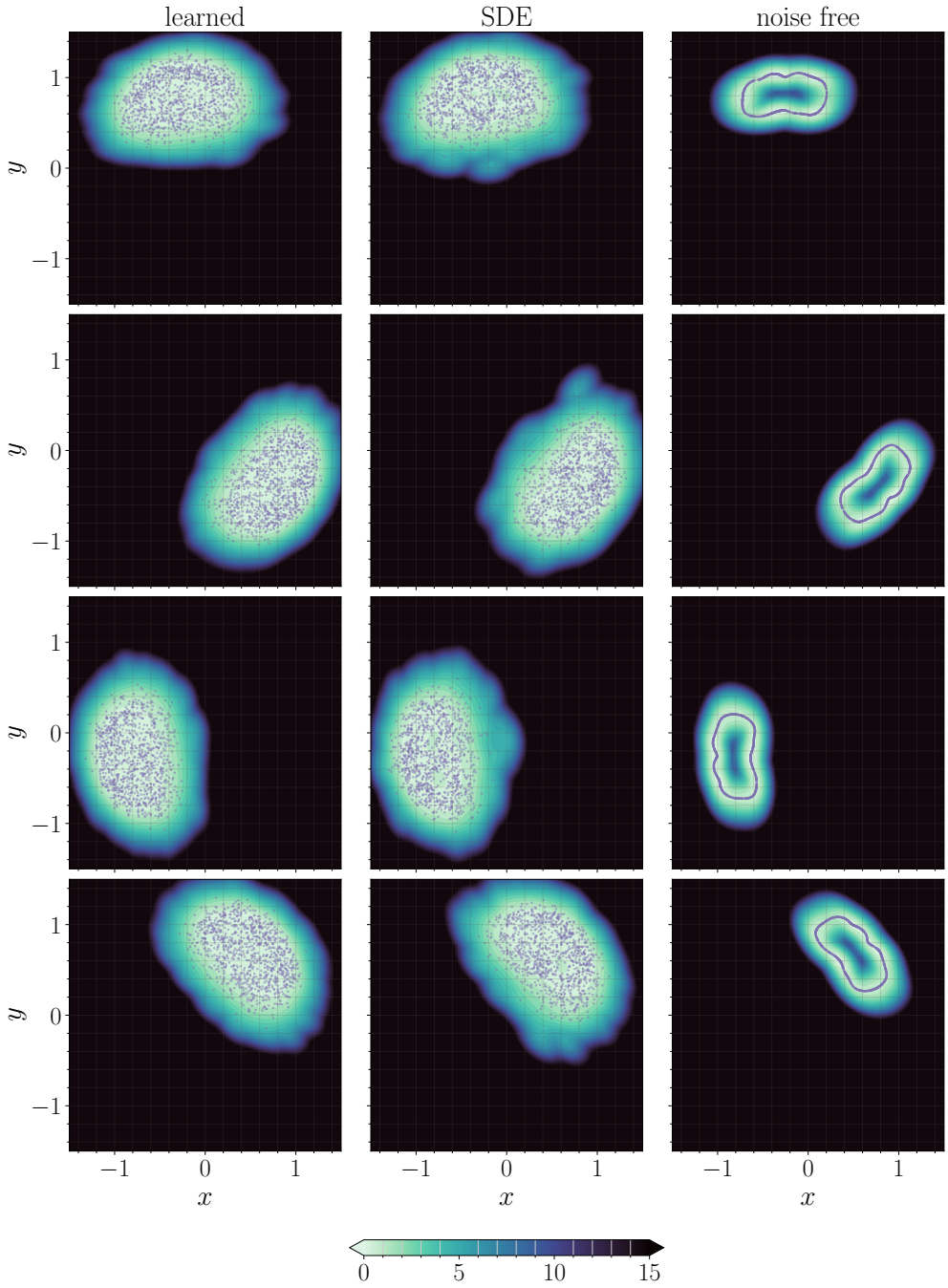

Figure D.3: *A system of $N = 5$ soft-sphere particles in an anharmonic trap: one-particle potential.* Cross sections of the one-particle potential field $U(x) = -\log \rho_{\text{KDE}}(x)$ where ρ_{KDE} denotes a kernel density estimate of the one-particle density obtained by pooling all particles and treating them as equivalent two-dimensional 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.

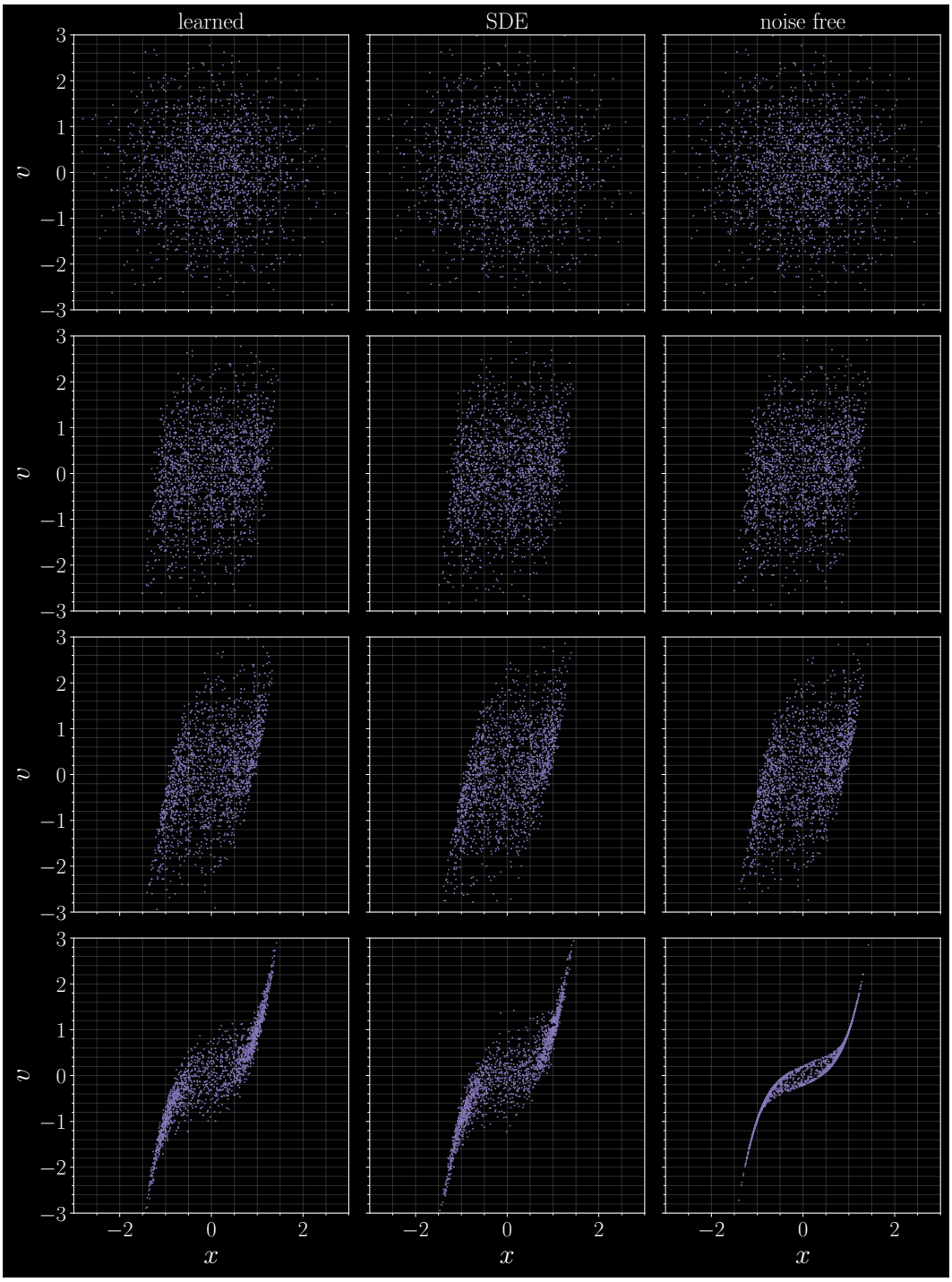

Figure D.4: *An active swimmer: sample trajectories.* Samples in the xv plane. Columns denote solution type and rows indicate snapshots in time ($t = 0, 0.25, 0.5, 3.0$, respectively). The learned and SDE systems develop bimodality while the noise free system collapses with time and does not correctly capture the variance.

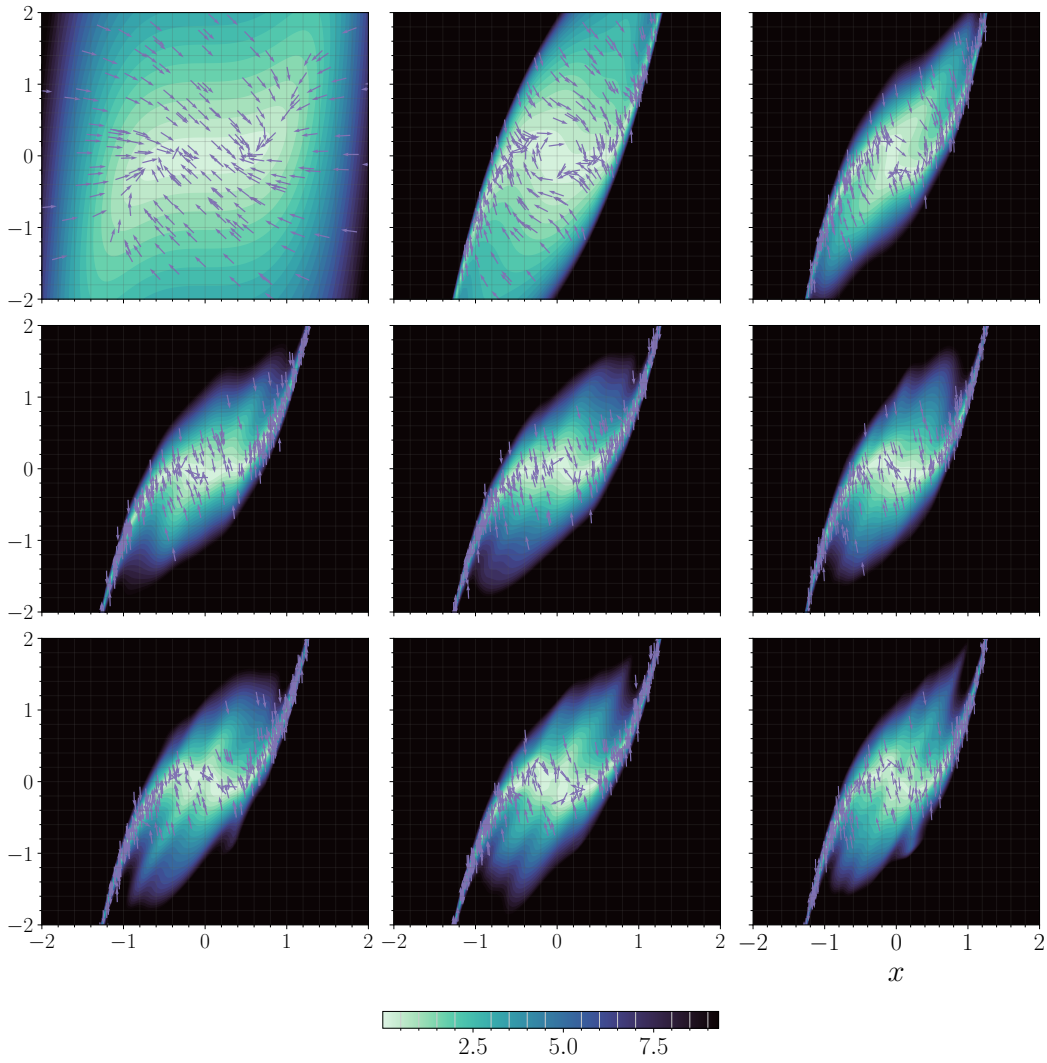

Figure D.5: *An active swimmer: learned velocity.* The learned velocity field (right-hand side of (4)) for the active swimmer example. Color indicates the magnitude of the velocity field computed on a grid, while arrows indicate the direction of the velocity field on samples. Time corresponds to progressing 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.

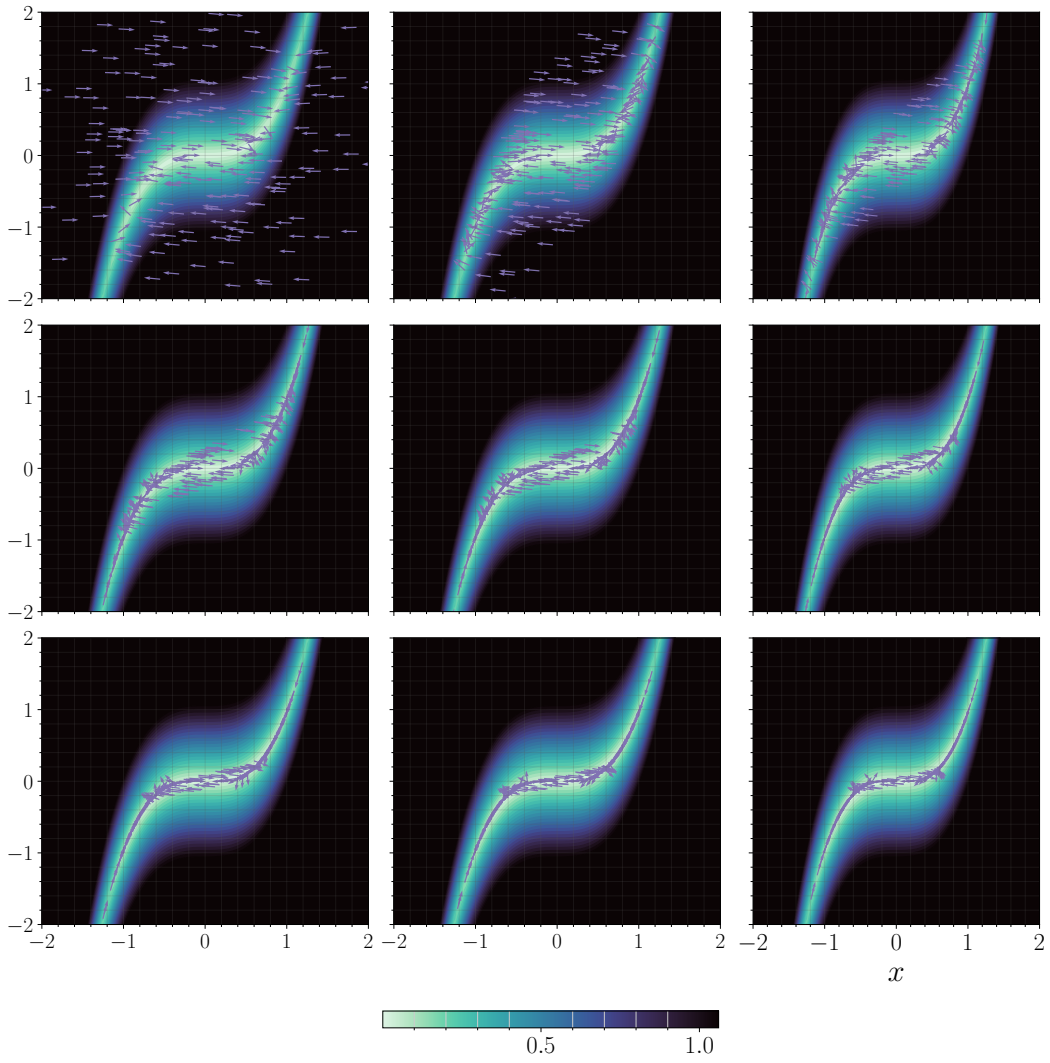

Figure D.6: *An active swimmer: noise free velocity.* Noise free velocity field. As in Figure D.5, color indicates the magnitude of the velocity field while arrows indicate the direction, and time corresponds to progressing in the grid along columns from the top-left to the bottom-right image ($t = k \times .75$ 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.

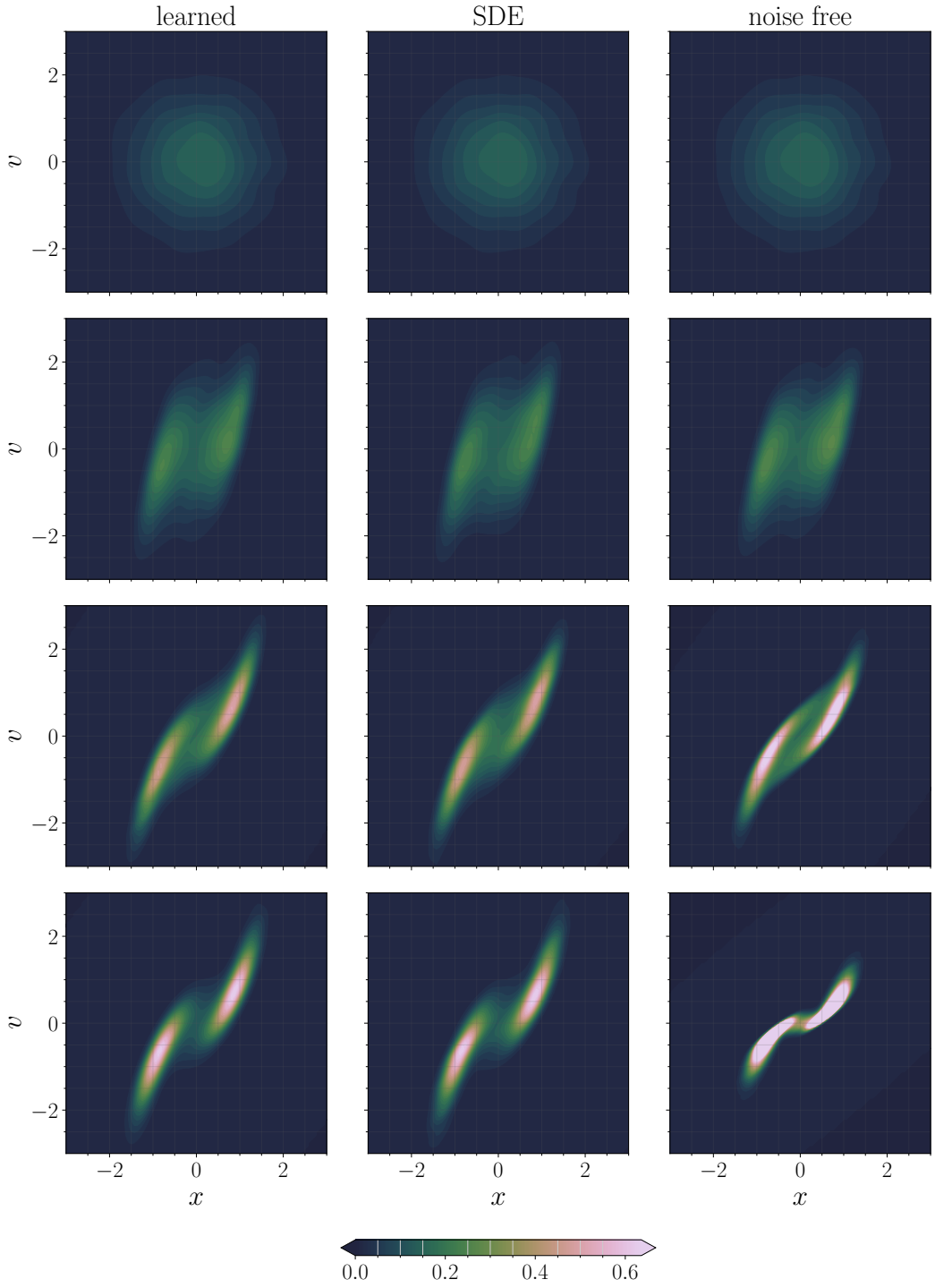

Figure D.7: *An active swimmer: density.* PDFs computed via kernel density estimation in the xv plane. Columns denote solution type and rows denote snapshots in time ($t = 0, 0.5, 1.5, 6.0$, respectively). Similar to the samples presented in Figure D.4, the KDE reveals bimodality in the probability density 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.