Closing the Approximation Gap in Simulation-free Latent SDEs

Henry D. Smith
Stanford University
smithhd@stanford.edu

Brian L. Trippe*
Stanford University
btrippe@stanford.edu

Scott W. Linderman*
Stanford University
swl1@stanford.edu

Abstract

Recovering dynamical systems from noisy observations is a recurring challenge across scientific domains, including neuroscience and physics. Latent stochastic differential equations (SDEs) address this by modeling the system as an unobserved state that evolves according to a learnable SDE and generates the observations. Variational inference (VI) provides a tractable objective for fitting latent SDEs. Traditional VI algorithms evaluate this objective by numerical simulation over a time discretization, trading fidelity for computational cost. A recent class of algorithms, simulation-free VI, sidesteps this tradeoff by parameterizing the posterior through its instantaneous marginals rather than its drift. In this work, we show that the efficiency of existing simulation-free VI algorithms comes at a price: their parameterizations restrict the approximate posterior to a subset of the SDEs available to simulation-based methods, degrading posterior inference and parameter learning. We propose Helmholtz-SDE, a simulation-free VI algorithm that closes this gap by optimizing over path laws compatible with a prescribed collection of marginals. Helmholtz-SDE recovers dynamics more faithfully than prior simulation-free methods, with the largest gains under high posterior uncertainty. It further matches the performance of simulation-based VI at a fraction of the runtime. Code is available at https://github.com/lindermanlab/helmholtz-sde.

1 Introduction

Latent stochastic differential equations (SDEs) enable the unsupervised discovery of dynamical systems from noisy, irregularly sampled time series, a setting common across scientific domains. They posit an unobserved latent state that evolves continuously in time and generates observations at a finite set of time points [1, 2]. Latent SDEs have been applied to neuroscience [3, 4], disease modeling [5], and reduced-order fluid dynamics [6]. Fitting these models requires both learning the latent dynamics and approximating the posterior over continuous trajectories.

Variational inference (VI) translates the problem of fitting latent SDEs into an optimization problem. However, evaluating the variational objective (ELBO) requires computing an integral over the full time horizon (a path-space KL divergence). Simulation-based VI algorithms estimate this term by sweeping across a time discretization at each gradient step, either by integrating an ODE to compute the posterior marginals or by simulating sample trajectories from the variational SDE. A recent line of work, simulation-free VI [6–8], sidesteps sequential simulation using stochastic VI. Rather than integrating, they compute an unbiased estimate of the objective by sampling a random time together with a state at that time. They make this estimate efficient by parameterizing the posterior through its one-time marginal distributions, i.e., the distribution of the latent state at each time, allowing states to be sampled directly without simulating the preceding trajectory.

In this work, we show that the family of approximate posterior distributions over which simulation-free VI optimizes is *strictly smaller* than that available to simulation-based VI. One-time marginals

*Equal advising

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

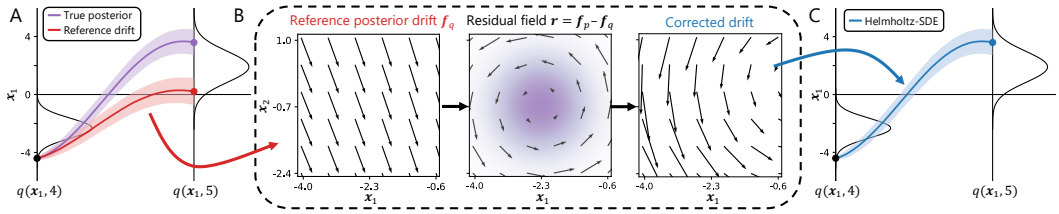


Figure 1: Overview of Helmholtz-SDE. **A**: Existing simulation-free VI algorithms commit to a single drift (red) for each set of one-time marginals (black). This choice induces transition distributions that can differ from those of the true posterior (purple), even when the one-time marginals agree. **B**: Helmholtz-SDE takes a drift consistent with a set of one-time marginals, then corrects it by a vector field that preserves the marginals. **C**: The resulting approximation (blue) better matches the posterior.

do not determine a path law. Existing simulation-free methods choose a single compatible drift from many possibilities, implicitly restricting the variational posterior. When observations are sparse or noisy, this arbitrary choice can substantially degrade both posterior inference and learned dynamics.

We propose *Helmholtz-SDE*, a simulation-free VI algorithm that removes this restriction. Starting from any drift compatible with the prescribed marginals, we compute a Helmholtz correction that preserves those marginals while optimizing the ELBO. The resulting posterior family remains simulation-free but spans a larger class of path laws, closing the gap to simulation-based VI.

2 Problem Statement

Model. We study *latent stochastic differential equation* (SDE) models, in which a K-dimensional latent state $\boldsymbol{x}(t)$ evolves over the interval $0 \leq t \leq T$ and gives rise to D-dimensional observations $\boldsymbol{y}(t_n)$ at times $\{t_n\}_{n=1}^N$. Let $\boldsymbol{\theta}$ denote the full collection of model parameters, including the parameters of the prior and the likelihood. The prior over the latent state is the solution to an SDE

$$\mathbb{P}_{\boldsymbol{\theta}}\colon\ d\boldsymbol{x}(t) = \boldsymbol{f}_p(\boldsymbol{x}(t)|\boldsymbol{\theta})dt + \boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}d\boldsymbol{w}(t), \quad \boldsymbol{x}(0) \sim p(\boldsymbol{x}, 0|\boldsymbol{\theta}), \quad 0 \leq t \leq T, \tag{1}$$

where $\boldsymbol{f}_p(\cdot|\boldsymbol{\theta})\colon \mathbb{R}^K \to \mathbb{R}^K$ is the drift function, $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}$ is the diffusion coefficient, $(\boldsymbol{w}(t))_{0\leq t\leq T}$ is a standard K-dimensional Brownian motion, and $p(\boldsymbol{x}, 0|\boldsymbol{\theta})$ is the initial state distribution. We denote the marginal density of $\mathbb{P}_{\boldsymbol{\theta}}$ at time t by $p(\boldsymbol{x}, t|\boldsymbol{\theta})$. For simplicity of presentation, we restrict notation to a time-independent drift and a state- and time-independent diffusion coefficient.

We observe data $\mathcal{D} = \{(t_n, \boldsymbol{y}(t_n))\}_{n=1}^N$ representing noisy readouts of the latent state. These observations are conditionally independent given the latent state: $p(\boldsymbol{y}|(\boldsymbol{x}(t))_{0\leq t\leq T}, \boldsymbol{\theta}) = \prod_{n=1}^N p(\boldsymbol{y}(t_n)|\boldsymbol{x}(t_n), \boldsymbol{\theta})$, where $\boldsymbol{y} = \{\boldsymbol{y}(t_n)\}_{n=1}^N$. We describe our technical assumptions on the generative model in Appendix A.1.

There are two goals in latent SDE models: (i) to learn the parameters $\boldsymbol{\theta}$ that maximize $\log p(\boldsymbol{y}|\boldsymbol{\theta})$; and (ii) to infer an approximate posterior over the latent trajectory $(\boldsymbol{x}(t))_{0\leq t\leq T}$ given the data $\mathcal{D}$ and parameters $\boldsymbol{\theta}$. In practice, one may have multiple trials $\{\boldsymbol{y}^{(m)}\}_{m=1}^M$, each representing an independent sample from the probabilistic model. In this case, the model parameters $\boldsymbol{\theta}$ are learned using all trials, and the posterior factorizes over trials. For simplicity, we suppress the trial index m.

Intractability. Both goals of learning and inference are generally intractable. For learning, evaluating the marginal likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ involves marginalizing over the prior on latent trajectories (1); this integral cannot be evaluated in closed form. For inference, the posterior law is the solution to an SDE with the same diffusion coefficient $\boldsymbol{\Sigma}^{1/2}$ as the prior and drift obtained by solving a linear partial differential equation [9–12]. In nonlinear or non-Gaussian settings, the posterior drift is unavailable in closed form, and solving the associated PDE can be computationally prohibitive even for small K.

3 Background

Variational inference (VI) provides a tractable framework for joint learning and inference in latent SDEs. We first describe the variational objective and survey existing VI algorithms. We then identify a class of stochastic VI algorithms that avoid sequential simulation, the setting of our method.

3.1 Variational inference for latent SDE models

Variational inference (VI) approximates the intractable posterior with a distribution belonging to a tractable family, known as a *variational family* [13, 14]. For latent SDEs, we consider the family $\mathcal{Q}$ comprised of diffusion processes

$$\mathbb{Q}_{\boldsymbol{\phi}}:\ d\boldsymbol{x}(t) = \boldsymbol{f}_q(\boldsymbol{x}(t), t|\boldsymbol{\phi})dt + \boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}d\boldsymbol{w}(t), \quad \boldsymbol{x}(0) \sim q(\boldsymbol{x}, 0|\boldsymbol{\phi}), \quad 0 \le t \le T, \tag{2}$$

where the variational parameters $\boldsymbol{\phi}$ specify the posterior drift $\boldsymbol{f}_q(\boldsymbol{x}, t|\boldsymbol{\phi})$ and the initial state distribution $q(\boldsymbol{x}, 0|\boldsymbol{\phi})$. This family is chosen because, when $\boldsymbol{f}_q(\boldsymbol{x}, t|\boldsymbol{\phi})$ and $q(\boldsymbol{x}, 0|\boldsymbol{\phi})$ are sufficiently expressive, the exact posterior will belong to $\mathcal{Q}$ [see, e.g., 15].

We minimize the negative evidence lower bound (ELBO) over distributions in $\mathcal{Q}$,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \underbrace{\sum_{n=1}^{N} -\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}} \log p(\boldsymbol{y}(t_n)|\boldsymbol{x}(t_n), \boldsymbol{\theta})}_{\mathcal{L}_{\text{rec}}} + \underbrace{\mathrm{D}_{\mathrm{KL}}\left(q(\boldsymbol{x}, 0|\boldsymbol{\phi}) \parallel p(\boldsymbol{x}, 0|\boldsymbol{\theta})\right)}_{\mathcal{L}_{\text{init}}} + \\ \underbrace{\frac{1}{2}\int_0^T \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}} \|\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1/2}\{\boldsymbol{f}_q(\boldsymbol{x}(t), t|\boldsymbol{\phi}) - \boldsymbol{f}_p(\boldsymbol{x}(t)|\boldsymbol{\theta})\}\|_2^2 dt}_{\mathcal{L}_{\text{path}}},\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction loss, $\mathcal{L}_{\text{init}}$ is the KL divergence to the prior initial distribution, and $\mathcal{L}_{\text{path}}$ is the path-space KL divergence to the prior.

The negative ELBO is an upper bound on the negative marginal log likelihood $-\log p(\boldsymbol{y}|\boldsymbol{\theta})$, and the gap between them is the KL divergence between $\mathbb{Q}_{\boldsymbol{\phi}}$ and the exact posterior. Therefore, for fixed $\boldsymbol{\theta}$, when the exact posterior is contained in $\mathcal{Q}$ it is the unique minimizer of eq. (3). Appendix A.2 gives a derivation. However, unlike the negative log likelihood, the ELBO provides a tractable objective for learning model parameters, which we seek to estimate by jointly optimizing with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$.

Existing VI algorithms for latent SDEs fall into two broad classes based on how they evaluate the negative ELBO eq. (3): simulation-based and simulation-free. We describe each in turn.

3.2 Simulation-based VI

Simulation-based algorithms evaluate the path KL $\mathcal{L}_{\text{path}}$ over a discretization of the interval $[0, T]$ at each gradient step. When $\mathbb{Q}_{\boldsymbol{\phi}}$ is restricted to a tractable family, such as a Gaussian process with affine $\boldsymbol{f}_q$ and state-independent $\boldsymbol{\Sigma}$, the marginals of $\mathbb{Q}_{\boldsymbol{\phi}}$ can be computed by integrating an ODE [1, 15–22]. When $\boldsymbol{f}_q$ is parameterized more flexibly, e.g., by a neural network [2, 23–26], samples $\boldsymbol{x}(t) \sim \mathbb{Q}_{\boldsymbol{\phi}}$ are obtained by simulating the variational SDE (2). In both cases, the time complexity per gradient step scales with the discretization of $[0, T]$, introducing a tradeoff between approximation fidelity and computational cost.

3.3 Simulation-free VI

Simulation-free VI algorithms introduce two ideas to avoid discretizing the variational SDE and accelerate optimization of $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi})$. The first is to optimize a Monte Carlo estimate of eq. (3)

$$\widehat{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = -N \log p(\boldsymbol{y}(t_n)|\boldsymbol{x}(t_n), \boldsymbol{\theta}) + \log \frac{q(\boldsymbol{x}(0), 0|\boldsymbol{\phi})}{p(\boldsymbol{x}(0), 0|\boldsymbol{\theta})} + \frac{T}{2}\|\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1/2}\{\boldsymbol{f}_q(\boldsymbol{x}(t), t|\boldsymbol{\phi}) - \boldsymbol{f}_p(\boldsymbol{x}(t)|\boldsymbol{\theta})\}\|_2^2.$$

Here, $n \sim \mathcal{U}(\{1, \ldots, N\})$ is a random observation index, $t \sim \mathcal{U}([0, T])$ is a random time, and $\boldsymbol{x}(0), \boldsymbol{x}(t_n), \boldsymbol{x}(t) \sim \mathbb{Q}_{\boldsymbol{\phi}}$ are posterior samples. The estimate $\widehat{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ avoids the sum over all observations in $\mathcal{L}_{\text{rec}}$, the integral over the full trajectory in $\mathcal{L}_{\text{path}}$, and the expectations over $\mathbb{Q}_{\boldsymbol{\phi}}$ in all three terms of $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi})$. Compared to optimizing the negative ELBO, this approach introduces variance that is managed by averaging across Monte Carlo samples. Because the estimate samples random times rather than full paths, optimization amounts to a stochastic VI algorithm [27].

The second idea is to parameterize the drift $\boldsymbol{f}_q$ that defines $\mathbb{Q}_{\boldsymbol{\phi}}$ implicitly through its one-time marginal distributions. In order for optimizing $\widehat{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ to provide runtime benefits over optimizing the negative ELBO, we must be able to draw samples $\boldsymbol{x}(t) \sim \mathbb{Q}_{\boldsymbol{\phi}}$ without simulating the variational

SDE eq. (2). This is achieved by reversing the usual construction of $\mathbb{Q}_{\boldsymbol{\phi}}$: we first prescribe the marginal distribution of the posterior at any single time $\{q(\boldsymbol{x}, t|\boldsymbol{\phi})\}_{0\leq t\leq T}$ and then choose a drift $\boldsymbol{f}_q$ compatible with those marginals. One can sample $\boldsymbol{x}(t)$ from the posterior with time complexity that does not depend on t.

For example, SDE Matching (Bartosh et al. [8]) parameterizes the posterior as having Gaussian marginals, $q(\boldsymbol{x}, t|\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{m}(t|\boldsymbol{\phi}), \boldsymbol{S}(t|\boldsymbol{\phi}))$ and shows that a time-dependent, affine drift $\boldsymbol{f}_q(\boldsymbol{x}, t|\boldsymbol{\phi}) = \boldsymbol{A}_q(t|\boldsymbol{\phi})\boldsymbol{x} + \boldsymbol{b}_q(t|\boldsymbol{\phi})$ defines a variational SDE with precisely these marginals.[2] A compatible drift is readily computed from $\boldsymbol{m}(t|\boldsymbol{\phi})$ and $\boldsymbol{S}(t|\boldsymbol{\phi})$ as

$$\boldsymbol{f}_q(\boldsymbol{x}, t|\boldsymbol{\phi}) = \left[\left(\frac{d}{dt}\boldsymbol{S}(t|\boldsymbol{\phi})^{1/2}\right)\boldsymbol{S}(t|\boldsymbol{\phi})^{-1/2} - \frac{1}{2}\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{S}(t|\boldsymbol{\phi})^{-1}\right][\boldsymbol{x} - \boldsymbol{m}(t|\boldsymbol{\phi})] + \frac{d}{dt}\boldsymbol{m}(t|\boldsymbol{\phi}), \quad (4)$$

where $\boldsymbol{S}(t|\boldsymbol{\phi})^{1/2}$ is the symmetric square root of $\boldsymbol{S}(t|\boldsymbol{\phi})$. SVISE (Course and Nair [6, 7]) also considers Gaussian marginals and a time-dependent affine SDE drift, but they instead choose $\boldsymbol{A}_q(t|\boldsymbol{\phi})$ to be symmetric. We provide background on both algorithms in Appendix A.5. Once we have chosen a drift $\boldsymbol{f}_q(\boldsymbol{x}, t|\boldsymbol{\phi})$, unbiased gradient estimates of $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi})$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can be obtained via the reparameterization trick, without backpropagating through a simulated trajectory [28].

4 Helmholtz-SDE: Augmented simulation-free variational inference

The simulation-free VI framework we described in Section 3.3 leaves an important ambiguity unresolved: the one-time marginal distributions $\{q(\boldsymbol{x}, t|\boldsymbol{\phi})\}_{0\leq t\leq T}$ do not uniquely determine the variational path law $\mathbb{Q}_{\boldsymbol{\phi}}$. In general, many drifts $\boldsymbol{f}_q$ realize the same collection of marginals, and existing simulation-free VI algorithms resolve this ambiguity by selecting a particular compatible drift for each set of marginals, independently of the variational objective in eq. (3). This choice is not a benign implementation detail: drifts with identical one-time marginals can induce different path laws, including different multi-time joint distributions $(\boldsymbol{x}(s), \boldsymbol{x}(t))$ for $s \neq t$ (Figure 1). When posterior uncertainty is high, we argue that fixing a single compatible drift can substantially degrade both posterior inference and learned dynamics.

We introduce Helmholtz-SDE, a simulation-free VI algorithm that removes this restriction by optimizing over path laws sharing the same one-time marginals. We first characterize the path-law ambiguity left unresolved by marginal-based parameterizations: drifts realizing the same marginals differ by divergence-free vector fields, which can be added without altering those marginals (Section 4.1). We then show that the variational objective selects an optimal drift among this class (Section 4.2). Finally, we turn this characterization into a practical algorithm for Gaussian one-time marginals and state-independent diffusion (Section 4.3).

4.1 The Fokker–Planck equation and divergence-free vector fields

The Fokker–Planck equation characterizes the ambiguity in the posterior path-law that arises from parameterizing the one-time marginals. Let $\mathbb{Q}_{\boldsymbol{\phi}} \in \mathcal{Q}$ be a variational distribution with drift $\boldsymbol{f}_q$ and one-time marginals $\{q(\boldsymbol{x}, t|\boldsymbol{\phi})\}_{0\leq t\leq T}$. The marginals evolve according to the PDE

$$\frac{\partial}{\partial t}q(\boldsymbol{x}, t|\boldsymbol{\phi}) = -\nabla_{\boldsymbol{x}} \cdot \left[q(\boldsymbol{x}, t|\boldsymbol{\phi})\boldsymbol{f}_q(\boldsymbol{x}, t|\boldsymbol{\phi})\right] + \frac{1}{2}\sum_{i,j}\boldsymbol{\Sigma}_{ij}(\boldsymbol{\theta})\partial_{\boldsymbol{x}_i\boldsymbol{x}_j}q(\boldsymbol{x}, t|\boldsymbol{\phi}), \quad 0 \leq t \leq T, \quad (5)$$

known as the *Fokker–Planck equation*, where $\nabla_{\boldsymbol{x}} \cdot \boldsymbol{f} = \sum_i \partial_{\boldsymbol{x}_i}\boldsymbol{f}_i(\boldsymbol{x})$ is the divergence operator.

Equation (5) depends on the variational drift $\boldsymbol{f}_q$ only through the divergence of $q(\boldsymbol{x}, t|\boldsymbol{\phi})\boldsymbol{f}_q(\boldsymbol{x}, t|\boldsymbol{\phi})$. Consequently, any vector field $\boldsymbol{v}$ satisfying $\nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x}, t|\boldsymbol{\phi})\boldsymbol{v}(\boldsymbol{x}, t)] = 0$ for all $(\boldsymbol{x}, t)$ leaves the marginals unchanged when added to $\boldsymbol{f}_q$. We call such vector fields *q-divergence-free*. The following proposition shows that divergence-free vector fields are the only source of non-uniqueness.

Proposition 1: Consider two diffusions with drifts $\boldsymbol{f}$ and $\tilde{\boldsymbol{f}}$ that share the same diffusion coefficient and initial distribution $q(\boldsymbol{x}, 0)$. They induce the same one-time marginals $\{q(\boldsymbol{x}, t)\}_{0\leq t\leq T}$ iff $\boldsymbol{f} - \tilde{\boldsymbol{f}}$ is q-divergence-free.

Proof. One analyzes the Fokker–Planck equations for the two diffusions. See Appendix A.3. □

[2] The framework proposed by Bartosh et al. [8] can in principle support non-Gaussian one-time marginals, but the method is instantiated and evaluated only with Gaussian marginals. See Appendix A.5 for further discussion.

4.2 Optimizing over path laws at fixed marginals

In Section 4.1, we identified the family of drifts compatible with a given set of one-time marginal distributions. Existing simulation-free VI algorithms commit to a single member, independent of the ELBO. We show next that optimizing the ELBO over this family produces an optimal drift.

Problem formulation. Fix a collection of one-time marginal distributions $\{q(\boldsymbol{x}, t|\boldsymbol{\phi})\}_{0\leq t\leq T}$. We call a drift $\boldsymbol{f}_q(\boldsymbol{x}, t|\boldsymbol{\phi})$ a *reference drift* for these marginals if the SDE with drift $\boldsymbol{f}_q$ and diffusion coefficient $\boldsymbol{\Sigma}^{1/2}$ induces one-time marginals $\{q(\boldsymbol{x}, t|\boldsymbol{\phi})\}_{0\leq t\leq T}$. For Gaussian one-time marginals, eq. (4) provides an example of a reference drift. By Proposition 1, every other drift with the same one-time marginals can be written as $\boldsymbol{f}_q + \boldsymbol{v}$, where $\boldsymbol{v}(\boldsymbol{x}, t)$ is q-divergence-free. Thus, optimizing over all path laws compatible with these marginals is equivalent to optimizing over all divergence-free perturbations to the reference drift.

In the negative ELBO, the reconstruction loss and the initial-time KL are unchanged by $\boldsymbol{v}$, so the only dependence on the path law is through the path-space KL $\mathcal{L}_{\text{path}}$. Moreover, since $\boldsymbol{v}$ is q-divergence-free, modifying it at time s does not affect the marginal distribution at time t. Therefore, the problem decouples over times $0 \leq t \leq T$ and the optimal compatible drift solves

$$\inf_{\boldsymbol{v}} \ \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}} \|\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1/2}\{\boldsymbol{f}_p(\boldsymbol{x}(t)|\boldsymbol{\theta}) - \boldsymbol{f}_q(\boldsymbol{x}(t), t|\boldsymbol{\phi}) - \boldsymbol{v}(\boldsymbol{x}(t), t)\}\|_2^2, \tag{6}$$

subject to $\nabla_{\boldsymbol{x}} \cdot [\boldsymbol{v}(\boldsymbol{x}, t)q(\boldsymbol{x}, t|\boldsymbol{\phi})] = 0, \ (\boldsymbol{x}, t) \in \mathbb{R}^K \times [0, T]$.

Solution via the Helmholtz decomposition. Define the *residual vector field* $\boldsymbol{r}(\boldsymbol{x}, t|\boldsymbol{\theta}, \boldsymbol{\phi}) = \boldsymbol{f}_p(\boldsymbol{x}|\boldsymbol{\theta}) - \boldsymbol{f}_q(\boldsymbol{x}, t|\boldsymbol{\phi})$ which measures the gap between the prior and the reference posterior drifts. Intuitively, we would like to choose $\boldsymbol{v} = \boldsymbol{r}$ so that eq. (6) becomes zero. However, the divergence-free constraint prevents us from adding the part of $\boldsymbol{r}$ that would alter the marginals.

To isolate the part of $\boldsymbol{r}$ that alters the marginals, we appeal to the *q-weighted Helmholtz decomposition*

$$\boldsymbol{r}(\boldsymbol{x}, t|\boldsymbol{\theta}, \boldsymbol{\phi}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t) + \boldsymbol{h}(\boldsymbol{x}, t), \tag{7}$$

where $\nabla_{\boldsymbol{x}}\psi$ is a gradient field and $\boldsymbol{h}$ is a q-divergence-free vector field [29]. $\boldsymbol{\Sigma}\nabla_{\boldsymbol{x}}\psi$ points in directions that alter the one-time marginals, while $\boldsymbol{h}$ points in directions that leave them invariant. The q-weighted Helmholtz decomposition thus separates $\boldsymbol{r}$ into exactly the two components we care about: the part we must leave alone to preserve marginals, and the part we are free to add.

Substituting $\boldsymbol{r} = \boldsymbol{f}_p - \boldsymbol{f}_q$ into eq. (6) and using the q-weighted Helmholtz decomposition yields

$$\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}} \|\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1/2}\{\boldsymbol{h}(\boldsymbol{x}(t), t) - \boldsymbol{v}(\boldsymbol{x}(t), t)\}\|_2^2 + \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}} \|\boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}(t), t)\|_2^2. \tag{8}$$

The cross term vanishes by integration by parts, using that $\boldsymbol{h}$ and $\boldsymbol{v}$ are q-divergence-free. The second term in eq. (8) does not depend on $\boldsymbol{v}$, so the optimum of eq. (6) is achieved by $\boldsymbol{v}^\star = \boldsymbol{h}$.

In other words, among all diffusions sharing the prescribed marginals, the negative ELBO is minimized by adding the q-divergence-free component of $\boldsymbol{r}(\boldsymbol{x}, t|\boldsymbol{\theta}, \boldsymbol{\phi})$ to the reference drift. While previous simulation-free VI algorithms commit to a single drift consistent with a set of marginals, we have shown that it is possible, in theory, to optimize over *all* such drifts.

4.3 A variational inference algorithm for Gaussian one-time marginals

Next, we translate our insights into a simulation-free VI algorithm that optimizes over multiple path laws for each collection of one-time marginals. We follow prior works [6–8] and take the variational family to have Gaussian one-time marginals $q(\boldsymbol{x}, t|\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{m}(t|\boldsymbol{\phi}), \boldsymbol{S}(t|\boldsymbol{\phi}))$ with $\boldsymbol{m}(t|\boldsymbol{\phi})$ and $\boldsymbol{S}(t|\boldsymbol{\phi})$ continuously differentiable in t and $\boldsymbol{S}(t|\boldsymbol{\phi})$ positive definite. We further restrict to state-independent diffusion; we present the time-independent case $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}$ for simplicity.

To leverage the simulation-free stochastic VI framework from Section 3.3, we must compute a variational drift for each collection of marginals. By Section 4.2, the optimal drift is given by the q-weighted Helmholtz decomposition of the residual $\boldsymbol{r} = \boldsymbol{f}_p - \boldsymbol{f}_q$, where we choose the reference drift to be that in eq. (4). However, even for Gaussian marginals, the weighted Helmholtz decomposition eq. (7) is generally not closed-form. We describe a tractable approximation to the optimal variational drift and then state the resulting theoretical guarantee. Algorithm 1 summarizes Helmholtz-SDE.

Algorithm 1: Helmholtz-SDE training

Require: data $\mathcal{D}$; prior drift $\boldsymbol{f}_p$; Gaussian marginals $(\boldsymbol{m}, \boldsymbol{S})$; reference drift $\boldsymbol{f}_q$; degree ℓ
for *gradient iterations* **do**
- ▷ Sample observation times, states to estimate $\widehat{\mathcal{L}}_{\text{rec}}$; compute closed-form Gaussian KL $\mathcal{L}_{\text{init}}$
- ▷ Sample KL times $s_j \sim \text{Unif}[0, T]$ and states $\boldsymbol{x}_{j,k} \sim \mathcal{N}(\boldsymbol{m}(s_j|\boldsymbol{\phi}), \boldsymbol{S}(s_j|\boldsymbol{\phi}))$
- ▷ Compute reference drift $(\boldsymbol{f}_q)_{j,k} \leftarrow \boldsymbol{f}_q(\boldsymbol{x}_{j,k}, s_j|\boldsymbol{\phi})$ and degree ℓ Helmholtz correction $\widehat{\boldsymbol{h}}^{(\ell)}_{j,k} \leftarrow \widehat{\boldsymbol{h}}^{(\ell)}(\boldsymbol{x}_{j,k}, s_j)$ via Algorithm 2
- ▷ Set $(\boldsymbol{f}_{q,\text{full}})_{j,k} \leftarrow (\boldsymbol{f}_q)_{j,k} + \widehat{\boldsymbol{h}}^{(\ell)}_{j,k}$ and estimate $\widehat{\mathcal{L}}_{\text{path}}$
- ▷ Update $(\boldsymbol{\theta}, \boldsymbol{\phi})$ by a gradient step

end

Polynomial approximation. For $K \geq 2$, the vector space of q-divergence-free vector fields is infinite-dimensional. To obtain a tractable approximation to $\boldsymbol{h}$ in the weighted Helmholtz decomposition of $\boldsymbol{r}$, we restrict the optimization problem in eq. (6) to the finite-dimensional subspace of q-divergence-free vector fields whose components are polynomials of degree at most ℓ.

To do so, we first introduce a change of variable $\tilde{\boldsymbol{x}}$ such that the one-time marginal at time t is standard Gaussian. We then expand the residual vector field $\boldsymbol{r}(\boldsymbol{x}, t)$, transformed into the new coordinates as $\boldsymbol{\rho}(\tilde{\boldsymbol{x}}, t)$, in the basis of Hermite polynomials:

$$\boldsymbol{\rho}_i(\tilde{\boldsymbol{x}}, t) = \sum_{\boldsymbol{\alpha}} \widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha},i}(t) \hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}}), \quad \widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha},i}(t) = \mathbb{E}_{\tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_K)} \left[\boldsymbol{\rho}_i(\tilde{\boldsymbol{x}}, t) \hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}}) \right], \quad 1 \leq i \leq K. \tag{9}$$

The functions $\hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}})$ are the normalized multivariate probabilists' Hermite polynomials.

We also expand the q-divergence-free vector field, written in the new coordinates as $\boldsymbol{\nu}(\tilde{\boldsymbol{x}}, t)$, in the same basis, and we truncate this expansion to basis functions of degree at most ℓ. By orthogonality of the Hermite basis, the terms in eq. (6) that depend on the truncated vector field $\boldsymbol{\nu}^{(\ell)}(\tilde{\boldsymbol{x}}, t)$ involve only the coefficients of $\boldsymbol{\rho}(\tilde{\boldsymbol{x}}, t)$ up to degree ℓ. Therefore, we can obtain the optimal q-divergence-free vector field $\boldsymbol{h}^{(\ell)}$ within this subspace by computing the coefficients $\widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha},i}(t)$ up to degree ℓ, solving the objective for $\boldsymbol{\nu}^{(\ell)}(\tilde{\boldsymbol{x}}, t)$, and transforming back to $\boldsymbol{x}$. See Appendix B.3 for details.

In general, the Hermite coefficients in eq. (9) cannot be evaluated in closed-form. One can approximate $\widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha},i}(t)$ using Gauss–Hermite quadrature in low latent dimensions or with Monte Carlo. The resulting vector field $\widehat{\boldsymbol{h}}^{(\ell)}$ is q-divergence-free but not generally equal to $\boldsymbol{h}^{(\ell)}$. In Appendix B.3, we show that the complexity of computing $\widehat{\boldsymbol{h}}^{(\ell)}$ at a point $\boldsymbol{x}$ is $O(K^{\max(\ell+1,3)})$; for linear and quadratic approximations, the complexity is the same as with the fixed reference drift. Corrections with $\ell \geq 2$ can produce posteriors with Gaussian one-time marginals but non-Gaussian joint distributions.

Optimality of Helmholtz-SDE. When the prior is linear, the initial distribution is Gaussian, and the observation model is linear-Gaussian, the true posterior is a Gaussian process given by the Kalman–Bucy smoother [30]. Existing simulation-free VI algorithms commit to a single posterior drift for each $(\boldsymbol{m}(t|\boldsymbol{\phi}), \boldsymbol{S}(t|\boldsymbol{\phi}))_{0 \leq t \leq T}$ and, as we show in Appendix C.1, can incur an *unbounded KL gap to the true posterior* even in this conjugate setting. Helmholtz-SDE provably closes the gap:

Theorem 1 (Optimality of Helmholtz-SDE in the conjugate setting): Suppose the prior $\mathbb{P}_{\boldsymbol{\theta}}$ is linear with a Gaussian initial law and the observation model is linear-Gaussian. Let $\mathcal{Q}_{(1)}$ consist of laws from Helmholtz-SDE with ℓ=1 applied to continuously differentiable $\boldsymbol{m}(t), \boldsymbol{S}(t)$ with $\boldsymbol{S}(t)$ positive definite. Then $\inf_{\mathbb{Q} \in \mathcal{Q}_{(1)}} \mathrm{D}_{\mathrm{KL}}(\mathbb{Q} \parallel \mathbb{Q}^\star) = 0$, where $\mathbb{Q}^\star$ is the posterior.

Proof. In the conjugate setting, the residual $\boldsymbol{r}(\boldsymbol{x}, t)$ is affine. Gauss–Hermite quadrature with two nodes per latent dimension computes the degree-one Hermite coefficients, and hence the optimal correction, exactly; see Appendix B.5. □

Outside the conjugate setting, the true posterior generally does not have Gaussian one-time marginals, so restricting the variational family to diffusion processes with Gaussian one-time marginals incurs a

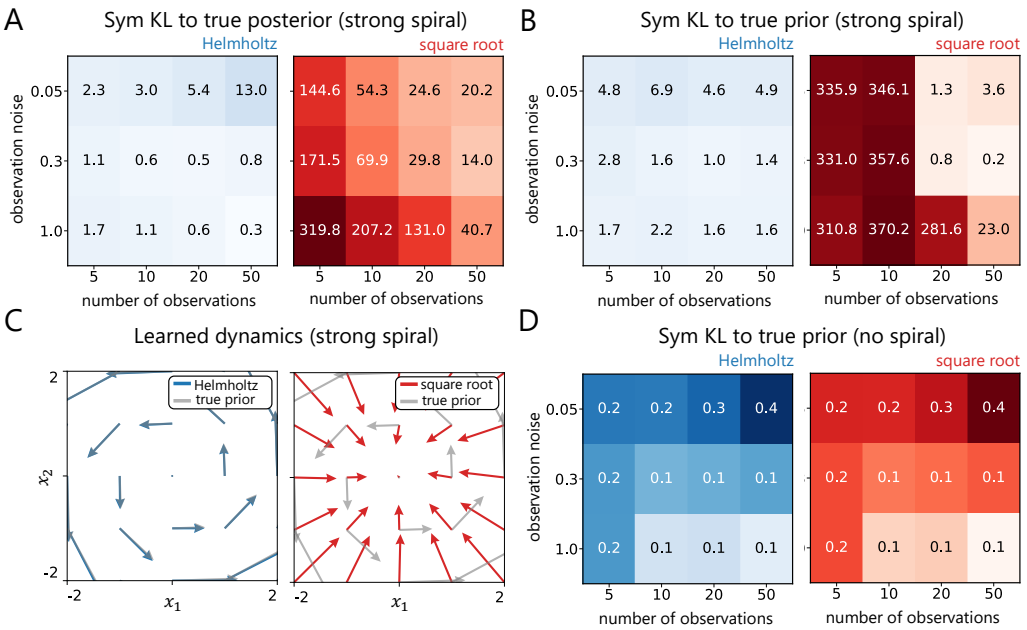


Figure 2: Comparison of Helmholtz-SDE and the square root gauge on two OU spiral datasets: strong spiral, $\omega{=}2\pi$ (**A**-**C**), and no spiral, $\omega{=}0$ (**D**). **A**: The symmetric KL distance (nats) to the posterior across observation noise and number of observations. **B**: The symmetric KL distance to the true prior. **C**: The dynamics learned by Helmholtz-SDE and the square root gauge against the true prior drift at noise $\sigma = 0.3$ and $N = 10$ observations. **D**: Same as **B**, but for the no spiral dataset.

nonzero KL gap. Nonetheless, when the Hermite coefficients are evaluated exactly, Helmholtz-SDE is guaranteed to perform at least as well as using the reference drift alone. Appendix B.5.2 provides an estimable lower bound on the resulting improvement.

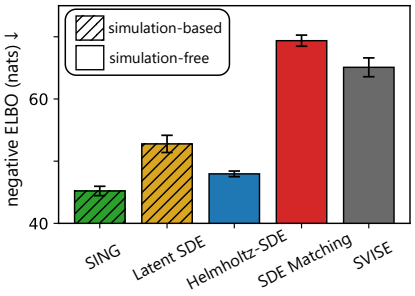
5 Related Work

Dynamical systems models. Beyond the VI algorithms for latent SDEs discussed in Section 3, several adjacent works model continuous-time, irregularly sampled data. Neklyudov et al. [31] and Zhang et al. [32] propose simulation-free methods inspired by score-based generative modeling [33–36], but model the process directly in observation space rather than via a latent SDE. Daems et al. [37] study latent SDEs driven by fractional Brownian motion. Kiyohara et al. [38] develop a VI algorithm permitting simulation-free sampling from both prior and posterior but with a mean-field variational family. A separate line of work replaces the SDE with a latent ODE [39–42]. None of these works address the path-law non-uniqueness arising from marginal-based posterior parameterizations.

Divergence-free constraints. Divergence-free vector fields and Helmholtz decompositions appear across machine learning in problems with physical or geometric structure: as parameterizations encoding continuity [43] or incompressibility [44, 45], and in methods that model or decompose non-gradient vector-field structure [46–50]. Our use is different: rather than imposing a divergence-free condition on the dynamics, we use the Helmholtz decomposition to characterize, at fixed one-time marginals, the family of compatible path laws and the optimal correction to a reference drift.

6 Experiments

We validate experimentally that Helmholtz-SDE shrinks the expressivity gap between simulation-free and simulation-based VI, with downstream improvements in dynamics learning. We begin with a linear, conjugate setting where the posterior is tractable and the gap can be quantified directly (Section 6.1), then move to nonlinear synthetic dynamics (Section 6.2) and two real-world systems: predator-prey cycles (Section 6.3) and reduced order fluid dynamics (Section 6.4).



	predictive LL ↑	sym KL ↓	autocorr error ↓
SING	-0.32 ± 0.03	0.24 ± 0.05	0.045 ± 0.005
Latent SDE	-0.23 ± 0.03	1.37 ± 0.24	0.151 ± 0.006
Helmholtz–SDE	$\mathbf{-0.09 \pm 0.00}$	$\mathbf{0.06 \pm 0.00}$	$\mathbf{0.024 \pm 0.002}$
SDE Matching	-0.28 ± 0.01	0.26 ± 0.01	0.057 ± 0.001
SVISE	-0.33 ± 0.01	0.35 ± 0.02	0.052 ± 0.002

Figure 3: Helmholtz-SDE closes the gap between simulation-based and simulation-free VI algorithms on the noisy Lorenz attractor. nELBO and posterior predictive LL at unobserved times are averaged over trials; sym KL is averaged over one-time marginals. Autocorr error is the entrywise RMSE of the global 3×3 autocorrelation matrix, over lags $\tau \in [0, 2]$. Errors are $\pm$2SE over 10 replicates.

Throughout, we match implementation choices across methods. In particular, existing VI algorithms differ in how they amortize the posterior approximation: across time, by representing it as a function of t, and across trials, by using an encoder to produce trial-specific parameters. Both strategies reduce memory at some cost to expressivity. We use the same amortization scheme whenever a method supports it. For Helmholtz-SDE we use the degree $\ell = 1$ polynomial approximation in all experiments; we provide ablations of the polynomial degree in Appendix D. Full experimental details are in Appendices C and D. We provide runtime comparisons against all baselines in Appendix E.

6.1 Ornstein–Uhlenbeck (OU) spiral

We first study the impact of augmenting the variational family in the conjugate setting, where the suboptimality of both the variational posterior and the learned prior can be analyzed directly. We consider the two-dimensional OU spiral

$$\mathbb{P}\colon\ d\boldsymbol{x}(t) = (-\alpha\mathbb{I}_2 + \omega\boldsymbol{J})\boldsymbol{x}(t)dt + d\boldsymbol{w}(t), \quad p(\boldsymbol{x}, 0) = \mathcal{N}\left(\mathbf{0}, \tfrac{1}{2\alpha}\mathbb{I}_2\right), \quad \boldsymbol{J} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right) \tag{10}$$

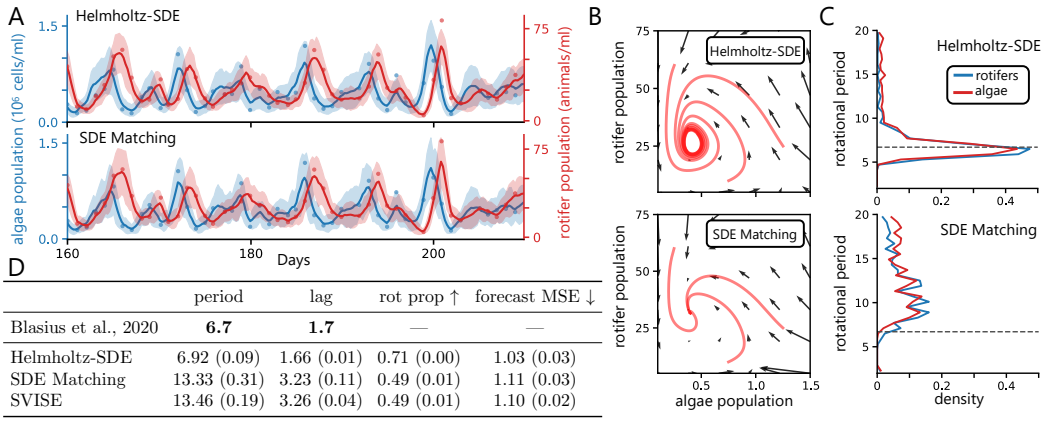
whose drift decomposes into a contracting gradient field $-\alpha\boldsymbol{x}$ and a p-divergence-free rotational field $\omega\boldsymbol{J}\boldsymbol{x}$. Observations follow the linear-Gaussian model $p(\boldsymbol{y}(t_i)|\boldsymbol{x}(t_i)) = \mathcal{N}(\boldsymbol{C}\boldsymbol{x}(t_i) + \boldsymbol{d}, \sigma^2\boldsymbol{C}\boldsymbol{C}^\top)$.

Suboptimality of inference. Theorem 1 establishes that Helmholtz-SDE has zero KL gap to the posterior in this setting. In Appendix C.1, we show that the choices of variational drift made by SDE Matching [8] and SVISE [6, 7], which we refer to as the "square root" and "symmetric" gauges, respectively, incur a KL gap that grows at least linearly in ω. We validate this in simulation. Fixing $T_{\max} = 5$ and $\alpha = 0.2$, we vary observation noise σ, number of observations N, and rotational speed ω. Figure 2A shows that when prior rotation is strong, Helmholtz-SDE achieves a variational approximation orders of magnitude closer in symmetric KL to the true posterior than the square root gauge, with the largest gains at sparse observations and large noise. When $\omega = 0$ (Figure 10), the posterior drift contains no divergence-free component, in which case the square root gauge can represent it exactly. Results for the symmetric gauge and $\omega \in \{\pi/4, \pi/2, \pi\}$ are in Appendix C.2.

Degradation in learning. We next test whether posterior misspecification degrades learning. Using the same grid of (σ, N, ω) values, we jointly learn the prior (parameterized by a neural network drift) and observation model from 1024 independent trials. Figure 2B and D show that the regimes where the square root gauge yields the worst posterior approximation are also where prior learning degrades most. In many cases, the divergence-free correction is the difference between recovering the correct drift and learning an incorrect one (Figure 2C). However, consistent with previous work [51], better variational approximations do not translate monotonically to improvements in learning.

6.2 Noisy Lorenz attractor

In the following three subsections, we demonstrate that the advantages from Section 6.1 extend to nonlinear dynamics. We first consider Lorenz attractor dynamics, a classical chaotic system. The standard setup of Li et al. [2] uses small process noise and dense, low-noise observations; in this regime the posterior concentrates tightly around the true latent trajectory (Figure 11A). To simulate a



D

	period	lag	rot prop ↑	forecast MSE ↓
Blasius et al., 2020	**6.7**	**1.7**	—	—
Helmholtz-SDE	6.92 (0.09)	1.66 (0.01)	0.71 (0.00)	1.03 (0.03)
SDE Matching	13.33 (0.31)	3.23 (0.11)	0.49 (0.01)	1.11 (0.03)
SVISE	13.46 (0.19)	3.26 (0.04)	0.49 (0.01)	1.10 (0.02)

Figure 4: Helmholtz-SDE recovers predator-prey dynamics more faithfully than SDE Matching. **A**: Posterior mean ± 2 SD over a representative window. **B**: Phase portraits of learned dynamics: Helmholtz-SDE recovers a limit cycle, SDE Matching decays to a fixed point. **C**: Peak-period distribution across 100-day forward simulations from the learned prior; dashed line marks the empirical period of 6.7 days. **D**: Rotation period, predator lag, and proportion of time in coherent rotation computed per replicate from forward simulations using the wavelet analysis of Blasius et al. [52]; forecast MSE, in log coordinates. Values are median (IQR) across 10 replicates.

setting with non-trivial posterior uncertainty, we sample latents on $[0, 5]$ with increased process noise and observe them every $\Delta t = 0.25$ (Figure 11B). We perform joint inference and learning using 1024 trials with $K{=}4$ latent dimensions, evaluating both one-time (symmetric KL to the true prior) and multi-time (lagged autocorrelation error) statistics of the prior.

We compare Helmholtz-SDE to simulation-free (SDE Matching [8], SVISE [6, 7]) and simulation-based (SING [15], Latent SDE [2]) baselines (Figure 3). Among simulation-free methods, Helmholtz-SDE achieves a $\sim$20 nat smaller negative ELBO than the alternatives, reflecting the inability of a pre-specified drift to represent the posterior path law. As a result, only Helmholtz-SDE accurately learns the outward-spiraling dynamics within each lobe of the attractor (Figure 12). Helmholtz-SDE also matches or exceeds simulation-based methods in dynamics recovery. SING attains a slightly tighter nELBO but at substantially higher cost: on an H100 GPU, Helmholtz-SDE reaches per-trial nELBO below 50 nat in 123.7 ± 3.4 s, whereas SING requires 3695.6 ± 503.8 s, over $25\times$ slower.

6.3 Predator-prey system

Predator-prey systems exhibit cyclic dynamics in which the predator population lags the prey, and recovering this lag structure is important to ecological interpretation [53–57]. Because the lag is a multi-time statistic, it directly probes the path law approximation made by simulation-free algorithms.

We study the dataset of Blasius et al. [52], who cultivate rotifers (*B. calyciflorus*, a freshwater zooplankton) feeding on algae in a controlled environment and measure both populations daily for up to 374 days. The system exhibits persistent cycles with a mean period of 6.7 days, in which the rotifer population lags the algae by approximately 1.7 days. Rather than imposing a biological model on dynamics [58], we aim to learn a generative model of the dynamics directly from data.

To this end, we consider the longest trial and restrict to windows of coherent oscillation following Blasius et al. [52], yielding 255 observations across four subtrials. We hold out the last eight observations of each ($\sim$1 cycle) for forecasting. We then fit Helmholtz-SDE, SDE Matching, and SVISE on the log-transformed rotifer and algae populations with $K{=}2$ latent dimensions.

All methods produce plausible posterior approximations over the observed populations (Figure 4A). However, the learned generative dynamics differ qualitatively: Helmholtz-SDE yields a stable limit cycle, while SDE Matching decays to a fixed point (Figure 4B). Moreover, the rotational dynamics learned by Helmholtz-SDE closely match the true biological system, with median period 6.92 days and predator-prey lag 1.66 days, both close to the estimates from Blasius et al. [52] (Figure 4C,D). SDE Matching, in contrast, recovers a period twice as long (13.33 days) and a correspondingly inflated lag (3.23 days). We provide details on our wavelet analysis in Appendix D.2.

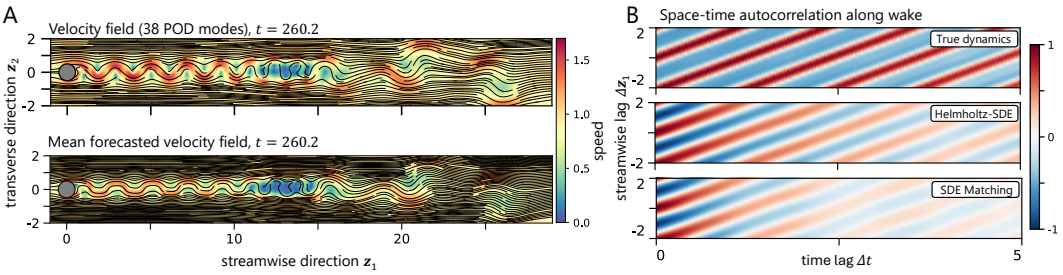


Figure 5: Helmholtz-SDE learns coherent wake dynamics in high-dimensional fluid flow past a 2D cylinder. **A**: True and forecasted mean velocity fields at 4 time units into the forecast window. **B**: Joint space–time autocorrelation along the streamwise direction in the vortex region $[1, 10] \times [-0.1, 0.1]$, from prior samples at the true t=0 initial condition. Diagonal banding indicates vortices being advected downstream at a coherent speed; rapid decay reflects loss of temporal coherence.

6.4 Reduced order models of fluid dynamics

As a final case study, we apply Helmholtz-SDE to a problem in computational fluid dynamics: modeling a two-dimensional incompressible flow past a circular cylinder at Reynolds number $2 \cdot 10^3$, a regime with strong, persistent vortices in the wake (Figure 5A). We use the dataset of Course and Nair [6], in which the Navier–Stokes equations are discretized and solved on a grid of nearly $6 \cdot 10^5$ points using the immersed boundary projection method [59].

Reduced order modeling (ROM) provides a toolkit for representing such high-dimensional systems with low-dimensional surrogates. Classical projection-based approaches require explicit access to the discretized PDE [60]. Data-driven approaches instead learn reduced dynamics directly from snapshots of the time series, sidestepping projection of the PDE. Following Course and Nair [6], we apply dimensionality reduction (POD, [61, 62]) to velocity snapshots and retain $K = 38$ modes capturing 90% of the kinetic energy. We observe each POD coefficient, perturbed by Gaussian noise, every $\Delta t = 0.3$ time units over an interval of length 195.2. We hold out 10 time units for forecasting.

Helmholtz-SDE successfully learns a ROM of the cylinder dynamics (Figure 5A). The space–time autocorrelation along the wake reveals that, compared to SDE Matching, Helmholtz-SDE better recovers the traveling-wave structure of vortex advection (Figure 5B). This translates to an approximately 10% reduction in cumulative MSE over SDE Matching and SVISE over the forecast window, at modest additional training cost (approx. $3\times$ the per-iteration cost of SVISE on H100 GPU). This example demonstrates that even for moderate latent dimensions K, our approximation to the Helmholtz decomposition remains tractable and yields improvements in dynamics learning.

7 Discussion

Simulation-free variational inference promises the accuracy of simulation-based algorithms at a fraction of the cost, but only when the variational family is rich enough to represent the posterior. We have shown that under high posterior uncertainty, existing simulation-free algorithms fall short of their promise: by committing to an arbitrary drift consistent with their one-time marginals, they restrict the posterior to a subset of the path laws available to simulation-based methods. Helmholtz-SDE removes this restriction. Starting from any reference drift, it adds a closed-form correction that preserves the prescribed marginals while optimizing over a larger family of path laws. The result is a simulation-free algorithm that closes the expressivity gap to simulation-based VI and yields more faithful recovery of the underlying dynamics.

Two limitations point to natural extensions. First, the analytical form for the Helmholtz correction depends on the state independence of the diffusion coefficient and Gaussianity of the marginals. Relaxing either assumption, for example by learning the divergence-free correction, would broaden the algorithm's applicability. Second, higher-order polynomial corrections to the residual scale unfavorably as the latent dimension grows; in our experiments we did not observe gains beyond the linear and quadratic approximations, but more efficient parameterizations may be beneficial in settings we have not explored. More broadly, we view simulation-free VI as a promising path toward scalable unsupervised discovery of dynamical systems, and Helmholtz-SDE as a step toward making that path competitive with, and ultimately preferable to, its simulation-based counterpart.

Acknowledgments and Disclosure of Funding

HDS is supported by the NSF Graduate Research Fellowship (DGE-2146755) and the Knight–Hennessy graduate fellowship. HDS and SWL are supported by grants from the NIH BRAIN Initiative (U19NS113201, R01NS131987, R01NS113119, & RF1MH133778), the NSF/NIH CRCNS Program (R01NS130789), and the Sloan, Simons, and McKnight Foundations. The authors have no competing interests to declare.

We thank Amber Hu, Cole Citrenbaum, Hannah Lee, Joshua Lunger, and Renato Berlinghieri for their helpful feedback on the manuscript. We thank Peter Pao–Huang for discussions at the beginning of the project regarding divergence-free vector fields.

References

[1] Cédric Archambeau, Dan Cornford, Manfred Opper, and John Shawe-Taylor. Gaussian process approximations of stochastic differential equations. In *Gaussian Processes in Practice*, 2007.

[2] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David K Duvenaud. Scalable gradients and variational inference for stochastic differential equations. In *Symposium on Advances in Approximate Bayesian Inference*, 2020.

[3] Ahmed ElGazzar and Marcel van Gerven. Universal differential equations as a common modeling language for neuroscience. *arXiv preprint arXiv:2403.14510*, 2024.

[4] Amber Hu, David Zoltowski, Aditya Nair, David Anderson, Lea Duncker, and Scott Linderman. Modeling latent neural dynamics with Gaussian process switching linear dynamical systems. *Advances in Neural Information Processing Systems*, 2024.

[5] Joshua Durso-Finley, Berardino Barile, Jean-Pierre Falet, Douglas L Arnold, Nick Pawlowski, and Tal Arbel. Probabilistic temporal prediction of continuous disease trajectories and treatment effects using neural SDEs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024.

[6] Kevin Course and Prasanth B Nair. State estimation of a physical system with unknown governing equations. *Nature*, 2023.

[7] Kevin Course and Prasanth Nair. Amortized reparametrization: Efficient and scalable variational inference for latent SDEs. *Advances in Neural Information Processing Systems*, 2023.

[8] Grigory Bartosh, Dmitry Vetrov, and Christian A Naesseth. SDE Matching: scalable and simulation-free training of latent stochastic differential equations. In *International Conference on Machine Learning*, 2025.

[9] Harold J Kushner. On the differential equations satisfied by conditional probability densities of Markov processes, with applications. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1964.

[10] Ruslan Leont'evich Stratonovich. Conditional Markov processes. In *Non-linear Transformations of Stochastic Processes*. Elsevier, 1965.

[11] Etienne Pardoux. Équations du filtrage non linéaire de la prédiction et du lissage. *Stochastics*, 1982.

[12] Cédric Archambeau and Manfred Opper. Approximate inference for continuous-time Markov processes. *Bayesian Time Series Models*, 2011.

[13] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 1999.

[14] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 2017.

[15] Amber Hu, Henry Smith, and Scott Linderman. SING: SDE inference via natural gradients. *Advances in Neural Information Processing Systems*, 2025.

[16] Cédric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, and John Shawe-Taylor. Variational inference for diffusion processes. *Advances in Neural Information Processing Systems*, 2007.

[17] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable continuous-time models of latent stochastic dynamical systems. In *International Conference on Machine Learning*, 2019.

[18] Prakhar Verma, Vincent Adam, and Arno Solin. Variational Gaussian process diffusion processes. In *International Conference on Artificial Intelligence and Statistics*, 2024.

[19] Lukas Köhs, Bastian Alt, and Heinz Koeppl. Variational inference for continuous-time switching dynamical systems. *Advances in Neural Information Processing Systems*, 2021.

[20] Christian Wildner and Heinz Koeppl. Moment-based variational inference for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, 2021.

[21] Byoungwoo Park, Hyungi Lee, and Juho Lee. Amortized control of continuous state space Feynman-Kac model for irregular time series. In *International Conference on Learning Representations*, 2025.

[22] Joonhyeong Park, Byoungwoo Park, Chang-Bae Bang, Jungwon Choi, Hyungjin Chung, Byung-Hoon Kim, and Juho Lee. Stochastic optimal control for continuous-time fMRI representation learning. In *International Conference on Learning Representations*, 2026.

[23] Tom Ryder, Andrew Golightly, A Stephen McGough, and Dennis Prangle. Black-box variational inference for stochastic differential equations. In *International Conference on Machine Learning*, 2018.

[24] Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent Gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019.

[25] Patrick Kidger, James Foster, Xuechen (Chen) Li, and Terry Lyons. Efficient and accurate gradients for neural SDEs. *Advances in Neural Information Processing Systems*, 2021.

[26] YongKyung Oh, Dong-Young Lim, and Sungil Kim. Stable neural stochastic differential equations in analyzing irregular time series data. In *International Conference on Learning Representations*, 2024.

[27] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

[28] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[29] Harsh Bhatia, Gregory Norgard, Valerio Pascucci, and Peer-Timo Bremer. The Helmholtz-Hodge decomposition—a survey. *IEEE Transactions on Visualization and Computer Graphics*, 2013.

[30] Rudolph E Kalman and Richard S Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 1961.

[31] Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *International Conference on Machine Learning*, 2023.

[32] Xi Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis L Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modelling. *Advances in Neural Information Processing Systems*, 2024.

[33] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2005.

[34] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[35] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

[36] Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 2025.

[37] Rembert Daems, Manfred Opper, Guillaume Crevecoeur, and Tolga Birdal. Variational inference for SDEs driven by fractional noise. In *International Conference on Learning Representations*, 2024.

[38] Naoki Kiyohara, Edward Johns, and Yingzhen Li. Neural stochastic flows: Solver-free modelling and inference for SDE solutions. *Advances in Neural Information Processing Systems*, 2025.

[39] Markus Heinonen, Cagatay Yildiz, Henrik Mannerström, Jukka Intosalmi, and Harri Lähdesmäki. Learning unknown ODE models with Gaussian processes. In *International Conference on Machine Learning*, 2018.

[40] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 2018.

[41] Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in Neural Information Processing Systems*, 2019.

[42] Cagatay Yildiz, Markus Heinonen, and Harri Lahdesmaki. ODE^{2}VAE: Deep generative second order ODEs with Bayesian neural networks. *Advances in Neural Information Processing Systems*, 2019.

[43] Jack Richter-Powell, Yaron Lipman, and Ricky TQ Chen. Neural conservation laws: A divergence-free perspective. *Advances in Neural Information Processing Systems*, 35, 2022.

[44] Jonathan Tompson, Kristofer Schlachter, Pablo Sprechmann, and Ken Perlin. Accelerating Eulerian fluid simulation with convolutional networks. In *International Conference on Machine Learning*, 2017.

[45] Xigui Li, Hongwei Zhang, Ruoxi Jiang, Deshu Chen, Chensen Lin, Limei Han, Yuan Qi, Xin Guo, and Yuan Cheng. Project and generate: Divergence-free neural operators for incompressible flows. *arXiv preprint arXiv:2603.24500*, 2026.

[46] Renato Berlinghieri, Brian L Trippe, David R Burt, Ryan James Giordano, Kaushik Srinivasan, Tamay Özgökmen, Junfei Xia, and Tamara Broderick. Gaussian processes at the Helm(holtz): A more fluid model for ocean currents. In *International Conference on Machine Learning*, 2023.

[47] Hao Xu and Jia Pan. HHD-GP: Incorporating Helmholtz-Hodge decomposition into Gaussian processes for learning dynamical systems. *Advances in Neural Information Processing Systems*, 37, 2024.

[48] Miao Qi, Ramzi Idoughi, and Wolfgang Heidrich. HDNet: physics-inspired neural network for flow estimation based on Helmholtz decomposition. *arXiv preprint arXiv:2406.08570*, 2024.

[49] Christian Horvat and Jean-Pascal Pfister. On gauge freedom, conservativity and intrinsic dimensionality estimation in diffusion models. In *International Conference on Learning Representations*, 2024.

[50] Katarina Petrović, Lazar Atanackovic, Viggo Moro, Kacper Kapuśniak, Ismail Ilkan Ceylan, Michael Bronstein, Avishek Joey Bose, and Alexander Tong. Curly flow matching for learning non-gradient field dynamics. *Advances in Neural Information Processing Systems*, 2025.

[51] Richard Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. *Bayesian Time Series Models*, 2011.

[52] Bernd Blasius, Lars Rudolf, Guntram Weithoff, Ursula Gaedke, and Gregor F Fussmann. Long-term cyclic persistence in an experimental predator–prey system. *Nature*, 2020.

[53] Alfred James Lotka. *Elements of physical biology*. Williams & Wilkins, 1925.

[54] Vito Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 1926.

[55] Charles Elton and Mary Nicholson. The ten-year cycle in numbers of the lynx in Canada. *The Journal of Animal Ecology*, 1942.

[56] Syunro Utida. Cyclic fluctuations of population density intrinsic to the host-parasite system. *Ecology*, 1957.

[57] Alan Berryman. *Population cycles: the case for trophic interactions*. Oxford University Press, 2002.

[58] Edward McCauley, Roger M Nisbet, Andre M De Roos, William W Murdoch, and William SC Gurney. Structured population models of herbivorous zooplankton. *Ecological Monographs*, 1996.

[59] Kunihiko Taira and Tim Colonius. The immersed boundary method: a projection approach. *Journal of Computational Physics*, 2007.

[60] Steven L Brunton, Bernd R Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics*, 2020.

[61] Lawrence Sirovich. Turbulence and the dynamics of coherent structures. Parts I–III. *Quarterly of Applied Mathematics*, 1987.

[62] Gal Berkooz, Philip Holmes, and John L Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 1993.

[63] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer Science & Business Media, 2013.

[64] Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*. Springer, 2014.

[65] Andrew D Barbour. Stein's method for diffusion approximations. *Probability Theory and Related Fields*, 1990.

[66] Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 2019.

[67] Qiang Liu, Lester Mackey, and Chris Oates. Probabilistic inference and learning with Stein's method. *arXiv preprint arXiv:2603.07467*, 2026.

[68] Etienne Pardoux and A Yu Veretennikov. On the Poisson equation and diffusion approximation. *The Annals of Probability*, 2005.

[69] David Gilbarg and Neil S Trudinger. *Elliptic partial differential equations of second order*. Springer, 1998.

[70] Rafail Khasminskii. *Stochastic stability of differential equations*. Springer Science & Business Media, 2012.

[71] Dario Trevisan. Well-posedness of multidimensional diffusion processes with weakly differentiable coefficients. *Electronic Journal of Probability*, 2016.

[72] Simo Särkkä and Lennart Svensson. *Bayesian filtering and smoothing*. Cambridge University Press, 2023.

[73] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[74] Xuerong Mao, Glenn Marion, and Eric Renshaw. Environmental Brownian noise suppresses explosions in population dynamics. *Stochastic Processes and their Applications*, 2002.

[75] Christopher Torrence and Gilbert P Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 1998.

Appendix

Appendix A: Background on variational inference for latent SDEs

- A.1: Technical assumptions
- A.2: ELBO derivation
- A.3: Divergence-free vector fields
- A.4: Gaussian one-time marginals
- A.5: SVISE (Course and Nair [6]) and SDE-Matching (Bartosh et al. [8])

Appendix B: Helmholtz-SDE

- B.1: Computing the Helmholtz decomposition
- B.2: The Helmholtz decomposition for Gaussian marginals
- B.3: Polynomial approximation to the optimal correction
- B.4: Verification of assumptions for the Helmholtz-SDE variational family
- B.5: Optimality of Helmholtz-SDE
  - B.5.1: Conjugate setting
  - B.5.2: Nonconjugate setting
- B.6: Algorithm summary

Appendix C: Linear SDE experiments: Ornstein–Uhlenbeck spiral

- C.1: Theoretical results for OU spiral
- C.2: Simulation experiment details

Appendix D: Nonlinear SDE experiments

- D.1: Noisy Lorenz attractor
- D.2: Experimental predator-prey system
- D.3: Reduced order models for fluid dynamics

Appendix E: Runtime comparisons

Notation. Throughout the Appendix, for ease of notation, we drop the dependence of the prior and likelihood on the model parameters $\boldsymbol{\theta}$, as well as the dependence of the variational posterior on its parameters $\boldsymbol{\phi}$.

A Background on variational inference for latent SDEs

In this section, we provide background on variational inference for latent SDEs. First, in Appendix A.1 we describe our formal assumptions necessary for the probabilistic model and variational inference objective to be well-posed. In Appendix A.2 we derive the evidence lower bound (ELBO) on the observed data log likelihood. In Appendix A.3 we provide further discussion of divergence-free vector fields and prove Proposition 1. In Appendix A.4 we characterize the affine divergence-free vector fields for Gaussian one-time marginals (Proposition 2). We conclude in Appendix A.5 by discussing two simulation-free VI algorithms, SVISE (Course and Nair [6, 7]) and SDE Matching (Bartosh et al. [8]). For each algorithm we derive the choice of posterior drift and discuss how this choice relates to Helmholtz-SDE.

For full generality, we allow the prior drift $\boldsymbol{f}_p$ to be time-dependent and for the diffusion coefficient $\boldsymbol{\Sigma}^{1/2}$ to be both space- and time-dependent.

A.1 Technical Assumptions

In this subsection, we detail our formal assumptions on the prior (1) and posterior (2) processes. The assumptions are stated abstractly here and are verified for the variational family considered by Helmholtz-SDE in Appendix B.4. They are required for every admissible choice of model parameters $\boldsymbol{\theta}$ and variational parameters $\boldsymbol{\phi}$. Here, $\boldsymbol{f}_q$ denotes the *full* posterior drift in eq. (2); for Helmholtz-SDE, this is the sum of the reference drift and the Helmholtz correction.

Existence and uniqueness. We first require that the prior and posterior SDEs, eq. (1) and eq. (2), respectively, admit unique strong solutions.

Assumption (A1) (SDE existence and uniqueness): The prior and posterior SDEs admit unique strong solutions on $[0, T]$.

For the prior SDE, we follow the classical route of imposing Lipschitz continuity and linear growth on the drift and diffusion coefficient (see, e.g., Øksendal [63, Theorem 5.2.1]): there exist constants $C_1, C_2 > 0$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^K$ and $t \in [0, T]$

$$\|\boldsymbol{f}_p(\boldsymbol{x}, t) - \boldsymbol{f}_p(\boldsymbol{y}, t)\|_2 + \|\boldsymbol{\Sigma}(\boldsymbol{x}, t)^{1/2} - \boldsymbol{\Sigma}(\boldsymbol{y}, t)^{1/2}\|_F \leq C_1 \|\boldsymbol{x} - \boldsymbol{y}\|_2, \qquad \text{(Lipschitz)}$$

$$\|\boldsymbol{f}_p(\boldsymbol{x}, t)\|_2 + \|\boldsymbol{\Sigma}(\boldsymbol{x}, t)^{1/2}\|_F \leq C_2(1 + \|\boldsymbol{x}\|_2). \qquad \text{(linear growth)}$$

For example, if $\boldsymbol{f}_p$ is a neural network that takes as input $\boldsymbol{x}$ and t, then Lipschitz continuity and linear growth can be ensured by restricting the operator norm of the weight matrices and choosing (globally) Lipschitz activation functions.

For the posterior SDE, the polynomial approximation to the Helmholtz correction need not satisfy linear growth when the polynomial degree is $\ell > 1$. We will therefore pursue an alternate route to ensure existence and uniqueness.

Densities. We additionally require that, for every $0 \leq t \leq T$, the one-time marginals of $\mathbb{P}$ and of any $\mathbb{Q} \in \mathcal{Q}$ are absolutely continuous with respect to the K-dimensional Lebesgue measure. The corresponding densities are written as $\{p(\boldsymbol{x}, t)\}_{0 \leq t \leq T}$ and $\{q(\boldsymbol{x}, t)\}_{0 \leq t \leq T}$, respectively.

Assumption (A2) (Marginal densities): For every $t \in [0, T]$, the one-time marginals of $\mathbb{P}$ and $\mathbb{Q}$ admit Lebesgue densities. Moreover, $\boldsymbol{\Sigma}(\boldsymbol{x}, t)$ is positive-definite for all $(\boldsymbol{x}, t) \in \mathbb{R}^K \times [0, T]$.

Absolute continuity. In order for the upper bound to the negative marginal log likelihood (negative ELBO) to be well-defined, we need the approximate posterior SDE $\mathbb{Q} \in \mathcal{Q}$ to possess a probability density with respect to the prior $\mathbb{P}$ and for the KL divergence between $\mathbb{Q}$ and $\mathbb{P}$ to be finite. These two requirements are the subjects of the following two assumptions.

Assumption (A3) (Mutual absolute continuity): The prior $\mathbb{P}$ and the variational posterior eq. (2) path laws are mutually absolutely continuous. Denote the density by $d\mathbb{Q}/d\mathbb{P}(\boldsymbol{x})$.

Assumption (A4) (Finite relative entropy): The initial distributions satisfy

$$\mathrm{D}_{\mathrm{KL}}\left(q(\boldsymbol{x},0)\parallel p(\boldsymbol{x},0)\right)<\infty$$

and the difference between the prior and posterior drifts satisfies

$$\int_0^T \mathbb{E}_{\mathbb{Q}}\|\boldsymbol{\Sigma}(\boldsymbol{x}(t),t)^{-1/2}\{\boldsymbol{f}_p(\boldsymbol{x}(t),t)-\boldsymbol{f}_q(\boldsymbol{x}(t),t)\}\|_2^2 dt<\infty.$$

Under these assumptions, we can obtain an expression for the probability density of $\mathbb{Q}$ with respect to $\mathbb{P}$ via Girsanov's theorem (see e.g., Karatzas and Shreve [64, Corollary 3.5.13]).

Theorem 2 (Girsanov's theorem for SDEs): Suppose Assumptions (A1) to (A4) hold. Then the probability density of $\mathbb{Q}\in\mathcal{Q}$ with respect to $\mathbb{P}$ is

$$\frac{d\mathbb{Q}}{d\mathbb{P}}(\boldsymbol{x})=\frac{q(\boldsymbol{x}(0),0)}{p(\boldsymbol{x}(0),0)}\exp\left\{\int_0^T\sum_{i=1}^K\left[\boldsymbol{\Sigma}^{-1/2}\boldsymbol{b}(\boldsymbol{x}(t),t)\right]_i d\boldsymbol{w}_i(t)-\frac{1}{2}\int_0^T\sum_{i=1}^K\left[\boldsymbol{\Sigma}^{-1/2}\boldsymbol{b}(\boldsymbol{x}(t),t)\right]_i^2 dt\right\},$$

$\boldsymbol{b}(\boldsymbol{x},t)=\boldsymbol{f}_q(\boldsymbol{x},t)-\boldsymbol{f}_p(\boldsymbol{x},t)$,

where $(\boldsymbol{w}(t))_{0\le t\le T}$ is a K-dimensional Brownian motion with respect to $\mathbb{P}$.

In particular, for any bounded, measurable function Φ defined on K-dimensional continuous functions $\boldsymbol{x}$ having domain $[0,T]$,

$$\mathbb{E}_{\mathbb{Q}}[\Phi(\boldsymbol{x})]=\mathbb{E}_{\mathbb{P}}\left[\Phi(\boldsymbol{x})\frac{d\mathbb{Q}}{d\mathbb{P}}(\boldsymbol{x})\right].$$

As a direct corollary of Theorem 2, we obtain an expression for the KL divergence between the posterior and prior

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}\left(\mathbb{Q}\parallel\mathbb{P}\right)&=\mathbb{E}_{\mathbb{Q}}\left[\log\left(\frac{d\mathbb{Q}}{d\mathbb{P}}(\boldsymbol{x})\right)\right]\\&=\mathrm{D}_{\mathrm{KL}}\left(q(\boldsymbol{x},0)\parallel p(\boldsymbol{x},0)\right)+\frac{1}{2}\int_0^T\mathbb{E}_{\mathbb{Q}}\|\boldsymbol{\Sigma}(\boldsymbol{x}(t),t)^{-1/2}\{\boldsymbol{f}_p(\boldsymbol{x}(t),t)-\boldsymbol{f}_q(\boldsymbol{x}(t),t)\}\|_2^2 dt.\end{aligned}\tag{11}$$

The stochastic integral term vanishes because, when expressed using the $\mathbb{Q}$-Brownian motion, it is a mean-zero martingale.

Fokker–Planck equation. For a drift $\boldsymbol{f}$ and initial density $\rho(\boldsymbol{x},0)$, we call $\{\rho(\boldsymbol{x},t)\}_{0\le t\le T}$ a *weak solution* to the Fokker–Planck equation if for every infinitely differentiable function $\varphi:\mathbb{R}^K\to\mathbb{R}$ having compact support,

$$t\longmapsto\int\varphi(\boldsymbol{x})\rho(d\boldsymbol{x},t)$$

is absolutely continuous and

$$\frac{d}{dt}\int\varphi(\boldsymbol{x})\rho(d\boldsymbol{x},t)=\int\left(\boldsymbol{f}(\boldsymbol{x},t)\cdot\nabla_{\boldsymbol{x}}\varphi(\boldsymbol{x})+\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}(\boldsymbol{x},t)\nabla_{\boldsymbol{x}}^2\varphi(\boldsymbol{x}))\right)\rho(d\boldsymbol{x},t),\quad\text{a.e. }t.\tag{12}$$

Since $\mathbb{Q}\in\mathcal{Q}$ is a solution to the SDE eq. (2), we know that $\{q(\boldsymbol{x},t)\}_{0\le t\le T}$ is a weak Fokker–Planck solution. However, in order to argue that two variational SDEs whose drifts differ by a divergence-free vector field induce the same marginals, we need the solution to the Fokker–Planck equation to be unique within a suitable regularity class.

Assumption (A5) (Uniqueness of the Fokker–Planck solution): For every member $\mathbb{Q}\in\mathcal{Q}$ of the variational family having drift $\boldsymbol{f}_q$, the marginal densities belong to the class

$$\int_0^T\int\left(\|\boldsymbol{f}_q(\boldsymbol{x},t)\|_2+\|\boldsymbol{\Sigma}(\boldsymbol{x},t)\|_F\right)\rho(d\boldsymbol{x},t)dt<\infty.\tag{13}$$

The associated weak Fokker–Planck equation with initial law $q(\boldsymbol{x},0)$ has at most one solution belonging to the class eq. (13).

A.2 ELBO derivation

In this subsection, we derive the evidence lower bound (ELBO) in the latent SDE model.

Adopting the notation from Theorem 2, we have

$$\begin{aligned}
&\log p(\boldsymbol{y}) = \log \int p(\boldsymbol{y}|\boldsymbol{x})\mathbb{P}(d\boldsymbol{x}) \\
&= \log \int p(\boldsymbol{y}|\boldsymbol{x}) \frac{d\mathbb{P}}{d\mathbb{Q}}(\boldsymbol{x})\mathbb{Q}(d\boldsymbol{x}) \\
&\overset{(i)}{\geq} \int \log \left(p(\boldsymbol{y}|\boldsymbol{x}) \frac{d\mathbb{P}}{d\mathbb{Q}}(\boldsymbol{x}) \right) \mathbb{Q}(d\boldsymbol{x}) \\
&= \mathbb{E}_{\mathbb{Q}}[\log p(\boldsymbol{y}|\boldsymbol{x})] - \mathrm{D}_{\mathrm{KL}}\left(\mathbb{Q} \parallel \mathbb{P}\right) \\
&\overset{(ii)}{=} \mathbb{E}_{\mathbb{Q}}[\log p(\boldsymbol{y}|\boldsymbol{x})] - \mathrm{D}_{\mathrm{KL}}\left(q(\boldsymbol{x},0) \parallel p(\boldsymbol{x},0)\right) - \frac{1}{2}\int_0^T \mathbb{E}_{\mathbb{Q}} \|\boldsymbol{\Sigma}^{-1/2}\{\boldsymbol{f}_p(\boldsymbol{x}(t),t) - \boldsymbol{f}_q(\boldsymbol{x}(t),t)\}\|_2^2 dt.
\end{aligned}$$

Inequality (i) is an application of Jensen's inequality and equality (ii) is an application of Girsanov's theorem. Taking the negative of this expression yields the negative ELBO eq. (3).

Notice that each term in the negative ELBO depends both on the parameters of the generative model $\boldsymbol{\theta}$ and on the parameters of the variational posterior $\boldsymbol{\phi}$.

A.3 Divergence-free vector fields

In this subsection, we discuss the role of q-divergence-free vector fields in determining the expressivity of the variational family $\mathcal{Q}$. We prove Proposition 1, which characterizes the non-uniqueness in path law that arises from marginal-based parameterizations of the variational posterior.

In eq. (5), the Fokker–Planck equation is stated as a pointwise equality for $(\boldsymbol{x}, t) \in \mathbb{R}^K \times [0, T]$; this is only valid under regularity assumptions on the posterior SDE beyond those in Appendix A.1. Still, we can state the divergence-free condition in its *weak form*, i.e., using the weak form of the Fokker–Planck equation eq. (12)

$$\int \boldsymbol{v}(\boldsymbol{x},t) \cdot \nabla_{\boldsymbol{x}} \varphi(\boldsymbol{x}) q(d\boldsymbol{x},t) = 0 \quad \text{a.e. } t \tag{14}$$

for all infinitely differentiable test functions φ having compact support. Indeed, if $\boldsymbol{v}(\boldsymbol{x}, t)$ and $q(\boldsymbol{x}, t)$ are smooth then integration by parts implies

$$0 = \int \boldsymbol{v}(\boldsymbol{x},t) \cdot \nabla_{\boldsymbol{x}} \varphi(\boldsymbol{x}) q(d\boldsymbol{x},t) = -\int \nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x},t)\boldsymbol{v}(\boldsymbol{x},t)]\varphi(\boldsymbol{x}) d\boldsymbol{x} \quad \text{a.e. } t$$

for all test functions φ. By Fubini's theorem, this implies $\nabla \cdot [q(\boldsymbol{x}, t)\boldsymbol{v}(\boldsymbol{x}, t)] = 0$ a.e. $(\boldsymbol{x}, t)$.

Having stated the divergence-free condition in its weak form, we are prepared to prove Proposition 1, which states that at fixed one-time marginals, among diffusion processes having a shared diffusion coefficient $\boldsymbol{\Sigma}^{1/2}$, the only degree of freedom is the choice of divergence-free vector field.

Proof of Proposition 1. Let $\mathbb{Q}, \widetilde{\mathbb{Q}} \in \mathcal{Q}$ have drifts $\boldsymbol{f}$ and $\widetilde{\boldsymbol{f}}$, respectively, common diffusion coefficient $\boldsymbol{\Sigma}^{1/2}$, and common initial marginal density $q(\boldsymbol{x}, 0)$. Let $\{q(\boldsymbol{x}, t)\}_{0 \leq t \leq T}$ denote the marginal densities of $\mathbb{Q}$.

Suppose first that $\mathbb{Q}$ and $\widetilde{\mathbb{Q}}$ have the same one-time marginals. Since these common marginals solve the weak Fokker–Planck equation for both drifts, and the diffusion coefficients are equal, subtracting the two equations gives

$$\int \left(\boldsymbol{f}(\boldsymbol{x},t) - \widetilde{\boldsymbol{f}}(\boldsymbol{x},t) \right) \cdot \nabla \varphi(\boldsymbol{x}) q(d\boldsymbol{x},t) = 0, \qquad \text{a.e. } t, \tag{15}$$

for every test function φ. This is precisely the statement that $\boldsymbol{f} - \widetilde{\boldsymbol{f}}$ is q-divergence-free.

Conversely, suppose that $\boldsymbol{f} - \widetilde{\boldsymbol{f}}$ is q-divergence-free. Since q solves the weak Fokker–Planck equation associated with $\boldsymbol{f}$, eq. (15) implies that it also solves the weak Fokker–Planck equation associated

with $\widetilde{\boldsymbol{f}}$ and the same initial density. By Assumption (A5), this solution is unique. It therefore coincides with the marginal densities of $\widetilde{\mathbb{Q}}$, so $\mathbb{Q}$ and $\widetilde{\mathbb{Q}}$ have the same one-time marginals. □

There are also diffusion processes with different diffusion coefficients that share the same one-time marginals. This fact is leveraged by Bartosh et al. [8] to learn the shared prior and posterior diffusion coefficient. However, for a fixed prior SDE eq. (1), the diffusion coefficient of the posterior eq. (2) must be equal to that of the prior. This is because, by Girsanov's theorem (Theorem 2), diffusion processes whose diffusion coefficients differ are generally not mutually absolutely continuous. In other words, the diffusion coefficient must be viewed as a parameter of the generative model, rather than a degree of freedom in the posterior process. On the other hand, the prior and variational posterior path laws are mutually absolutely continuous for all admissible choices of variational drift.

A.4 Gaussian one-time marginals

In the case of Gaussian one-time marginals, the non-uniqueness of the path law described in Appendix A.3 can be made explicit. In this subsection, we provide a characterization of the affine vector fields that are q-divergence free.

Let $(\boldsymbol{m}(t), \boldsymbol{S}(t))_{0\le t\le T}$ be continuously differentiable and $\boldsymbol{S}(t)$ be positive definite. For a matrix-valued function $\boldsymbol{K}(t)$, define the affine velocity field

$$\boldsymbol{v}_{\boldsymbol{K}}(\boldsymbol{x}, t) = \frac{d}{dt}\boldsymbol{m}(t) + \boldsymbol{K}(t)(\boldsymbol{x} - \boldsymbol{m}(t)). \tag{16}$$

And consider the SDE having diffusion coefficient $\boldsymbol{\Sigma}(\boldsymbol{x}, t)^{1/2}$ and drift

$$\begin{aligned}\boldsymbol{f}_q(\boldsymbol{x}, t) &= \boldsymbol{v}_{\boldsymbol{K}}(\boldsymbol{x}, t) + \frac{1}{2}\boldsymbol{\Sigma}(\boldsymbol{x}, t)\nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}, t) + \frac{1}{2}\sum_j \partial_{x_j}\boldsymbol{\Sigma}_{\cdot,j}(\boldsymbol{x}, t) \\ &= \frac{d}{dt}\boldsymbol{m}(t) + \boldsymbol{K}(t)(\boldsymbol{x} - \boldsymbol{m}(t)) - \frac{1}{2}\boldsymbol{\Sigma}(\boldsymbol{x}, t)\boldsymbol{S}(t)^{-1}(\boldsymbol{x} - \boldsymbol{m}(t)) + \frac{1}{2}\sum_j \partial_{x_j}\boldsymbol{\Sigma}_{\cdot,j}(\boldsymbol{x}, t).\end{aligned} \tag{17}$$

In Proposition 2, we characterize the choices of $\boldsymbol{K}(t)$ for which this SDE produces one-time marginals $q(\boldsymbol{x}, t) = \mathcal{N}(\boldsymbol{m}(t), \boldsymbol{S}(t))$.

Proposition 2: Fix any continuously differentiable square-root factorization $\boldsymbol{L}(t)$ satisfying $\boldsymbol{L}(t)\boldsymbol{L}(t)^\top = \boldsymbol{S}(t)$. Then the following statements are equivalent:

(i) The SDE

$$d\boldsymbol{x}(t) = \boldsymbol{f}_q(\boldsymbol{x}(t), t)dt + \boldsymbol{\Sigma}(\boldsymbol{x}(t), t)^{1/2}d\boldsymbol{w}(t), \quad q(\boldsymbol{x}, 0) = \mathcal{N}(\boldsymbol{m}(0), \boldsymbol{S}(0))$$

has one-time marginals $q(\boldsymbol{x}, t) = \mathcal{N}(\boldsymbol{m}(t), \boldsymbol{S}(t))$.

(ii) $\boldsymbol{K}(t)$ satisfies the matrix Lyapunov equation

$$\frac{d}{dt}\boldsymbol{S}(t) = \boldsymbol{K}(t)\boldsymbol{S}(t) + \boldsymbol{S}(t)\boldsymbol{K}(t)^\top, \qquad 0 \le t \le T. \tag{18}$$

(iii) There exists a unique skew-symmetric matrix-valued function $\boldsymbol{\Omega}(t) = -\boldsymbol{\Omega}(t)^\top$ such that

$$\boldsymbol{K}(t) = \left(\frac{d}{dt}\boldsymbol{L}(t)\right)\boldsymbol{L}(t)^{-1} + \boldsymbol{L}(t)\boldsymbol{\Omega}(t)\boldsymbol{L}(t)^{-1}, \qquad 0 \le t \le T. \tag{19}$$

Proof. Before proving the equivalence of (i), (ii), and (iii), we show that the Fokker–Planck equation associated with the drift (17) and diffusion coefficient $\boldsymbol{\Sigma}(\boldsymbol{x}, t)^{1/2}$ reduces to a continuity equation. Using $q(\boldsymbol{x}, t)\nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}, t) = \nabla_{\boldsymbol{x}} q(\boldsymbol{x}, t)$,

$$\begin{aligned}q(\boldsymbol{x}, t)\boldsymbol{f}_q(\boldsymbol{x}, t) &= q(\boldsymbol{x}, t)\boldsymbol{v}_{\boldsymbol{K}}(\boldsymbol{x}, t) + \frac{1}{2}\boldsymbol{\Sigma}(\boldsymbol{x}, t)\nabla_{\boldsymbol{x}} q(\boldsymbol{x}, t) + \frac{1}{2}q(\boldsymbol{x}, t)\sum_j \partial_{x_j}\boldsymbol{\Sigma}_{\cdot,j}(\boldsymbol{x}, t) \\ &= q(\boldsymbol{x}, t)\boldsymbol{v}_{\boldsymbol{K}}(\boldsymbol{x}, t) + \frac{1}{2}\sum_j \partial_{x_j}\left[q(\boldsymbol{x}, t)\boldsymbol{\Sigma}_{\cdot,j}(\boldsymbol{x}, t)\right].\end{aligned}$$

Consequently, we can rewrite the Fokker–Planck equation as

$$\begin{aligned}\frac{\partial}{\partial t}q(\boldsymbol{x},t) &= -\nabla_{\boldsymbol{x}}\cdot\left[q(\boldsymbol{x},t)\boldsymbol{f}_q(\boldsymbol{x},t)\right] + \frac{1}{2}\sum_{i,j}\partial_{x_i,x_j}\left[q(\boldsymbol{x},t)\boldsymbol{\Sigma}_{ij}(\boldsymbol{x},t)\right] \\ &= -\nabla_{\boldsymbol{x}}\cdot\left[q(\boldsymbol{x},t)\boldsymbol{v}_{\boldsymbol{K}}(\boldsymbol{x},t)\right]. \end{aligned} \tag{20}$$

This is the continuity equation corresponding to the affine flow

$$\frac{d}{dt}\boldsymbol{x}(t) = \boldsymbol{v}_{\boldsymbol{K}}(\boldsymbol{x}(t),t). \tag{21}$$

Because the flow map is affine, it maps the Gaussian initial distribution $q(\boldsymbol{x},0) = \mathcal{N}(\boldsymbol{m}(0),\boldsymbol{S}(0))$ to a Gaussian distribution at every time.

To prove the equivalence of (i) and (ii), we must show that $q(\boldsymbol{x},t) = \mathcal{N}(\boldsymbol{m}(t),\boldsymbol{S}(t))$ solves eq. (20) if and only if $\boldsymbol{K}(t)$ satisfies the Lyapunov equation eq. (18).

Let $\widetilde{\boldsymbol{m}}(t) = \mathbb{E}[\boldsymbol{x}(t)]$ denote the mean induced by the affine flow eq. (21). By the weak form of the continuity equation eq. (20), $\widetilde{\boldsymbol{m}}(t)$ satisfies the ODE

$$\frac{d}{dt}\widetilde{\boldsymbol{m}}(t) = \frac{d}{dt}\boldsymbol{m}(t) + \boldsymbol{K}(t)\left(\widetilde{\boldsymbol{m}}(t) - \boldsymbol{m}(t)\right), \quad \widetilde{\boldsymbol{m}}(0) = \boldsymbol{m}(0).$$

It is immediate that $\boldsymbol{m}(t)$ satisfies the ODE for the mean of $\boldsymbol{x}(t)$. For the covariance, we define the centered state $\boldsymbol{z}(t) = \boldsymbol{x}(t) - \boldsymbol{m}(t)$, which satisfies the affine flow

$$\frac{d}{dt}\boldsymbol{z}(t) = \boldsymbol{K}(t)\boldsymbol{z}(t).$$

Therefore, if $\widetilde{\boldsymbol{S}}(t) = \mathbb{E}\left[\boldsymbol{z}(t)\boldsymbol{z}(t)^\top\right]$ denotes the covariance induced by the affine flow, the weak form of the continuity equation yields the ODE

$$\frac{d}{dt}\widetilde{\boldsymbol{S}}(t) = \mathbb{E}\left[\boldsymbol{K}(t)\boldsymbol{z}(t)\boldsymbol{z}(t)^\top + \boldsymbol{z}(t)\boldsymbol{z}(t)^\top\boldsymbol{K}(t)^\top\right] = \boldsymbol{K}(t)\widetilde{\boldsymbol{S}}(t) + \widetilde{\boldsymbol{S}}(t)\boldsymbol{K}(t)^\top, \quad \widetilde{\boldsymbol{S}}(0) = \boldsymbol{S}(0).$$

If $\boldsymbol{K}(t)$ satisfies the Lyapunov equation eq. (18), uniqueness of solutions to the covariance ODE implies $\widetilde{\boldsymbol{S}}(t) = \boldsymbol{S}(t)$.

Conversely, if $q(\boldsymbol{x},t) = \mathcal{N}(\boldsymbol{m}(t),\boldsymbol{S}(t))$ solves eq. (20), then its covariance must satisfy the same equation, so

$$\frac{d}{dt}\boldsymbol{S}(t) = \boldsymbol{K}(t)\boldsymbol{S}(t) + \boldsymbol{S}(t)\boldsymbol{K}(t)^\top.$$

This proves the equivalence of (i) and (ii).

Suppose next that $\boldsymbol{K}(t)$ has the representation in (19). Since $\boldsymbol{S}(t) = \boldsymbol{L}(t)\boldsymbol{L}(t)^\top$, we obtain

$$\begin{aligned}\boldsymbol{K}(t)\boldsymbol{S}(t) + \boldsymbol{S}(t)\boldsymbol{K}(t)^\top &= \left(\frac{d}{dt}\boldsymbol{L}(t)\right)\boldsymbol{L}(t)^\top + \boldsymbol{L}(t)\left(\frac{d}{dt}\boldsymbol{L}(t)\right)^\top + \boldsymbol{L}(t)\left(\boldsymbol{\Omega}(t) + \boldsymbol{\Omega}(t)^\top\right)\boldsymbol{L}(t)^\top \\ &= \frac{d}{dt}\left[\boldsymbol{L}(t)\boldsymbol{L}(t)^\top\right] = \frac{d}{dt}\boldsymbol{S}(t).\end{aligned}$$

Thus, (iii) implies (ii).

Conversely, suppose that $\boldsymbol{K}(t)$ satisfies (18) and define

$$\boldsymbol{\Omega}(t) = \boldsymbol{L}(t)^{-1}\boldsymbol{K}(t)\boldsymbol{L}(t) - \boldsymbol{L}(t)^{-1}\frac{d}{dt}\boldsymbol{L}(t).$$

Then

$$\boldsymbol{\Omega}(t) + \boldsymbol{\Omega}(t)^\top = \boldsymbol{L}(t)^{-1}\left[\boldsymbol{K}(t)\boldsymbol{S}(t) + \boldsymbol{S}(t)\boldsymbol{K}(t)^\top - \frac{d}{dt}\boldsymbol{S}(t)\right]\boldsymbol{L}(t)^{-\top} = \boldsymbol{0}.$$

Hence, $\boldsymbol{\Omega}(t)$ is skew-symmetric. Its definition also shows that it is unique for the fixed square root $\boldsymbol{L}(t)$, proving the equivalence of (ii) and (iii). □

For a fixed square root $\boldsymbol{L}(t)$, the skew-symmetric matrix $\boldsymbol{\Omega}(t)$ parameterizes the freedom among compatible affine velocities. In particular, changing $\boldsymbol{\Omega}(t)$ adds the vector field

$$\boldsymbol{h}_{\boldsymbol{\Omega}}(\boldsymbol{x}, t) = \boldsymbol{L}(t)\boldsymbol{\Omega}(t)\boldsymbol{L}(t)^{-1}(\boldsymbol{x} - \boldsymbol{m}(t)),$$

which satisfies

$$\nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x}, t)\boldsymbol{h}_{\boldsymbol{\Omega}}(\boldsymbol{x}, t)] = 0.$$

Thus, $\boldsymbol{h}_{\boldsymbol{\Omega}}$ changes the path law without changing the one-time marginals. When $K \geq 2$, the space of skew-symmetric matrices has dimension $\binom{K}{2}$.

The representation in (19) depends on the chosen square root $\boldsymbol{L}(t)$. Replacing $\boldsymbol{L}(t)$ by another factor of the covariance $\boldsymbol{S}(t)$ reparameterizes $\boldsymbol{\Omega}(t)$ but does not change the set of compatible $\boldsymbol{K}(t)$.

Finally, Proposition 2 characterizes only the subfamily of affine q-divergence-free vector fields. There also exist nonlinear q-divergence-free vector fields that preserve Gaussian one-time marginals while inducing non-Gaussian multi-time distributions.

A.5 SVISE (Course and Nair [6]) and SDE Matching (Bartosh et al. [8])

In this subsection, we explain how existing simulation-free VI methods SVISE [6, 7] and SDE Matching [8] select a single drift function that realizes a given collection of one-time Gaussian marginal distributions $q(\boldsymbol{x}, t) = \mathcal{N}(\boldsymbol{m}(t), \boldsymbol{S}(t))$, where $\boldsymbol{m}(t), \boldsymbol{S}(t)$ are continuously differentiable on $[0, T]$ and $\boldsymbol{S}(t)$ is positive definite. We refer to these choices as the symmetric and square root gauges, respectively.

Before describing each choice of posterior drift, we recall Proposition 2 from Appendix A.4, which tells us that

$$\boldsymbol{f}_q(\boldsymbol{x}, t) = \frac{d}{dt}\boldsymbol{m}(t) + \boldsymbol{K}(t)(\boldsymbol{x} - \boldsymbol{m}(t)) - \frac{1}{2}\boldsymbol{\Sigma}(\boldsymbol{x}, t)\boldsymbol{S}(t)^{-1}(\boldsymbol{x} - \boldsymbol{m}(t)) + \frac{1}{2}\sum_j \partial_{x_j}\boldsymbol{\Sigma}_{\cdot,j}(\boldsymbol{x}, t) \quad (22)$$

attains the marginals $q(\boldsymbol{x}, t)$ whenever $\boldsymbol{K}(t)$ is a solution to the Lyapunov equation eq. (18). As we will see, SVISE and SDE Matching select the variational drift eq. (22) using different choices of $\boldsymbol{K}$.

Symmetric gauge. Suppose $\boldsymbol{\Sigma}(t) = \boldsymbol{\Sigma}(\boldsymbol{x}, t)$ is state-independent and $\boldsymbol{f}_q$ is defined as in Proposition 2. Then $\boldsymbol{A}_q(t) = \boldsymbol{K}(t) - \frac{1}{2}\boldsymbol{\Sigma}(t)\boldsymbol{S}(t)^{-1}$ also satisfies a Lyapunov equation

$$\frac{d}{dt}\boldsymbol{S}(t) - \boldsymbol{\Sigma}(t) = \boldsymbol{A}_q(t)\boldsymbol{S}(t) + \boldsymbol{S}(t)\boldsymbol{A}_q(t)^\top, \quad 0 \leq t \leq T. \quad (23)$$

In SVISE [6, 7], the authors choose the posterior drift $\boldsymbol{A}_q(t)$ to be the unique symmetric solution to eq. (23). They write this solution in closed-form using the Kronecker sum.

Unlike the square root gauge, the symmetric gauge does not naturally extend to the case of a state-dependent diffusion coefficient. This is because, in general, the matrix $\boldsymbol{K}$ for which $\boldsymbol{A}_q(\boldsymbol{x}, t) = \boldsymbol{K} - \frac{1}{2}\boldsymbol{\Sigma}(\boldsymbol{x}, t)\boldsymbol{S}(t)^{-1}$ is symmetric cannot be chosen to be independent of the state $\boldsymbol{x}$.

Square root gauge. Bartosh et al. [8] instead identify the one-time Gaussian marginals with the drift function $\boldsymbol{f}_q$ defined as in Proposition 2 for $\boldsymbol{K}(t) = \left(\frac{d}{dt}\boldsymbol{S}(t)^{1/2}\right)\boldsymbol{S}(t)^{-1/2}$. Here, $\boldsymbol{S}(t)^{1/2}$ denotes the symmetric square root of $\boldsymbol{S}(t)$. It is straightforward to verify that this choice of $\boldsymbol{K}$ satisfies the matrix Lyapunov equation (18). This choice of $\boldsymbol{K}$ is valid even for a state-dependent diffusion coefficient.

The SDE Matching framework, in theory, allows for any square root $\boldsymbol{\sigma}(t)$ such that $\boldsymbol{S}(t) = \boldsymbol{\sigma}(t)\boldsymbol{\sigma}(t)^\top$, to be chosen. However, in their implementation, $\boldsymbol{\sigma}(t)$ is chosen to be diagonal, which yields the symmetric square root $\boldsymbol{S}(t)^{1/2}$. At the end of this section, we discuss an extension of SDE Matching that represents all square roots simultaneously.

The authors arrive at the choice of variational drift by converting from a probability flow ODE [34] to an SDE. In particular, suppose $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_K)$ and $F(\boldsymbol{\varepsilon}, t) = \boldsymbol{m}(t) + \boldsymbol{S}(t)^{1/2}\boldsymbol{\varepsilon}$. F can be viewed as a map that takes as input a sample $\boldsymbol{\varepsilon}$ as well as a time t and produces a sample from $q(\boldsymbol{x}, t)$. If

$\tilde{\boldsymbol{x}}(t) = F(\boldsymbol{\varepsilon}, t)$, then each particle $\tilde{\boldsymbol{x}}(t)$ evolves according to the ODE

$$\begin{aligned}\frac{d}{dt}\tilde{\boldsymbol{x}}(t) &= \frac{d}{dt}\boldsymbol{m}(t) + \left(\frac{d}{dt}\boldsymbol{S}(t)^{1/2}\right)\boldsymbol{\varepsilon} \\ &= \frac{d}{dt}\boldsymbol{m}(t) + \left(\frac{d}{dt}\boldsymbol{S}(t)^{1/2}\right)\{\boldsymbol{S}(t)^{-1/2}(\tilde{\boldsymbol{x}}(t) - \boldsymbol{m}(t))\}, \quad 0 \le t \le T\end{aligned} \tag{24}$$

with $\tilde{\boldsymbol{x}}(0) \overset{d}{=} F(\boldsymbol{\varepsilon}, 0)$.

From the ODE (24), the authors construct an SDE with diffusion coefficient $\boldsymbol{\Sigma}(\boldsymbol{x}, t)^{1/2}$. A straightforward manipulation of the Fokker–Planck equation (see e.g., Song et al. [34, Appendix D.1]) shows that the solution to the SDE

$$\begin{aligned}d\boldsymbol{x}(t) = \Bigg(&\frac{d}{dt}\boldsymbol{m}(t) + \left(\frac{d}{dt}\boldsymbol{S}(t)^{1/2}\right)\{\boldsymbol{S}(t)^{-1/2}(\boldsymbol{x}(t) - \boldsymbol{m}(t))\} + \frac{1}{2}\boldsymbol{\Sigma}(\boldsymbol{x}(t), t)\nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}(t), t) \\ &+ \frac{1}{2}\sum_j \partial_{\boldsymbol{x}_j}\boldsymbol{\Sigma}_{\cdot,j}(\boldsymbol{x}(t), t)\Bigg)dt + \boldsymbol{\Sigma}(\boldsymbol{x}(t), t)^{1/2}d\boldsymbol{w}(t), \quad \boldsymbol{x}(0) \sim q(\boldsymbol{x}, 0)\end{aligned}$$

has the same one-time marginal distributions as $\tilde{\boldsymbol{x}}(t)$. Substituting in the Gaussian score and grouping terms yields the drift

$$\begin{aligned}&\left\{\left(\frac{d}{dt}\boldsymbol{S}(t)^{1/2}\right)\boldsymbol{S}(t)^{-1/2} - \frac{1}{2}\boldsymbol{\Sigma}(\boldsymbol{x}(t), t)\boldsymbol{S}(t)^{-1}\right\}\boldsymbol{x}(t) + \\ &\frac{d}{dt}\boldsymbol{m}(t) - \left\{\left(\frac{d}{dt}\boldsymbol{S}(t)^{1/2}\right)\boldsymbol{S}(t)^{-1/2} - \frac{1}{2}\boldsymbol{\Sigma}(\boldsymbol{x}(t), t)\boldsymbol{S}(t)^{-1}\right\}\boldsymbol{m}(t) + \frac{1}{2}\sum_j \partial_{\boldsymbol{x}_j}\boldsymbol{\Sigma}_{\cdot,j}(\boldsymbol{x}(t), t).\end{aligned}$$

Once we identify $\boldsymbol{K}(t) = \left(\frac{d}{dt}\boldsymbol{S}(t)^{1/2}\right)\boldsymbol{S}(t)^{-1/2}$, the resulting drift is of the form given in eq. (17).

Although the construction in Bartosh et al. [8] allows the marginals $q(\boldsymbol{x}, t)$ to be non-Gaussian (e.g., by parameterizing F as a normalizing flow), the paper itself evaluates the method only in the Gaussian marginal setting. We therefore restrict our discussion to that case.

An extension of SDE Matching. The framework of SDE Matching [8] can in principle be extended to span every solution to the Lyapunov equation (18). This is not discussed or implemented by the authors. As described in the previous paragraph, their parameterization $\boldsymbol{S}(t) = \boldsymbol{\sigma}(t)\boldsymbol{\sigma}(t)^\top$ induces the linear coefficient $\boldsymbol{K}(t) = (\frac{d}{dt}\boldsymbol{\sigma}(t))\boldsymbol{\sigma}(t)^{-1}$. Fixing $\boldsymbol{\sigma}$ to a specific square root selects a single path law for each set of one-time marginals. By Proposition 2, the full set of solutions to the Lyapunov equation is parameterized by an additional skew-symmetric matrix $\boldsymbol{\Omega}(t)$ via

$$\boldsymbol{K}(t) = \left(\frac{d}{dt}\boldsymbol{\sigma}(t)\right)\boldsymbol{\sigma}(t)^{-1} + \boldsymbol{\sigma}(t)\boldsymbol{\Omega}(t)\boldsymbol{\sigma}(t)^{-1}, \quad 0 \le t \le T.$$

Augmenting SDE Matching with a learned $\boldsymbol{\Omega}(t)$ would in principle allow the posterior to span at each time the full $\binom{K}{2}$-dimensional affine space of compatible linear drift coefficients.

This extension has two limitations. First, it requires introducing and optimizing over $\boldsymbol{\Omega}(t)$, sacrificing the closed-form structure of SDE Matching's original parameterization. Helmholtz-SDE, by contrast, *computes the optimal correction in closed form* via the Helmholtz decomposition of the residual vector field (Section 4). Second, for a state-independent diffusion coefficient, the resulting path law remains a Gaussian process. Proposition 2 characterizes only diffusions with affine drift that produce the prescribed Gaussian marginals; non-linear drifts can also yield Gaussian one-time marginals while inducing non-Gaussian multi-time joint distributions. Helmholtz-SDE can produce posterior approximations that lie outside the Gaussian-process family by using higher-order polynomial approximations to the optimal drift (Appendix B.3).

B Helmholtz-SDE

In this section, we provide further explanation of our algorithm, Helmholtz-SDE.

First, in Appendix B.1 we describe how the q-weighted Helmholtz decomposition can be obtained for arbitrary smooth, strictly positive marginal densities q by inverting a Stein operator. For the remainder of Appendix B, we then specialize to the setting of Gaussian one-time marginals. In Appendix B.2, we provide sufficient conditions under which the Helmholtz decomposition exists uniquely. In Appendix B.3, we translate this theoretical result into a practical formula using an order-ℓ polynomial approximation to the optimal divergence-free correction. In Appendix B.4, we verify that the resulting variational drift, obtained by adding the polynomial correction to the reference drift, produces an SDE satisfying our assumptions in Appendix A.1. In Appendix B.5, we discuss the optimality of Helmholtz-SDE in the conjugate and non-conjugate settings. In Appendix B.6 we provide pseudocode of our algorithm.

B.1 Computing the Helmholtz decomposition

We first discuss how the Helmholtz decomposition can be computed for an arbitrary smooth, strictly positive density q as the inverse of a Stein operator. The resulting probabilistic expression can be approximated via Monte Carlo. For ease of notation, we consider a state-independent diffusion coefficient, although the same construction holds in the state-dependent setting.

The Helmholtz decomposition and diffusion Stein operator. As described in Section 4.2, the optimal correction to the reference drift $\boldsymbol{f}_q(\boldsymbol{x}, t)$ is obtained by computing the q-weighted Helmholtz decomposition of the residual vector field $\boldsymbol{r}(\boldsymbol{x}, t) = \boldsymbol{f}_p(\boldsymbol{x}, t) - \boldsymbol{f}_q(\boldsymbol{x}, t)$ and taking the q-divergence-free vector field. In particular, we want to find a decomposition of $\boldsymbol{r}$ as

$$\boldsymbol{r}(\boldsymbol{x}, t) = \boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t) + \boldsymbol{h}(\boldsymbol{x}, t), \quad (\boldsymbol{x}, t) \in \mathbb{R}^K \times [0, T], \tag{25}$$

where $\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t)$ is a gradient field and $\boldsymbol{h}(\boldsymbol{x}, t)$ is a q-divergence-free vector field.

Suppose $q(\boldsymbol{x}, t)$ is a smooth, strictly positive density. To obtain such a decomposition, observe that if eq. (25) holds, then multiplying both sides by $q(\boldsymbol{x}, t)$ and computing the divergence yields

$$\nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x}, t)\boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t)] = \nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x}, t)\boldsymbol{r}(\boldsymbol{x}, t)], \quad (\boldsymbol{x}, t) \in \mathbb{R}^K \times [0, T], \tag{26}$$

since $\boldsymbol{h}$ is q-divergence-free. Dividing both sides by q and expanding the left-hand side yields

$$\operatorname{tr}(\boldsymbol{\Sigma}(t)\nabla^2_{\boldsymbol{x}}\psi(\boldsymbol{x}, t)) + \nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}, t) \cdot \boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t) = \frac{1}{q(\boldsymbol{x}, t)}\nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x}, t)\boldsymbol{r}(\boldsymbol{x}, t)]. \tag{27}$$

From eq. (27), we make two observations. First, the left-hand side can be written as $[\mathcal{L}_t^{\boldsymbol{\Sigma}}\nabla_{\boldsymbol{x}}\psi(\cdot, t)](\boldsymbol{x})$, where $\mathcal{L}_t^{\boldsymbol{\Sigma}}$ is the differential operator

$$[\mathcal{L}_t^{\boldsymbol{\Sigma}}\boldsymbol{f}](\boldsymbol{x}) = \operatorname{tr}(\boldsymbol{\Sigma}(t)D_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x}, t)) + \boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}, t) \cdot \boldsymbol{f}(\boldsymbol{x}, t) \tag{28}$$

defined for differentiable vector fields $\boldsymbol{f} : \mathbb{R}^K \to \mathbb{R}^K$, where $D_{\boldsymbol{x}}\boldsymbol{f}$ represents the Jacobian of $\boldsymbol{f}$. This is the *diffusion Stein operator* corresponding to the density $q(\boldsymbol{x}, t)$ and constant covariance coefficient $\boldsymbol{\Sigma}(t)$ [65, 66]. When $\boldsymbol{\Sigma}(t) = \mathbb{I}_K$, it reduces to the *Langevin Stein* operator

$$[\mathcal{L}_t\boldsymbol{f}](\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \cdot \boldsymbol{f}(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}, t) \cdot \boldsymbol{f}(\boldsymbol{x}).$$

We refer the reader to Gorham et al. [66] for background on diffusion Stein operators.

Second, whenever $\boldsymbol{r}(\boldsymbol{x}, t)$ and $\mathcal{L}_t[\boldsymbol{r}(\cdot, t)](\boldsymbol{x})$ are integrable, the right-hand side of eq. (27) has expectation

$$\mathbb{E}_{\mathbb{Q}}\left[\frac{1}{q(\boldsymbol{x}, t)}\nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x}, t)\boldsymbol{r}(\boldsymbol{x}, t)]\right] = \int \nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x}, t)\boldsymbol{r}(\boldsymbol{x}, t)]d\boldsymbol{x} = 0$$

The final equality holds by the divergence theorem [67, Theorem 3.1].

Overall, we have reduced computing the Helmholtz decomposition to solving

$$\begin{aligned} &[\mathcal{L}_t^{\boldsymbol{\Sigma}}\nabla_{\boldsymbol{x}}\psi(\cdot, t)](\boldsymbol{x}) = \eta(\boldsymbol{x}, t), \quad (\boldsymbol{x}, t) \in \mathbb{R}^K \times [0, T], \\ &\eta(\boldsymbol{x}, t) = \frac{1}{q(\boldsymbol{x}, t)}\nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x}, t)\boldsymbol{r}(\boldsymbol{x}, t)] \end{aligned} \tag{29}$$

where $\eta(\boldsymbol{x}, t)$ has mean zero under $q(\boldsymbol{x}, t)$. Intuitively, in order to solve eq. (29), we need to find an inverse of the diffusion Stein operator $\mathcal{L}_t^{\boldsymbol{\Sigma}}$. Doing so has been the subject of prior literature, namely Barbour [65], Pardoux and Veretennikov [68], Gorham et al. [66].

Inverting the diffusion Stein operator. Fix $t \in [0, T]$. Consider the time-homogeneous diffusion in an auxiliary time variable s with diffusion coefficient $\boldsymbol{\Sigma}(t)^{1/2}$:

$$d\boldsymbol{z}^{(t)}(s) = \frac{1}{2}\boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}} \log q(\boldsymbol{z}^{(t)}(s), t)ds + \boldsymbol{\Sigma}(t)^{1/2}d\boldsymbol{w}(s). \tag{30}$$

When $\boldsymbol{\Sigma}(t) = \mathbb{I}_K$, eq. (30) is recognizable as the overdamped Langevin diffusion. In particular, eq. (30) has invariant density $q(\boldsymbol{x}, t)$. Its *infinitesimal generator* is

$$\begin{aligned}[\mathcal{A}_t u](\boldsymbol{z}) &= \frac{1}{2q(\boldsymbol{z}, t)}\nabla_{\boldsymbol{z}} \cdot [q(\boldsymbol{z}, t)\boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{z}} u(\boldsymbol{z})] \\ &= \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}(t)\nabla^2_{\boldsymbol{z}} u(\boldsymbol{z})) + \frac{1}{2}\boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{z}} \log q(\boldsymbol{z}, t) \cdot \nabla_{\boldsymbol{z}} u(\boldsymbol{z})\end{aligned} \tag{31}$$

for $u : \mathbb{R}^K \to \mathbb{R}$ twice continuously differentiable and satisfying integrability conditions [64]. This implies, by eq. (28), $\mathcal{A}_t u = \mathcal{L}_t^{\boldsymbol{\Sigma}}(\frac{1}{2}\nabla_{\boldsymbol{x}} u)$. Hence, solving eq. (29) is equivalent to solving

$$[\mathcal{A}_t u(\cdot, t)](\boldsymbol{x}) = \eta(\boldsymbol{x}, t), \quad (\boldsymbol{x}, t) \in \mathbb{R}^K \times [0, T] \tag{32}$$

and then setting $\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t) = \frac{1}{2}\nabla_{\boldsymbol{x}} u(\boldsymbol{x}, t)$. Equation (32) is known as the *Poisson equation* [68, 69].

Assume that, for each fixed t, $\eta(\boldsymbol{x}, t)$ and the coefficients of the above diffusion satisfy the regularity conditions of Gorham et al. [66, Theorem 5]. Then Gorham et al. [66, Theorem 5] implies that

$$u(\boldsymbol{x}, t) = -\int_0^\infty \mathbb{E}[\eta(\boldsymbol{z}^{(t)}(s), t)|\boldsymbol{z}^{(t)}(0) = \boldsymbol{x}]ds \tag{33}$$

solves the Poisson equation. The integrand in eq. (33) is computed by starting the diffusion eq. (30) from $\boldsymbol{x}$ and then running it forward until time s.

Therefore,

$$\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t) = -\frac{1}{2}\nabla_{\boldsymbol{x}} \int_0^\infty \mathbb{E}[\eta(\boldsymbol{z}^{(t)}(s), t)|\boldsymbol{z}^{(t)}(0) = \boldsymbol{x}]ds \tag{34}$$

solves eq. (29). Consequently, defining $\boldsymbol{h}(\boldsymbol{x}, t) = \boldsymbol{r}(\boldsymbol{x}, t) - \boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t)$ gives $\nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x}, t)\boldsymbol{h}(\boldsymbol{x}, t)] = 0$. Thus $\boldsymbol{h}(\cdot, t)$ is the q-divergence-free component of the residual vector field.

B.2 The Helmholtz decomposition for Gaussian marginals

For the remainder of Appendix B, we specialize to the setting of Gaussian one-time marginals and a state-independent diffusion coefficient. In this subsection, we establish the existence, uniqueness, and regularity of the weighted Helmholtz decomposition in this setting.

Since we are considering only the existence and uniqueness of the weighted Helmholtz decomposition, and not the implied variational SDE, we do not require the assumptions from Appendix A.1.

First, we show that under minimal assumptions, the q-divergence-free vector field $\boldsymbol{h}$ that minimizes eq. (6) exists. It is the orthogonal projection onto the subspace of q-divergence-free vector fields.

Theorem 3 (Existence and uniqueness of weak Helmholtz decomposition): Suppose that $\boldsymbol{r}(\boldsymbol{x}, t)$ has polynomial growth and that $\boldsymbol{\Sigma}(t)$ is positive definite. Let $\mathcal{L}^2(q(\boldsymbol{x}, t), \mathbb{R}^K, \boldsymbol{\Sigma}(t))^K$ be the Hilbert space of vector fields $\boldsymbol{f} : \mathbb{R}^K \to \mathbb{R}^K$ satisfying $\mathbb{E}_{q(\boldsymbol{x},t)}\|\boldsymbol{f}(\boldsymbol{x})\|^2 < \infty$ equipped with the inner product

$$\langle \boldsymbol{f}, \boldsymbol{g}\rangle_{q,t} = \mathbb{E}_{q(\boldsymbol{x},t)}[\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\Sigma}(t)^{-1}\boldsymbol{g}(\boldsymbol{x})], \qquad \|\boldsymbol{f}\|_{q,t} = \langle \boldsymbol{f}, \boldsymbol{f}\rangle_{q,t}^{1/2} \tag{35}$$

Then $\boldsymbol{r}(\boldsymbol{x}, t) \in \mathcal{L}^2(q(\boldsymbol{x}, t), \mathbb{R}^K, \boldsymbol{\Sigma}(t))^K$ admits the decomposition

$$\boldsymbol{r}(\boldsymbol{x}, t) = \boldsymbol{g}(\boldsymbol{x}, t) + \boldsymbol{h}(\boldsymbol{x}, t) \tag{36}$$

where $\boldsymbol{h}, \boldsymbol{g} \in \mathcal{L}^2(q(\boldsymbol{x}, t), \mathbb{R}^K, \boldsymbol{\Sigma}(t))^K$, $\boldsymbol{h}$ is q-divergence-free in the weak sense eq. (14), and $\boldsymbol{g}$ is orthogonal to all q-divergence-free vector fields. Moreover, this decomposition is unique in $\mathcal{L}^2(q(\boldsymbol{x}, t), \mathbb{R}^K, \boldsymbol{\Sigma}(t))^K$.

Proof. Consider the set $\mathcal{V}_t$ of all vector fields that are divergence-free in the weak sense eq. (14):

$$\mathcal{V}_t = \left\{ \boldsymbol{v} \in \mathcal{L}^2(q(\boldsymbol{x},t), \mathbb{R}^K, \boldsymbol{\Sigma}(t))^K \;\middle|\; \int \boldsymbol{v}(\boldsymbol{x},t)^\top \nabla_{\boldsymbol{x}} \varphi(\boldsymbol{x}) q(d\boldsymbol{x},t) = 0, \quad \forall \varphi \in \mathcal{C} \right\}, \tag{37}$$

where $\mathcal{C}$ is the collection of test functions that are infinitely differentiable and have compact support.

Since $\boldsymbol{r}(\boldsymbol{x},t)$ has polynomial growth, it belongs to $\mathcal{L}^2(q(\boldsymbol{x},t), \mathbb{R}^K, \boldsymbol{\Sigma}(t))^K$. By the Hilbert projection theorem, the result follows once we establish that $\mathcal{V}_t$ is a closed subspace of $\mathcal{L}^2(q(\boldsymbol{x},t), \mathbb{R}^K, \boldsymbol{\Sigma}(t))^K$.

The fact that $\mathcal{V}_t$ is a subspace is immediate from the definition. To show that it is closed, fix $\varphi \in \mathcal{C}$ and define

$$\mathcal{L}_\varphi(\boldsymbol{v}) = \int \boldsymbol{v}(\boldsymbol{x})^\top \nabla_{\boldsymbol{x}} \varphi(\boldsymbol{x}) q(d\boldsymbol{x},t).$$

Since $\mathcal{L}_\varphi(\boldsymbol{v}) = \langle \boldsymbol{v}, \boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\varphi \rangle_{q,t}$, the Cauchy–Schwarz inequality gives

$$|\mathcal{L}_\varphi(\boldsymbol{v})| \le \|\boldsymbol{v}\|_{q,t} \, \|\boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\varphi\|_{q,t}\,.$$

Consequently, $\mathcal{L}_\varphi$ is a continuous linear functional. Therefore, $\mathcal{V}_t = \bigcap_{\varphi \in \mathcal{C}} \ker \mathcal{L}_\varphi$ is closed. □

The preceding theorem tells us that, among the vector fields that are q-divergence-free in the weak sense eq. (14), $\boldsymbol{h}(\boldsymbol{x},t)$ uniquely minimizes eq. (6). We call eq. (36) the *weak Helmholtz decomposition*. Theorem 3 does *not*, however, provide a Helmholtz decomposition as in eq. (7). Specifically, it neither establishes that $\boldsymbol{h}$ satisfies the q-divergence-free condition pointwise, nor identifies $\boldsymbol{g}$ with a gradient field of the form $\boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\psi$, nor verifies the regularity of $\boldsymbol{g}$, $\boldsymbol{h}$, or ψ.

The following result supplies this representation and regularity for the Gaussian marginals used in Helmholtz-SDE.

Assumption (A6): Assume that the following hold:

(i) $(\boldsymbol{m}(t), \boldsymbol{S}(t))_{0 \le t \le T}$ is continuously differentiable in t, and $\boldsymbol{S}(t)$ is positive definite.

(ii) $\boldsymbol{r}$ is twice differentiable in $\boldsymbol{x}$, and its derivatives are continuous in $(\boldsymbol{x}, t)$.

(iii) $\boldsymbol{r}$, $D_{\boldsymbol{x}}\boldsymbol{r}$, and $D_{\boldsymbol{x}}^{(2)}\boldsymbol{r}$ have polynomial growth in $\boldsymbol{x}$, uniformly in t.

(iv) $\boldsymbol{\Sigma}(t)$ is continuous and positive definite.

$D_{\boldsymbol{x}}\boldsymbol{r} \in \mathbb{R}^{K\times K}$ and $D_{\boldsymbol{x}}^{(2)}\boldsymbol{r} \in \mathbb{R}^{K\times K\times K}$ represent the derivative tensors of the residual vector field $\boldsymbol{r} : \mathbb{R}^K \to \mathbb{R}^K$.

Theorem 4 (Regularity of the Helmholtz decomposition): Suppose Assumption (A6) holds. For every $t \in [0,T]$, the weak Helmholtz decomposition of $\boldsymbol{r}(\boldsymbol{x},t)$ in eq. (36) has the pointwise representation

$$\boldsymbol{r}(\boldsymbol{x},t) = \boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x},t) + \boldsymbol{h}(\boldsymbol{x},t),$$

where $\boldsymbol{h}$ is q-divergence-free in the strong sense

$$\nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x},t)\boldsymbol{h}(\boldsymbol{x},t)] = 0, \quad (\boldsymbol{x},t) \in \mathbb{R}^K \times [0,T].$$

ψ is twice differentiable in $\boldsymbol{x}$, $\boldsymbol{h}$ is once differentiable in $\boldsymbol{x}$, and their derivatives are continuous in $(\boldsymbol{x},t)$. Within the stated regularity class, $\boldsymbol{h}$ is unique pointwise, and $\psi(\cdot,t)$ is unique up to an additive constant, which may depend on t.

Proof. Fix $t \in [0,T]$. Since $\boldsymbol{S}(t)$ is positive definite, $q(\boldsymbol{x},t)$ is a smooth, strictly positive density on $\mathbb{R}^K$. Recall from Appendix B.1 that solving the Stein operator equation eq. (29) for $\nabla_{\boldsymbol{x}}\psi$ reduces, via the identity $\mathcal{A}_t u = \mathcal{L}_t^{\boldsymbol{\Sigma}}(\frac{1}{2}\nabla_{\boldsymbol{x}} u)$, to solving the Poisson equation eq. (32) for u:

$$[\mathcal{A}_t u(\cdot,t)](\boldsymbol{x}) \;=\; \frac{1}{2q(\boldsymbol{x},t)}\nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x},t)\boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}} u(\boldsymbol{x},t)] \;=\; \eta(\boldsymbol{x},t), \tag{38}$$

where $\mathcal{A}_t$ is the infinitesimal generator of the Langevin diffusion eq. (30) and $\eta(\boldsymbol{x},t)$ is the source term in eq. (29). The gradient component of the Helmholtz decomposition is then recovered as $\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x},t) = \frac{1}{2}\nabla_{\boldsymbol{x}} u(\boldsymbol{x},t)$.

Our goal is to show that the candidate $u(\boldsymbol{x}, t)$ in eq. (33) satisfies eq. (38) pointwise, and that u is twice continuously differentiable in $\boldsymbol{x}$ with derivatives continuous in $(\boldsymbol{x}, t)$. Once eq. (38) is established, defining

$$\boldsymbol{h}(\boldsymbol{x}, t) = \boldsymbol{r}(\boldsymbol{x}, t) - \boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t) \tag{39}$$

gives $\nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x}, t)\boldsymbol{h}(\boldsymbol{x}, t)] = 0$, yielding the desired Helmholtz decomposition.

Our proof proceeds in four main steps:

(i) Reexpress the candidate u in whitened and rotated coordinates, where the associated Langevin diffusion becomes a product of one-dimensional Ornstein–Uhlenbeck (OU) processes.

(ii) Prove that the candidate u is well-defined and satisfies the desired regularity conditions.

(iii) Verify that u satisfies the Poisson equation eq. (38) pointwise.

(iv) Identify the constructed vector fields with the unique weak Helmholtz decomposition from Theorem 3.

Step (i): The candidate u proposed in Appendix B.1 (eq. (33)) is expressed as an integral of expectations along the Langevin diffusion eq. (30) with invariant law $q(\boldsymbol{x}, t)$. We use the same whitening and rotation as in Appendix B.3. Define

$$\begin{aligned} \boldsymbol{C}(t) &= \boldsymbol{S}(t)^{-1/2}\boldsymbol{\Sigma}(t)\boldsymbol{S}(t)^{-1/2} = \boldsymbol{Q}(t)\boldsymbol{\Lambda}(t)\boldsymbol{Q}(t)^\top, \\ \boldsymbol{\Lambda}(t) &= \operatorname{diag}(\lambda_1(t), \ldots, \lambda_K(t)), \qquad \boldsymbol{Q}(t)^\top\boldsymbol{Q}(t) = \mathbb{I}_K, \end{aligned}$$

and introduce the coordinates

$$\tilde{\boldsymbol{x}} = \boldsymbol{Q}(t)^\top\boldsymbol{S}(t)^{-1/2}(\boldsymbol{x} - \boldsymbol{m}(t)), \qquad \chi(\tilde{\boldsymbol{x}}, t) = u(\boldsymbol{m}(t) + \boldsymbol{S}(t)^{1/2}\boldsymbol{Q}(t)\tilde{\boldsymbol{x}}, t).$$

Transform the residual in the same way:

$$\widetilde{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}}, t) = \boldsymbol{Q}(t)^\top\boldsymbol{S}(t)^{-1/2}\boldsymbol{r}(\boldsymbol{m}(t) + \boldsymbol{S}(t)^{1/2}\boldsymbol{Q}(t)\tilde{\boldsymbol{x}}, t).$$

Then $q(\tilde{\boldsymbol{x}}, t) = \mathcal{N}(\boldsymbol{0}, \mathbb{I}_K)$. Since $\boldsymbol{S}(t)$ and $\boldsymbol{\Sigma}(t)$ are positive definite, every $\lambda_i(t)$ is strictly positive.

After an orthogonal transformation of the driving Brownian motion, the Langevin diffusion eq. (30) becomes

$$d\tilde{\boldsymbol{z}}^{(t)}(s) = -\tfrac{1}{2}\boldsymbol{\Lambda}(t)\tilde{\boldsymbol{z}}^{(t)}(s)ds + \boldsymbol{\Lambda}(t)^{1/2}d\boldsymbol{w}(s), \qquad \tilde{\boldsymbol{z}}^{(t)}(0) = \tilde{\boldsymbol{x}}. \tag{40}$$

Thus its coordinates are independent one-dimensional OU processes with invariant law $\mathcal{N}(0, 1)$. Equivalently, each coordinate is a standard OU process after rescaling the artificial time s by $\lambda_i(t)$. The joint invariant law is $\mathcal{N}(\boldsymbol{0}, \mathbb{I}_K)$, and the infinitesimal generator is

$$(\tilde{\mathcal{A}}_t f)(\tilde{\boldsymbol{x}}) = \tfrac{1}{2}\operatorname{tr}\left(\boldsymbol{\Lambda}(t)\nabla^2_{\tilde{\boldsymbol{x}}} f(\tilde{\boldsymbol{x}})\right) - \tfrac{1}{2}\tilde{\boldsymbol{x}}^\top\boldsymbol{\Lambda}(t)\nabla_{\tilde{\boldsymbol{x}}} f(\tilde{\boldsymbol{x}})$$

for $f : \mathbb{R}^K \to \mathbb{R}$ twice continuously differentiable with derivatives having polynomial growth. Define the transformed source

$$g(\tilde{\boldsymbol{x}}, t) = \nabla_{\tilde{\boldsymbol{x}}} \cdot \widetilde{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}}, t) - \tilde{\boldsymbol{x}}^\top\widetilde{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}}, t). \tag{41}$$

A direct computation shows that

$$g(\tilde{\boldsymbol{x}}, t) = \eta(\boldsymbol{m}(t) + \boldsymbol{S}(t)^{1/2}\boldsymbol{Q}(t)\tilde{\boldsymbol{x}}, t),$$

where η is the source term in eq. (29). Overall, in the $\tilde{\boldsymbol{x}}$ coordinates, the Poisson equation becomes

$$(\tilde{\mathcal{A}}_t\chi(\cdot, t))(\tilde{\boldsymbol{x}}) = g(\tilde{\boldsymbol{x}}, t). \tag{42}$$

By Assumption (A6), $\boldsymbol{r}$ is differentiable in $\boldsymbol{x}$, and both $\boldsymbol{r}$ and $D_{\boldsymbol{x}}\boldsymbol{r}$ have polynomial growth; the same therefore holds of $\widetilde{\boldsymbol{\rho}}(\cdot, t)$ and $D_{\tilde{\boldsymbol{x}}}\widetilde{\boldsymbol{\rho}}(\cdot, t)$. Stein's identity gives

$$\mathbb{E}_{\tilde{\boldsymbol{x}}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)}\left[\nabla_{\tilde{\boldsymbol{x}}} \cdot \widetilde{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}}, t)\right] = \mathbb{E}_{\tilde{\boldsymbol{x}}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)}\left[\tilde{\boldsymbol{x}}^\top\widetilde{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}}, t)\right],$$

so $g(\cdot, t)$ has mean zero under the OU invariant law.

In these whitened coordinates, the candidate from eq. (33) takes the form

$$\chi(\tilde{\boldsymbol{x}}, t) = -\int_0^\infty (P_s^t g)(\tilde{\boldsymbol{x}})\, ds, \qquad (P_s^t g)(\tilde{\boldsymbol{x}}) = \mathbb{E}\left[g\left(\tilde{\boldsymbol{z}}^{(t)}(s), t\right) \mid \tilde{\boldsymbol{z}}^{(t)}(0) = \tilde{\boldsymbol{x}} \right], \tag{43}$$

where $(P_s^t)_{s\geq 0}$ is the OU semigroup associated with eq. (40). Equivalently,

$$u(\boldsymbol{x}, t) = \chi\left(\boldsymbol{Q}(t)^\top \boldsymbol{S}(t)^{-1/2} (\boldsymbol{x} - \boldsymbol{m}(t)), t \right).$$

Step (ii): We first justify that eq. (43) is well-defined. For all times $0 \leq s < \infty$, the marginal distribution of the OU process (40) satisfies

$$\tilde{\boldsymbol{z}}^{(t)}(s) \overset{d}{=} \boldsymbol{A}_t(s)\tilde{\boldsymbol{x}} + \boldsymbol{B}_t(s)\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_K),$$

where

$$\boldsymbol{A}_t(s) = e^{-s\boldsymbol{\Lambda}(t)/2}, \qquad \boldsymbol{B}_t(s) = \left(\mathbb{I}_K - e^{-s\boldsymbol{\Lambda}(t)} \right)^{1/2}.$$

Thus

$$(P_s^t g)(\tilde{\boldsymbol{x}}) = \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)} \left[g\left(\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}} + \boldsymbol{B}_t(s)\boldsymbol{\xi}, t \right) \right].$$

Using $\mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)}[g(\boldsymbol{\xi}, t)] = 0$, we may write

$$(P_s^t g)(\tilde{\boldsymbol{x}}) = \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)} \left[g\left(\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}} + \boldsymbol{B}_t(s)\boldsymbol{\xi}, t \right) - g(\boldsymbol{\xi}, t) \right].$$

Hence,

$$\left| (P_s^t g)(\tilde{\boldsymbol{x}}) \right| \leq \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)} \left| g\left(\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}} + \boldsymbol{B}_t(s)\boldsymbol{\xi}, t \right) - g(\boldsymbol{\xi}, t) \right|.$$

Now decompose

$$\begin{aligned} g\left(\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}} + \boldsymbol{B}_t(s)\boldsymbol{\xi}, t \right) - g(\boldsymbol{\xi}, t) = &\left(g\left(\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}} + \boldsymbol{B}_t(s)\boldsymbol{\xi}, t \right) - g\left(\boldsymbol{B}_t(s)\boldsymbol{\xi}, t \right) \right) \\ &+ \left(g\left(\boldsymbol{B}_t(s)\boldsymbol{\xi}, t \right) - g(\boldsymbol{\xi}, t) \right). \end{aligned}$$

Applying the mean value theorem to each term and using that $\nabla_{\tilde{\boldsymbol{x}}} g$ has at most polynomial growth, there exist constants $C, m > 0$ such that

$$\|\nabla_{\tilde{\boldsymbol{x}}} g(\tilde{\boldsymbol{x}}, t)\|_2 \leq C(1 + \|\tilde{\boldsymbol{x}}\|_2^m).$$

Since $\boldsymbol{\xi}$ is Gaussian, all of its moments are finite. Moreover,

$$\|\boldsymbol{A}_t(s)\|_{\mathrm{op}} \leq C_t e^{-c_t s}, \qquad \|\boldsymbol{B}_t(s) - \mathbb{I}_K\|_{\mathrm{op}} \leq C_t e^{-c_t s},$$

for some constants $C_t, c_t > 0$. It follows that

$$\left| \mathbb{E}\left[g\left(\tilde{\boldsymbol{z}}^{(t)}(s), t \right) \mid \tilde{\boldsymbol{z}}^{(t)}(0) = \tilde{\boldsymbol{x}} \right] \right| \leq C_t'(1 + \|\tilde{\boldsymbol{x}}\|_2^{m+1}) e^{-c_t s},$$

for some constant $C_t' > 0$. In particular, eq. (43) is finite for every $(\tilde{\boldsymbol{x}}, t)$.

We next show that, for every fixed t, $\chi(\cdot, t)$ is twice continuously differentiable in $\tilde{\boldsymbol{x}}$. We then establish joint continuity in $(\boldsymbol{x}, t)$ after transforming these derivatives back to the original coordinates. The estimates below also justify passage of the spatial derivatives, and hence of $\tilde{\mathcal{A}}_t$, under the integral sign in step (iii).

Using the notation introduced above, we can rewrite eq. (43) as

$$\chi(\tilde{\boldsymbol{x}}, t) = -\int_0^\infty \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)} \left[g\left(\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}} + \boldsymbol{B}_t(s)\boldsymbol{\xi}, t \right) \right] ds. \tag{44}$$

We first study the gradient of χ. To do so, we must justify the exchange of the gradient and integral in eq. (44). Applying Stein's identity to the integrand of eq. (44) yields

$$F(\tilde{\boldsymbol{x}}, t, s) = \nabla_{\tilde{\boldsymbol{x}}} (P_s^t g)(\tilde{\boldsymbol{x}}) = \boldsymbol{A}_t(s)\boldsymbol{B}_t(s)^{-1} \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(0,\mathbb{I}_K)} \left[\boldsymbol{\xi}\, g\left(\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}} + \boldsymbol{B}_t(s)\boldsymbol{\xi}, t \right) \right].$$

The polynomial growth of g implies

$$\begin{aligned} \|F(\tilde{\boldsymbol{x}}, t, s)\|_2 &\leq \|\boldsymbol{A}_t(s)\boldsymbol{B}_t(s)^{-1}\|_{\mathrm{op}} \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(0,\mathbb{I}_K)} \left[\|\boldsymbol{\xi}\|_2 \left| g\left(\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}} + \boldsymbol{B}_t(s)\boldsymbol{\xi}, t \right) \right| \right] \\ &\leq C \|\boldsymbol{A}_t(s)\boldsymbol{B}_t(s)^{-1}\|_{\mathrm{op}} \left(1 + \|\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}}\|_2^m \right), \end{aligned}$$

where we used that $\boldsymbol{\xi}$ is Gaussian and $\boldsymbol{B}_t(s)$ is bounded. Now, as $s \downarrow 0$, $\|\boldsymbol{A}_t(s)\boldsymbol{B}_t(s)^{-1}\|_{\mathrm{op}} = O(s^{-1/2})$, while as $s \to \infty$,

$$\|\boldsymbol{A}_t(s)\|_{\mathrm{op}} \le C_t e^{-c_t s}, \quad \|\boldsymbol{B}_t(s)^{-1}\|_{\mathrm{op}} \le C_t.$$

Therefore

$$\|F(\tilde{\boldsymbol{x}}, t, s)\|_2 \le \begin{cases} C(1 + \|\tilde{\boldsymbol{x}}\|_2^m) s^{-1/2}, & 0 < s \le 1, \\ C_t(1 + \|\tilde{\boldsymbol{x}}\|_2^m) e^{-c_t s}, & s \ge 1, \end{cases}$$

and the right-hand side is integrable on $(0, \infty)$. Moreover, for every compact set $K \subset \mathbb{R}^K \times [0, T]$, this integrable bound can be chosen uniformly for $(\tilde{\boldsymbol{x}}, t) \in K$. By dominated convergence we deduce

$$\nabla_{\tilde{\boldsymbol{x}}} \chi(\tilde{\boldsymbol{x}}, t) = -\int_0^\infty F(\tilde{\boldsymbol{x}}, t, s) ds.$$

We next study the Hessian of χ. Since g is continuously differentiable in $\tilde{\boldsymbol{x}}$, for each fixed $s > 0$ the map $(\tilde{\boldsymbol{x}}, t) \mapsto F(\tilde{\boldsymbol{x}}, t, s)$ is continuously differentiable in $\tilde{\boldsymbol{x}}$, and

$$\nabla_{\tilde{\boldsymbol{x}}} F(\tilde{\boldsymbol{x}}, t, s) = \nabla^2_{\tilde{\boldsymbol{x}}} (P_s^t g)(\tilde{\boldsymbol{x}}) = \boldsymbol{A}_t(s)\boldsymbol{B}_t(s)^{-1} \mathbb{E}_{\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbb{I}_K)} \left[\boldsymbol{\xi} \left(\nabla_{\tilde{\boldsymbol{x}}} g \left(\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}} + \boldsymbol{B}_t(s)\boldsymbol{\xi}, t \right) \right)^\top \right] \boldsymbol{A}_t(s).$$

Using that $\nabla_{\tilde{\boldsymbol{x}}} g$ has at most polynomial growth, we obtain

$$\begin{aligned} \|\nabla_{\tilde{\boldsymbol{x}}} F(\tilde{\boldsymbol{x}}, t, s)\|_2 &\le \|\boldsymbol{A}_t(s)\boldsymbol{B}_t(s)^{-1}\|_{\mathrm{op}} \|\boldsymbol{A}_t(s)\|_{\mathrm{op}} \mathbb{E}_{\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbb{I}_K)} \left[\|\boldsymbol{\xi}\|_2 \|\nabla_{\tilde{\boldsymbol{x}}} g \left(\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}} + \boldsymbol{B}_t(s)\boldsymbol{\xi}, t \right) \|_2 \right] \\ &\le C \|\boldsymbol{A}_t(s)\boldsymbol{B}_t(s)^{-1}\|_{\mathrm{op}} \|\boldsymbol{A}_t(s)\|_{\mathrm{op}} \left(1 + \|\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}}\|_2^m\right). \end{aligned}$$

Hence,

$$\|\nabla_{\tilde{\boldsymbol{x}}} F(\tilde{\boldsymbol{x}}, t, s)\|_2 \le \begin{cases} C(1 + \|\tilde{\boldsymbol{x}}\|_2^m) s^{-1/2}, & 0 < s \le 1, \\ C_t(1 + \|\tilde{\boldsymbol{x}}\|_2^m) e^{-c_t s}, & s \ge 1, \end{cases}$$

which is again integrable on $(0, \infty)$. Again, for every compact set $K \subset \mathbb{R}^K \times [0, T]$, this integrable bound may be chosen uniformly for $(\tilde{\boldsymbol{x}}, t) \in K$, and dominated convergence yields

$$\nabla^2_{\tilde{\boldsymbol{x}}} \chi(\tilde{\boldsymbol{x}}, t) = -\int_0^\infty \nabla_{\tilde{\boldsymbol{x}}} F(\tilde{\boldsymbol{x}}, t, s)\, ds.$$

Hence, $\chi(\cdot, t)$ is twice continuously differentiable in $\tilde{\boldsymbol{x}}$ for every t.

It remains to establish continuity after transforming back to the original coordinates. A continuous choice of the eigenvector matrix $\boldsymbol{Q}(t)$ is not required. Indeed,

$$\boldsymbol{Q}(t)\boldsymbol{A}_t(s)\boldsymbol{Q}(t)^\top = e^{-s\boldsymbol{C}(t)/2}, \quad \boldsymbol{Q}(t)\boldsymbol{B}_t(s)\boldsymbol{Q}(t)^\top = \left(\mathbb{I}_K - e^{-s\boldsymbol{C}(t)}\right)^{1/2}.$$

Consequently, after undoing the rotation, the expressions for the derivatives of u depend on t through $\boldsymbol{m}(t)$, $\boldsymbol{S}(t)$, $\boldsymbol{r}$, and the continuous matrix functions of $\boldsymbol{C}(t)$ displayed above, rather than through the particular choice of $\boldsymbol{Q}(t)$.

Continuity and positive definiteness of $\boldsymbol{S}(t)$ and $\boldsymbol{\Sigma}(t)$ imply uniform upper and lower spectral bounds for $\boldsymbol{C}(t)$ on $[0, T]$. Therefore the bounds above may be chosen locally uniformly in $(\boldsymbol{x}, t)$. Dominated convergence in the Gaussian expectation and in s then gives joint continuity of the first and second spatial derivatives in $\boldsymbol{x}$ coordinates.

Step (iii): We now verify that u satisfies the Poisson equation eq. (38) pointwise. By step (i), it suffices to show that χ satisfies the Poisson equation in the transformed coordinates (42).

The estimates established in step (ii) show that $(P_s^t g)(\tilde{\boldsymbol{x}})$ is twice continuously differentiable in $\tilde{\boldsymbol{x}}$. For each $s > 0$, $(P_s^t g)$ is also continuously differentiable in s and satisfies the backward Kolmogorov equation

$$\frac{\partial}{\partial s}(P_s^t g)(\tilde{\boldsymbol{x}}) = (\tilde{\mathcal{A}}_t P_s^t g)(\tilde{\boldsymbol{x}}), \qquad s > 0.$$

Moreover, the spatial derivatives of $(P_s^t g)$ admit integrable bounds in s, locally uniformly in $\tilde{\boldsymbol{x}}$ for fixed t. Therefore the operator $\tilde{\mathcal{A}}_t$ may be passed under the integral sign in eq. (43), yielding

$$\begin{aligned} (\tilde{\mathcal{A}}_t \chi(\cdot, t))(\tilde{\boldsymbol{x}}) &= -\int_0^\infty (\tilde{\mathcal{A}}_t P_s^t g)(\tilde{\boldsymbol{x}}) ds \\ &= -\int_0^\infty \frac{\partial}{\partial s}(P_s^t g)(\tilde{\boldsymbol{x}}) ds \\ &= g(\tilde{\boldsymbol{x}}, t) - \lim_{R \to \infty} (P_R^t g)(\tilde{\boldsymbol{x}}). \end{aligned}$$

By the tail estimate established in step (ii),

$$|(P_R^t g)(\tilde{\boldsymbol{x}})| \le C_t'(1+\|\tilde{\boldsymbol{x}}\|_2^{m+1})e^{-c_t R} \to 0 \qquad \text{as } R \to \infty,$$

for each $(\tilde{\boldsymbol{x}}, t)$. Hence $(\tilde{\mathcal{A}}_t \chi(\cdot, t))(\tilde{\boldsymbol{x}}) = g(\tilde{\boldsymbol{x}}, t)$ for every $(\tilde{\boldsymbol{x}}, t)$, and consequently

$$u(\boldsymbol{x}, t) = \chi\left(\boldsymbol{Q}(t)^\top \boldsymbol{S}(t)^{-1/2}(\boldsymbol{x} - \boldsymbol{m}(t)), t\right)$$

satisfies eq. (38) pointwise. The gradient component of the Helmholtz decomposition is then $\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t) = \frac{1}{2}\nabla_{\boldsymbol{x}} u(\boldsymbol{x}, t)$.

Step (iv): It remains to identify the constructed vector fields with the components of the weak Helmholtz decomposition in Theorem 3. The estimates in step (ii) show that ψ, $\nabla_{\boldsymbol{x}}\psi$, and $\boldsymbol{h}$ have polynomial growth. Since $q(\boldsymbol{x}, t)$ is Gaussian, $\boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t)$ and $\boldsymbol{h}(\boldsymbol{x}, t)$ belong to $\mathcal{L}^2(q(\boldsymbol{x}, t), \mathbb{R}^K, \boldsymbol{\Sigma}(t))^K$.

First, by appealing to integration by parts, the pointwise identity $\nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x}, t)\boldsymbol{h}(\boldsymbol{x}, t)] = 0$ implies that $\boldsymbol{h}(\boldsymbol{x}, t)$ is q-divergence-free in the weak sense eq. (14). We next show that $\boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\psi(\cdot, t)$ is orthogonal to all weakly q-divergence-free vector fields.

Fix $\boldsymbol{v} \in \mathcal{V}_t$, and define an infinitely differentiable function ζ_R satisfying $\zeta_R = 1$ on the ball of radius R, $\zeta_R = 0$ outside the ball of radius $3R/2$, and $\|\nabla_{\boldsymbol{x}}\zeta_R\|_2 \le \frac{C}{R}$. $\zeta_R(\boldsymbol{x})\psi(\boldsymbol{x}, t)$ is twice continuously differentiable and compactly supported on the ball of radius $3R/2$. We may therefore choose $\{\varphi_{R,k}\}_{k=1}^\infty$ infinitely differentiable and compactly supported on the ball of radius $2R$ such that

$$\nabla_{\boldsymbol{x}}\varphi_{R,k} \to \nabla_{\boldsymbol{x}}\left[\zeta_R \psi(\cdot, t)\right]$$

uniformly. Since $\boldsymbol{v} \in \mathcal{V}_t$,

$$\int \boldsymbol{v}(\boldsymbol{x})^\top \nabla_{\boldsymbol{x}}\varphi_{R,k}(\boldsymbol{x}) q(\boldsymbol{x}, t) d\boldsymbol{x} = 0$$

for every k. Cauchy–Schwarz allows us to pass to the limit as $k \to \infty$, producing

$$\int \boldsymbol{v}(\boldsymbol{x})^\top \nabla_{\boldsymbol{x}}\left[\zeta_R(\boldsymbol{x})\psi(\boldsymbol{x}, t)\right] q(\boldsymbol{x}, t) d\boldsymbol{x} = 0.$$

The polynomial growth of ψ and $\nabla_{\boldsymbol{x}}\psi$, together with the Gaussian tails of $q(\boldsymbol{x}, t)$, gives

$$\left\|\boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\left[\zeta_R(\boldsymbol{x})\psi(\boldsymbol{x}, t)\right] - \boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t)\right\|_{q,t} \to 0$$

as $R \to \infty$. Taking the limit in the preceding identity therefore gives $\langle \boldsymbol{v}(\boldsymbol{x}), \boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t)\rangle_{q,t} = 0$. Since this holds for every $\boldsymbol{v} \in \mathcal{V}_t$, we have the desired result.

Thus,

$$\boldsymbol{r}(\boldsymbol{x}, t) = \boldsymbol{\Sigma}(t)\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t) + \boldsymbol{h}(\boldsymbol{x}, t)$$

is precisely the unique orthogonal decomposition supplied by Theorem 3. In particular, the two vector fields are unique up to $q(\boldsymbol{x}, t)$-almost-everywhere equality among all square-integrable weak Helmholtz decompositions. Because $q(\boldsymbol{x}, t) > 0$ for every $\boldsymbol{x}$ and the constructed vector fields are continuous, almost-everywhere uniqueness implies pointwise uniqueness. □

Theorem 4 implies that for Gaussian marginals and a state-independent diffusion coefficient, the Helmholtz decomposition exists and admits the probabilistic solution

$$\begin{aligned}
\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}, t) &= \frac{1}{2}\boldsymbol{S}(t)^{-1/2}\boldsymbol{Q}(t)\nabla_{\tilde{\boldsymbol{x}}}\chi\left(\boldsymbol{Q}(t)^\top \boldsymbol{S}(t)^{-1/2}(\boldsymbol{x} - \boldsymbol{m}(t)), t\right), \\
\nabla_{\tilde{\boldsymbol{x}}}\chi(\tilde{\boldsymbol{x}}, t) &= -\int_0^\infty \boldsymbol{A}_t(s)\boldsymbol{B}_t(s)^{-1}\mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\boldsymbol{0}, \mathbb{I}_K)}\left[\boldsymbol{\xi}\, g\left(\boldsymbol{A}_t(s)\tilde{\boldsymbol{x}} + \boldsymbol{B}_t(s)\boldsymbol{\xi}, t\right)\right] ds, \\
\boldsymbol{A}_t(s) &= e^{-s\boldsymbol{\Lambda}(t)/2}, \qquad \boldsymbol{B}_t(s) = \left(\mathbb{I}_K - e^{-s\boldsymbol{\Lambda}(t)}\right)^{1/2}, \\
\boldsymbol{C}(t) &= \boldsymbol{S}(t)^{-1/2}\boldsymbol{\Sigma}(t)\boldsymbol{S}(t)^{-1/2} = \boldsymbol{Q}(t)\boldsymbol{\Lambda}(t)\boldsymbol{Q}(t)^\top, \\
g(\tilde{\boldsymbol{x}}, t) &= \nabla_{\tilde{\boldsymbol{x}}} \cdot \tilde{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}}, t) - \tilde{\boldsymbol{x}}^\top \tilde{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}}, t), \\
\tilde{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}}, t) &= \boldsymbol{Q}(t)^\top \boldsymbol{S}(t)^{-1/2}\boldsymbol{r}\left(\boldsymbol{m}(t) + \boldsymbol{S}(t)^{1/2}\boldsymbol{Q}(t)\tilde{\boldsymbol{x}}, t\right).
\end{aligned} \tag{45}$$

B.3 Polynomial approximation of the optimal correction

In Section 4.2 we showed that at fixed marginals $\{q(\boldsymbol{x},t)\}_{0\le t\le T}$, the q-divergence-free component $\boldsymbol{h}$ of the weighted Helmholtz decomposition is the vector field $\boldsymbol{v}$ that minimizes

$$\int_0^T \mathbb{E}_{\mathbb{Q}} \left\| \boldsymbol{\Sigma}(t)^{-1/2}\{\boldsymbol{r}(\boldsymbol{x}(t),t) - \boldsymbol{v}(\boldsymbol{x}(t),t)\} \right\|^2 dt. \tag{46}$$

However, even when $q(\boldsymbol{x},t)$ is Gaussian, the correction generally admits only the probabilistic solution eq. (34). In our preliminary experiments, we found that estimating eq. (34) directly has prohibitively high variance. We make the problem tractable by restricting the divergence-free correction $\boldsymbol{v}$ to a finite-dimensional subspace comprised of polynomials of degree at most ℓ. Within this subspace, the minimizer of eq. (46) can be computed coefficient by coefficient in a Hermite polynomial basis.

Throughout this subsection, we assume only that $(\boldsymbol{m}(t), \boldsymbol{S}(t))_{0\le t\le T}$, $\boldsymbol{r}(\boldsymbol{x},t)$, $\boldsymbol{\Sigma}(t)$ satisfy the regularity conditions in Assumption (A6).

Whiten and rotate. First, we define the whitened coordinates $\tilde{\boldsymbol{x}}$ just as we did in the proof of Theorem 4:

$$\tilde{\boldsymbol{x}} = \boldsymbol{Q}(t)^\top \boldsymbol{S}(t)^{-1/2}(\boldsymbol{x} - \boldsymbol{m}(t)), \quad \widetilde{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}},t) = \boldsymbol{Q}(t)^\top \boldsymbol{S}(t)^{-1/2}\boldsymbol{r}(\boldsymbol{m}(t) + \boldsymbol{S}(t)^{1/2}\boldsymbol{Q}(t)\tilde{\boldsymbol{x}},t), \tag{47}$$

where $\boldsymbol{Q}(t)\boldsymbol{\Lambda}(t)\boldsymbol{Q}(t)^\top$ is the eigendecomposition of $\boldsymbol{C}(t) = \boldsymbol{S}(t)^{-1/2}\boldsymbol{\Sigma}(t)\boldsymbol{S}(t)^{-1/2}$. In the whitened coordinates, $\tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_K)$.

Suppose $\boldsymbol{\nu}(\tilde{\boldsymbol{x}},t)$ is divergence-free with respect to the standard Gaussian density. In the original coordinates, this vector field is

$$\boldsymbol{v}(\boldsymbol{x},t) = \boldsymbol{S}(t)^{1/2}\boldsymbol{Q}(t)\boldsymbol{\nu}\left(\boldsymbol{Q}(t)^\top \boldsymbol{S}(t)^{-1/2}(\boldsymbol{x} - \boldsymbol{m}(t)), t\right). \tag{48}$$

The objective and constraint become

$$\mathbb{E}_{\tilde{\boldsymbol{x}}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)}\left[(\widetilde{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}},t) - \boldsymbol{\nu}(\tilde{\boldsymbol{x}},t))^\top \boldsymbol{\Lambda}(t)^{-1}(\widetilde{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}},t) - \boldsymbol{\nu}(\tilde{\boldsymbol{x}},t))\right] \tag{49}$$

$$\nabla_{\tilde{\boldsymbol{x}}} \cdot \boldsymbol{\nu}(\tilde{\boldsymbol{x}},t) - \tilde{\boldsymbol{x}}^\top \boldsymbol{\nu}(\tilde{\boldsymbol{x}},t) = 0. \tag{50}$$

Indeed, eq. (50) is exactly $\nabla_{\boldsymbol{x}} \cdot [q(\boldsymbol{x},t)\boldsymbol{v}(\boldsymbol{x},t)] = 0$ after the change of variables.

Closed-form Hermite projection. Let $\mathrm{He}_d(u)$, $d \in \mathbb{N}$, be the dth unnormalized probabilists' Hermite polynomial. The collection $\{\mathrm{He}_d(u)/\sqrt{d!}\}_{d=0}^{\infty}$ constitutes an orthonormal basis of the Hilbert space $\mathcal{L}^2(\mathcal{N}(0,1);\mathbb{R})$, which is comprised of functions $f:\mathbb{R}\to\mathbb{R}$ satisfying $\mathbb{E}_{u\sim\mathcal{N}(0,1)}[f(u)^2] < \infty$ with the inner product $\langle f, g\rangle = \mathbb{E}_{u\sim\mathcal{N}(0,1)}[f(u)g(u)]$.

We can extend the univariate Hermite polynomial basis to an orthonormal basis of $\mathcal{L}^2(\mathcal{N}(\boldsymbol{0},\mathbb{I}_K);\mathbb{R})$, comprised of functions $f:\mathbb{R}^K\to\mathbb{R}$ satisfying $\mathbb{E}_{\tilde{\boldsymbol{x}}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)}[f(\tilde{\boldsymbol{x}})^2] < \infty$ with the analogous inner product. Define the collection of multi-indices

$$\mathcal{N}_{K,\ell} = \left\{ \boldsymbol{\alpha} \in \{0,1,\dots,\ell\}^K \,\middle|\, \sum_{j=1}^K \boldsymbol{\alpha}_j \le \ell \right\}, \quad \mathcal{N}_{K,+\infty} = \{0,1,\dots\}^K, \quad \ell \ge 0.$$

For a multi-index $\boldsymbol{\alpha} \in \mathcal{N}_{K,+\infty}$, we define the corresponding normalized multivariate probabilists' Hermite polynomial

$$\hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}}) = \prod_{i=1}^K \frac{\mathrm{He}_{\boldsymbol{\alpha}_i}(\tilde{\boldsymbol{x}}_i)}{\sqrt{\boldsymbol{\alpha}_i!}}. \tag{51}$$

$\{\hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}})\}_{\boldsymbol{\alpha}\in\mathcal{N}_{K,+\infty}}$ forms an orthonormal basis of $\mathcal{L}^2(\mathcal{N}(\boldsymbol{0},\mathbb{I}_K);\mathbb{R})$.

By Assumption (A6), $\widetilde{\boldsymbol{\rho}}_i(\tilde{\boldsymbol{x}},t)$, $1 \le i \le K$, has polynomial growth and hence is in $\mathcal{L}^2(\mathcal{N}(\boldsymbol{0},\mathbb{I}_K);\mathbb{R})$. Therefore, we can expand the rotated residual in the multivariate Hermite basis as

$$\widetilde{\boldsymbol{\rho}}_i(\tilde{\boldsymbol{x}},t) = \sum_{\boldsymbol{\alpha}\in\mathcal{N}_{K,+\infty}} \widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha},i}(t)\hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}}), \qquad \widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha},i}(t) = \mathbb{E}_{\tilde{\boldsymbol{x}}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)}\left[\widetilde{\boldsymbol{\rho}}_i(\tilde{\boldsymbol{x}},t)\hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}})\right]. \tag{52}$$

We choose this basis expansion because of the identity

$$(\partial_{\tilde{\boldsymbol{x}}_i} - \tilde{\boldsymbol{x}}_i)\hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}}) = -\sqrt{\boldsymbol{\alpha}_i + 1}\,\hat{H}_{\boldsymbol{\alpha}+\boldsymbol{e}_i}(\tilde{\boldsymbol{x}}). \tag{53}$$

The Gaussian weighted divergence raises the Hermite degree by one and creates no lower-order terms. Rather than keeping the infinitely many basis vectors in the expansion of $\widetilde{\boldsymbol{\rho}}_i(\tilde{\boldsymbol{x}}, t)$, we truncate the expansion eq. (52) to $\{\hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}})\}_{\boldsymbol{\alpha}\in\mathcal{N}_{K,\ell}}$, which span the $\binom{K+\ell}{\ell}$-dimensional subspace of polynomials having degree at most ℓ. In Proposition 3, we show that by doing so, we can obtain the optimal correction to the reference posterior drift restricted to this subspace.

Proposition 3 (Hermite projection): Among all polynomial fields $\boldsymbol{\nu}(\tilde{\boldsymbol{x}}, t)$ of degree at most ℓ satisfying eq. (50), the unique minimizer of eq. (49) is

$$\begin{aligned}
\mu_{\boldsymbol{\beta}}(t) &= \frac{\sum_{i=1}^K \sqrt{\boldsymbol{\beta}_i}\,\widetilde{\boldsymbol{b}}_{\boldsymbol{\beta}-\boldsymbol{e}_i,i}(t)}{\sum_{i=1}^K \lambda_i(t)\boldsymbol{\beta}_i}, && \boldsymbol{\beta} \in \mathcal{N}_{K,\ell+1} \setminus \mathcal{N}_{K,0}, \\
\widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha},i}(t) &= \widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha},i}(t) - \lambda_i(t)\sqrt{\boldsymbol{\alpha}_i + 1}\,\mu_{\boldsymbol{\alpha}+\boldsymbol{e}_i}(t), && \boldsymbol{\alpha} \in \mathcal{N}_{K,\ell}, \\
\boldsymbol{\nu}^{(\ell)}(\tilde{\boldsymbol{x}}, t) &= \sum_{\boldsymbol{\alpha}\in\mathcal{N}_{K,\ell}} \widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}(t)\hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}}),
\end{aligned} \tag{54}$$

where a coefficient with a negative multi-index is zero. The optimal correction $\boldsymbol{h}^{(\ell)}(\boldsymbol{x}, t)$ in the original coordinates is obtained from $\boldsymbol{\nu}^{(\ell)}(\tilde{\boldsymbol{x}}, t)$ using eq. (48).

Proof. Note that since $\boldsymbol{\nu}$ is a polynomial field, its components are in $\mathcal{L}^2(\mathcal{N}(\mathbf{0}, \mathbb{I}_K); \mathbb{R})$. Expanding $\boldsymbol{\nu}$ in the multivariate Hermite basis $\boldsymbol{\nu}_i(\tilde{\boldsymbol{x}}, t) = \sum_{|\boldsymbol{\alpha}|\le\ell} \widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha},i}(t)\hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}})$ and applying eq. (53) to $\boldsymbol{\nu}(\tilde{\boldsymbol{x}}, t)$ gives

$$\nabla_{\tilde{\boldsymbol{x}}} \cdot \boldsymbol{\nu}(\tilde{\boldsymbol{x}}, t) - \tilde{\boldsymbol{x}}^\top \boldsymbol{\nu}(\tilde{\boldsymbol{x}}, t) = -\sum_{1\le|\boldsymbol{\beta}|\le\ell+1} \left(\sum_{i=1}^K \sqrt{\boldsymbol{\beta}_i}\,\widetilde{\boldsymbol{c}}_{\boldsymbol{\beta}-\boldsymbol{e}_i,i}(t)\right) \hat{H}_{\boldsymbol{\beta}}(\tilde{\boldsymbol{x}}). \tag{55}$$

Thus divergence freedom gives one scalar constraint for each $\boldsymbol{\beta}$. Moreover, every coefficient $\widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha},i}(t)$ occurs in exactly one constraint, namely $\boldsymbol{\beta} = \boldsymbol{\alpha} + \boldsymbol{e}_i$.

By Parseval's identity, the part of the objective that depends on the correction is

$$\sum_{|\boldsymbol{\alpha}|\le\ell} (\widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha}}(t) - \widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}(t))^\top \boldsymbol{\Lambda}(t)^{-1} (\widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha}}(t) - \widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}(t)).$$

For a fixed $\boldsymbol{\beta}$, the relevant coefficients solve

$$\min_{\{c_i\}} \frac{1}{2} \sum_{i:\boldsymbol{\beta}_i>0} \frac{1}{\lambda_i(t)} \left(\widetilde{\boldsymbol{b}}_{\boldsymbol{\beta}-\boldsymbol{e}_i,i}(t) - c_i\right)^2 \quad \text{subject to} \quad \sum_{i:\boldsymbol{\beta}_i>0} \sqrt{\boldsymbol{\beta}_i}\,c_i = 0.$$

Introducing the Lagrange multiplier $\mu_{\boldsymbol{\beta}}(t)$ and differentiating the Lagrangian yields the first-order conditions

$$c_i = \widetilde{\boldsymbol{b}}_{\boldsymbol{\beta}-\boldsymbol{e}_i,i}(t) - \lambda_i(t)\sqrt{\boldsymbol{\beta}_i}\,\mu_{\boldsymbol{\beta}}(t).$$

Substituting this expression into the constraint yields

$$\begin{aligned}
\sum_i \sqrt{\boldsymbol{\beta}_i} c_i &= \sum_i \sqrt{\boldsymbol{\beta}_i}\widetilde{\boldsymbol{b}}_{\boldsymbol{\beta}-\boldsymbol{e}_i,i}(t) - \mu_{\boldsymbol{\beta}}(t) \sum_i \lambda_i(t)\boldsymbol{\beta}_i = 0, \\
\implies \mu_{\boldsymbol{\beta}}(t) &= \frac{\sum_{i=1}^K \sqrt{\boldsymbol{\beta}_i}\widetilde{\boldsymbol{b}}_{\boldsymbol{\beta}-\boldsymbol{e}_i,i}(t)}{\sum_{i=1}^K \lambda_i(t)\boldsymbol{\beta}_i}.
\end{aligned}$$

Because every $\lambda_i(t) > 0$, the objective is strictly convex. Consequently, this feasible stationary point is the unique minimizer. □

Recall from Appendix B.1 that the weighted Helmholtz decomposition can be obtained by solving a Poisson equation eq. (32) defined by the infinitesimal generator of the Langevin diffusion eq. (30). In the whitened coordinates, the Langevin diffusion is a time-rescaled Ornstein–Uhlenbeck process

eq. (40), and the eigenfunctions of the infinitesimal generator are the Hermite polynomials. Indeed, evaluating eq. (53) at $\boldsymbol{\alpha} - \boldsymbol{e}_i$ and combining it with $\partial_{\tilde{\boldsymbol{x}}_i} \hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}}) = \sqrt{\boldsymbol{\alpha}_i} \hat{H}_{\boldsymbol{\alpha}-\boldsymbol{e}_i}(\tilde{\boldsymbol{x}})$ produces

$$(\partial_{\tilde{\boldsymbol{x}}_i} - \tilde{\boldsymbol{x}}_i)\partial_{\tilde{\boldsymbol{x}}_i} \hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}}) = \sqrt{\boldsymbol{\alpha}_i}(\partial_{\tilde{\boldsymbol{x}}_i} - \tilde{\boldsymbol{x}}_i)\hat{H}_{\boldsymbol{\alpha}-\boldsymbol{e}_i}(\tilde{\boldsymbol{x}}) = -\boldsymbol{\alpha}_i \hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}}).$$

Hence, Proposition 3 can be understood as expanding the left and right-hand sides of the Poisson equation in the orthonormal basis of eigenfunctions and then solving for the coefficients belonging to $\mathcal{N}_{K,\ell+1}$ (indexed by $\boldsymbol{\beta}$). We then differentiate the result to obtain the gradient field $\nabla_{\tilde{\boldsymbol{x}}} \chi^{(\ell)}(\tilde{\boldsymbol{x}}, t)$.

Two useful special cases. When $\ell = 0$, $\boldsymbol{\beta} = \boldsymbol{e}_i$, eq. (54) gives $\widetilde{\boldsymbol{c}}_{\boldsymbol{0},i}(t) = 0$, so a compatible correction has no constant term and cannot change the prescribed mean path. For $\ell = 1$, write $\boldsymbol{\nu}^{(1)}(\tilde{\boldsymbol{x}}, t) = \widetilde{\boldsymbol{A}}(t)\tilde{\boldsymbol{x}}$ and set

$$\widetilde{\boldsymbol{B}}_{ij}(t) = \widetilde{\boldsymbol{b}}_{\boldsymbol{e}_j,i}(t) = \mathbb{E}_{\tilde{\boldsymbol{x}}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)}\left[\widetilde{\boldsymbol{\rho}}_i(\tilde{\boldsymbol{x}}, t)\tilde{\boldsymbol{x}}_j\right].$$

By Stein's identity, $\widetilde{\boldsymbol{B}}(t) = \mathbb{E}_{\tilde{\boldsymbol{x}}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)}[D_{\tilde{\boldsymbol{x}}}\widetilde{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}}, t)]$. The solution reduces to

$$\widetilde{\boldsymbol{A}}_{ij}(t) = \frac{\lambda_j(t)\widetilde{\boldsymbol{B}}_{ij}(t) - \lambda_i(t)\widetilde{\boldsymbol{B}}_{ji}(t)}{\lambda_i(t) + \lambda_j(t)}, \qquad \widetilde{\boldsymbol{A}}_{ii}(t) = 0. \tag{56}$$

Hence the linear correction is skew-symmetric in whitened coordinates. If $\boldsymbol{\Lambda}(t)$ is proportional to the identity, it is simply the antisymmetric part of the averaged Jacobian.

Monte Carlo. Computing the polynomial approximation to the optimal divergence-free vector field requires computing the coefficients of $\widetilde{\boldsymbol{\rho}}_i(\tilde{\boldsymbol{x}}, t)$ in the Hermite basis. Generally, these coefficients cannot be computed in closed-form. However, they do have the unbiased estimator

$$\widehat{\widetilde{\boldsymbol{b}}}_{\boldsymbol{\alpha},i}(t) = \frac{1}{N_{\mathrm{mc}}} \sum_{n=1}^{N_{\mathrm{mc}}} \widetilde{\boldsymbol{\rho}}_i(\tilde{\boldsymbol{x}}^{(n)}, t)\hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}}^{(n)}), \qquad \tilde{\boldsymbol{x}}^{(n)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \mathbb{I}_K). \tag{57}$$

Substituting any estimated coefficients into eq. (54) still makes eq. (55) vanish identically. Thus, the estimated field is q-divergence-free, although it will generally not be the optimal degree ℓ polynomial q-divergence-free vector field. In the stochastic VI algorithm, the batch of points $\boldsymbol{x}_{j,k} \overset{\text{i.i.d.}}{\sim} q(\boldsymbol{x}, t_j)$ used to evaluate the path KL $\mathcal{L}_{\mathrm{path}}$ and the batch used to compute the correction to the reference drift at these points should be independent.

Gauss–Hermite quadrature. Alternatively, for low dimensional latent spaces, we can use Gauss–Hermite quadrature to estimate the Hermite coefficients of the residual vector field. If N_{quad} is the number of quadrature points along each latent dimension, the estimator is

$$\widehat{\widetilde{\boldsymbol{b}}}_{\boldsymbol{\alpha},i}(t) = \sum_{n=1}^{N_{\mathrm{quad}}^K} w_n \widetilde{\boldsymbol{\rho}}_i(\tilde{\boldsymbol{x}}_n, t)\hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}}_n), \qquad \sum_{n=1}^{N_{\mathrm{quad}}^K} w_n = 1, \tag{58}$$

where the sum is taken over the Cartesian product of one-dimensional quadrature points $\tilde{\boldsymbol{x}}_n$, and w_n is the product of the corresponding one-dimensional quadrature weights. When $\widetilde{\boldsymbol{\rho}}_i(\tilde{\boldsymbol{x}}, t)\hat{H}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{x}})$ is a polynomial of degree at most $2N_{\mathrm{quad}} - 1$, Gauss–Hermite quadrature computes the Hermite coefficient $\widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha},i}(t)$ exactly.

Stein estimand. If derivatives of the residual are available, Stein's identity provides an alternate estimator for $\widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha},i}(t)$. For any multi-index $\boldsymbol{\kappa}$ such that $\boldsymbol{\kappa}_j \leq \boldsymbol{\alpha}_j, 1 \leq j \leq K$, repeated integration by parts gives

$$\widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha},i}(t) = \sqrt{\frac{(\boldsymbol{\alpha}-\boldsymbol{\kappa})!}{\boldsymbol{\alpha}!}}\, \mathbb{E}_{\tilde{\boldsymbol{x}}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)}\left[\partial_{\tilde{\boldsymbol{x}}}^{\boldsymbol{\kappa}} \widetilde{\boldsymbol{\rho}}_i(\tilde{\boldsymbol{x}}, t)\hat{H}_{\boldsymbol{\alpha}-\boldsymbol{\kappa}}(\tilde{\boldsymbol{x}})\right]. \tag{59}$$

The expectation in eq. (59) can be approximated using either Monte Carlo or Gauss–Hermite quadrature. The choice $\boldsymbol{\kappa} = \boldsymbol{0}$ recovers eq. (57). If derivatives through order $|\boldsymbol{\alpha}|$ are available, choosing $\boldsymbol{\kappa} = \boldsymbol{\alpha}$ removes the Hermite polynomial entirely:

$$\widetilde{\boldsymbol{b}}_{\boldsymbol{\alpha},i}(t) = \frac{1}{\sqrt{\boldsymbol{\alpha}!}} \mathbb{E}_{\tilde{\boldsymbol{x}}\sim\mathcal{N}(\boldsymbol{0},\mathbb{I}_K)}\left[\partial_{\tilde{\boldsymbol{x}}}^{\boldsymbol{\alpha}} \widetilde{\boldsymbol{\rho}}_i(\tilde{\boldsymbol{x}}, t)\right].$$

Under the twice-differentiability condition in Assumption (A6), one may use any $\boldsymbol{\kappa} \in \mathcal{N}_{K,2}$ with $\boldsymbol{\kappa} \leq \boldsymbol{\alpha}$ componentwise. The derivatives with respect to $\tilde{\boldsymbol{x}}$ can be obtained from spatial derivatives of $\boldsymbol{r}$ through the affine change of coordinates in eq. (47) and evaluated using automatic differentiation.

Although the normalized Hermite polynomials have unit second moments, their higher moments grow rapidly with their degree. Integration by parts can therefore reduce the variance of the coefficient estimates by replacing a high-degree Hermite factor with derivatives of the residual. This reduction is not guaranteed, since the derivatives may themselves have high variance. The choice of $\boldsymbol{\kappa}$ trades the degree of the remaining Hermite factor against the cost and variability of higher-order derivatives. Regardless of this choice, substituting the resulting coefficient estimates into eq. (54) preserves the divergence-free constraint exactly.

Computational complexity. There are $K\binom{K+\ell}{\ell}$ scalar coefficients. Once they have been estimated, eq. (54) touches each coefficient a constant number of times and costs $O(K\binom{K+\ell}{\ell})$, rather than $O(\binom{K+\ell+1}{\ell+1}^3)$ for a generic dense solve of the polynomial Poisson equation. Including residual evaluations, Hermite features, and eigendecomposition, the cost at one time point is

$$O\left(N_{\text{est}} C_{\boldsymbol{r}} + N_{\text{est}} K \binom{K+\ell}{\ell} + K^3\right),$$

where $C_{\boldsymbol{r}}$ is the cost of one residual evaluation and N_{est} is the number of Monte Carlo samples or quadrature points used to approximate the Hermite coefficients. When ℓ is fixed, $\binom{K+\ell}{\ell} = \Theta(K^\ell)$, so the coefficient-dependent term scales as $K^{\ell+1}$.

The Stein estimand eq. (59) additionally requires derivatives of the transformed residual. For the full-order choice $\boldsymbol{\kappa} = \boldsymbol{\alpha}$, the highest-order derivative tensor has K^ℓ entries for each of the K output components, and hence $K^{\ell+1}$ entries in total. Using automatic differentiation, computing these derivatives costs $O(K^{\ell+1} C_{\boldsymbol{r}})$.

Relationship to Taylor approximation. In an earlier version of this work, we constructed the polynomial correction by approximating the residual $\boldsymbol{r}$ with a multivariate Taylor expansion about the posterior mean $\boldsymbol{m}(t)$ and then computing the weighted Helmholtz decomposition of the resulting polynomial vector field. The present version also computes the weighted Helmholtz decomposition of a polynomial vector field, but its coefficients are obtained by expanding the residual in the truncated multivariate Hermite polynomial basis.

The distinction between these two approaches is their approximation criterion: a Taylor expansion matches local derivatives at the posterior mean, whereas a truncated Hermite expansion with exact coefficients gives the orthogonal projection of the residual onto the space of polynomial vector fields having degree at most ℓ, with respect to the Gaussian marginal. As shown above, the weighted Helmholtz decomposition of this projection yields the optimal degree ℓ polynomial q-divergence-free correction at fixed marginals. In practice, the Hermite coefficients are generally estimated numerically, and the resulting correction need not attain this optimum. Nonetheless, both the Taylor and Hermite constructions preserve the q-divergence-free constraint exactly.

Moreover, the two approximations are closely related through the Stein representation. For $\ell = 1$, the Taylor approximation uses the Jacobian of the residual evaluated at the posterior mean, while the Hermite construction uses the expectation of the Jacobian under the Gaussian marginal. Consequently, approximating this expectation with one-node Gauss–Hermite quadrature recovers the degree-one Taylor correction.

B.4 Verification of assumptions for the Helmholtz-SDE variational family

In Appendices B.2 and B.3 we discussed how to obtain a correction $\boldsymbol{h}^{(\ell)}$ to the reference posterior drift from the weighted Helmholtz decomposition of the residual vector field $\boldsymbol{r} = \boldsymbol{f}_p - \boldsymbol{f}_q$. It remains to establish that this correction produces a valid variational SDE. In this subsection, we verify that the Gaussian variational family used by Helmholtz-SDE satisfies the assumptions in Appendix A.1.

Let $q(\boldsymbol{x}, t) = \mathcal{N}(\boldsymbol{m}(t), \boldsymbol{S}(t))$. We assume that the prior SDE admits a unique strong solution on $[0, T]$ and that $(\boldsymbol{m}(t), \boldsymbol{S}(t))_{0 \leq t \leq T}$, $\boldsymbol{r}(\boldsymbol{x}, t)$, and $\boldsymbol{\Sigma}(t)$ satisfy the regularity conditions in Assumption (A6). Assumptions (A1) to (A5) are verified under these conditions for $0 \leq \ell \leq 3$.

Helmholtz-SDE uses the reference drift

$$\boldsymbol{f}_q(\boldsymbol{x},t) = \frac{d}{dt}\boldsymbol{m}(t) + \left\{\left(\frac{d}{dt}\boldsymbol{L}(t)\right)\boldsymbol{L}(t)^{-1} - \frac{1}{2}\boldsymbol{\Sigma}(t)\boldsymbol{S}(t)^{-1}\right\}(\boldsymbol{x}-\boldsymbol{m}(t)),$$

where $\boldsymbol{L}(t)$ is the symmetric square root of $\boldsymbol{S}(t)$. The full posterior drift is $\boldsymbol{f}_q(\boldsymbol{x},t) + \boldsymbol{h}^{(\ell)}(\boldsymbol{x},t)$, where $\boldsymbol{h}^{(\ell)}$ is the degree ℓ correction constructed in Appendix B.3. The posterior initial density is $q(\boldsymbol{x},0)$, and the prior initial distribution is $\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{V}_0)$, with $\boldsymbol{V}_0$ positive definite.

Because $[0,T]$ is compact, the eigenvalues of $\boldsymbol{S}(t)$ and $\boldsymbol{\Sigma}(t)$ are bounded above and bounded away from zero. In unrotated whitened coordinates, the polynomial coefficients defining $\boldsymbol{h}^{(\ell)}$ are continuous, and hence bounded, in t. Indeed, continuity follows from Assumption (A6), dominated convergence under the Gaussian marginals, and the finite-dimensional projection in a fixed basis in Proposition 3.

Existence and uniqueness. Introduce the whitened variable $\boldsymbol{z}$

$$\boldsymbol{z} = \boldsymbol{L}(t)^{-1}(\boldsymbol{x}-\boldsymbol{m}(t)), \qquad \boldsymbol{\nu}^{(\ell)}(\boldsymbol{z},t) = \boldsymbol{L}(t)^{-1}\boldsymbol{h}^{(\ell)}(\boldsymbol{m}(t)+\boldsymbol{L}(t)\boldsymbol{z},t).$$

For the symmetric-square-root reference drift, the terms involving $d\boldsymbol{L}(t)/dt$ cancel under this change of variables. Consequently, the posterior SDE becomes

$$\begin{aligned} d\boldsymbol{z}(t) &= \left\{-\frac{1}{2}\boldsymbol{C}(t)\boldsymbol{z}(t) + \boldsymbol{\nu}^{(\ell)}(\boldsymbol{z}(t),t)\right\}dt + \boldsymbol{L}(t)^{-1}\boldsymbol{\Sigma}(t)^{1/2}d\boldsymbol{w}(t), \\ \boldsymbol{C}(t) &= \boldsymbol{S}(t)^{-1/2}\boldsymbol{\Sigma}(t)\boldsymbol{S}(t)^{-1/2}. \end{aligned} \tag{60}$$

The drift in eq. (60) is polynomial in $\boldsymbol{z}$, and is therefore locally Lipschitz. This gives a pathwise-unique strong solution up to its explosion time.

It remains to rule out explosion. The correction is Gaussian divergence-free, so eq. (50) gives

$$\boldsymbol{z}^\top\boldsymbol{\nu}^{(\ell)}(\boldsymbol{z},t) = \nabla_{\boldsymbol{z}}\cdot\boldsymbol{\nu}^{(\ell)}(\boldsymbol{z},t).$$

For $\ell = 0$, the correction is zero. For $1 \le \ell \le 3$, the right-hand side is a polynomial of degree at most $\ell - 1$, and hence for some constant $C > 0$

$$\left|\boldsymbol{z}^\top\boldsymbol{\nu}^{(\ell)}(\boldsymbol{z},t)\right| \le CV(\boldsymbol{z}), \qquad (\boldsymbol{z},t) \in \mathbb{R}^K \times [0,T],$$

where V is the Lyapunov function $V(\boldsymbol{z}) = 1 + \|\boldsymbol{z}\|^2$. The infinitesimal generator of eq. (60) satisfies

$$\mathcal{L}_t V(\boldsymbol{z}) = -\boldsymbol{z}^\top\boldsymbol{C}(t)\boldsymbol{z} + 2\boldsymbol{z}^\top\boldsymbol{\nu}^{(\ell)}(\boldsymbol{z},t) + \mathrm{tr}(\boldsymbol{C}(t)),$$

and is therefore bounded above by $CV(\boldsymbol{z})$. Moreover,

$$\lim_{R\to\infty}\inf_{\|\boldsymbol{z}\|>R} V(\boldsymbol{z}) = \infty.$$

The non-explosion criterion from Khasminskii [70, Theorem 3.5] therefore implies that a pathwise-unique, strong solution is defined on all of $[0,T]$. This proves Assumption (A1) for the posterior SDE. Notice that our argument does not require the posterior drift to have linear growth.

Fokker–Planck equation and Gaussian marginals. By construction, $\{q(\boldsymbol{x},t)\}_{0\le t\le T}$ is a weak solution to the Fokker–Planck equation for the reference drift $\boldsymbol{f}_q$. Moreover,

$$\nabla_{\boldsymbol{x}}\cdot\left[q(\boldsymbol{x},t)\boldsymbol{h}^{(\ell)}(\boldsymbol{x},t)\right] = 0, \quad (\boldsymbol{x},t) \in \mathbb{R}^K \times [0,T]$$

so the same $q(\boldsymbol{x},t)$ also solves the weak Fokker–Planck equation for the corrected drift $\boldsymbol{f}_q + \boldsymbol{h}^{(\ell)}$. The corrected drift is polynomial, $\boldsymbol{\Sigma}(t)$ is bounded, and Gaussian distributions have moments of every finite order. Hence q belongs to the regularity class eq. (13).

It remains to show that, within the regularity class eq. (13), the Fokker–Planck equation has a unique solution. The superposition principle (Trevisan [71, Theorem 2.5]) states that any weak Fokker–Planck solution $\{\rho(\boldsymbol{x},t)\}_{0\le t\le T}$ can be lifted to a probability distribution over continuous paths

whose one-time marginals are $\{\rho(\boldsymbol{x},t)\}_{0\le t\le T}$ and which is a weak solution to the SDE with drift $\boldsymbol{f}_q+\boldsymbol{h}^{(\ell)}$ and diffusion coefficient $\boldsymbol{\Sigma}^{1/2}$.

Suppose $\{q(\boldsymbol{x},t)\}_{0\le t\le T}$ and $\{\tilde{q}(\boldsymbol{x},t)\}_{0\le t\le T}$ are two weak solutions to the Fokker–Planck equation with drift $\boldsymbol{f}_q+\boldsymbol{h}^{(\ell)}$ and diffusion coefficient $\boldsymbol{\Sigma}^{1/2}$. By the superposition principle, we obtain two probability distributions having one-time marginals $\{q(\boldsymbol{x},t)\}_{0\le t\le T}$ and $\{\tilde{q}(\boldsymbol{x},t)\}_{0\le t\le T}$ that are weak solutions to the variational SDE with drift $\boldsymbol{f}_q+\boldsymbol{h}^{(\ell)}$ and diffusion coefficient $\boldsymbol{\Sigma}^{1/2}$. The strong existence and pathwise uniqueness established above imply that this SDE is unique in law. Therefore, the two weak solutions have the same path law and hence the same one-time marginals. Thus $q(\boldsymbol{x},t)=\tilde{q}(\boldsymbol{x},t)$ almost everywhere in $\boldsymbol{x}$ for every $t\in[0,T]$. This verifies Assumption (A5) for the posterior drift.

Densities. The posterior marginals are non-degenerate Gaussians and hence have smooth, strictly positive Lebesgue densities. This verifies the posterior part of Assumption (A2).

Absolute continuity. Define the first exit time from the ball of radius n

$$\tau_n=\inf\left\{t\in[0,T]\mid\|\boldsymbol{x}(t)\|_2\ge n\right\},\qquad \inf\varnothing=\infty.$$

Since the prior and posterior drifts are locally bounded and $\boldsymbol{\Sigma}(t)^{-1/2}$ is bounded on $[0,T]$, there exists $K_n<\infty$ such that

$$\sup_{\substack{\|\boldsymbol{y}\|_2\le n\\0\le t\le T}}\left\|\boldsymbol{\Sigma}(t)^{-1/2}\left[\boldsymbol{f}_q(\boldsymbol{y},t)+\boldsymbol{h}^{(\ell)}(\boldsymbol{y},t)-\boldsymbol{f}_p(\boldsymbol{y},t)\right]\right\|_2\le K_n.$$

Consequently,

$$\mathbb{E}_{\mathbb{P}}\left[\exp\left\{\frac{1}{2}\int_0^{T\wedge\tau_n}\left\|\boldsymbol{\Sigma}(t)^{-1/2}\left[\boldsymbol{f}_q(\boldsymbol{x}(t),t)+\boldsymbol{h}^{(\ell)}(\boldsymbol{x}(t),t)-\boldsymbol{f}_p(\boldsymbol{x}(t),t)\right]\right\|_2^2dt\right\}\right]\le\exp\left\{\frac{1}{2}K_n^2T\right\}<\infty.$$

Thus, Novikov's criterion (Karatzas and Shreve [64, Corollary 3.5.13]) holds for the stopped process.

Both initial densities are non-degenerate Gaussian densities and are therefore strictly positive on $\mathbb{R}^K$. Define

$$\begin{aligned}\mathcal{Z}_T^{(n)}=\frac{q(\boldsymbol{x}(0),0)}{p(\boldsymbol{x}(0),0)}\exp\Bigg\{&\sum_{i=1}^K\int_0^{T\wedge\tau_n}\left[\boldsymbol{\Sigma}(t)^{-1/2}\left\{\boldsymbol{f}_q(\boldsymbol{x}(t),t)+\boldsymbol{h}^{(\ell)}(\boldsymbol{x}(t),t)-\boldsymbol{f}_p(\boldsymbol{x}(t),t)\right\}\right]_i d\boldsymbol{w}_i(t)\\&-\frac{1}{2}\int_0^{T\wedge\tau_n}\left\|\boldsymbol{\Sigma}(t)^{-1/2}\left[\boldsymbol{f}_q(\boldsymbol{x}(t),t)+\boldsymbol{h}^{(\ell)}(\boldsymbol{x}(t),t)-\boldsymbol{f}_p(\boldsymbol{x}(t),t)\right]\right\|_2^2dt\Bigg\},\end{aligned}$$

where $(\boldsymbol{w}(t))_{0\le t\le T}$ is a K-dimensional Brownian motion with respect to $\mathbb{P}$.

Novikov's criterion implies that the stochastic exponential appearing in $\mathcal{Z}_T^{(n)}$ is a martingale. Conditionally on $\boldsymbol{x}(0)$, it has expectation one, and therefore

$$\mathbb{E}_{\mathbb{P}}\left[\mathcal{Z}_T^{(n)}\right]=\mathbb{E}_{\mathbb{P}}\left[\frac{q(\boldsymbol{x}(0),0)}{p(\boldsymbol{x}(0),0)}\right]=\int\frac{q(\boldsymbol{x},0)}{p(\boldsymbol{x},0)}p(\boldsymbol{x},0)\,d\boldsymbol{x}=1.$$

We may therefore define a probability measure $\widehat{\mathbb{Q}}^{(n)}$ on $\mathcal{F}_T$ by

$$\frac{d\widehat{\mathbb{Q}}^{(n)}}{d\mathbb{P}}=\mathcal{Z}_T^{(n)}.$$

Under $\widehat{\mathbb{Q}}^{(n)}$, Girsanov's theorem shows that

$$\widehat{\boldsymbol{w}}^{(n)}(t)=\boldsymbol{w}(t)-\int_0^{t\wedge\tau_n}\boldsymbol{\Sigma}(s)^{-1/2}\left\{\boldsymbol{f}_q(\boldsymbol{x}(s),s)+\boldsymbol{h}^{(\ell)}(\boldsymbol{x}(s),s)-\boldsymbol{f}_p(\boldsymbol{x}(s),s)\right\}ds$$

is a Brownian motion. Hence, up to τ_n, the coordinate process satisfies

$$d\boldsymbol{x}(t)=\left[\boldsymbol{f}_q(\boldsymbol{x}(t),t)+\boldsymbol{h}^{(\ell)}(\boldsymbol{x}(t),t)\right]dt+\boldsymbol{\Sigma}(t)^{1/2}d\widehat{\boldsymbol{w}}^{(n)}(t),$$

and its initial density is $q(\boldsymbol{x},0)$.

Let $\mathcal{G}_n = \sigma\{\boldsymbol{x}(t \wedge \tau_n) : 0 \le t \le T\}$ be the sigma-algebra generated by the stopped coordinate process. Under both $\widehat{\mathbb{Q}}^{(n)}$ and $\mathbb{Q}$, the stopped coordinate process has initial density $q(\boldsymbol{x}, 0)$ and satisfies the same stopped posterior SDE. Uniqueness in law therefore gives

$$\widehat{\mathbb{Q}}^{(n)}\Big|_{\mathcal{G}_n} = \mathbb{Q}|_{\mathcal{G}_n}.$$

Because $\boldsymbol{\Sigma}(t)$ is invertible, the prior Brownian motion up to τ_n is determined by the stopped coordinate process. Hence $\mathcal{Z}_T^{(n)}$ is $\mathcal{G}_n$-measurable. It follows that, for every $A \in \mathcal{G}_n$,

$$\mathbb{Q}(A) = \widehat{\mathbb{Q}}^{(n)}(A) = \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}_A \mathcal{Z}_T^{(n)}\right].$$

Define the corresponding unstopped stochastic exponential by

$$\mathcal{Z}_T = \frac{q(\boldsymbol{x}(0), 0)}{p(\boldsymbol{x}(0), 0)} \exp\left\{ \sum_{i=1}^{K} \int_0^T \left[\boldsymbol{\Sigma}(t)^{-1/2}\left\{\boldsymbol{f}_q(\boldsymbol{x}(t), t) + \boldsymbol{h}^{(\ell)}(\boldsymbol{x}(t), t) - \boldsymbol{f}_p(\boldsymbol{x}(t), t)\right\}\right]_i d\boldsymbol{w}_i(t) \right.$$
$$\left. - \frac{1}{2}\int_0^T \left\|\boldsymbol{\Sigma}(t)^{-1/2}\left[\boldsymbol{f}_q(\boldsymbol{x}(t), t) + \boldsymbol{h}^{(\ell)}(\boldsymbol{x}(t), t) - \boldsymbol{f}_p(\boldsymbol{x}(t), t)\right]\right\|_2^2 dt \right\}.$$

This random variable is well defined $\mathbb{P}$-almost surely by local boundedness and non-explosion.

To remove the stopping time, let $E_n = \{\tau_n > T\}$. Since the prior and posterior processes have strong solutions defined on all of $[0, T]$, their paths are continuous and bounded on $[0, T]$. Therefore, $E_n \uparrow \Omega$ both $\mathbb{P}$-almost surely and $\mathbb{Q}$-almost surely. Moreover, by definition, $\mathcal{Z}_T^{(n)} = \mathcal{Z}_T$ on E_n. For every A measurable with respect to $\sigma(\boldsymbol{x}(t) \mid 0 \le t \le T)$, the event $A \cap E_n$ belongs to $\mathcal{G}_n$, and hence

$$\mathbb{Q}(A \cap E_n) = \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}_{A \cap E_n} \mathcal{Z}_T^{(n)}\right] = \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}_{A \cap E_n} \mathcal{Z}_T\right].$$

Continuity from below under $\mathbb{Q}$ and monotone convergence under $\mathbb{P}$ now give

$$\mathbb{Q}(A) = \lim_{n \to \infty} \mathbb{Q}(A \cap E_n) = \lim_{n \to \infty} \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}_{A \cap E_n} \mathcal{Z}_T\right] = \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}_A \mathcal{Z}_T\right].$$

Therefore,

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \mathcal{Z}_T.$$

In particular, $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$. Since $\mathcal{Z}_T > 0$ $\mathbb{P}$-almost surely, the reverse absolute continuity also holds. Therefore, the prior and posterior path laws are mutually absolutely continuous.

Finite KL divergence. The initial-distribution term is finite because both initial distributions are non-degenerate Gaussians.

For the path-space term, observe that

$$\boldsymbol{f}_p(\boldsymbol{x}, t) - \left\{\boldsymbol{f}_q(\boldsymbol{x}, t) + \boldsymbol{h}^{(\ell)}(\boldsymbol{x}, t)\right\} = \boldsymbol{r}(\boldsymbol{x}, t) - \boldsymbol{h}^{(\ell)}(\boldsymbol{x}, t).$$

Both terms on the right have polynomial growth in $\boldsymbol{x}$, uniformly for $t \in [0, T]$, while $\boldsymbol{\Sigma}(t)^{-1}$ is uniformly bounded. Since the posterior marginals are Gaussian and their means and covariances are bounded on $[0, T]$, they have uniformly bounded moments of every finite order. Consequently,

$$\int_0^T \mathbb{E}_{\mathbb{Q}}\left\|\boldsymbol{\Sigma}(t)^{-1/2}\left[\boldsymbol{f}_p(\boldsymbol{x}(t), t) - \left\{\boldsymbol{f}_q(\boldsymbol{x}(t), t) + \boldsymbol{h}^{(\ell)}(\boldsymbol{x}(t), t)\right\}\right]\right\|^2 dt < \infty.$$

Thus Assumption (A4) holds, and the path-space KL identity eq. (11) is finite.

We have therefore verified the assumptions of Appendix A.1 for the Helmholtz-SDE variational family with $0 \le \ell \le 3$, conditional on the stated assumptions on the prior. For $\ell \ge 4$, the divergence-free identity still preserves the prescribed Fokker–Planck equation, but the Lyapunov estimate above does not establish non-explosion. Global well-posedness must be verified separately.

B.5 Optimality of Helmholtz-SDE

Now that we have established that Helmholtz-SDE produces a valid variational SDE, we study its optimality. First, we prove Theorem 1, which shows Helmholtz-SDE approximates the posterior arbitrarily closely in the conjugate setting. We then study the nonconjugate setting. We quantify the suboptimality resulting from restricting the q-divergence-free correction to a polynomial vector field of degree ℓ. As a result, we get a lower bound on the improvement over the variational family defined by the reference drift.

Throughout this subsection, we let $\boldsymbol{h}^{(\ell)}$ denote the degree ℓ correction obtained from the weighted Helmholtz decomposition applied to the polynomial vector field computed from the *exact* Hermite coefficients of the residual. In the conjugate setting of Appendix B.5.1, the residual vector field is affine, so the exact Hermite coefficients can be computed using Gauss–Hermite quadrature with $N_{\text{quad}} = 2$ nodes per latent dimension. In the nonconjugate setting of Appendix B.5.2, this is generally not the case; our analysis does not include coefficient estimation error from Monte Carlo or numerical quadrature.

B.5.1 Conjugate setting

Proof of Theorem 1. Suppose that the prior SDE is linear with a nondegenerate Gaussian initial law and that the observation model is nondegenerate linear-Gaussian. The posterior path law $\mathbb{Q}^\star$ is then Gauss–Markov, its drift $\boldsymbol{f}^\star(\boldsymbol{x}, t)$ is affine in $\boldsymbol{x}$, and its one-time marginals have the form $q^\star(\boldsymbol{x}, t) = \mathcal{N}(\boldsymbol{m}^\star(t), \boldsymbol{S}^\star(t))$ [30, 72]. The covariance $\boldsymbol{S}^\star(t)$ is positive definite. The functions $\boldsymbol{m}^\star(t)$ and $\boldsymbol{S}^\star(t)$ are continuous and piecewise continuously differentiable, with possible derivative jumps at the observation times [15, Appendix J].

For any continuously differentiable mean and covariance paths $(\boldsymbol{m}(t), \boldsymbol{S}(t))_{0 \le t \le T}$, Helmholtz-SDE uses the affine reference drift from Proposition 2 and adds the degree-1 correction obtained from Proposition 3. Because the prior and reference drifts are affine, the residual vector field is affine. Moreover, the identity eq. (53) shows that the divergence-free projection preserves the Hermite degree. Consequently, the projection onto the subspace $\mathcal{V}_{t,1}$ of q-divergence-free vector fields that are polynomials of degree at most one is equal to the projection onto the subspace $\mathcal{V}_t$ of all q-divergence-free vector fields. Thus, at any prescribed Gaussian marginals, Helmholtz-SDE selects the drift that minimizes the path-space KL divergence to the prior among all admissible drifts having those marginals.

The paths $\boldsymbol{m}^\star$ and $\boldsymbol{S}^\star$ need not be continuously differentiable at the observation times, so $\mathbb{Q}^\star$ need not itself belong to the variational family $\mathcal{Q}_{(1)}$. We instead argue that $\mathbb{Q}^\star$ can be approximated arbitrarily closely in KL by $\mathcal{Q}_{(1)}$. Fix $\delta > 0$, and let $U_\delta \subset [0, T]$ be a union of neighborhoods of the observation times with Lebesgue measure at most δ. The mean and covariance paths can be smoothed inside U_δ to obtain continuously differentiable functions $\boldsymbol{m}^\delta$ and $\boldsymbol{S}^\delta$ that agree with $\boldsymbol{m}^\star$ and $\boldsymbol{S}^\star$ outside U_δ and converge uniformly to them as $\delta \to 0$. The smoothing may be chosen so that the first derivatives remain uniformly bounded in δ. For sufficiently small δ, $\boldsymbol{S}^\delta(t)$ remains positive definite because $\lambda_{\min}(\boldsymbol{S}^\star(t))$ is bounded away from zero on $[0, T]$.

Let $\mathbb{Q}^\delta$ be the path law produced by Helmholtz-SDE with the degree-1 correction applied to $\boldsymbol{m}^\delta$ and $\boldsymbol{S}^\delta$. Outside U_δ, these marginal paths and their derivatives agree with those of $\mathbb{Q}^\star$. The corresponding Helmholtz-SDE drift therefore minimizes the fixed-marginal path-space KL objective at $q^\star$. The posterior drift $\boldsymbol{f}^\star$ minimizes the same objective, since replacing it by a drift with the same marginals would leave the expected log likelihood and the initial KL term unchanged while changing only the path-space KL term. Strict convexity of the fixed-marginal objective gives

$$\boldsymbol{f}_q^\delta(\boldsymbol{x}, t) + \boldsymbol{h}^{(1),\delta}(\boldsymbol{x}, t) = \boldsymbol{f}^\star(\boldsymbol{x}, t), \qquad t \notin U_\delta,$$

up to $q^\star(\boldsymbol{x}, t)$-almost-everywhere equality.

The drifts $\boldsymbol{f}_q^\delta + \boldsymbol{h}^{(1),\delta}$ and $\boldsymbol{f}^\star$ are affine in $\boldsymbol{x}$. The uniform bounds on the smoothed marginal paths and their derivatives imply that their affine coefficients are uniformly bounded on U_δ. The Gaussian marginals of $\mathbb{Q}^\delta$ also have uniformly bounded second moments. Since $\boldsymbol{\Sigma}(t)^{-1}$ is bounded on $[0, T]$, there is a constant $C < \infty$, independent of δ, such that

$$\mathbb{E}_{\mathbb{Q}^\delta} \left\| \boldsymbol{\Sigma}(t)^{-1/2} \left\{ \boldsymbol{f}_q^\delta(\boldsymbol{x}(t), t) + \boldsymbol{h}^{(1),\delta}(\boldsymbol{x}(t), t) - \boldsymbol{f}^\star(\boldsymbol{x}(t), t) \right\} \right\|_2^2 \le C, \qquad t \in U_\delta.$$

Applying the path-space KL identity to $\mathbb{Q}^\delta$ and $\mathbb{Q}^\star$ gives

$$
\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}\left(\mathbb{Q}^\delta \parallel \mathbb{Q}^\star\right) \\
&= \mathrm{D}_{\mathrm{KL}}\left(q^\delta(\boldsymbol{x},0) \parallel q^\star(\boldsymbol{x},0)\right) + \frac{1}{2}\int_{U_\delta} \mathbb{E}_{\mathbb{Q}^\delta}\left\|\boldsymbol{\Sigma}(t)^{-1/2}\left\{\boldsymbol{f}_q^\delta(\boldsymbol{x}(t),t) + \boldsymbol{h}^{(1),\delta}(\boldsymbol{x}(t),t) - \boldsymbol{f}^\star(\boldsymbol{x}(t),t)\right\}\right\|_2^2 dt \\
&\leq \mathrm{D}_{\mathrm{KL}}\left(q^\delta(\boldsymbol{x},0) \parallel q^\star(\boldsymbol{x},0)\right) + \frac{C\delta}{2}.
\end{aligned}
$$

The Gaussian initial KL converges to zero by uniform convergence of the mean and covariance paths. Hence $\mathrm{D}_{\mathrm{KL}}\left(\mathbb{Q}^\delta \parallel \mathbb{Q}^\star\right) \to 0$ as $\delta \to 0$, and therefore $\inf_{\mathbb{Q}\in\mathcal{Q}_{(1)}} \mathrm{D}_{\mathrm{KL}}\left(\mathbb{Q} \parallel \mathbb{Q}^\star\right) = 0$. □

B.5.2 Nonconjugate setting

In the nonconjugate setting, the posterior generally does not have Gaussian one-time marginals. Simulation-free variational inference algorithms that impose Gaussian one-time marginals therefore share a common, irreducible approximation error. Helmholtz-SDE further restricts the variational family by considering a subset of divergence-free vector fields.

As in Theorem 3, we consider the Hilbert space $\mathcal{L}^2(q(\boldsymbol{x},t), \mathbb{R}^K, \boldsymbol{\Sigma}(t))^K$ of vector fields satisfying $\mathbb{E}_{q(\boldsymbol{x},t)}\|\boldsymbol{f}(\boldsymbol{x})\|^2 < \infty$ equipped with the inner product defined in eq. (35). Let $\mathcal{V}_t$ be the closed subspace of weakly q-divergence-free vector fields, let $\mathcal{P}_\ell$ be the subspace of polynomial vector fields of degree at most ℓ, and let $\mathcal{V}_{t,\ell} = \mathcal{V}_t \cap \mathcal{P}_\ell$.

Let $\Pi_{\mathcal{S}}$ be the orthogonal projection onto the subspace $\mathcal{S}$. Then the degree ℓ correction is $\boldsymbol{h}^{(\ell)}(\boldsymbol{x},t) = [\Pi_{\mathcal{V}_{t,\ell}}\boldsymbol{r}(\cdot,t)](\boldsymbol{x})$. First, we quantify the reduction in negative ELBO at *fixed marginals* from increasing the degree of the polynomial correction.

Proposition 4 (Improvement at fixed marginals): Suppose Assumption (A6) holds, and fix one-time marginals $\{q(\boldsymbol{x},t)\}_{0\leq t\leq T}$. Let $\ell \geq 0$ and $k \geq 1$ be integers for which the degree ℓ and $\ell + k$ corrections induce admissible path laws.
The reduction in negative ELBO from using the degree $\ell + k$ correction in place of the degree ℓ correction is

$$
\begin{aligned}
\Delta_{\ell,\ell+k}(q) &= \frac{1}{2}\int_0^T \left\{\|\boldsymbol{r}(\boldsymbol{x},t) - \boldsymbol{h}^{(\ell)}(\boldsymbol{x},t)\|_{q,t}^2 - \|\boldsymbol{r}(\boldsymbol{x},t) - \boldsymbol{h}^{(\ell+k)}(\boldsymbol{x},t)\|_{q,t}^2\right\} dt \\
&= \frac{1}{2}\int_0^T \|\boldsymbol{h}^{(\ell+k)}(\boldsymbol{x},t) - \boldsymbol{h}^{(\ell)}(\boldsymbol{x},t)\|_{q,t}^2 dt.
\end{aligned} \tag{61}
$$

Equivalently,

$$
\Delta_{\ell,\ell+k}(q) = \frac{1}{2}\sum_{\ell<|\boldsymbol{\alpha}|\leq\ell+k}\int_0^T \widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}(t)^\top \boldsymbol{\Lambda}(t)^{-1}\widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}(t)dt. \tag{62}
$$

$\widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}(t)$ are the coefficients of the q-divergence-free vector field in the Hermite basis obtained from Proposition 3.

Proof. Since the path laws obtained from $\boldsymbol{h}^{(\ell)}$ and $\boldsymbol{h}^{(\ell+k)}$ possess one-time marginals $\{q(\boldsymbol{x},t)\}_{0\leq t\leq T}$, their reconstruction losses and initial KL divergences are equal. Therefore, the reduction in negative ELBO is the difference between the path-space KL divergences to the prior. The second equality in eq. (61) follows from the Pythagorean identity for projections onto the nested subspaces $\mathcal{V}_{t,\ell} \subseteq \mathcal{V}_{t,\ell+k}$.

The identity eq. (53) maps each Hermite degree to the next degree and does not couple different degrees. Consequently, $\mathcal{V}_t$ is the orthogonal direct sum of its Hermite-degree components, and the degree ℓ and $\ell + k$ corrections have identical Hermite coefficients through degree ℓ. Orthonormality of the Hermite basis gives eq. (62). □

Proposition 4 compares polynomial degrees at fixed marginals. We now consider optimizing within each variational family. Define $\mathcal{Q}_{(\ell)}$ to be the variational family obtained by applying Helmholtz-SDE with the degree ℓ correction to the Gaussian one-time marginals whose mean and covariance satisfy

Assumption (A6). From eq. (62), we obtain a lower bound on the improvement from using $\mathcal{Q}_{(\ell+k)}$ over $\mathcal{Q}_{(\ell)}$.

Proposition 5 (Suboptimality of variational families): Let $\ell \geq 0$ and $k \geq 1$ be integers for which the degree ℓ and $\ell + k$ corrections induce admissible path laws. For any $\epsilon > 0$, suppose that $\mathbb{Q}_\epsilon \in \mathcal{Q}_{(\ell)}$ has marginals $q_\epsilon(\boldsymbol{x}, t)$ and satisfies

$$\mathrm{D_{KL}}\left(\mathbb{Q}_\epsilon \parallel \mathbb{Q}^\star\right) \leq \inf_{\mathbb{Q} \in \mathcal{Q}_{(\ell)}} \mathrm{D_{KL}}\left(\mathbb{Q} \parallel \mathbb{Q}^\star\right) + \epsilon.$$

Let $\boldsymbol{h}_\epsilon^{(\ell)}$ and $\boldsymbol{h}_\epsilon^{(\ell+k)}$ be the degree ℓ and $\ell + k$ corrections computed at $q_\epsilon(\boldsymbol{x}, t)$. Then

$$\begin{aligned} &\inf_{\mathbb{Q} \in \mathcal{Q}_{(\ell)}} \mathrm{D_{KL}}\left(\mathbb{Q} \parallel \mathbb{Q}^\star\right) - \inf_{\mathbb{Q} \in \mathcal{Q}_{(\ell+k)}} \mathrm{D_{KL}}\left(\mathbb{Q} \parallel \mathbb{Q}^\star\right) \\ &\geq \frac{1}{2} \int_0^T \|\boldsymbol{h}_\epsilon^{(\ell+k)}(\boldsymbol{x}, t) - \boldsymbol{h}_\epsilon^{(\ell)}(\boldsymbol{x}, t)\|_{q_\epsilon, t}^2 dt - \epsilon. \end{aligned} \tag{63}$$

Proof. Let $\mathbb{Q}_\epsilon^{(\ell+k)}$ be the path law obtained by adding $\boldsymbol{h}_\epsilon^{(\ell+k)}$ to the reference drift of $\mathbb{Q}_\epsilon$. Then $\mathbb{Q}_\epsilon^{(\ell+k)}$ and $\mathbb{Q}_\epsilon$ have the same one-time marginals because $\boldsymbol{h}_\epsilon^{(\ell+k)}$ is q_ϵ-divergence-free. Their reconstruction losses and initial KL divergences are equal. The difference between their KL divergences to $\mathbb{Q}^\star$ is consequently the difference between their path-space KL divergences to the prior. By Proposition 4 and the ϵ-optimality of $\mathbb{Q}_\epsilon$,

$$\begin{aligned} &\inf_{\mathbb{Q} \in \mathcal{Q}_{(\ell)}} \mathrm{D_{KL}}\left(\mathbb{Q} \parallel \mathbb{Q}^\star\right) - \inf_{\mathbb{Q} \in \mathcal{Q}_{(\ell+k)}} \mathrm{D_{KL}}\left(\mathbb{Q} \parallel \mathbb{Q}^\star\right) \\ &\geq \mathrm{D_{KL}}\left(\mathbb{Q}_\epsilon \parallel \mathbb{Q}^\star\right) - \mathrm{D_{KL}}\left(\mathbb{Q}_\epsilon^{(\ell+k)} \parallel \mathbb{Q}^\star\right) - \epsilon \\ &= \Delta_{\ell,\ell+k}(q_\epsilon) - \epsilon = \frac{1}{2} \int_0^T \|\boldsymbol{h}_\epsilon^{(\ell+k)}(\boldsymbol{x}, t) - \boldsymbol{h}_\epsilon^{(\ell)}(\boldsymbol{x}, t)\|_{q_\epsilon, t}^2 dt - \epsilon. \end{aligned}$$

□

In particular, by choosing $\ell = 0$, Proposition 5 provides a lower bound on the improvement over the variational family defined by the reference drift (e.g., SVISE [6, 7] and SDE Matching [8]).

Proposition 5 provides a heuristic for estimating the improvement of $\mathcal{Q}_{(\ell+k)}$ over $\mathcal{Q}_{(\ell)}$. After fitting Helmholtz-SDE with a degree ℓ correction, we hold the fitted marginals and generative model parameters fixed, compute the Hermite coefficients through degree $\ell + k$, and evaluate eq. (63). We use the resulting quantity as an estimate of the improvement, taking the unknown optimization error ϵ to be zero. With an optimal degree ℓ fit and exact coefficient evaluation, this quantity is a lower bound on the improvement between the variational families. We validate this heuristic for the stochastic Lorenz attractor in Appendix D.1.

B.6 Algorithm summary

We summarize the Hermite implementation of the polynomial correction in Algorithm 2 and the resulting Helmholtz-SDE training procedure in Algorithm 3.

The coefficient samples used by `HelmCorrection` are independent of the state samples used to estimate the path-space KL. The coefficient-estimation configuration $\mathcal{E}$ specifies the integration rule (Monte Carlo or Gauss–Hermite quadrature), the sample count or number of quadrature nodes per dimension, and the Stein derivative order (zero for the direct estimator). The corresponding coefficient estimators are described in Appendix B.3.

Algorithm 2: `HelmCorrection`: Hermite approximation to the divergence-free correction

Require: polynomial order ℓ; prior drift $\boldsymbol{f}_p$; reference drift $\boldsymbol{f}_q$; diffusion covariance $\boldsymbol{\Sigma}$; posterior mean $\boldsymbol{m}$ and covariance $\boldsymbol{S}$ at time t; coefficient-estimation configuration $\mathcal{E}$

Whiten and rotate

▷ Compute the symmetric square root $\boldsymbol{S}^{1/2}$ and its inverse $\boldsymbol{S}^{-1/2}$

▷ Form $\boldsymbol{C} \leftarrow \boldsymbol{S}^{-1/2}\boldsymbol{\Sigma}(t)\boldsymbol{S}^{-1/2}$ and compute $\boldsymbol{C} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$

Estimate the Hermite coefficients of the rotated residual

▷ Define $\widetilde{\boldsymbol{\rho}}(\boldsymbol{v}, t) \leftarrow \boldsymbol{Q}^\top \boldsymbol{S}^{-1/2} \left\{ \boldsymbol{f}_p(\boldsymbol{m} + \boldsymbol{S}^{1/2}\boldsymbol{Q}\boldsymbol{v}, t) - \boldsymbol{f}_q(\boldsymbol{m} + \boldsymbol{S}^{1/2}\boldsymbol{Q}\boldsymbol{v}, t) \right\}$

for $\boldsymbol{\alpha} \in \mathcal{N}_{K,\ell}$ *and* $i = 1, \ldots, K$ **do**

  ▷ Estimate $\widehat{\widetilde{\boldsymbol{b}}}_{\boldsymbol{\alpha},i}$ using the coefficient estimator specified by $\mathcal{E}$ (Appendix B.3)

end

Project coefficient by coefficient onto the divergence-free subspace

for $\boldsymbol{\beta} \in \mathcal{N}_{K,\ell+1} \setminus \mathcal{N}_{K,0}$ **do**

  ▷ $\widehat{\mu}_{\boldsymbol{\beta}} \leftarrow \left\{ \sum_{i=1}^K \sqrt{\boldsymbol{\beta}_i}\, \widehat{\widetilde{\boldsymbol{b}}}_{\boldsymbol{\beta}-\boldsymbol{e}_i,i} \right\} / \left\{ \sum_{i=1}^K \lambda_i \boldsymbol{\beta}_i \right\}$

end

for $\boldsymbol{\alpha} \in \mathcal{N}_{K,\ell}$ *and* $i = 1, \ldots, K$ **do**

  ▷ $\widehat{\widetilde{\boldsymbol{c}}}_{\boldsymbol{\alpha},i} \leftarrow \widehat{\widetilde{\boldsymbol{b}}}_{\boldsymbol{\alpha},i} - \lambda_i \sqrt{\boldsymbol{\alpha}_i + 1}\, \widehat{\mu}_{\boldsymbol{\alpha}+\boldsymbol{e}_i}$

end

Return the correction in the original coordinates

▷ Define $\widehat{\boldsymbol{\nu}}^{(\ell)}(\boldsymbol{v}, t) \leftarrow \sum_{\boldsymbol{\alpha} \in \mathcal{N}_{K,\ell}} \widehat{\widetilde{\boldsymbol{c}}}_{\boldsymbol{\alpha}} \hat{H}_{\boldsymbol{\alpha}}(\boldsymbol{v})$

▷ Define $\widehat{\boldsymbol{h}}^{(\ell)}(\boldsymbol{x}, t) \leftarrow \boldsymbol{S}^{1/2}\boldsymbol{Q}\widehat{\boldsymbol{\nu}}^{(\ell)}\left(\boldsymbol{Q}^\top \boldsymbol{S}^{-1/2}(\boldsymbol{x} - \boldsymbol{m}), t\right)$

return $\widehat{\boldsymbol{h}}^{(\ell)}(\boldsymbol{x}, t)$

Algorithm 3: Helmholtz-SDE training

Require: dataset $\mathcal{D} = \{\boldsymbol{y}^{(m)}(t_n)\}_{n=1,m=1}^{N,M}$; prior drift $\boldsymbol{f}_p(\boldsymbol{x}, t \mid \boldsymbol{\theta})$; diffusion covariance $\boldsymbol{\Sigma}(t \mid \boldsymbol{\theta})$; observation model $p(\boldsymbol{y}(t_n) \mid \boldsymbol{x}(t_n), \boldsymbol{\theta})$; model parameters $\boldsymbol{\theta}$; parameters $\boldsymbol{\phi}$ of the posterior marginal paths $(\boldsymbol{m}^{(m)}(t \mid \boldsymbol{\phi}), \boldsymbol{S}^{(m)}(t \mid \boldsymbol{\phi}))_{0 \le t \le T}$; reference drift $\boldsymbol{f}_q$; polynomial order ℓ; sample sizes N_{time} and N_{space}; coefficient-estimation configuration $\mathcal{E}$

for *learning iterations* **do**

Compute the initial KL term

▷ $\mathcal{L}_{\text{init}} \leftarrow \sum_{m=1}^{M} \mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(\boldsymbol{m}^{(m)}(0 \mid \boldsymbol{\phi}), \boldsymbol{S}^{(m)}(0 \mid \boldsymbol{\phi})) \,\|\, \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{V}_0)\right)$

Estimate the reconstruction term

▷ Sample $s_j \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{t_1, \ldots, t_N\}$ for $j = 1, \ldots, N_{\text{time}}$

▷ Sample $\boldsymbol{x}_{m,j,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{m}^{(m)}(s_j \mid \boldsymbol{\phi}), \boldsymbol{S}^{(m)}(s_j \mid \boldsymbol{\phi}))$ for $k = 1, \ldots, N_{\text{space}}$

▷ $\widehat{\mathcal{L}}_{\text{rec}} \leftarrow -\frac{N}{N_{\text{time}} N_{\text{space}}} \sum_{m=1}^{M} \sum_{j=1}^{N_{\text{time}}} \sum_{k=1}^{N_{\text{space}}} \log p(\boldsymbol{y}^{(m)}(s_j) \mid \boldsymbol{x}_{m,j,k}, \boldsymbol{\theta})$

Estimate the path-space KL term

▷ Sample $s_j \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0, T]$ for $j = 1, \ldots, N_{\text{time}}$

for $m = 1, \ldots, M$ *and* $j = 1, \ldots, N_{\text{time}}$ **do**

▷ $\widehat{\boldsymbol{h}}_{m,j}^{(\ell)}(\cdot) \leftarrow \texttt{HelmCorrection}(\ell, \boldsymbol{f}_p, \boldsymbol{f}_q, \boldsymbol{\Sigma}, \boldsymbol{m}^{(m)}(s_j \mid \boldsymbol{\phi}), \boldsymbol{S}^{(m)}(s_j \mid \boldsymbol{\phi}), s_j, \mathcal{E})$

end

▷ Independently sample $\boldsymbol{x}_{m,j,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{m}^{(m)}(s_j \mid \boldsymbol{\phi}), \boldsymbol{S}^{(m)}(s_j \mid \boldsymbol{\phi}))$ for $k = 1, \ldots, N_{\text{space}}$

▷ $\widehat{\mathcal{L}}_{\text{path}} \leftarrow \frac{T}{2 N_{\text{time}} N_{\text{space}}} \sum_{m=1}^{M} \sum_{j=1}^{N_{\text{time}}} \sum_{k=1}^{N_{\text{space}}}$
$\left\| \boldsymbol{\Sigma}(s_j \mid \boldsymbol{\theta})^{-1/2} \left\{ \boldsymbol{f}_p(\boldsymbol{x}_{m,j,k}, s_j \mid \boldsymbol{\theta}) - \boldsymbol{f}_q(\boldsymbol{x}_{m,j,k}, s_j) - \widehat{\boldsymbol{h}}_{m,j}^{(\ell)}(\boldsymbol{x}_{m,j,k}) \right\} \right\|_2^2$

Update the model and variational parameters

▷ $\mathcal{L} \leftarrow \widehat{\mathcal{L}}_{\text{rec}} + \mathcal{L}_{\text{init}} + \widehat{\mathcal{L}}_{\text{path}}$

▷ $\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow \texttt{gradient-step}(\mathcal{L}, \boldsymbol{\theta}, \boldsymbol{\phi})$

end

C Linear SDE experiments: the Ornstein–Uhlenbeck spiral

In this section, we study the linear SDE prior

$$\mathbb{P} : d\boldsymbol{x}(t) = (-\alpha \mathbb{I}_2 + \omega \boldsymbol{J})\boldsymbol{x}(t)dt + d\boldsymbol{w}(t), \quad 0 \le t \le T, \quad p(\boldsymbol{x}, 0) = \mathcal{N}(\boldsymbol{\mu}(0), \boldsymbol{V}(0)), \tag{64}$$

and the linear, Gaussian observation model $p(\boldsymbol{y}(t_i)|\boldsymbol{x}(t_i)) = \mathcal{N}(\boldsymbol{C}\boldsymbol{x}(t_i) + \boldsymbol{d}, \boldsymbol{R})$. We refer to this model as the *Ornstein–Uhlenbeck (OU) spiral*.

The prior SDE $\mathbb{P}$ has stationary distribution $\mathcal{N}(\boldsymbol{0}, 1/(2\alpha)\mathbb{I}_2)$. Moreover, the drift function of $\mathbb{P}$ is the sum of a gradient field $-\alpha\mathbb{I}_2\boldsymbol{x}$ as well as a vector field $\omega\boldsymbol{J}\boldsymbol{x}$ that is divergence-free with respect to the stationary distribution.

In the observation model, $\boldsymbol{C} \in \mathbb{R}^{2\times 2}$ is a linear transformation having determinant 1, $\boldsymbol{d} \in \mathbb{R}^2$ is an offset vector, and $\boldsymbol{R} = \sigma^2\boldsymbol{C}\boldsymbol{C}^\top$ is the observation noise, where $\sigma^2 > 0$.

First, in Appendix C.1 we demonstrate that neither SVISE [6, 7] nor SDE-Matching [8] can be expected to recover the true posterior in this model, even though it is tractable. Then, in Appendix C.2 we provide additional simulation results that validate our theoretical claim.

C.1 Theoretical results for OU spiral

In this subsection, we theoretically characterize the suboptimality of the symmetric and square root gauges when the prior and likelihood models are known and fixed:

$$d^\star_{\mathcal{Q}} = \inf_{\mathbb{Q}\in\mathcal{Q}} \mathrm{D}_{\mathrm{KL}}\left(\mathbb{Q} \parallel \mathbb{Q}^\star\right), \tag{65}$$

where $\mathbb{Q}^\star$ is the true posterior for the latent SDE model (Section 3.1). In the OU spiral model, since the prior SDE is linear and the likelihood model is linear and Gaussian, $\mathbb{Q}^\star$ is a Gaussian process.

Even though exact inference is possible, we show that *existing simulation-free VI algorithms cannot be expected to recover the posterior*. Specifically, consider prior initial distribution $p(\boldsymbol{x}, 0)$ equal to the stationary distribution $\mathcal{N}(\boldsymbol{0}, 1/(2\alpha)\mathbb{I}_2)$ and all continuously differentiable marginal parameters $(\boldsymbol{m}(t), \boldsymbol{S}(t))_{0\le t\le T}$ with $\boldsymbol{S}(t)$ positive definite. We show that $d^\star_{\mathcal{Q}}$ grows at least linearly with ω for the corresponding variational families defined by the symmetric and square root gauges. We denote them by $\mathcal{Q}_{\mathrm{sym}}$ and $\mathcal{Q}_{\mathrm{sqrt}}$, respectively.

Theorem 5 (Suboptimality of simulation-free VI algorithms in the OU spiral): Consider the OU spiral model eq. (64), and let $\mathbb{Q}^\star$ denote the (true) posterior SDE.
Fix any $c_0 > 1$ and $\omega \ge c_0\alpha$. Then every distribution $\mathbb{Q} \in \mathcal{Q}_{\mathrm{sym}} \cup \mathcal{Q}_{\mathrm{sqrt}}$ satisfies

$$\mathrm{D}_{\mathrm{KL}}\left(\mathbb{Q} \parallel \mathbb{Q}^\star\right) \ge \min\left\{\frac{1}{8}, \frac{1 - \frac{1}{2c_0}}{2\sqrt{2}}\right\} T\omega - \frac{1}{2\sqrt{2}} + C_{\mathrm{OU}}$$

where

$$C_{\mathrm{OU}}(\omega, \alpha, \boldsymbol{C}, \boldsymbol{d}, \sigma^2, T, \boldsymbol{y}) = \log p(\boldsymbol{y}) + N\log(2\pi\sigma^2). \tag{66}$$

Moreover, if $p(\boldsymbol{x}, 0) = \mathcal{N}(\boldsymbol{0}, \frac{1}{2\alpha}\mathbb{I}_2)$, then C_{OU} is lower bounded by a constant $C_{\mathrm{OU-stationary}}$ that does not depend on ω.

Proof. Our proof consists of five key steps:

(i) Express the KL divergence to $\mathbb{Q}^\star$ at any pair $(\boldsymbol{m}(t), \boldsymbol{S}(t))_{0\le t\le T}$ as a sum of two functions $f_{\boldsymbol{m}}$ and $f_{\boldsymbol{S}}$, where $f_{\boldsymbol{m}}$ is a function of the marginal means $(\boldsymbol{m}(t))_{0\le t\le T}$ and $f_{\boldsymbol{S}}$ is a function of the marginal covariances $(\boldsymbol{S}(t))_{0\le t\le T}$.

(ii) Justify that $f_{\boldsymbol{m}}$ is minimized at the posterior mean $(\boldsymbol{m}^\star(t))_{0\le t\le T}$.

(iii) Rewrite $f_{\boldsymbol{S}}$ in terms of the eigendecomposition of $\boldsymbol{S}(t)$ at each time $0 \le t \le T$.

(iv) Use an occupation-time argument to show that either $\lambda_1(t)$ is often bounded away from zero, in which case the rotational mismatch term is large, or else $\lambda_1(t)$ is often close to zero, in which case keeping it small requires a large negative control and hence large control cost.

(v) Show that when $p(\boldsymbol{x}, 0)$ is the stationary distribution of the prior (64), $\log p(\boldsymbol{y})$ is lower bounded by a constant not depending on ω. This is done using uniform (in ω) spectral bounds on the observation covariance matrix.

Step (i): Let $\mathbb{Q}$ be an SDE with diffusion coefficient $\mathbb{I}_2$ and Gaussian marginals $q(\boldsymbol{x}, t) = \mathcal{N}(\boldsymbol{m}(t), \boldsymbol{S}(t))$, where $(\boldsymbol{m}(t))_{0 \le t \le T}$ is a differentiable mean function and $(\boldsymbol{S}(t))_{0 \le t \le T}$ is a differentiable and symmetric, positive definite covariance function. Moreover, suppose that $\mathbb{Q}$ has drift function of the form

$$\boldsymbol{f}_q(\boldsymbol{x}, t) = \boldsymbol{A}_q(t)\boldsymbol{x} + \frac{d}{dt}\boldsymbol{m}(t) - \boldsymbol{A}_q(t)\boldsymbol{m}(t), \quad \boldsymbol{A}_q(t) = \boldsymbol{K}(t) - \frac{1}{2}\boldsymbol{S}(t)^{-1},$$

where $\boldsymbol{K}(t)$ is a solution to the Lyapunov equation eq. (18). In Proposition 2, we showed that this drift function induces the marginals $q(\boldsymbol{x}, t)$.

Using Girsanov's Theorem (Theorem 2), the KL divergence to the posterior decomposes as

$$\mathrm{D}_{\mathrm{KL}}\left(\mathbb{Q} \parallel \mathbb{Q}^\star\right) = f_{\boldsymbol{m}}(\boldsymbol{m}) + f_{\boldsymbol{S}}(\boldsymbol{S}) + C_{\mathrm{OU}}(\omega, \alpha, \boldsymbol{C}, \boldsymbol{d}, \sigma^2, T, \boldsymbol{y}),$$

where $f_{\boldsymbol{m}}$ depends only on the mean path, $f_{\boldsymbol{S}}$ depends only on the covariance path, and C_{OU} is a constant defined in eq. (66) not depending on $\mathbb{Q}$.

These functions are given by

$$\begin{aligned}
f_{\boldsymbol{m}}(\boldsymbol{m}) &= \frac{1}{2}(\boldsymbol{m}(0) - \boldsymbol{\mu}(0))^\top \boldsymbol{V}(0)^{-1}(\boldsymbol{m}(0) - \boldsymbol{\mu}(0)) + \frac{1}{2}\int_0^T \left\| \frac{d}{dt}\boldsymbol{m}(t) - (-\alpha\mathbb{I}_2 + \omega\boldsymbol{J})\boldsymbol{m}(t) \right\|_2^2 dt \\
&+ \frac{1}{2}\sum_{i=1}^N \left(\boldsymbol{C}\boldsymbol{m}(t_i) + \boldsymbol{d} - \boldsymbol{y}(t_i)\right)^\top \boldsymbol{R}^{-1}\left(\boldsymbol{C}\boldsymbol{m}(t_i) + \boldsymbol{d} - \boldsymbol{y}(t_i)\right), \\
f_{\boldsymbol{S}}(\boldsymbol{S}) &= \frac{1}{2}\left[\mathrm{tr}\left(\boldsymbol{V}(0)^{-1}\boldsymbol{S}(0)\right) - \log\frac{\det\left(\boldsymbol{S}(0)\right)}{\det\left(\boldsymbol{V}(0)\right)} - 2\right] + \\
&\frac{1}{2}\int_0^T \mathrm{tr}\left(\left((\boldsymbol{K}(t) - \frac{1}{2}\boldsymbol{S}(t)^{-1}) - (-\alpha\mathbb{I}_2 + \omega\boldsymbol{J})\right)\boldsymbol{S}(t)\left((\boldsymbol{K}(t) - \frac{1}{2}\boldsymbol{S}(t)^{-1}) - (-\alpha\mathbb{I}_2 + \omega\boldsymbol{J})\right)^\top\right) dt \\
&+ \frac{1}{2}\sum_{n=1}^N \mathrm{tr}\left(\boldsymbol{R}^{-1}\boldsymbol{C}\boldsymbol{S}(t_n)\boldsymbol{C}^\top\right). \qquad (67)
\end{aligned}$$

Notice that $f_{\boldsymbol{S}}$ does not depend on the observations $\{\boldsymbol{y}(t_n)\}_{n=1}^N$.

Since $\boldsymbol{R} = \sigma^2\boldsymbol{C}\boldsymbol{C}^\top$, we have $\boldsymbol{C}^\top\boldsymbol{R}^{-1}\boldsymbol{C} = \sigma^{-2}\mathbb{I}_2$, and therefore

$$\frac{1}{2}\sum_{i=1}^N \mathrm{tr}\left(\boldsymbol{R}^{-1}\boldsymbol{C}\boldsymbol{S}(t_i)\boldsymbol{C}^\top\right) = \frac{1}{2\sigma^2}\sum_{i=1}^N \mathrm{tr}\left(\boldsymbol{S}(t_i)\right) \ge 0.$$

The first term of $f_{\boldsymbol{S}}$ is also nonnegative since the matrix $\boldsymbol{A} = \boldsymbol{V}(0)^{-1/2}\boldsymbol{S}(0)\boldsymbol{V}(0)^{-1/2}$ is positive definite and

$$\mathrm{tr}\left(\boldsymbol{V}(0)^{-1}\boldsymbol{S}(0)\right) - \log\frac{\det\left(\boldsymbol{S}(0)\right)}{\det\left(\boldsymbol{V}(0)\right)} - 2 = \sum_{i=1}^2 \left(\lambda_i(\boldsymbol{A}) - \log\lambda_i(\boldsymbol{A}) - 1\right) \ge 0.$$

We will drop these two terms from $f_{\boldsymbol{S}}$ when constructing our lower bound. The single remaining term is the path KL term.

Step (ii): Because the posterior path law is Gaussian, the map $\boldsymbol{m} \mapsto f_{\boldsymbol{m}}(\boldsymbol{m})$ is a strictly convex quadratic functional. Its unique minimizer is the posterior mean path $(\boldsymbol{m}^\star(t))_{0 \le t \le T}$. Hence

$$f_{\boldsymbol{m}}(\boldsymbol{m}) \ge f_{\boldsymbol{m}}(\boldsymbol{m}^\star),$$

with equality if and only if $\boldsymbol{m} = \boldsymbol{m}^\star$. The optimum is ≥ 0 since each term in $f_{\boldsymbol{m}}$ is nonnegative. Therefore, the suboptimality of the family is controlled entirely by the covariance term $f_{\boldsymbol{S}}$.

Step (iii): Let $\lambda_1(t) \le \lambda_2(t)$ denote the ordered eigenvalues of $\boldsymbol{S}(t)$. Since $\boldsymbol{S}(\cdot)$ is continuously differentiable and symmetric positive definite, Weyl's perturbation inequality implies that λ_1, λ_2 are Lipschitz, hence absolutely continuous and differentiable almost everywhere.

Define the sets

$$A = \{t \in [0,T] : \lambda_1(t) < \lambda_2(t)\}, \qquad E = \{t \in [0,T] : \lambda_1(t) = \lambda_2(t)\}.$$

We will decompose the integral in $f_{\boldsymbol{S}}$ over each of these sets individually.

We will start with A, the more involved case. A is an open subset of $[0,T]$, and thus can be written as a countable union of disjoint open intervals. On each such interval, we can choose a continuously differentiable rotation matrix $\boldsymbol{U}(t)$ such that

$$\boldsymbol{S}(t) = \boldsymbol{U}(t)\boldsymbol{\Lambda}(t)\boldsymbol{U}(t)^\top,\ \boldsymbol{U}(t) = \begin{pmatrix} \cos\theta(t) & -\sin\theta(t) \\ \sin\theta(t) & \cos\theta(t) \end{pmatrix},\ \boldsymbol{\Lambda}(t) = \operatorname{diag}\big(\lambda_1(t), \lambda_2(t)\big).$$

Using this eigendecomposition, we rewrite the integrand appearing in $f_{\boldsymbol{S}}$ as

$$\operatorname{tr}\left(\widetilde{\boldsymbol{\Delta}}(t)\boldsymbol{\Lambda}(t)\widetilde{\boldsymbol{\Delta}}(t)^\top\right), \quad \widetilde{\boldsymbol{\Delta}}(t) = \boldsymbol{U}(t)^\top\big((\boldsymbol{K}(t) - \frac{1}{2}\boldsymbol{S}(t)^{-1}) - (-\alpha\mathbb{I}_2 + \omega\boldsymbol{J})\big)\boldsymbol{U}(t).$$

Since $\boldsymbol{U}(t)^\top(-\alpha\mathbb{I}_2 + \omega\boldsymbol{J})\boldsymbol{U}(t) = -\alpha\mathbb{I}_2 + \omega\boldsymbol{J}$, we need to compute $\widetilde{\boldsymbol{A}}_q(t) = \boldsymbol{U}(t)^\top(\boldsymbol{K}(t) - \frac{1}{2}\boldsymbol{S}(t)^{-1})\boldsymbol{U}(t)$ for the symmetric and square root gauges.

(a) **Symmetric.** Define the skew-symmetric matrix

$$\boldsymbol{\Omega}(t) = \boldsymbol{U}(t)^\top \frac{d}{dt}\boldsymbol{U}(t) = \begin{pmatrix} 0 & -w(t) \\ w(t) & 0 \end{pmatrix}, \quad w(t) = \frac{d}{dt}\theta(t).$$

By definition of the symmetric gauge, $\boldsymbol{A}_q(t) = \boldsymbol{K}(t) - \frac{1}{2}\boldsymbol{S}(t)^{-1}$ is the unique symmetric solution to

$$\frac{d}{dt}\boldsymbol{S}(t) - \mathbb{I}_2 = \boldsymbol{A}_q(t)\boldsymbol{S}(t) + \boldsymbol{S}(t)\boldsymbol{A}_q(t)^\top, \quad 0 \le t \le T. \tag{68}$$

A simple algebraic calculation yields

$$\boldsymbol{U}(t)^\top \frac{d}{dt}\boldsymbol{S}(t)\boldsymbol{U}(t) = \frac{d}{dt}\boldsymbol{\Lambda}(t) + \boldsymbol{\Omega}(t)\boldsymbol{\Lambda}(t) + \boldsymbol{\Lambda}(t)\boldsymbol{\Omega}(t)^\top. \tag{69}$$

Multiplying eq. (68) by $\boldsymbol{U}(t)^\top$ on the left and $\boldsymbol{U}(t)$ on the right, and then substituting eq. (69) into eq. (68) yields

$$\frac{d}{dt}\boldsymbol{\Lambda}(t) + \boldsymbol{\Omega}(t)\boldsymbol{\Lambda}(t) + \boldsymbol{\Lambda}(t)\boldsymbol{\Omega}(t)^\top - \mathbb{I}_2 = \widetilde{\boldsymbol{A}}_q\boldsymbol{\Lambda}(t) + \boldsymbol{\Lambda}(t)\widetilde{\boldsymbol{A}}_q^\top, \quad 0 \le t \le T.$$

Hence, the entries of $\widetilde{\boldsymbol{A}}_q(t)$ are

$$(\widetilde{\boldsymbol{A}}_q(t))_{ij} = \begin{cases} (\frac{d}{dt}\lambda_i(t) - 1)/(2\lambda_i(t)) & \text{if } i = j \\ \boldsymbol{\Omega}_{ij}(t)(\lambda_j(t) - \lambda_i(t))/(\lambda_i(t) + \lambda_j(t)) & \text{if } i \ne j \end{cases}. \tag{70}$$

(b) **Square root.** The square root gauge sets $\boldsymbol{K}(t) = (\frac{d}{dt}\boldsymbol{S}(t)^{1/2})\boldsymbol{S}(t)^{-1/2}$, where $\boldsymbol{S}(t)^{1/2} = \boldsymbol{U}(t)\boldsymbol{\Lambda}(t)^{1/2}\boldsymbol{U}(t)^\top$ is the symmetric square root. Rewriting $\boldsymbol{K}(t)$ in terms of $\boldsymbol{U}(t)$ and $\boldsymbol{\Lambda}(t)$ yields

$$\begin{aligned}\left(\frac{d}{dt}\boldsymbol{S}(t)^{1/2}\right)\boldsymbol{S}(t)^{-1/2} &= \frac{d}{dt}\boldsymbol{U}(t)\boldsymbol{U}(t)^\top + (1/2)\boldsymbol{U}(t)\frac{d}{dt}\boldsymbol{\Lambda}(t)\boldsymbol{\Lambda}(t)^{-1}\boldsymbol{U}(t)^\top + \\ \boldsymbol{U}(t)\boldsymbol{\Lambda}(t)^{1/2}\boldsymbol{\Omega}(t)^\top\boldsymbol{\Lambda}(t)^{-1/2}\boldsymbol{U}(t)^\top.\end{aligned} \tag{71}$$

Multiplying eq. (71) by $\boldsymbol{U}(t)^\top$ on the left and $\boldsymbol{U}(t)$ on the right yields

$$\widetilde{\boldsymbol{A}}_q(t) = \boldsymbol{\Omega}(t) + (1/2)\frac{d}{dt}\boldsymbol{\Lambda}(t)\boldsymbol{\Lambda}(t)^{-1} - \boldsymbol{\Lambda}(t)^{1/2}\boldsymbol{\Omega}(t)\boldsymbol{\Lambda}(t)^{-1/2} - (1/2)\boldsymbol{\Lambda}(t)^{-1}.$$

Hence, the entries of $\widetilde{\boldsymbol{A}}_q(t)$ are

$$(\widetilde{\boldsymbol{A}}_q(t))_{ij} = \begin{cases} (\frac{d}{dt}\lambda_i(t) - 1)/(2\lambda_i(t)) & \text{if } i = j \\ \boldsymbol{\Omega}_{ij}(t)(1 - \sqrt{\lambda_i(t)/\lambda_j(t)}) & \text{if } i \ne j \end{cases}. \tag{72}$$

Overall, the integrand is equal to

$$\operatorname{tr}\left(\widetilde{\boldsymbol{\Delta}}(t)\boldsymbol{\Lambda}(t)\widetilde{\boldsymbol{\Delta}}(t)^\top\right) = \lambda_1(t)(\widetilde{\boldsymbol{\Delta}}(t)_{11}^2 + \widetilde{\boldsymbol{\Delta}}(t)_{21}^2) + \lambda_2(t)(\widetilde{\boldsymbol{\Delta}}(t)_{12}^2 + \widetilde{\boldsymbol{\Delta}}(t)_{22}^2), \tag{73}$$

where we have now computed $\widetilde{\boldsymbol{\Delta}}(t)$ for the symmetric and square root gauges.

Next, we explicitly minimize eq. (73) with respect to $w(t)$, for each of the symmetric and square root gauges. Note that only the off-diagonal entries of $\widetilde{\boldsymbol{\Delta}}(t)$ depend on $w(t)$. In particular let $\widetilde{\boldsymbol{\Delta}}(t)_{12} = aw(t) + \omega$ and $\widetilde{\boldsymbol{\Delta}}(t)_{21} = bw(t) - \omega$ for $a, b \in \mathbb{R}$. Then the contribution of the off-diagonal terms to the integrand can be written as

$$\lambda_1(t)\widetilde{\boldsymbol{\Delta}}(t)_{21}^2+\lambda_2(t)\widetilde{\boldsymbol{\Delta}}(t)_{12}^2 = (\lambda_1(t)b^2+\lambda_2(t)a^2)w(t)^2+2\omega(\lambda_2(t)a-\lambda_1(t)b)w(t)+(\lambda_1(t)+\lambda_2(t))\omega^2.$$

Completing the square yields the optimal value of $w(t)$

$$w(t)^\star = -\omega(\lambda_2(t)a - \lambda_1(t)b)/(\lambda_1(t)b^2 + \lambda_2(t)a^2).$$

The contribution of the off-diagonal terms at this optimum is

$$\begin{aligned}&(\lambda_1(t) + \lambda_2(t))\omega^2 - \omega^2(\lambda_2(t)a - \lambda_1(t)b)^2/(\lambda_1(t)b^2 + \lambda_2(t)a^2)\\ =&\omega^2\lambda_1(t)\lambda_2(t)(a+b)^2/(\lambda_1(t)b^2 + \lambda_2(t)a^2).\end{aligned} \tag{74}$$

For the symmetric gauge, $a = b = -(\lambda_2(t) - \lambda_1(t))/(\lambda_1(t) + \lambda_2(t))$, so

$$\lambda_1(t)\widetilde{\boldsymbol{\Delta}}(t)_{21}^2 + \lambda_2(t)\widetilde{\boldsymbol{\Delta}}(t)_{12}^2 \geq 4\omega^2\lambda_1(t)\lambda_2(t)/(\lambda_1(t) + \lambda_2(t)) \geq 2\omega^2\lambda_1(t).$$

In the last inequality, we used the ordering of the eigenvalues $\lambda_2(t) \geq \lambda_1(t)$.

And for the square root gauge, we have $a = r(t) - 1$, $b = 1 - r(t)^{-1}$ where we have defined $r(t) = \sqrt{\lambda_1(t)/\lambda_2(t)}$. This implies

$$\begin{aligned}&(a+b)^2 = (r(t)^2-1)^2/r(t)^2,\\ &\lambda_1(t)b^2 + \lambda_2(t)a^2 = (r(t)-1)^2(\lambda_1(t)/r(t)^2 + \lambda_2(t)) = 2\lambda_2(t)(r(t)-1)^2\\ &\lambda_1(t)\lambda_2(t)(a+b)^2/(\lambda_1(t)b^2 + \lambda_2(t)a^2) = \frac{1}{2}\lambda_2(t)(r(t)+1)^2 = \frac{1}{2}(\sqrt{\lambda_1(t)} + \sqrt{\lambda_2(t)})^2.\end{aligned}$$

Hence,

$$\lambda_1(t)\widetilde{\boldsymbol{\Delta}}(t)_{21}^2 + \lambda_2(t)\widetilde{\boldsymbol{\Delta}}(t)_{12}^2 \geq \frac{\omega^2}{2}(\sqrt{\lambda_1(t)} + \sqrt{\lambda_2(t)})^2 \geq 2\omega^2\sqrt{\lambda_1(t)\lambda_2(t)} \geq 2\omega^2\lambda_1(t).$$

In both cases, we have proven that for each open interval that makes up A,

$$\min_{w(t)}\left\{\lambda_1(t)\widetilde{\Delta}_{21}(t)^2 + \lambda_2(t)\widetilde{\Delta}_{12}(t)^2\right\} \geq 2\omega^2\lambda_1(t), \quad \text{for a.e. } t.$$

Therefore, for almost every $t \in A$,

$$\operatorname{tr}\left(\widetilde{\boldsymbol{\Delta}}(t)\boldsymbol{\Lambda}(t)\widetilde{\boldsymbol{\Delta}}(t)^\top\right) \geq \frac{\left(\frac{d}{dt}\lambda_1(t) - 1 + 2\alpha\lambda_1(t)\right)^2}{4\lambda_1(t)} + 2\omega^2\lambda_1(t).$$

We now proceed to the second set E. On this set, the covariance is isotropic: $\boldsymbol{S}(t) = \lambda(t)\mathbb{I}_2, \lambda(t) = \lambda_1(t) = \lambda_2(t)$. No eigenbasis is needed here. Since $\boldsymbol{S} - \lambda\mathbb{I}_2 = 0$ on E and both terms are absolutely continuous, we have

$$\frac{d}{dt}\boldsymbol{S}(t) = \frac{d}{dt}\lambda(t)\mathbb{I}_2$$

for almost every $t \in E$. For the symmetric gauge, the linear drift coefficient $\boldsymbol{A}_q(t)$ solves

$$\frac{d}{dt}\lambda(t)\mathbb{I}_2 = 2\lambda(t)\boldsymbol{A}_q(t) + \mathbb{I}_2,$$

and in the symmetric square root gauge,

$$\boldsymbol{A}_q(t) = \left(\frac{d}{dt}\boldsymbol{S}(t)^{1/2}\right)\boldsymbol{S}(t)^{-1/2} - \frac{1}{2}\boldsymbol{S}(t)^{-1} = \frac{\frac{d}{dt}\lambda(t) - 1}{2\lambda(t)}\mathbb{I}_2.$$

Thus, in either gauge,

$$\boldsymbol{A}_q(t) = \frac{\frac{d}{dt}\lambda(t) - 1}{2\lambda(t)}\mathbb{I}_2 \qquad \text{for a.e. } t \in E.$$

A direct calculation then yields

$$\operatorname{tr}\Big(\big(\boldsymbol{A}_q(t) - (-\alpha\mathbb{I}_2 + \omega\boldsymbol{J})\big)\boldsymbol{S}(t)\big(\boldsymbol{A}_q(t) - (-\alpha\mathbb{I}_2 + \omega\boldsymbol{J})\big)^\top\Big) = \frac{\big(\frac{d}{dt}\lambda(t) - 1 + 2\alpha\lambda(t)\big)^2}{2\lambda(t)} + 2\omega^2\lambda(t),$$

which is stronger than the preceding bound.

Combining both cases, we conclude that for almost every $t \in [0, T]$,

$$\operatorname{tr}\Big(\big(\boldsymbol{A}_q(t) - (-\alpha\mathbb{I}_2 + \omega\boldsymbol{J})\big)\boldsymbol{S}(t)\big(\boldsymbol{A}_q(t) - (-\alpha\mathbb{I}_2 + \omega\boldsymbol{J})\big)^\top\Big) \geq \frac{u(t)^2}{4\lambda_1(t)} + 2\omega^2\lambda_1(t)$$

$$u(t) = \frac{d}{dt}\lambda_1(t) - 1 + 2\alpha\lambda_1(t).$$

Therefore,

$$f_{\boldsymbol{S}}(\boldsymbol{S}) \geq I = \int_0^T \left[\frac{u(t)^2}{8\lambda_1(t)} + \omega^2\lambda_1(t)\right] \mathrm{d}t.$$

Step (iv): Fix $c_0 > 1$ and assume $\omega \geq c_0\alpha$. Define the sets E_δ and F_δ that track the time $\lambda_1(t)$ spends above and below $\delta > 0$, respectively:

$$\delta = \frac{1}{4\omega}, \qquad E_\delta = \{t \in [0, T] : \lambda_1(t) \geq \delta\}, \qquad F_\delta = [0, T] \setminus E_\delta.$$

We will argue that if either E_δ or F_δ is large, then I must be large.

If $|E_\delta| \geq T/2$, then

$$I \geq \int_{E_\delta} \omega^2\lambda_1(t)\ \mathrm{d}t \geq |E_\delta|\,\omega^2\delta \geq \frac{T}{8}\,\omega.$$

Suppose instead that $|F_\delta| > T/2$. Since $\lambda_1(t) < \delta$ on F_δ,

$$2\alpha\lambda_1(t) \leq 2\alpha\delta = \frac{\alpha}{2\omega} \leq \frac{1}{2c_0}.$$

Define

$$a_0 = 1 - \frac{1}{2c_0} > 0.$$

Then, for all $t \in F_\delta$,

$$\frac{d\lambda_1(t)}{dt} = 1 - 2\alpha\lambda_1(t) + u(t) \geq a_0 + u(t).$$

Since F_δ is open in $[0, T]$, we can write it as a countable disjoint union of open intervals (a_j, b_j). Integrating over a single interval,

$$\lambda_1(b_j) - \lambda_1(a_j) = \int_{a_j}^{b_j} \frac{d\lambda_1(t)}{dt} dt \geq a_0(b_j - a_j) + \int_{a_j}^{b_j} u(t)\ \mathrm{d}t.$$

Rearranging yields

$$-\int_{a_j}^{b_j} u(t)\ \mathrm{d}t \geq a_0(b_j - a_j) - \big(\lambda_1(b_j) - \lambda_1(a_j)\big).$$

If (a_j, b_j) is an interior component, maximality and continuity imply

$$\lambda_1(a_j) = \lambda_1(b_j) = \delta,$$

so the boundary term vanishes. If (a_j, b_j) touches 0 or T, then both endpoint values lie in $[0, \delta]$, and therefore

$$\big|\lambda_1(b_j) - \lambda_1(a_j)\big| \leq \delta.$$

Since there are at most two boundary components, summing over all components gives

$$-\int_{F_\delta} u(t)\ \mathrm{d}t \geq a_0|F_\delta| - 2\delta.$$

By the AM-GM inequality,

$$\frac{u^2}{8\lambda} + \omega^2\lambda \geq \frac{\omega}{\sqrt{2}}|u| \implies I \geq \frac{\omega}{\sqrt{2}}\int_0^T |u(t)|\ \mathrm{d}t.$$

Therefore, we deduce

$$\int_0^T |u(t)|\ \mathrm{d}t \geq \int_{F_\delta} |u(t)|\ \mathrm{d}t \geq -\int_{F_\delta} u(t)\ \mathrm{d}t \geq a_0|F_\delta| - 2\delta > \frac{a_0 T}{2} - 2\delta$$
$$\implies I \geq \frac{\omega}{\sqrt{2}}\left(\frac{a_0 T}{2} - 2\delta\right) = \frac{1 - \frac{1}{2c_0}}{2\sqrt{2}} T\omega - \frac{1}{2\sqrt{2}}.$$

Combining the two cases and the definition of $f_{\boldsymbol{S}}$ in eq. (67), we conclude

$$f_{\boldsymbol{S}}(\boldsymbol{S}) \geq \min\left\{\frac{1}{8},\ \frac{1 - \frac{1}{2c_0}}{2\sqrt{2}}\right\} T\omega - \frac{1}{2\sqrt{2}}.$$

Step (v): Lastly, we show that when $p(\boldsymbol{x}, 0) = \mathcal{N}\left(\mathbf{0}, \frac{1}{2\alpha}\mathbb{I}_2\right)$, then C_{OU} can be lower bounded by a constant independent of ω.

Consider the vector of stacked observations

$$\boldsymbol{y}_{1:N} = \left(\boldsymbol{y}(t_1)^\top, \ldots, \boldsymbol{y}(t_N)^\top\right)^\top \in \mathbb{R}^{2N}.$$

Under the stationary OU spiral prior, the latent process is mean-zero Gaussian and

$$\mathrm{Cov}(\boldsymbol{x}(t_i), \boldsymbol{x}(t_j)) = \frac{1}{2\alpha} e^{-\alpha|t_i - t_j|} \boldsymbol{R}_{\omega(t_i - t_j)},$$

where $\boldsymbol{R}_\theta \in \mathbb{R}^{2\times 2}$ represents the 2×2 rotation matrix by angle θ.

Therefore the stacked observation vector $\boldsymbol{y}_{1:N}$ is Gaussian,

$$\boldsymbol{y}_{1:N} \sim \mathcal{N}\big(\bar{\boldsymbol{d}}, \boldsymbol{\Sigma}_\omega\big),$$

with block covariance matrix $\boldsymbol{\Sigma}_\omega \in \mathbb{R}^{2N\times 2N}$

$$(\boldsymbol{\Sigma}_\omega)_{ij} = \frac{1}{2\alpha} e^{-\alpha|t_i - t_j|} \boldsymbol{C}\boldsymbol{R}_{\omega(t_i - t_j)}\boldsymbol{C}^\top + \mathbf{1}\{i = j\}\, \sigma^2 \boldsymbol{C}\boldsymbol{C}^\top,$$

and mean

$$\bar{\boldsymbol{d}} = \left(\boldsymbol{d}^\top, \ldots, \boldsymbol{d}^\top\right)^\top \in \mathbb{R}^{2N}.$$

We first derive an ω-uniform lower spectral bound. Since $\boldsymbol{\Sigma}_\omega$ is a covariance matrix, its latent contribution is positive semidefinite. Hence,

$$\boldsymbol{\Sigma}_\omega \succeq \sigma^2\, \mathrm{diag}\big(\boldsymbol{C}\boldsymbol{C}^\top, \ldots, \boldsymbol{C}\boldsymbol{C}^\top\big).$$

It follows that

$$\lambda_{\min}(\boldsymbol{\Sigma}_\omega) \geq \sigma^2 \lambda_{\min}(\boldsymbol{C}\boldsymbol{C}^\top) = m_* > 0,$$

uniformly in ω.

Next we derive an ω-uniform upper spectral bound. Let

$$\boldsymbol{z} = \left(\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_N^\top\right)^\top \in \mathbb{R}^{2N}, \qquad \boldsymbol{z}_i \in \mathbb{R}^2.$$

Then

$$\boldsymbol{z}^\top \boldsymbol{\Sigma}_\omega \boldsymbol{z} = \sigma^2 \sum_{i=1}^N \boldsymbol{z}_i^\top \boldsymbol{C}\boldsymbol{C}^\top \boldsymbol{z}_i + \frac{1}{2\alpha} \sum_{i,j=1}^N e^{-\alpha|t_i - t_j|} \boldsymbol{z}_i^\top \boldsymbol{C}\boldsymbol{R}_{\omega(t_i - t_j)}\boldsymbol{C}^\top \boldsymbol{z}_j.$$

Since $\boldsymbol{R}_{\omega(t_i-t_j)}$ is orthogonal, $\|\boldsymbol{C}\boldsymbol{R}_{\omega(t_i-t_j)}\boldsymbol{C}^\top\|_{\mathrm{op}} \le \|\boldsymbol{C}\|_{\mathrm{op}}^2$. Therefore

$$\boldsymbol{z}^\top \boldsymbol{\Sigma}_\omega \boldsymbol{z} \le \sigma^2 \|\boldsymbol{C}\|_{\mathrm{op}}^2 \sum_{i=1}^N \|\boldsymbol{z}_i\|_2^2 + \frac{\|\boldsymbol{C}\|_{\mathrm{op}}^2}{2\alpha} \sum_{i,j=1}^N \|\boldsymbol{z}_i\|_2 \|\boldsymbol{z}_j\|_2.$$

Using

$$\sum_{i,j=1}^N \|\boldsymbol{z}_i\|_2 \, \|\boldsymbol{z}_j\|_2 = \left(\sum_{i=1}^N \|\boldsymbol{z}_i\|_2\right)^2 \le N \sum_{i=1}^N \|\boldsymbol{z}_i\|_2^2,$$

we obtain

$$\boldsymbol{z}^\top \boldsymbol{\Sigma}_\omega \boldsymbol{z} \le \|\boldsymbol{C}\|_{\mathrm{op}}^2 \left(\sigma^2 + \frac{N}{2\alpha}\right) \sum_{i=1}^N \|\boldsymbol{z}_i\|_2^2.$$

Hence

$$\lambda_{\max}(\boldsymbol{\Sigma}_\omega) \le \|\boldsymbol{C}\|_{\mathrm{op}}^2 \left(\sigma^2 + \frac{N}{2\alpha}\right) = M_*,$$

again uniformly in ω.

Thus, for every ω,

$$m_* \mathbb{I}_{2N} \preceq \boldsymbol{\Sigma}_\omega \preceq M_* \mathbb{I}_{2N}.$$

Since $\boldsymbol{y}_{1:N} \sim \mathcal{N}(\bar{\boldsymbol{d}}, \boldsymbol{\Sigma}_\omega)$, its density is

$$p_\omega(\boldsymbol{y}_{1:N}) = (2\pi)^{-N} \det(\boldsymbol{\Sigma}_\omega)^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y}_{1:N} - \bar{\boldsymbol{d}})^\top \boldsymbol{\Sigma}_\omega^{-1} (\boldsymbol{y}_{1:N} - \bar{\boldsymbol{d}})\right).$$

Using the spectral bounds above,

$$(\boldsymbol{y}_{1:N} - \bar{\boldsymbol{d}})^\top \boldsymbol{\Sigma}_\omega^{-1} (\boldsymbol{y}_{1:N} - \bar{\boldsymbol{d}}) \le \frac{\|\boldsymbol{y}_{1:N} - \bar{\boldsymbol{d}}\|_2^2}{m_*}, \quad \det(\boldsymbol{\Sigma}_\omega) \le M_*^{2N}.$$

Therefore,

$$p_\omega(\boldsymbol{y}_{1:N}) \ge (2\pi)^{-N} M_*^{-N} \exp\left(-\frac{\|\boldsymbol{y}_{1:N} - \bar{\boldsymbol{d}}\|_2^2}{2m_*}\right).$$

Taking logarithms yields

$$\log p_\omega(\boldsymbol{y}_{1:N}) \ge -N \log(2\pi) - N \log M_* - \frac{\|\boldsymbol{y}_{1:N} - \bar{\boldsymbol{d}}\|_2^2}{2m_*}.$$

Thus the claimed bound holds with

$$C_{\mathrm{OU-stationary}} = N \log \sigma^2 - N \log M_* - \frac{\|\boldsymbol{y}_{1:N} - \bar{\boldsymbol{d}}\|_2^2}{2m_*},$$

which is independent of ω. □

C.2 Simulation experiment details

In this subsection, we provide details regarding our OU spiral experiments. These experiments support the theoretical results in Appendix B.5 and Appendix C.1. In summary, these results state that Helmholtz-SDE achieves zero KL gap to the true posterior for all rotation speeds ω, while the symmetric [6, 7] and square root [8] gauges incur a KL gap that grows at least linearly in ω. These results characterize the variational families themselves, not the optimization problem solved in practice: with finite-capacity parameterizations of $\boldsymbol{m}(t)$ and $\boldsymbol{S}(t)$, and with stochastic gradient-based optimization, neither family attains its theoretical optimum. Our experiments test whether the theoretical separation between Helmholtz-SDE and the gauges translates into practical differences.

For both our inference and learning experiments, we use the Adam optimizer [73] with initial learning rate 10^{-3}, final learning rate 10^{-4}, and exponential learning rate decay. We run experiments on an NVIDIA A100 or H100 GPU.

We generate the OU dataset by sampling latents according to the initial distribution $p(\boldsymbol{x}, 0) = \mathcal{N}(\boldsymbol{0}, 1/(2\alpha)\mathbb{I}_2)$ and prior dynamics eq. (64) on $[0, 5]$. For all experiments we fix $\alpha = 0.2$. We

sample $(\boldsymbol{x}^{(m)}(t))_{0\le t\le 5}$ using Euler–Maruyama with stepsize $5 \cdot 10^{-4}$. To generate observations, we select random times $\{t_n^{(m)}\}_{n=1}^N$ on the interval $[0, 5]$ for each trial m, and then we draw $\boldsymbol{y}^{(m)}(t_n^{(m)}) \sim \mathcal{N}(\boldsymbol{C}_{\text{true}}\boldsymbol{x}^{(m)}(t_n^{(m)}) + \boldsymbol{d}_{\text{true}}, \sigma^2\boldsymbol{C}_{\text{true}}\boldsymbol{C}_{\text{true}}^\top)$. $\boldsymbol{C}_{\text{true}} \in \mathbb{R}^{2\times 2}$ and $\boldsymbol{d}_{\text{true}} \in \mathbb{R}^2$ are chosen to have standard normal entries, with $\boldsymbol{C}_{\text{true}}$ normalized to have determinant 1. In our experiments, we vary the rotation speed ω, the number of observations N, and the observation noise $\sigma > 0$.

Inference. We first investigate the gap in expressivity of the variational family when the parameters of the generative model are held fixed at their true values.

We do so by performing inference over 16 independent trials. For each trial, we can compute the true posterior by running the Kalman–Bucy smoother. We do so with the SING codebase, which Hu et al. [15] prove converges to the true posterior in a single step in the conjugate setting. We fix the time discretization of the true posterior to stepsize $5 \cdot 10^{-4}$.

To represent the variational posterior, we discretize $[0, 5]$ on a grid with constant separation $5 \cdot 10^{-3}$ (1001 total grid points). At each point we parameterize the posterior mean and covariance, and we interpolate linearly between the grid times. This yields approximately $6 \cdot 10^3$ variational parameters to learn, per trial. Moreover, note that for a randomly sampled time, the unbiased estimate of the path-space KL depends only on the two neighboring grid points. For this reason, the variational posterior is slow to converge using a single Monte Carlo sample. Instead, we use $N_{\text{time}} = 1001$ samples to compute each gradient update (i.e., on average, one sample per grid point). At each time point we draw $N_{\text{space}} = 1$ spatial sample. Increasing the sample count adds minimal computational overhead, since both sampling and gradient computation parallelize across samples. We compute the optimal q-divergence-free correction using the Stein estimand eq. (59) and $N_{\text{quad}} = 1$ Gauss–Hermite quadrature point. We perform training for $6 \cdot 10^5$ iterations.

To compute each direction (forward and backward KL divergence) of the symmetric KL divergence between the true and approximate posterior, we use Girsanov's formula eq. (11). In particular, we discretize the interval $[0, 5]$ with stepsize $5 \cdot 10^{-4}$ and at each time evaluate the integrand of the path KL $\mathcal{L}_{\text{path}}$. Since both the true and approximate posterior drifts are affine, this can be done in closed form using Gaussian expectations.

In the top row of Figures 6 to 10, we report the symmetric KL divergence between the approximate and true posterior for $\omega \in \{2\pi, \pi, \pi/2, \pi/4, 0\}$. Broadly, we see that the suboptimality of choosing a single posterior drift per one-time Gaussian marginals (either the symmetric or square root gauge) increases as the number of observations decreases, the observation noise increases, and the rotational speed increases. Even when the rotation speed is strong, choosing a single posterior drift suffices as long as posterior uncertainty is small, i.e., the observations are strongly informative of the latent trajectory. In settings where the posterior is weakly constrained by the data, we see order-of-magnitude differences between Helmholtz-SDE and the symmetric and square root gauges.

Joint inference and learning. Next, we study the impact of the constrained variational family imposed by prior simulation-free VI algorithms on dynamics learning.

To do so, we parameterize the prior drift with a shallow neural network
`(2)` $\to$ `Linear(2, 100)` $\to$ `Softplus` $\to$ `Linear(100, 2)` $\to$ `(2)`.
We learn prior dynamics together with the affine mapping $\{\boldsymbol{C}, \boldsymbol{d}\}$ and output noise $\boldsymbol{R}$ that define the Gaussian likelihood model.

We perform learning using 1024 trials. We average results over 5 replicates, each of which uses a different initialization of the prior and output parameters. For our experiments, we considered numerous parameterizations of the variational posterior, including a per-trial grid posterior as in our inference experiments and an encoder-based amortized posterior as in our Lorenz experiments. For the grid posterior, we observed very slow convergence, even with a large number of Monte Carlo samples N_{time}. We hypothesize that this is because the parameter space is high dimensional (1001 grid points $\times$ 1024 trials $\times$ 6 parameters per grid point per trial $\approx 6 \cdot 10^6$ total parameters), and that each Monte Carlo sample only provides gradient signal at two neighboring grid points; moreover, the grid posterior does not benefit from a smoothness inductive bias. On the other hand, training the neural network posterior was unstable; it required choosing a large hidden dimension and performing aggressive gradient clipping. We did not observe that any parameterization preferred a specific gauge, i.e., the choice of posterior parameterization was independent of the gauge.

We settled on the kernel-GLM posterior parameterization introduced by Course and Nair [6], which can be viewed as a grid posterior augmented with a smoothness inductive bias and which couples gradient updates across nearby times. At a high level, this parameterization chooses a grid of temporal inducing locations $\boldsymbol{\tau} = \{\tau_n\}_{n=1}^{N_{\text{induce}}}$ that partition $[0, T]$. It then defines the posterior mean $\boldsymbol{m}(t)$ at each time point $0 \leq t \leq T$ as $\boldsymbol{M}\boldsymbol{C}k(\boldsymbol{\tau}, t)$, where $k(\cdot, \cdot)$ is a kernel function, $k(\boldsymbol{\tau}, t) \in \mathbb{R}^{N_{\text{induce}} \times 1}$ is a Gram vector, $\boldsymbol{M} \in \mathbb{R}^{2 \times N_{\text{induce}}}$ contains the inducing values, and $\boldsymbol{C} \in \mathbb{R}^{N_{\text{induce}} \times N_{\text{induce}}}$ is a preconditioning matrix. If $\boldsymbol{C}$ were equal to the inverse Gram matrix $k(\boldsymbol{\tau}, \boldsymbol{\tau})^{-1}$, then $\boldsymbol{m}(t)$ would exactly recover the predictive mean of a multi-output Gaussian process regressor with pseudo-observations $\boldsymbol{M}$ at locations $\boldsymbol{\tau}$. However, this requires inverting an $N_{\text{induce}} \times N_{\text{induce}}$ matrix at each step; therefore, the authors fix it to a deterministic value at initialization (no gradient propagation). A similar parameterization is used for the covariance $\boldsymbol{S}(t)$, which is chosen to be dense. See [6] for details.

We use $N_{\text{induce}} = 200$ inducing values and choose the kernel $k(\cdot, \cdot)$ to be a Matérn 5/2 kernel, following Course and Nair [6]. The learnable parameters of the kernel are the output scale, length scale, and warping function parameters (Kumaraswamy warping). The kernel parameters (but not the inducing values) are shared across all trials.

We perform training for $3 \cdot 10^4$ iterations. At each iteration, we approximate the path KL for each trial using $N_{\text{time}} = 16$ samples of $[0, 5]$. At each sampled time point we use $N_{\text{space}} = 1$ spatial sample. To compute the degree $\ell = 1$ q-divergence-free correction, we use the Stein estimand eq. (59) and $N_{\text{quad}} = 1$ Gauss–Hermite quadrature point.

To compute the symmetric KL between the learned and true prior, we again use Girsanov's theorem eq. (11). However, unlike in our inference experiments, the integrand cannot be computed in closed form, since the neural network prior is nonlinear and does not admit closed-form marginal distributions. Instead, we draw 2048 trials from both the learned and true prior using Euler–Maruyama with stepsize $5 \cdot 10^{-4}$, and approximate the path KL using Monte Carlo.

In the bottom row of Figures 6 to 10, we report the symmetric KL divergence between the learned and true prior for $\omega \in \{2\pi, \pi, \pi/2, \pi/4, 0\}$. Generally, we observe that improved inference yields improved learning: the regimes in which the symmetric and square root gauges poorly approximate the posterior are the same regimes in which they poorly recover the prior dynamics. On the other hand, Helmholtz-SDE well approximates the posterior in regimes of high posterior uncertainty, which yields substantive improvements in dynamics recovery.

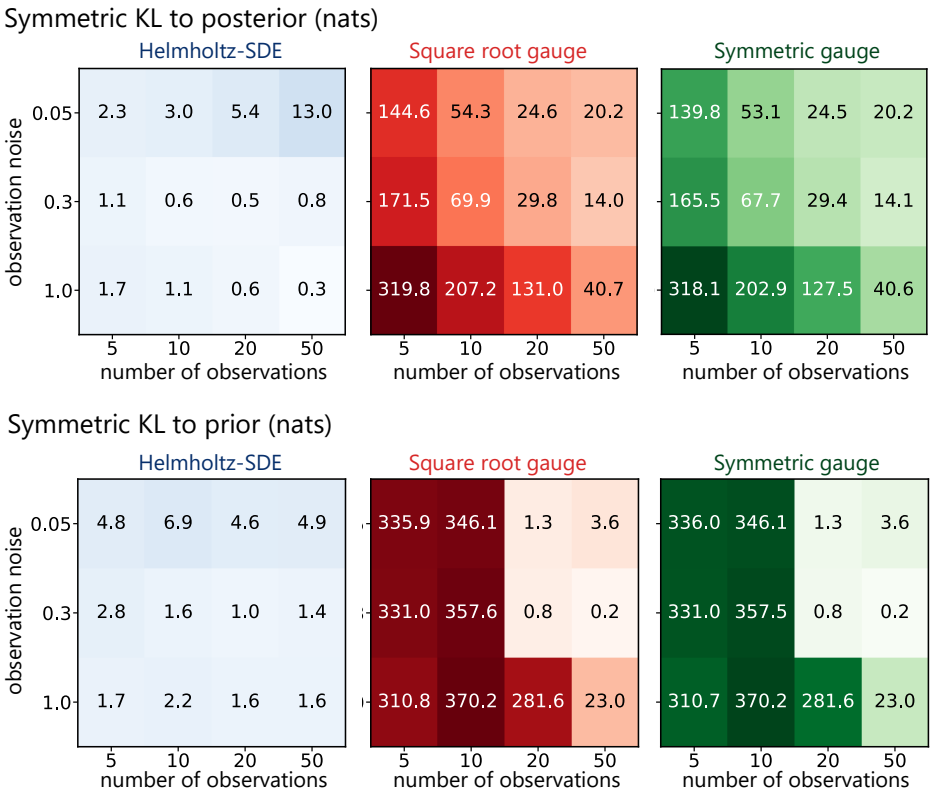


Figure 6: OU spiral with $\omega = 2\pi$; top: symmetric KL to posterior; bottom: symmetric KL to prior.

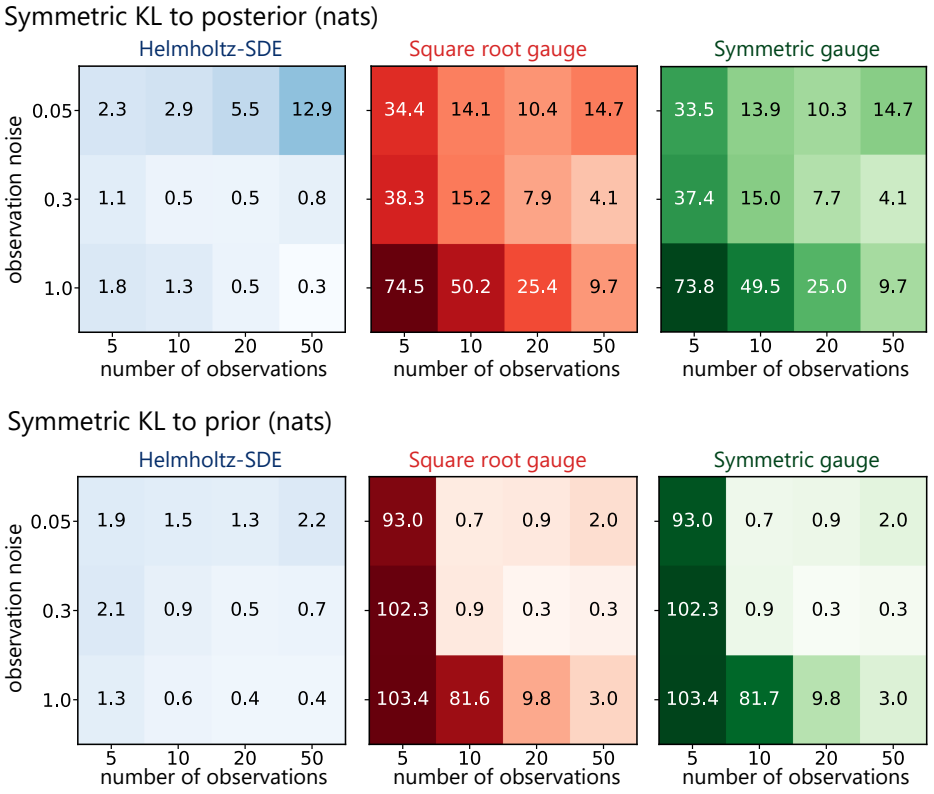


Figure 7: OU spiral with $\omega = \pi$; top: symmetric KL to posterior; bottom: symmetric KL to prior.

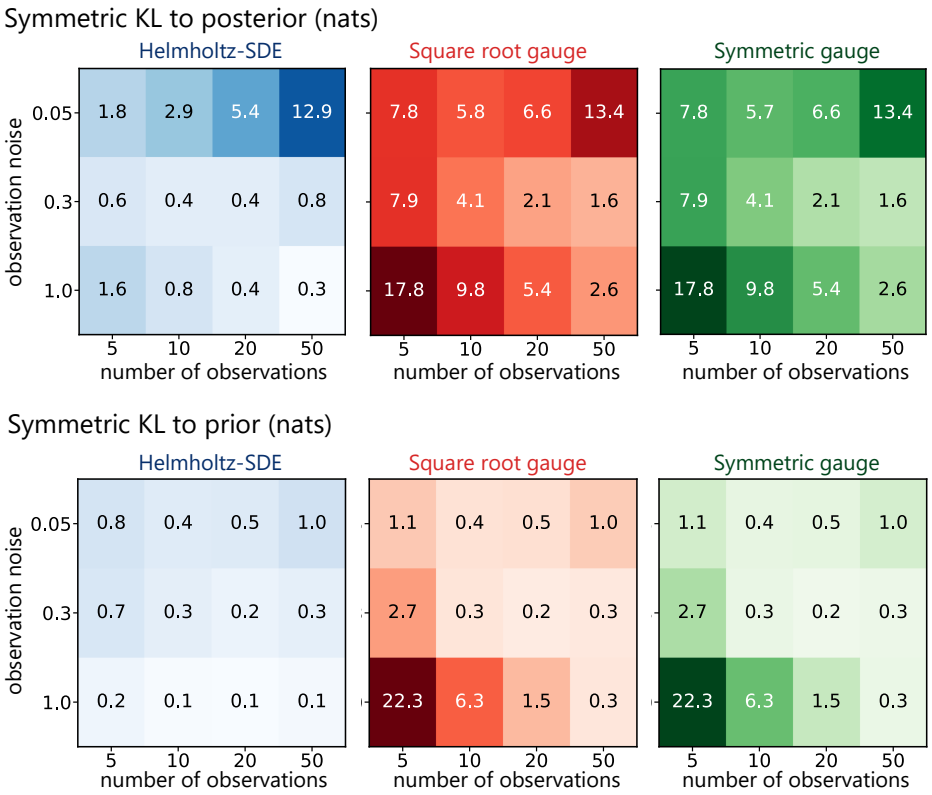


Figure 8: OU spiral with $\omega = \pi/2$; top: symmetric KL to posterior; bottom: symmetric KL to prior.

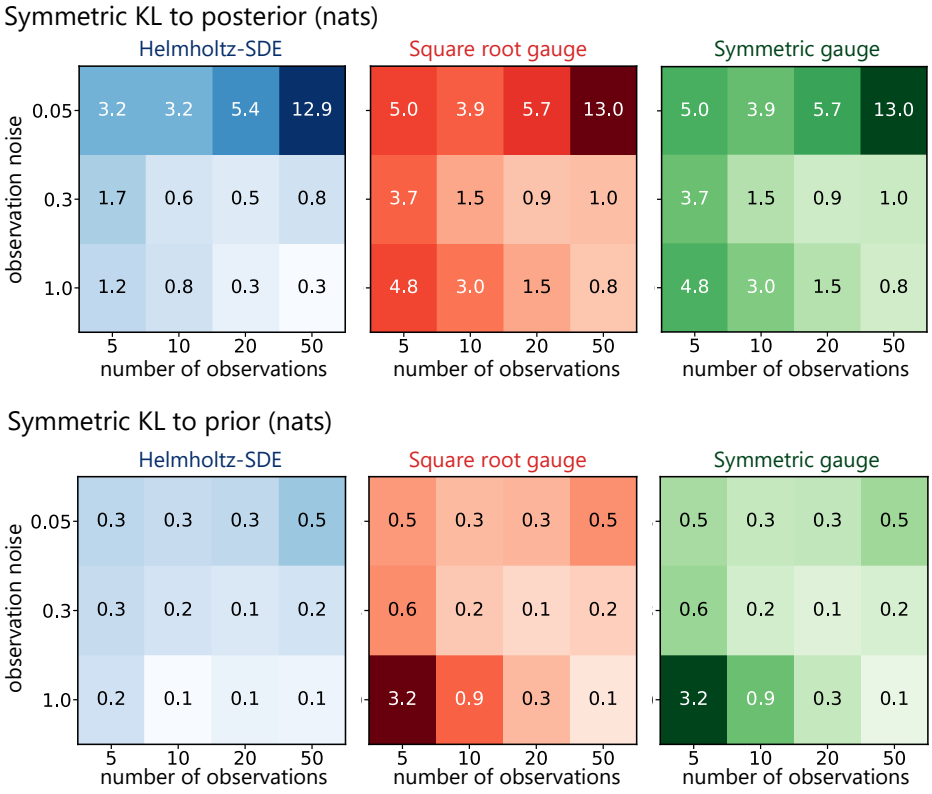


Figure 9: OU spiral with $\omega = \pi/4$; top: symmetric KL to posterior; bottom: symmetric KL to prior.

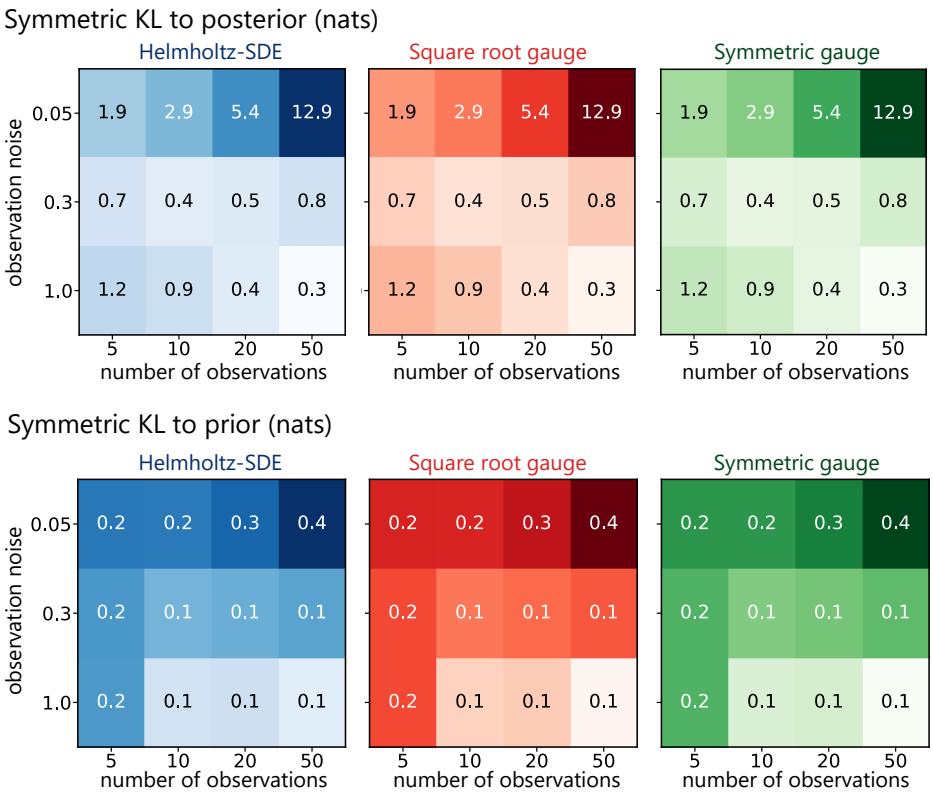


Figure 10: OU spiral with $\omega = 0$; top: symmetric KL to posterior; bottom: symmetric KL to prior.

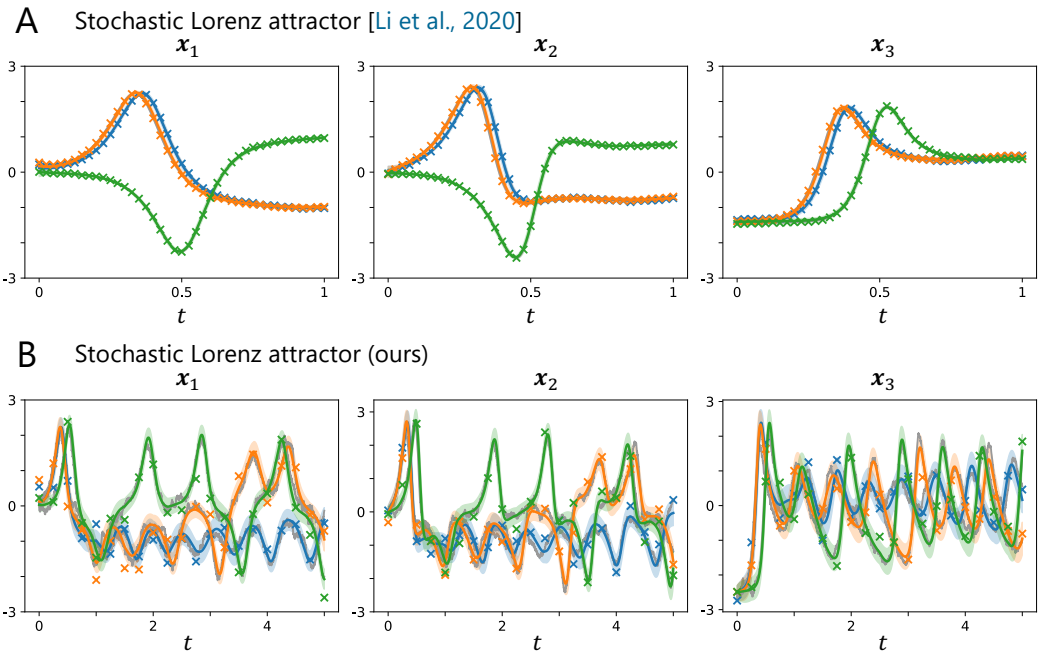


Figure 11: Visualization of three samples of the stochastic Lorenz attractor dataset from Li et al. [2] (**A**) and this paper (**B**). Grey curves represent the true latent trajectory and colored $\times$ represent regularly spaced observations. Colored curves represent the posterior mean and shading represents ± 2 posterior standard deviations obtained from the iterated extended Rauch–Tung–Striebel smoother (Särkkä and Svensson [72, Chapter 13.3]). The Bayesian posterior in Li et al. [2] collapses to the true latent trajectory, whereas our dataset displays nontrivial posterior variance.

D Nonlinear SDE experiments

In this section, we provide additional details on our three nonlinear SDE experiments: the stochastic Lorenz attractor (Appendix D.1), rotifer-algae predator-prey dynamics (Appendix D.2), and reduced order models for fluid dynamics (Appendix D.3).

For all experiments with the exception of the Latent SDE baseline [2], we use a Gaussian posterior approximation $q(\boldsymbol{x}, t) = \mathcal{N}(\boldsymbol{m}(t), \boldsymbol{S}(t))$ with dense $\boldsymbol{S} \in \mathbb{R}^{K \times K}$. We also use the Adam optimizer [73] with exponential learning rate decay. We run experiments on an NVIDIA A100 or H100 GPU.

D.1 Noisy Lorenz attractor

Data generation. We generate a dataset consisting of 1024 training samples and 2048 test samples according to the 3-dimensional Lorenz attractor dynamics

$$d\boldsymbol{x}(t) = \begin{pmatrix} \alpha(\boldsymbol{x}_2(t) - \boldsymbol{x}_1(t)) \\ \boldsymbol{x}_1(t)(\rho - \boldsymbol{x}_3(t)) - \boldsymbol{x}_2(t) \\ \boldsymbol{x}_1(t)\boldsymbol{x}_2(t) - \beta \boldsymbol{x}_3(t) \end{pmatrix} dt + \sigma d\boldsymbol{w}(t), \quad \boldsymbol{x}(0) \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_3), \quad 0 \le t \le T \tag{75}$$

where $(\alpha, \rho, \beta) = (10, 28, 8/3)$ and $\sigma > 0$ is the diffusion scaling. We do so using the Euler–Maruyama sampler with discretization 10^{-3}.

Following Li et al. [2], we make observations of the latent state every Δt according to $\boldsymbol{C}_{\text{true}}\boldsymbol{x}(t) + \boldsymbol{d}_{\text{true}} + \boldsymbol{\epsilon}(t)$, $\boldsymbol{\epsilon}(t) \sim \mathcal{N}(\boldsymbol{0}, \eta^2 \mathbb{I}_3)$, where $\boldsymbol{C}_{\text{true}} = \text{diag}(\boldsymbol{s}^{-1})$ and $\boldsymbol{d}_{\text{true}} = -\boldsymbol{m} \odot \boldsymbol{s}^{-1}$; $\boldsymbol{m}, \boldsymbol{s} \in \mathbb{R}^3$ represent the empirical mean and standard deviation of the training trajectories, respectively, along each dimension and pooled across time. $\eta > 0$ is the observation standard deviation.

In the standard stochastic Lorenz attractor, the authors choose $T = 1$ and parameters $(\Delta t, \sigma, \eta) = (2.5 \cdot 10^{-2}, 1.5 \cdot 10^{-1}, 10^{-2})$. In the observation coordinates, the effective diffusion coefficient is $\sigma \boldsymbol{s}^{-1} \approx (1.7 \cdot 10^{-2}, 1.3 \cdot 10^{-2}, 9.8 \cdot 10^{-3})$, which is exceedingly small relative to the scale on which the latents evolve. As a result of the small process noise, frequent observations, and small

measurement noise, we observe that the posterior collapses to the true latent trajectory (Figure 11). Methods that differ substantially in how they represent or sample posterior uncertainty are therefore indistinguishable on this benchmark.

We instead simulate the latents over a long interval $T = 5$ with larger process noise $\sigma = 5$, less frequent observations $\Delta t = 2.5 \cdot 10^{-1}$, and larger measurement noise $\eta = 0.3$. In this case, the effective diffusion coefficient in the observation space is $\sigma \boldsymbol{s}^{-1} \approx (6.0 \cdot 10^{-1}, 5.2 \cdot 10^{-1}, 5.2 \cdot 10^{-1})$. We find that this setting leads both to more interesting latent trajectories (frequent switching between lobes of the attractor, which is not demonstrated in the original dataset; Figure 12A) and nontrivial posterior uncertainty (Figure 11).

Generative model. Consistent with prior work, we learn a latent dynamical system with $K = 4$ dimensions representing the Lorenz attractor. The prior initial state is modeled as Gaussian $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{V}_0)$, $\boldsymbol{\mu}_0 \in \mathbb{R}^4$, $\boldsymbol{V}_0 \in \mathbb{R}^{4\times4}$; the prior drift function eq. (1) is a neural network; and the observation model is linear and Gaussian $\mathcal{N}(\boldsymbol{y}(t)|\boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{d}, \sigma_{\text{obs}}^2\mathbb{I})$, $\boldsymbol{C} \in \mathbb{R}^{3\times4}$, $\boldsymbol{d} \in \mathbb{R}^3$, $\sigma_{\text{obs}} > 0$. We learn the parameters of the prior SDE and observation model jointly with those of the posterior.

Baselines. We compare Helmholtz-SDE to two simulation-free VI algorithms, ARCTA [7] and SDE Matching [8], as well as two simulation-based VI algorithms, SING [15] and Latent SDE [2]. Note that 'ARCTA' simply describes an amortization strategy for SVISE [6]. In particular, the variational drift, the symmetric gauge, remains the same.

We aim to reproduce the Lorenz attractor experimental setup as faithfully as possible from the original papers, with our modified data generation procedure. Among these algorithms, there are two noteworthy differences: (1) SDE Matching [8] and Latent SDE [2] support state-dependent diffusion coefficients; and (2) SING does not support an amortized posterior approximation. For SING, we discretize the posterior on a grid of size 10^{-3} and use mini-batches of 500 trials; mini-batching is necessary since the posteriors for all 1024 trials cannot fit simultaneously into memory (on an 80 GB GPU). We performed a small sweep over the SING mini-batch size and grid size.

For Helmholtz-SDE, we adopt the same amortization strategy employed by SDE Matching [8] and Latent SDE [2] wherein for each trial, the sequence of observations is encoded by a GRU to produce an embedding. This embedding is then passed through an MLP to represent the posterior at each time. This strategy shares posterior parameters both across trials and across time.

We used the ARCTA (https://github.com/coursekevin/arlatentsde), SDE Matching (https://github.com/GrigoryBartosh/sde_matching), SING (https://github.com/lindermanlab/sing), and Latent SDE (https://github.com/google-research/torchsde) codebases, with modifications as described below.

- **Helmholtz-SDE**:
  - `latent_dim = 4`, `obs_dim = 3`, `time_dim = 1`, `hidden_dim = 100`
  - Prior drift:
    `(latent_dim)` → `Linear(latent_dim, hidden_dim)` → `Softplus` → `Linear(hidden_dim, latent_dim)` → `(latent_dim)`
  - Posterior encoder:
    `(n_obs, obs_dim)` → `GRU` → `(n_obs + 1, hidden_dim)`
  - Posterior MLP:
    for a specified time, aggregate embeddings from encoder using softmax,
    `(n_obs + 1, hidden_dim)`, `(time_dim)` → `softmax` → `(hidden_dim)`
    then map to posterior at that time
    `(hidden_dim + time_dim)` → `Linear(hidden_dim + time_dim, hidden_dim)` → `SiLU` → `Linear(hidden_dim, hidden_dim)` → `SiLU` → `Linear(hidden_dim, output_dim)` → `(output_dim)`
    where `output_dim` is `latent_dim` (mean) + `latent_dim` · `(latent_dim + 1) / 2` (covariance)
  - Diffusion network: fixed $\mathbb{I}_{\texttt{latent_dim}}$
  - Helmholtz correction: linear approximation $\ell = 1$, computed using the Stein estimand with $N_{\text{mc}} = 1$
  - Observation model: fixed covariance $\sigma_{\text{obs}}^2 = \eta^2$, learned $\boldsymbol{C}$, $\boldsymbol{d}$

- LR: 10^{-3} (constant)
- Gradient clipping: 5.0
- Iterations: $3 \cdot 10^4$
- Number of MC samples (per iteration): $N_{\text{time}} = 1$ time sample over $[0, T]$ to estimate the KL term and 1 sampled observation time to estimate the reconstruction loss, per trial

- **SDE Matching**:
  - `latent_dim = 4`, `obs_dim = 3`, `time_dim = 1`, `hidden_dim = 100`
  - Prior drift:
    `(latent_dim)` $\to$ `Linear(latent_dim, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, latent_dim)` $\to$ `(latent_dim)`
  - Posterior encoder:
    `(n_obs, obs_dim)` $\to$ `GRU` $\to$ `(n_obs + 1, hidden_dim)`
  - Posterior MLP:
    for a specified time, aggregate embeddings from encoder using softmax,
    `(n_obs + 1, hidden_dim), (time_dim)` $\to$ `softmax` $\to$ `(hidden_dim)`
    then map to posterior at that time
    `(hidden_dim + time_dim)` $\to$ `Linear(hidden_dim + time_dim, hidden_dim)` $\to$ `SiLU` $\to$ `Linear(hidden_dim, hidden_dim)` $\to$ `SiLU` $\to$ `Linear(hidden_dim, output_dim)` $\to$ `(output_dim)`
    where `output_dim` is `latent_dim` (mean) + `latent_dim` $\cdot$ `(latent_dim + 1) / 2` (covariance)
  - Diffusion network:
    state-dependent, diagonal diffusion coefficient; the input to each diagonal entry is a single latent coordinate
    `latent_dim` $\times$ `[(1)` $\to$ `Linear(1, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, 1)` $\to$ `Sigmoid` $\to$ `(1)]`
  - Observation model: fixed covariance $\sigma_{\text{obs}}^2 = \eta^2$, learned $\boldsymbol{C}$, $\boldsymbol{d}$
  - LR: 10^{-3} (constant)
  - Gradient clipping: 5.0
  - Iterations: $3 \cdot 10^4$
  - Number of MC samples (per iteration): $N_{\text{time}} = 1$ time sample over $[0, T]$ to estimate the KL term and 1 sampled observation time to estimate the reconstruction loss, per trial

- **ARCTA/SVISE**: We found that the Gaussian process parameterization of the posterior described by Course and Nair [7] (see Appendix D therein) led to poor empirical performance. This is because, at a high level, the posterior defines the mean and covariance at each time as a linear combination of means and covariances at observation times. Since in our setting the observations occur sparsely in time, interpolating in the space of means and covariances directly is inappropriate. Instead, we adopt the posterior parameterization from SDE Matching, wherein interpolation occurs in the embedding space [8].

  Moreover, Course and Nair [7] propose chunking the full time series of observations into non-overlapping windows. For each observation within a given window, they produce an embedding using a specified number of future observations, and then they aggregate embeddings within the window. We find that choosing a window size and a number of observations smaller than the full time series degrades performance. For this reason, we consider a window equal to the full time series and allow each embedding to depend on all future observations.

  - `latent_dim = 4`, `obs_dim = 3`, `time_dim = 1`, `hidden_dim = 100`
  - Prior drift:
    `(latent_dim)` $\to$ `Linear(latent_dim, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, latent_dim)` $\to$ `(latent_dim)`
  - Posterior encoder:
    `(n_obs, obs_dim)` $\to$ `GRU` $\to$ `(n_obs, hidden_dim)`

  - Posterior MLP:
    for a specified time, aggregate embeddings from encoder using softmax,
    `(n_obs, hidden_dim), (time_dim)` $\to$ `softmax` $\to$ `(hidden_dim)`
    then map to posterior at that time
    `(hidden_dim + time_dim)` $\to$ `Linear(hidden_dim + time_dim, hidden_dim)` $\to$ `SiLU` $\to$ `Linear(hidden_dim, hidden_dim)` $\to$ `SiLU` $\to$ `Linear(hidden_dim, output_dim)` $\to$ `(output_dim)`
    where `output_dim` is `latent_dim` (mean) + `latent_dim` $\cdot$ `(latent_dim + 1) / 2` (covariance)
  - Diffusion network: fixed $\mathbb{I}_{\texttt{latent_dim}}$
  - Observation model: fixed covariance $\sigma_{\text{obs}}^2 = \eta^2$, learned $\boldsymbol{C}$, $\boldsymbol{d}$
  - LR: 10^{-3} (constant)
  - Gradient clipping: 5.0
  - Iterations: $3 \cdot 10^4$
  - Number of MC samples (per iteration): 1 sample over $[0, T]$ to estimate the KL, per trial; use full set of per-trial observations at each step to estimate reconstruction loss (for consistency with ARCTA window algorithm)

- **SING**:
  - `latent_dim = 4`, `obs_dim = 3`, `hidden_dim = 100`, `grid_size =` $5 \cdot 10^3$ `+ 1`
  - Prior drift:
    `(latent_dim)` $\to$ `Linear(latent_dim, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, latent_dim)` $\to$ `(latent_dim)`
  - Posterior (non-amortized): tensors of shape `(grid_size, latent_dim)` (mean), `(grid_size, latent_dim, latent_dim)` (covariance), `(grid_size - 1, latent_dim, latent_dim)` (cross-covariance)
  - Diffusion network: fixed $\mathbb{I}_{\texttt{latent_dim}}$
  - Observation model: fixed covariance $\sigma_{\text{obs}}^2 = \eta^2$, learned $\boldsymbol{C}$, $\boldsymbol{d}$
  - Mini-batch size: 500
  - LR:
    * SING stepsize (i.e., natural gradient update): 10^{-3} to 10^{-2} on log-linear scale for first 10 iterations, constant after
    * M-step stepsize: $5 \cdot 10^{-3}$ (Adam)
  - Gradient clipping: 5.0
  - Iterations: for each iteration perform 1 SING step, 1 M-step gradient update; 10^4 total iterations
  - Number of MC samples (per iteration): 1 sample to compute expectations under the variational posterior

- **Latent SDE**:
  - `latent_dim = 4`, `obs_dim = 3`, `time_dim = 1`, `hidden_dim = 128`, `embedding_dim = 64`
  - Prior drift:
    `(latent_dim)` $\to$ `Linear(latent_dim, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, latent_dim)` $\to$ `(latent_dim)`
  - Posterior encoder:
    `(n_obs, obs_dim)` $\to$ `reverse` $\to$ `GRU` $\to$ `reverse` $\to$ `(n_obs, hidden_dim)` $\to$ `Linear(hidden_dim, embedding_dim)` $\to$ `(embedding_dim)`
  - Posterior drift network:
    for a specified latent state, use embedding to produce drift
    `(latent_dim + embedding_dim)` $\to$ `Linear(latent_dim + embedding_dim, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, latent_dim)` $\to$ `(latent_dim)`

- Posterior initial distribution:
  use embedding to produce distribution (Gaussian) over initial latent state
  `(embedding_dim)` $\to$ `Linear(embedding_dim, latent_dim + latent_dim)` $\to$ `(latent_dim + latent_dim)`
  first `latent_dim` outputs represent mean, second `latent_dim` outputs represent log standard deviation
- Diffusion network:
  state-dependent, diagonal diffusion coefficient; the input to each diagonal entry is a single latent coordinate
  `latent_dim` $\times$ `[(1)` $\to$ `Linear(1, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, 1)` $\to$ `Sigmoid` $\to$ `(1)]`
- Observation model: fixed covariance $\sigma_{\text{obs}}^2 = \eta^2$, learned $\boldsymbol{C}$, $\boldsymbol{d}$
- Training discretization: 10^{-2} (for posterior sampling)
- LR: initial 10^{-3}, final 10^{-4} (exponential LR decay)
- Gradient clipping: None
- Iterations: $3 \cdot 10^4$
- KL annealing iterations: 10^3
- Number of MC samples (per iteration): 1 path simulated from the posterior per trial

Metrics.
We evaluate the fidelity of the posterior and prior dynamics according to four metrics, which we explain below.

nELBO. Average per-trial negative ELBO (the variational objective). Smaller values are better but do not imply better posterior recovery or learning.

Posterior predictive LL. To evaluate posterior quality, we use the posterior predictive log density at the noise-free observations, averaged over trials and evaluation times. Let $\mathcal{I}$ denote the set of trial-time pairs (m, t) used for evaluation, excluding the observation times used for training in each trial. The metric is

$$\frac{1}{|\mathcal{I}|} \sum_{(m,t)\in\mathcal{I}} \log \bar{q}(\boldsymbol{y}^{(m)}(t), t). \tag{76}$$

Here, $\boldsymbol{y}^{(m)}(t) = \boldsymbol{C}_{\text{true}} \boldsymbol{x}_{\text{true}}^{(m)}(t) + \boldsymbol{d}_{\text{true}}$ is the noise-free observation obtained from the true 3-dimensional latent state generated according to eq. (75). $\bar{q}(\boldsymbol{y}, t)$ represents the posterior predictive density of $\bar{\boldsymbol{y}}(t) = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{d} + \boldsymbol{\epsilon}(t)$, where $\boldsymbol{\epsilon}(t) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$ and $\boldsymbol{x}(t)$ is distributed according to the learned variational posterior for trial m. We suppress the trial index on q, $\bar{q}$, and the posterior samples, which are conditional on the observations from the corresponding trial.

For all cases except Latent SDE [2], $q(\boldsymbol{x}, t) = \mathcal{N}(\boldsymbol{m}(t), \boldsymbol{S}(t))$ is Gaussian, in which case the posterior predictive density admits a closed form $\bar{q}(\boldsymbol{y}, t) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{C}\boldsymbol{m}(t) + \boldsymbol{d}, \boldsymbol{C}\boldsymbol{S}(t)\boldsymbol{C}^\top + \boldsymbol{R})$. For Latent SDE, we estimate $\bar{q}(\boldsymbol{y}^{(m)}(t), t)$ by drawing $N_{\text{density}} = 64$ samples $\boldsymbol{x}^{(k)}(t) \sim q(\boldsymbol{x}, t)$, computing $\bar{\boldsymbol{y}}^{(k)}(t) = \boldsymbol{C}\boldsymbol{x}^{(k)}(t) + \boldsymbol{d}$, and then evaluating

$$\frac{1}{|\mathcal{I}|} \sum_{(m,t)\in\mathcal{I}} \log \left(\frac{1}{N_{\text{density}}} \sum_{k=1}^{N_{\text{density}}} \mathcal{N}(\boldsymbol{y}^{(m)}(t) \mid \bar{\boldsymbol{y}}^{(k)}(t), \boldsymbol{R}) \right). \tag{77}$$

Note that eq. (77) is a sensible approximation to eq. (76) since

$$\begin{aligned} \bar{q}(\boldsymbol{y}, t) &= \int q(\boldsymbol{x}, t) \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{C}\boldsymbol{x} + \boldsymbol{d}, \boldsymbol{R}) d\boldsymbol{x} = \mathbb{E}_{q(\boldsymbol{x},t)}[\mathcal{N}(\boldsymbol{y} \mid \boldsymbol{C}\boldsymbol{x} + \boldsymbol{d}, \boldsymbol{R})] \\ &\approx \frac{1}{N_{\text{density}}} \sum_{k=1}^{N_{\text{density}}} \mathcal{N}(\boldsymbol{y} \mid \bar{\boldsymbol{y}}^{(k)}(t), \boldsymbol{R}). \end{aligned}$$

Prior learning. For evaluating the learned prior dynamics, we draw 2048 samples from the learned prior, starting from $p(\boldsymbol{x}, 0)$ and simulating using Euler–Maruyama with discretization 10^{-3}. We evaluate the learned output mapping at these samples, from which we compute the following metrics:

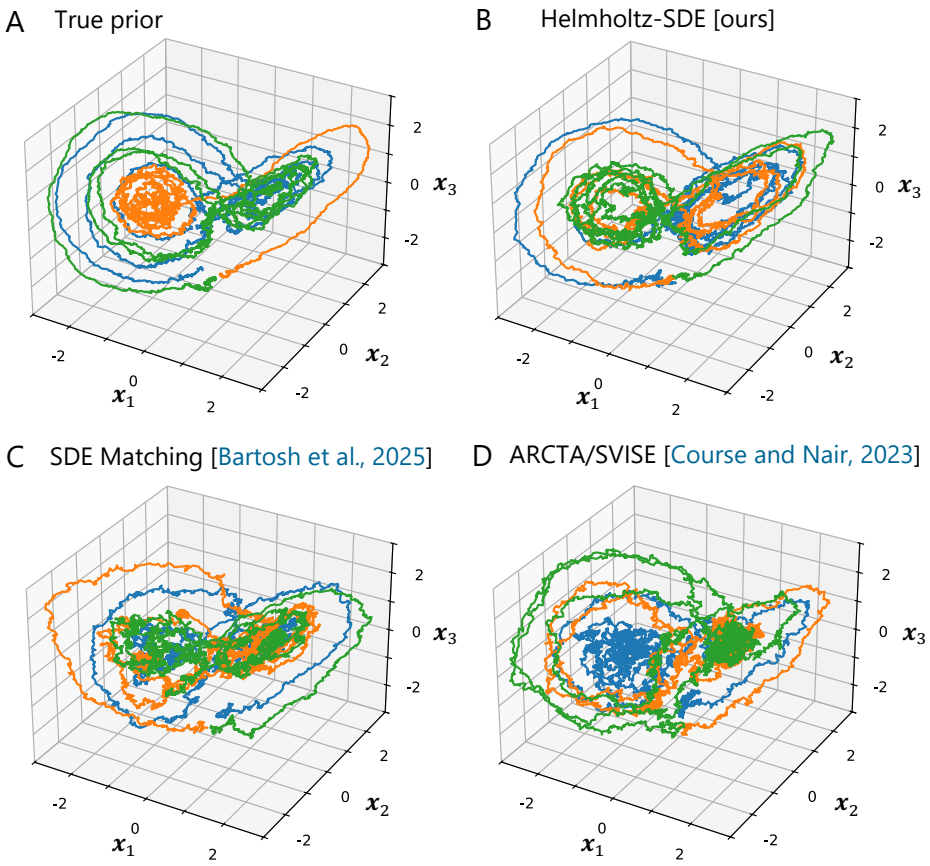


Figure 12: Three random samples from the true prior (**A**), the prior learned by Helmholtz-SDE (**B**), SDE Matching (**C**), and ARCTA/SVISE (**D**). Helmholtz-SDE more faithfully learns the dynamics, particularly the outward spiral dynamics within each lobe of the attractor.

(i) **One-time symmetric KL divergence**: symmetric KL divergence at each time in $[0, T]$, averaged over times. For samples from two distributions p and q, we estimate the KL divergence $\mathrm{D}_{\mathrm{KL}}(p \parallel q)$ by fitting two Gaussian kernel density estimates and computing a Monte Carlo average.

We do so using the following procedure, which mirrors that of Kiyohara et al. [38]:

(a) Using 3-fold CV, select the KDE bandwidth for q; fit the density estimate (on all samples) using the chosen bandwidth.

(b) Split the samples from p into 10 folds. Holding out a single fold at a time, repeat the following subprocedure:

i. Choose the kernel bandwidth for p using 3-fold CV; fit the density estimate (on all training folds) using the chosen bandwidth.

ii. Compute a Monte Carlo estimate of the KL divergence using the held-out fold.

The outer CV step evaluates each density estimate for p on held-out samples, avoiding reuse of its training samples for the Monte Carlo log-density average. We repeat this procedure with p and q exchanged to estimate the reverse KL divergence.

(ii) **Global autocorrelation error**: Using samples from the learned prior and true prior, we estimate lagged correlation matrices in the noise-free observation coordinates $\boldsymbol{u}(t) = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{d}$ and $\boldsymbol{u}_{\mathrm{true}}(t) = \boldsymbol{C}_{\mathrm{true}}\boldsymbol{x}_{\mathrm{true}}(t) + \boldsymbol{d}_{\mathrm{true}}$, respectively. For each lag τ, we estimate $\mathrm{Corr}(\boldsymbol{u}(t), \boldsymbol{u}(t+\tau))$ across samples at each time t such that $t, t+\tau$ lie on the sampling grid, and then average these matrices over all valid times. We do the same for the true prior. We compute the squared Frobenius norm difference between the two time-averaged

correlation matrices and average over lags τ on a linear grid of $[0, 2]$ with spacing $5 \cdot 10^{-2}$. Finally, we normalize by the total number of entries in the correlation matrix ($3 \times 3 = 9$) and take the square root of the result, yielding the entrywise RMSE.

Ablation of q-divergence-free correction. We report ablations over the q-divergence-free correction used by Helmholtz-SDE. In Table 1, we report the negative ELBO for polynomial degrees $\ell \in \{1, 2, 3\}$ and for different estimators of the polynomial divergence-free correction. For all estimators, we evaluate the final negative ELBO using $N_{\text{quad}} = 6$ quadrature points per dimension (i.e., the estimators differ during training but not evaluation). κ denotes the degree of the Stein estimand eq. (59); when κ is smaller than the polynomial degree ℓ, we take derivatives of the largest entries of $\boldsymbol{\alpha}$.

In general, we observe that the Stein estimand eq. (59) performs better than the original expectation, especially when it is estimated using Monte Carlo with few samples N_{mc}. Gauss–Hermite quadrature is less sensitive to the degree of the Stein estimand. Moreover, we see that increasing the polynomial degree from $\ell = 1$ to $\ell = 2$ reduces the negative ELBO by approximately 0.7 nats. We do not observe a significant improvement when increasing the degree from $\ell = 2$ to $\ell = 3$.

	Monte Carlo (N_{mc})					Quadrature (N_{quad})	
κ	1	4	16	64	1024	2	4
0	76.56 ± 2.80	55.74 ± 0.41	49.89 ± 0.18	48.05 ± 0.13	47.61 ± 0.30	47.44 ± 0.17	47.42 ± 0.14
1	47.96 ± 0.45	47.48 ± 0.17	47.43 ± 0.17	47.53 ± 0.23	47.60 ± 0.22	47.43 ± 0.15	47.50 ± 0.20

(a) $\ell = 1$

	Monte Carlo (N_{mc})					Quadrature (N_{quad})	
κ	1	4	16	64	1024	2	4
0	111.66 ± 2.08	93.88 ± 0.47	60.32 ± 0.53	51.71 ± 0.54	46.94 ± 0.12	46.82 ± 0.16	46.69 ± 0.16
1	68.32 ± 0.90	52.97 ± 1.12	47.89 ± 0.17	47.11 ± 0.20	46.95 ± 0.18	46.71 ± 0.13	47.03 ± 0.25
2	47.23 ± 0.41	46.85 ± 0.19	46.93 ± 0.16	46.80 ± 0.20	46.71 ± 0.12	46.86 ± 0.13	46.86 ± 0.17

(b) $\ell = 2$

	Monte Carlo (N_{mc})					Quadrature (N_{quad})	
κ	1	4	16	64	1024	2	4
0	172.69 ± 0.90	109.99 ± 1.94	94.06 ± 0.37	60.25 ± 0.39	47.84 ± 0.18	55.33 ± 0.40	46.84 ± 0.28
1	92.99 ± 1.28	59.86 ± 0.44	50.16 ± 0.68	49.25 ± 3.36	46.82 ± 0.23	46.79 ± 0.18	46.85 ± 0.19
2	47.19 ± 0.30	46.94 ± 0.22	46.83 ± 0.18	46.81 ± 0.16	46.79 ± 0.19	46.82 ± 0.21	46.82 ± 0.17
3	46.98 ± 0.23	47.04 ± 0.35	46.96 ± 0.28	46.65 ± 0.11	—	46.77 ± 0.16	—

(c) $\ell = 3$

Table 1: Ablation of the q-divergence-free correction used by Helmholtz-SDE. Dashes represent configurations that could not fit in memory on an 80 GB GPU. Errors are ± 2 SE over 10 replicates.

Next, we evaluate a strategy for estimating the improvement from higher order polynomial corrections that does *not* require retraining Helmholtz-SDE. In Appendix B.5.2, we derived a lower bound on the improvement in the negative ELBO from degree ℓ to $\ell + k$ that can be approximated at the marginals of the degree ℓ model. We train models with degree $\ell = 0$ (reference drift) and $\ell = 1$ polynomial corrections using the Stein estimand with $N_{\text{mc}} = 1$. After training, we approximate the lower bound by computing the Hermite coefficients of the residual with $N_{\text{quad}} = 6$ quadrature points.

In Table 2, we see that our estimated bound on the improvement from $\ell = 0$ to $\ell = 1$ is nearly 12 nats, which is smaller than the observed 20 nat improvement in Figure 3. The estimated bound on the improvement from $\ell = 1$ to $\ell = 2$ is 0.4 nats, smaller than the 0.7 nat improvement in Table 1.

D.2 Experimental predator-prey dynamics

In this subsection, we provide details regarding our experiments with the algae-rotifer predator-prey system from Blasius et al. [52].

	$\ell = 1$	$\ell = 2$	$\ell = 3$
$0 \to \ell$	11.72 ± 0.80	12.16 ± 0.77	12.18 ± 0.77
$1 \to \ell$	—	0.374 ± 0.006	0.378 ± 0.006

Table 2: $k \to \ell$ estimates a lower bound on the reduction in negative ELBO from using a degree ℓ polynomial over a degree k polynomial. See Appendix B.5.2 for details. Errors are ± 2 SE over 10 replicates.

Preprocessing. We obtain the dataset made public by the authors at https://figshare.com/articles/dataset/Time_series_of_long-term_experimental_predator-prey_cycles/10045976/1.

In our analysis, we consider only the first trial, which is the longest and spans 374 days. We choose not to consider multiple trials (4 total in the same experimental condition) simultaneously since there are apparent differences between them, for example, the mean and standard deviation of population sizes. Modeling the dynamics across multiple trials using, for example, a hierarchical Bayesian model is interesting future work.

Since we learn autonomous (time-independent) dynamics, we restrict each trial to periods of coherent oscillation as determined by Blasius et al. [52]. Otherwise, the latent dynamics must explain both coherent and non-coherent (no systematic phase lag) oscillations. For the trial of interest, this yields four subtrials $\{[19.5, 130.5], [159.5, 240.5], [274.5, 295.5], [314.5, 370.5]\}$.

We next log transform the data. The reason for this decision is that, in stochastic models of population dynamics, modeling the noise as multiplicative in the state is a popular choice, particularly due to the stability properties it affords [74]. Log transforming the data then allows us to model the noise as state-independent, which is the setting in which Helmholtz-SDE operates. For example, consider the Lotka–Volterra equations with state-dependent noise for predator y and prey x

$$\begin{aligned} dx(t) &= (\alpha x(t) - \beta x(t) y(t))dt + \sigma_x x d\boldsymbol{w}_1(t) \\ dy(t) &= (\delta x(t) y(t) - \gamma y(t))dt + \sigma_y y d\boldsymbol{w}_2(t) \end{aligned} \tag{78}$$

with fixed initial distribution and $\alpha, \beta, \delta, \gamma > 0$. Applying Ito's lemma to the map $(x(t), y(t)) \mapsto (\log x(t), \log y(t))$ yields the dynamics in log coordinates

$$\begin{aligned} d(\log x(t)) &= \left(\alpha - \beta y(t) - \frac{1}{2}\sigma_x^2\right) dt + \sigma_x d\boldsymbol{w}_1(t) \\ d(\log y(t)) &= \left(\delta x(t) - \gamma - \frac{1}{2}\sigma_y^2\right) dt + \sigma_y d\boldsymbol{w}_2(t) \end{aligned} . \tag{79}$$

Importantly, the diffusion coefficient appearing in eq. (79) does not depend on the state. Observe that if eq. (79) describes the prior of a latent SDE model and we choose a posterior approximation with Gaussian one-time marginals, then the one-time marginal distributions of $x(t)$ become log-normal.

We hold out the last 8 observations of each subtrial to compute the forecast MSE.

Generative model. We learn a $K = 2$ dimensional dynamical systems model of the predator-prey dynamics. The prior initial state is modeled as Gaussian $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{V}_0)$, $\boldsymbol{\mu}_0 \in \mathbb{R}^2$, $\boldsymbol{V}_0 \in \mathbb{R}^{2\times 2}$; the prior drift function eq. (1) is a neural network; and the observation model is linear and Gaussian $\mathcal{N}(\boldsymbol{y}(t) | \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{d}, \sigma_{\text{obs}}^2 \mathbb{I})$, $\boldsymbol{C} \in \mathbb{R}^{2\times 2}$, $\boldsymbol{d} \in \mathbb{R}^2$, $\sigma_{\text{obs}} > 0$. We perform joint inference and parameter learning.

Baselines. We compare Helmholtz-SDE to SDE Matching [8] and SVISE [6]. For this case study, we find that defining the posterior on a grid of size 0.1 days, as done in Appendix C.2, yields good performance. Other parameterizations, such as the kernel-GLM in SVISE [6], perform comparably. For SDE Matching, we consider modeling the diffusion coefficient as both state-dependent and state-independent, but we find no substantive difference between the results. Results reported are for a state-dependent diffusion coefficient.

- **Helmholtz-SDE**:
  - `latent_dim = 2`, `obs_dim = 2`, `time_dim = 1`, `hidden_dim = 64`

  - Prior drift:
    `(latent_dim)` $\to$ `Linear(latent_dim, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, latent_dim)` $\to$ `(latent_dim)`
  - Posterior: grid of means and covariances, with spacing 0.1 days; linear interpolation between grid points
  - Diffusion network: fixed $\mathbb{I}_{\texttt{latent_dim}}$
  - Helmholtz correction: linear approximation $\ell = 1$, computed using the Stein estimand with $N_{\text{mc}} = 1$
  - Observation model: learned $\boldsymbol{C}$, $\boldsymbol{d}$ and σ^2_{obs}
  - LR: 10^{-2} initial, 10^{-3} final (exponential LR decay)
  - Gradient clipping: 5.0
  - Iterations: $3 \cdot 10^4$
  - Number of MC samples (per iteration): $N_{\text{time}} = 10^4$ samples over $[0, T]$ to estimate each of the KL and reconstruction loss, per trial
- **SDE Matching**:
  - `latent_dim = 2`, `obs_dim = 2`, `time_dim = 1`, `hidden_dim = 64`
  - Prior drift:
    `(latent_dim)` $\to$ `Linear(latent_dim, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, latent_dim)` $\to$ `(latent_dim)`
  - Posterior: grid of means and covariances, with spacing 0.1 days; linear interpolation between grid points
  - Diffusion network:
    state-dependent, diagonal diffusion coefficient; the input to each diagonal entry is a single latent coordinate
    `latent_dim` $\times$ `[(1)` $\to$ `Linear(1, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, 1)` $\to$ `Sigmoid` $\to$ `(1)]`
  - Observation model: learned $\boldsymbol{C}$, $\boldsymbol{d}$ and σ^2_{obs}
  - LR: 10^{-2} initial, 10^{-3} final (exponential LR decay)
  - Gradient clipping: 5.0
  - Iterations: $3 \cdot 10^4$
  - Number of MC samples (per iteration): $N_{\text{time}} = 10^4$ samples over $[0, T]$ to estimate each of the KL and reconstruction loss, per trial
- **SVISE**:
  - `latent_dim = 2`, `obs_dim = 2`, `time_dim = 1`, `hidden_dim = 64`
  - Prior drift:
    `(latent_dim)` $\to$ `Linear(latent_dim, hidden_dim)` $\to$ `Softplus` $\to$ `Linear(hidden_dim, latent_dim)` $\to$ `(latent_dim)`
  - Posterior: grid of means and covariances, with spacing 0.1 days; linear interpolation between grid points
  - Diffusion network: fixed $\mathbb{I}_{\texttt{latent_dim}}$
  - Observation model: learned $\boldsymbol{C}$, $\boldsymbol{d}$ and σ^2_{obs}
  - LR: 10^{-2} initial, 10^{-3} final (exponential LR decay)
  - Gradient clipping: 5.0
  - Iterations: $3 \cdot 10^4$
  - Number of MC samples (per iteration): $N_{\text{time}} = 10^4$ samples over $[0, T]$ to estimate each of the KL and reconstruction loss, per trial

Wavelet analysis. We assess the quality of the learned prior dynamics using the wavelet analysis from Blasius et al. [52]. In particular, we draw 1024 samples from the learned prior with stepsize $2 \cdot 10^{-3}$ days, each of length 100 days. From each sample, we downsample to a grid of 1 day and extract the peak cyclic period (Figure 4C), the lag of predator behind prey (Figure 4D), and the rotational coherence (Figure 4D). We then report the median and IQR of these statistics across 10 replicates of fitting the latent SDE model.

We use the authors' public codebase https://github.com/berndblasius/WaveletAnalysis to perform our analyses. Next, we briefly discuss how we compute each of these statistics. We refer the interested reader to Blasius et al. [52, supplementary methods] as well as Torrence and Compo [75] for additional background on wavelet analysis.

1. The continuous wavelet transform.
The *continuous wavelet transform* of a one-dimensional signal $(x(t))_{0\le t\le T}$ is defined as the convolution of the signal with the complex-valued wavelet function, centered at time t and dilated by scale $s > 0$

$$W_x(s,t) = \frac{1}{\sqrt{s}} \int_0^T x(t')\psi^* \left(\frac{t'-t}{s} \right) dt'. \tag{80}$$

Here, ψ^* represents the complex conjugate of ψ. While there are many possible choices of wavelet function, we consider the Morlet wavelet with $\omega_0 = 6$:

$$\psi(\tau) = \pi^{-1/4} \exp\{i\omega_0\tau\} \exp\{-\tau^2/2\}.$$

To build intuition, consider what eq. (80) would compute if ψ were a pure complex exponential $\psi(\tau) = \exp\{2\pi i\tau\}$. Then $|W_x(s,t)|$ equals $s^{-1/2}$ times the magnitude of the Fourier transform of x at frequency $1/s$: it measures how strongly the signal oscillates at period s, but the result does not depend on t and so carries no information about when in the time series this oscillation occurs. The Morlet wavelet recovers time localization by multiplying the complex exponential by a Gaussian envelope $\exp\{-\tau^2/2\}$, which restricts the integral in eq. (80) to a neighborhood of t. The price of this localization is that ψ no longer corresponds to a single frequency; instead, each scale s is associated with a *Fourier period* by matching s to the scale at which the wavelet power of a sinusoid of that period is maximized. For the Morlet wavelet, this correspondence is

$$T(s) = \frac{4\pi s}{\omega_0 + \sqrt{2+\omega_0^2}}, \tag{81}$$

so that $T(s) \approx 1.03s$ when $\omega_0 = 6$ [75]. With this identification, $|W_x(s,t)|$ measures the local oscillatory content of the signal at time t and period $T(s)$.

Following Blasius et al. [52], we evaluate the wavelet transform on a logarithmic grid of scales. The smallest scale is $s_0 = 2$ days, and the largest scale is $s_0 2^J = 16$ days, where $J = 3$ is the number of octaves, and we use 100 grid points per octave.

2. Wavelet power spectrum and cross spectrum.
The local oscillatory content of a single signal is summarized by the *wavelet power spectrum* (WPS), $|W_x(s,t)|^2$. To compare two signals, x_1 (algae) and x_2 (rotifers), we form the *wavelet cross-spectrum* (WCS) $W_{x_1x_2}(s,t) = W_{x_1}(s,t)W_{x_2}^*(s,t)$, whose magnitude $|W_{x_1x_2}(s,t)|$ quantifies the joint oscillatory power at scale s and time t. Following Blasius et al. [52], we smooth both the WPS and WCS using a Gaussian temporal filter and a moving-average filter over scales; we denote this operator by $\mathcal{S}(\cdot)$. The argument of the smoothed WCS,

$$\phi(s,t) = \arg \mathcal{S}(W_{x_1x_2}(s,t)), \tag{82}$$

gives the local phase difference between the two signals at scale s and time t. A nonzero phase difference $\phi(s,t)$ indicates that one signal leads the other; converting to a time lag at the corresponding Fourier period yields $\Delta t(s,t) = \phi(s,t)T(s)/(2\pi)$.

A high value of $|W_{x_1x_2}(s,t)|$ does not by itself indicate that the two signals are oscillating coherently: it may simply reflect that both signals have large individual power at (s,t), even if their phases drift independently. To distinguish coherent oscillation from coincident power, we use the *wavelet coherence* (WCO), defined as the normalized, locally smoothed cross-spectrum

$$R(s,t) = \frac{|\mathcal{S}\left(W_{x_1x_2}(s,t)\right)|}{\sqrt{\mathcal{S}\left(|W_{x_1}(s,t)|^2\right) \cdot \mathcal{S}\left(|W_{x_2}(s,t)|^2\right)}}. \tag{83}$$

Coherence takes values in $[0,1]$ and is invariant to the marginal amplitudes of x_1 and x_2: $R^2(s,t) \approx 1$ at points (s,t) where the two signals oscillate with a stable phase relationship over a local neighborhood. As pointed out by Torrence and Compo [75], smoothing is necessary since otherwise the wavelet coherence is identically 1 for all times and scales.

3. Cone of influence.
Because the wavelet transform eq. (80) is computed on a finite signal, the integral is implicitly zero-padded outside $[0, T]$. This padding contaminates $W_x(s,t)$ near the boundaries $t = 0$ and $t = T$, with the contamination extending farther into the interior at larger scales s. The *cone of influence* (COI) is the region of the time-scale plane where this edge contamination is important; following Torrence and Compo [75], points (s,t) inside the COI are discarded. For the regular time grid $\mathcal{T}$, let $\mathcal{T}_{\mathrm{COI}}(s)$ denote the times inside the COI at scale s, and write $\mathcal{T}_{\mathrm{COI}}(s)^c = \mathcal{T} \setminus \mathcal{T}_{\mathrm{COI}}(s)$ for the retained times.

4. Extracted statistics. For each prior sample, we use the WCS, WCO, and per-signal WPS to extract three summary statistics matching those reported by Blasius et al. [52].

1. **Peak cyclic period**: The peak period of each individual population (algae, rotifers) is obtained by averaging the square root of the smoothed WPS outside the COI and selecting the scale at which this time-averaged amplitude is maximized:

$$s_i^\star = \arg\max_s \frac{1}{|\mathcal{T}_{\mathrm{COI}}(s)^c|} \sum_{t \in \mathcal{T}_{\mathrm{COI}}(s)^c} \mathcal{S}(|W_{x_i}(s,t)|^2)^{1/2}, \qquad T_i^\star = T(s_i^\star), \tag{84}$$

   where the maximization is over scales with at least one retained time, and the maximizing scale is mapped to a period via eq. (81). The peak period for a given sample is the average of the rotifer and algae peak periods. We report the median peak period across all samples, $T^\star_{\mathrm{ensemble}}$.
2. **Predator-prey phase lag**: Let s_{joint} be the scale of the dominant joint period, the scale at which the amplitude of the smoothed WCS, $|\mathcal{S}(W_{x_1x_2}(s,t))|$, averaged over the full series is maximal. Let $\mathcal{B}$ be the band of scales within 0.6 octaves of s_{joint}. At each time t, we locate the scale of maximal coherence within this band,

$$s^\star(t) = \arg\max_{s \in \mathcal{B}} R(s,t),$$

   and retain the time t if $(s^\star(t), t)$ lies outside the COI, the coherence $R(s^\star(t), t)$ is at least the significance threshold 0.83 used by Blasius et al. [52], and the maximum is not attained at an edge of the band (an edge maximum suggests that the coherence peak may lie outside the band). We denote this set of retained times by $\mathcal{T}^{(n)}_{\mathrm{coh}}$ for prior sample n. At each retained time we record the phase $\phi(s^\star(t), t)$ and summarize these phases by their circular mean

$$\bar{\phi}^{(n)} = \arg\left(\sum_{t \in \mathcal{T}^{(n)}_{\mathrm{coh}}} e^{i\phi(s^\star(t),t)} \right).$$

   The ensemble phase lag is the arithmetic mean of $\bar{\phi}^{(n)}$ across prior samples with at least one retained time. We convert this angle to a lag in time units by multiplying by $T^\star_{\mathrm{ensemble}}/(2\pi)$, where $T^\star_{\mathrm{ensemble}}$ is the median peak period from the previous statistic.
3. **Rotational coherence**: The rotational coherence of prior sample n is the fraction of time points that are retained in the phase analysis above,

$$\rho^{(n)}_{\mathrm{coh}} = \frac{|\mathcal{T}^{(n)}_{\mathrm{coh}}|}{|\mathcal{T}|},$$

   where $|\mathcal{T}|$ is the total number of time points on the regular grid, including those excluded from the phase analysis by the cone of influence. This statistic measures the proportion of the trajectory over which algae and rotifers oscillate with a stable phase relationship near the dominant period, and is a sample-level analogue of the "rotational fraction" reported by Blasius et al. [52]. We report the median across prior samples.

In addition to the wavelet analysis, we compute the forecast MSE on held-out observations. We do so by simulating 128 trajectories from the learned prior SDE, initialized from the posterior at the last training time, using Euler–Maruyama with stepsize 10^{-3}; squared errors are averaged over held-out observations, subtrials, and trajectories.

D.3 Reduced order models for fluid dynamics

In this subsection, we discuss our experiments modeling a two-dimensional incompressible flow past a circular cylinder. We obtain the public dataset from Course and Nair [6] at https://github.com/coursekevin/svise.

The data is simulated on the domain $[-0.98, 28.98] \times [-1.98, 1.98]$ with a circular cylinder of radius 0.5 placed at $(0, 0)$ (Figure 5). The velocity field is obtained by numerically solving the incompressible Navier–Stokes equation

$$\begin{aligned} &\frac{\partial \boldsymbol{u}_i(\boldsymbol{z}, t)}{\partial t} + \sum_j \boldsymbol{u}_j(\boldsymbol{z}, t) \frac{\partial \boldsymbol{u}_i(\boldsymbol{z}, t)}{\partial \boldsymbol{z}_j} = -\frac{\partial \bar{p}(\boldsymbol{z}, t)}{\partial \boldsymbol{z}_i} + \frac{1}{\mathrm{Re}} \sum_j \frac{\partial^2 \boldsymbol{u}_i(\boldsymbol{z}, t)}{\partial \boldsymbol{z}_j^2} \\ &\sum_i \frac{\partial \boldsymbol{u}_i(\boldsymbol{z}, t)}{\partial \boldsymbol{z}_i} = 0, \quad i \in \{1, 2\}, \quad (\boldsymbol{z}, t) \in \mathbb{R}^2 \times \mathbb{R} \end{aligned} \tag{85}$$

where $\boldsymbol{u} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2$ and $\bar{p} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ are the non-dimensionalized velocity and pressure, respectively, and Re is the Reynolds number. Re represents the ratio of inertial to viscous forces. For our experiments we use $\mathrm{Re} = 2 \cdot 10^3$, which yields strong vortices in the wake of the cylinder. Equation (85) is solved by discretizing the spatial domain on a grid of size 596602, and solving the Navier–Stokes equation on this grid (in time), subject to the boundary conditions defined by the outer domain and 2D cylinder (Figure 5A). The dataset simulates the dynamics for 305.1 time units with snapshots every 0.1 time units.

Preprocessing. We follow the same preprocessing steps as in Course and Nair [6]. Namely, we drop the first 20% of the snapshots representing transients in the system. We then use the interval $[61, 256.2]$ for training and the subsequent 10 time units $[256.2, 266.2]$ for forecasting. From the training data, we perform the proper orthogonal decomposition (POD) [62] and retain the leading spatial modes that together capture 90% of the time-averaged fluctuation kinetic energy; this amounts to 38 modes. We then project the snapshots onto these 38 POD modes to obtain a low-dimensional representation of the system. Lastly, we normalize the POD coefficients by the maximum standard deviation, computed along each dimension (i.e., normalization is by a scalar a).

Modeling the dynamics in the space of the top POD modes will yield a reduced order model of the fluid dynamics. Given an initial condition, one can project the velocity field on the grid into the space of POD modes, roll forward the (stochastic) dynamics, and reconstruct the trajectory on the original grid. Mathematically, this is equivalent to taking the dynamics $\boldsymbol{f}$ defined in the space of POD modes and defining the dynamics $\bar{\boldsymbol{f}} = a\boldsymbol{\Phi}\boldsymbol{f}(a^{-1}\boldsymbol{\Phi}^\top(\boldsymbol{z}_{\text{grid}} - \bar{\boldsymbol{z}}_{\text{grid}}))$ in the original data space, where $\bar{\boldsymbol{z}}_{\text{grid}}$ is the mean of the velocity measurements across time and $\boldsymbol{\Phi}$ is the matrix whose columns contain the top POD modes. Note that the dynamics $\bar{\boldsymbol{f}}$ live on a low-dimensional affine subspace.

To generate observations, we subsample the POD coefficients to a grid of size $\Delta t = 0.3$. We observe that many of the higher-order modes oscillate at a smaller scale than the lower-order modes. In order to add Gaussian noise in such a way that the higher order modes do not have low signal-to-noise ratio, we add noise proportional to the standard deviation along each dimension. And to demonstrate a setting with high posterior uncertainty, we choose noise standard deviation equal to 0.3 times the standard deviation along each dimension (signal-to-noise ratio 3.33). We observe that with more frequent observations (e.g., $\Delta t = 0.1$) and smaller noise scale (e.g., 0.1 times the standard deviation along each dimension), the discrepancy between the learned dynamics is less apparent.

Generative model. We learn a $K = 38$ dimensional ROM of fluid flow past the cylinder. The prior initial state is modeled as Gaussian $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{V}_0)$, $\boldsymbol{\mu}_0 \in \mathbb{R}^{38}$, $\boldsymbol{V}_0 \in \mathbb{R}^{38\times 38}$; the prior drift function eq. (1) is a neural network; and the observation model is Gaussian $\mathcal{N}(\boldsymbol{y}(t)|\boldsymbol{x}(t), \boldsymbol{R})$. $\boldsymbol{R}$ is the diagonal observation-noise covariance, which we fix to its true value.

Baselines. To match the original experimental setup in Course and Nair [6], we use the kernel-GLM posterior parameterization proposed by the authors. See Appendix C.2 for a discussion.

For SDE Matching [8], we saw in our preliminary experiments that learning a state-dependent diffusion coefficient degraded learning. In order to separate the prior parameterization from posterior expressivity, we drop the state dependence from the diffusion parameterization.

- **Helmholtz-SDE**:
  - `latent_dim = 38`, `obs_dim = 38`, `time_dim = 1`, `hidden_dim = 100`
  - Prior drift:
    `(latent_dim)` $\to$ `Linear(latent_dim, hidden_dim)` $\to$ `Tanh` $\to$ `Linear(hidden_dim, hidden_dim)` $\to$ `Tanh` $\to$ `Linear(hidden_dim, latent_dim)` $\to$ `(latent_dim)`
  - Kernel-GLM, with 500 evenly-spaced temporal inducing points; defined for both the posterior mean and covariance
  - Diffusion network: learned diagonal $\mathbf{\Sigma}$, state and time independent
  - Helmholtz correction: linear approximation $\ell = 1$, computed using the Stein estimand with $N_{\text{mc}} = 1$
  - Observation model: fixed
  - LR: 10^{-3} initial, 10^{-4} final (exponential LR decay)
  - Gradient clipping: None
  - Iterations: $3 \cdot 10^4$
  - Number of MC samples (per iteration): $N_{\text{time}} = 512$ samples over $[0, T]$ to estimate each of the KL and reconstruction loss, per trial
- **SDE Matching**:
  - `latent_dim = 38`, `obs_dim = 38`, `time_dim = 1`, `hidden_dim = 100`
  - Prior drift:
    `(latent_dim)` $\to$ `Linear(latent_dim, hidden_dim)` $\to$ `Tanh` $\to$ `Linear(hidden_dim, hidden_dim)` $\to$ `Tanh` $\to$ `Linear(hidden_dim, latent_dim)` $\to$ `(latent_dim)`
  - Kernel-GLM, with 500 evenly-spaced temporal inducing points; defined for both the posterior mean and covariance
  - Diffusion network: learned diagonal $\mathbf{\Sigma}$, state and time independent
  - Observation model: fixed
  - LR: 10^{-3} initial, 10^{-4} final (exponential LR decay)
  - Gradient clipping: None
  - Iterations: $3 \cdot 10^4$
  - Number of MC samples (per iteration): $N_{\text{time}} = 512$ samples over $[0, T]$ to estimate each of the KL and reconstruction loss, per trial
- **SVISE**:
  - `latent_dim = 38`, `obs_dim = 38`, `time_dim = 1`, `hidden_dim = 100`
  - Prior drift:
    `(latent_dim)` $\to$ `Linear(latent_dim, hidden_dim)` $\to$ `Tanh` $\to$ `Linear(hidden_dim, hidden_dim)` $\to$ `Tanh` $\to$ `Linear(hidden_dim, latent_dim)` $\to$ `(latent_dim)`
  - Kernel-GLM, with 500 evenly-spaced temporal inducing points; defined for both the posterior mean and covariance
  - Diffusion network: learned diagonal $\mathbf{\Sigma}$, state and time independent
  - Observation model: fixed
  - LR: 10^{-3} initial, 10^{-4} final (exponential LR decay)
  - Gradient clipping: None
  - Iterations: $3 \cdot 10^4$
  - Number of MC samples (per iteration): $N_{\text{time}} = 512$ samples over $[0, T]$ to estimate each of the KL and reconstruction loss, per trial

Metrics. In Figure 5B, we compute the space–time autocorrelation of streamwise velocity fluctuations along the wake centerline. We draw samples from the prior starting at the true initial condition $\boldsymbol{x}(0)$ at the beginning of the training window, and simulating forward for 10 time units using Euler–Maruyama with stepsize 10^{-4}. We observe that some trajectories begin to numerically diverge beginning at approximately $15 - 20$ time units, so we only analyze the first 10 units.

At each sampled trajectory, we reconstruct the velocity field on the original grid by projecting back through the POD modes, and we extract the streamwise velocity component $\boldsymbol{u}_1(\boldsymbol{z}, t)$ along the strip $[1, 10] \times [-0.1, 0.1]$ in the near wake, averaging across the transverse direction. For each trajectory, we compute the joint space–time autocorrelation

$$C(\Delta \boldsymbol{z}_1, \Delta t) = \frac{\mathbb{E}\left[\tilde{\boldsymbol{u}}_1(\boldsymbol{z}_1, t)\tilde{\boldsymbol{u}}_1(\boldsymbol{z}_1 + \Delta \boldsymbol{z}_1, t + \Delta t)\right]}{\mathbb{E}\left[\tilde{\boldsymbol{u}}_1(\boldsymbol{z}_1, t)^2\right]},$$

where $\tilde{\boldsymbol{u}}_1$ denotes the fluctuation about the temporal mean at each streamwise location within the trajectory. The numerator is averaged over valid reference pairs at each lag, and the denominator over all reference points in the trajectory. We average the resulting correlations across prior samples. The reference correlation is computed by concatenating the training and test trajectories.

There are two apparent patterns in the space–time autocorrelation: (1) oscillation in streamwise lag (from the periodic vortex pattern in the wake) and (2) diagonal banding in time (vortices being carried downstream). The failure mode of simulation-free VI is specifically a loss of temporal coherence. This is because temporal correlations are those that depend on the joint posterior path law which, in turn, degrades the prior path law. The spatial correlations, on the other hand, exist within a single snapshot of time.

We also report the cumulative MSE, summed over the 10-unit forecasting window at intervals of 0.1 time units. We compute this quantity by simulating 128 trajectories for 10 time units using Euler–Maruyama with stepsize 10^{-4}, starting from the ground truth POD coefficients at the end of the training window. We then transform the samples into the physical space by multiplying by the normalization constant and computing the linear combination with the top-38 POD vectors. Finally, we compute the MSE between the reconstructed velocity and the projection of the (true) velocity onto the top-38 POD modes, averaged over the 128 simulated trajectories. We normalize by the total number of velocity components on the grid. We report results for 10 training replicates with ± 2 SE.

	Helmholtz-SDE	SDE Matching	SVISE
forecast MSE $\downarrow$	$\mathbf{3.46 \pm 0.75}$	$\mathbf{3.88 \pm 0.53}$	$\mathbf{3.86 \pm 0.43}$
nELBO $\downarrow$	$\mathbf{-2.57 \pm 0.00}$	-2.11 ± 0.01	-2.13 ± 0.01

Table 3: Cumulative MSE computed over the 10-time-unit forecasting window. nELBO is reported in units of 10^4, e.g., Helmholtz-SDE has nELBO $-2.57 \cdot 10^4 \pm 0.00 \cdot 10^4$.

E Runtime comparisons

We compare the runtime of Helmholtz-SDE to all baseline VI algorithms considered.

We time all algorithms on a single NVIDIA H100 GPU, and we exclude JIT compilation time and ELBO evaluation time from the runtime calculation. For Helmholtz-SDE, SING, SDE Matching, and SVISE, we assess convergence using the negative ELBO eq. (3) calculated on a dense time grid. For Latent SDE, we approximate the negative ELBO using 10 samples per trial from the variational SDE.

Method	ms/iter	nELBO < 50		nELBO < 70	
		time (s)	iterations	time (s)	iterations
Helmholtz-SDE	8.56 ± 0.05	123.7 ± 3.4	14400 ± 427	27.7 ± 1.2	3200 ± 158
SDE Matching	8.08 ± 0.06	—	—	185.9 ± 18.7	22950 ± 2294
ARCTA/SVISE	39.58 ± 0.04	—	—	336.2 ± 32.6	8490 ± 823
SING	1157.8 ± 0.8	3695.6 ± 503.8	3192 ± 435	807.5 ± 117.8	697 ± 102
Latent SDE	1068.2 ± 42.7	—	—	9464.9 ± 930.4	8870 ± 893

(a) Lorenz system

Method	ms/iter	nELBO < 70		nELBO < 90	
		time (s)	iterations	time (s)	iterations
Helmholtz-SDE	21.55 ± 0.01	96.0 ± 7.6	4450 ± 352	16.5 ± 1.6	760 ± 74
SDE Matching	15.28 ± 0.01	—	—	40.2 ± 8.1	2620 ± 528
SVISE	12.80 ± 0.01	—	—	41.3 ± 7.7	3220 ± 604

(b) Predator–prey system

Method	ms/iter	nELBO $< -2.5 \cdot 10^4$		nELBO $< -2.0 \cdot 10^4$	
		time (s)	iterations	time (s)	iterations
Helmholtz-SDE	859.5 ± 9.6	17094.7 ± 314.0	19870 ± 313	5197.9 ± 72.8	6000 ± 42
SDE Matching	577.5 ± 7.0	—	—	9017.5 ± 252.2	15580 ± 421
SVISE	298.7 ± 3.2	—	—	4178.6 ± 106.0	13900 ± 336

(c) Two-dimensional cylinder

Table 4: Per-iteration runtime and the time and number of iterations required to reach the indicated nELBO thresholds. Errors are $\pm$ 2 standard errors over 10 replicates. A dash indicates that the threshold was not reached by every replicate within the training budget.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our paper tackles the problem of simulation-free variational inference (VI) for latent stochastic differential equations. We describe our theoretical and experimental contributions in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 7, we discuss the limitations of our manuscript and indicate directions for future research.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Our paper theoretically studies the limitations of simulation-free VI algorithms for latent SDEs. Abbreviated theorem statements are provided in the main text (Sections 4.1 and 4.3). Full assumptions and proofs are provided in Appendices A to C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In Appendices C and D, we describe all details necessary to reproduce our experiments. This includes dataset availability, dataset preprocessing, model formulation, optimization algorithms, baselines, and evaluation metrics.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide a link to our project repository in the abstract. This repository contains code to reproduce experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All details regarding our training and test procedures are provided in Appendices C and D. Evaluation metrics are clearly described therein.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Across our experiments, we report uncertainty estimates whenever appropriate. The sources of variability captured by these estimates are described in figure/table captions (Figures 3 and 4) and in relevant experimental appendices (Appendices C and D).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss the hardware used for our experiments in Appendices D and E as well as runtime in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and verified the compliance of the paper.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The authors discuss the potential for latent SDEs to advance scientific discovery across various domains (e.g., neuroscience, ecology, physics). There are no clear negative impacts of the work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not include data or models that have a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: For our comparisons to baseline algorithms in Section 6, we adapt the code-bases from the relevant papers and provide attribution to the authors. For the two real-data examples described in Section 6, we use open-source datasets and provide attribution to the owners. See also Appendix D.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code for reproducing experiments is well-documented in the repository.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not require review by an institutional review board.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used to develop the core methods or write large sections of the paper. They were only used for polishing writing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.