GAUSSIAN FLOW DYNAMICS: SIMULATION-FREE NEURAL SDE LEARNING BEYOND ONE-TIME MARGINALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Simulation-free training of latent Stochastic Differential Equations (SDEs) relies on a variational posterior process whose one-time marginals are tractable, typically Gaussian. Such marginals, however, do not determine the underlying dynamics: many processes share the same marginals while differing in their temporal structure, and existing parameterizations fix this structure implicitly, which restricts the posterior family and biases the learned model. We introduce Gaussian flow dynamics, which construct stochastic processes directly from smoothly evolving Gaussian marginals while making the marginal-preserving, or gauge, degrees of freedom explicit and parameterizable. The construction admits state-dependent diffusion coefficients and recovers every linear SDE with additive noise and a non-degenerate Gaussian initial distribution. Building on it, we propose Gauge Matching, a simulation-free method for latent SDE learning that combines Gaussian flow dynamics with the SDE Matching objective. Gauge Matching costs at most quadratically in the latent dimension per step, like SDE Matching, but learns the temporal structure of the posterior beyond its one-time marginals. It comes within a nat of Helmholtz-SDE, which computes the gauge from the prior Jacobian at cubic cost, on the linear benchmark where the exact posterior is known, matches it on nonlinear systems, and applies where Helmholtz-SDE does not, to state-dependent noise.

1 INTRODUCTION

Neural or latent Stochastic Differential Equations (SDEs) model a time series as noisy observations of an unobserved state that evolves according to a learnable SDE (Movellan et al., 2002; Tzen & Raginsky, 2019; Hodgkinson et al., 2020; Li et al., 2020), a natural choice for irregularly sampled, inherently stochastic dynamics, from neuroscience to fluid dynamics (Course & Nair, 2023b; Smith et al., 2026). Training them requires the posterior over latent trajectories, which Variational Inference (VI) approximates with another SDE, the *posterior process*, whose drift depends on the data. Simulation-based methods integrate this process numerically at every gradient step (Li et al., 2020); simulation-free methods (Course & Nair, 2023a; Bartosh et al., 2025) instead parameterize its one-time marginals, which can be sampled at any time, so that the objective becomes an expectation over a random time and state and a gradient step costs the same for any length of the series.

The price of this efficiency is that the marginals underdetermine the process, and a method that parameterizes marginals still commits to one of the many processes that share them, implicitly. For Gaussian marginals, Course & Nair (2023a;b) take the affine drift whose linear part is symmetric, and SDE Matching (Bartosh et al., 2025) takes the drift induced by a particular square root of the covariance. Smith et al. (2026) recently showed that this choice, the *gauge*, is not benign: a fixed gauge restricts the variational family to a strict subset of the processes available to simulation-based methods, which degrades both inference and the learned dynamics when the posterior is uncertain. Their remedy, Helmholtz-SDE, keeps the marginals of SDE Matching and adds to its drift the marginal-preserving component of the difference between the prior and posterior drifts; the correction needs the Jacobian of the prior drift, costs at least cubically in the latent dimension, and requires state-independent noise.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We make the gauge part of the parameterization. We introduce *Gaussian flow dynamics*, which build a stochastic process from a prescribed path of Gaussian marginals. Factorizing the square root of the covariance into a Cholesky factor and an orthogonal matrix separates the marginals from the gauge, which enters the dynamics only through a skew-symmetric generator that can be parameterized directly, at only quadratic cost and with state-dependent noise. We prove that the family is complete: it recovers every linear SDE with state-independent noise and a non-degenerate Gaussian initial distribution, and every marginal-preserving affine field is generated by such a skew-symmetric matrix. Building on this, we propose *Gauge Matching*, which combines Gaussian flow dynamics with the SDE Matching objective, so that the gauge is selected by the objective rather than by the parameterization. We characterize the selected gauge, in closed form for state-independent noise, and show that Helmholtz-SDE computes it for a linearized prior process with state-independent diffusion. Experimentally, the learned gauge removes most of the gap of the fixed gauges to the true posterior process in the linear-Gaussian (conjugate) setting, where that gap is available in closed form, matches Helmholtz-SDE on nonlinear systems at lower cost, improves on it when the prior drift is known and its linearization is poor, and extends to state-dependent noise, where Helmholtz-SDE does not apply. Section 6 discusses related work.

2 Background

2.1 Dynamics and Their Marginals

Let z_t solve the Ordinary Differential Equation (ODE) $\mathrm{d}z_t = b(t, z_t)\mathrm{d}t$ on $[0, T]$ with *vector field* $b : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$. Its *flow map* sends an initial condition z_0 to the solution z_t at time t; conversely, a continuously differentiable flow map $\Phi : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ with $\Phi(t, \cdot)$ invertible for every t determines its vector field: the trajectories $z_t = \Phi(t, z_0)$ solve the ODE with

$$b(t, z) = \partial_t \Phi(t, z_0)\big|_{z_0 = \Phi^{-1}(t,z)}, \tag{1}$$

where the inverse is taken with respect to the second argument (Lee, 2012, Fundamental Theorem on Time-Dependent Flows, Theorem 9.48). If the initial condition is random with density $p(0, z)$, then under regularity assumptions the solution has a density $p(t, z)$ satisfying the continuity equation

$$\partial_t p(t, z) = -\nabla \cdot [b(t, z)p(t, z)], \tag{2}$$

where $\nabla \cdot v = \sum_{i=1}^{d} \partial_{z^i} v^i$ is the divergence of a vector field $v = (v^1, \ldots, v^d)^\top$. Adding noise gives a diffusion process Z_t, defined by the SDE $\mathrm{d}Z_t = b(t, Z_t)\mathrm{d}t + g(t, Z_t)\mathrm{d}W_t$ with $Z_0 \sim p(0, z)$, *drift coefficient* b, *diffusion coefficient* $g : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{d \times d'}$ and a d'-dimensional Brownian motion W_t. Its marginal densities satisfy the Fokker–Planck–Kolmogorov (FPK) equation

$$\partial_t p(t, z) = -\nabla \cdot [b(t, z)p(t, z)] + \frac{1}{2}\nabla \cdot \nabla \cdot [g(t, z)g(t, z)^\top p(t, z)], \tag{3}$$

where the divergence of a matrix-valued function is taken column-wise.

2.2 Simulation-Free Latent SDE Learning

Latent SDEs A latent SDE model explains observations $X = (x_{t_1}, \ldots, x_{t_N})$ at times $0 \le t_1 < \cdots < t_N \le T$ through an unobserved latent process: the *prior process* is a diffusion, and the observations are conditionally independent given the latent path, $x_{t_i} \sim p_\theta(x \mid z_{t_i})$. Since the marginal likelihood $p_\theta(X)$ is intractable, the model is trained with VI, using a *posterior process* whose drift coefficient depends on the observations but whose diffusion coefficient is shared with the prior process,

$$\text{prior:} \quad \mathrm{d}Z_t = h_\theta(t, Z_t)\mathrm{d}t + g_\theta(t, Z_t)\mathrm{d}W_t, \quad Z_0 \sim p_\theta(0, z), \tag{4}$$

$$\text{posterior:} \quad \mathrm{d}Z_t = f_\phi(t, Z_t, X)\mathrm{d}t + g_\theta(t, Z_t)\mathrm{d}W_t, \quad Z_0 \sim q_\phi(0, z \mid X). \tag{5}$$

Writing $q_\phi(t, z \mid X)$ for the marginal density of the posterior process, the negative Evidence Lower Bound (ELBO) is

$$-\log p_\theta(X) \le \mathcal{L} = D_{\mathrm{KL}}\left(q_\phi(0, z \mid X) \,\|\, p_\theta(0, z)\right) + \frac{1}{2}\int_0^T \mathbb{E}_{q_\phi(t,z|X)}\left[\|r_{\theta,\phi}(t, Z_t, X)\|_2^2\right] \mathrm{d}t$$

$$- \sum_{i=1}^{N} \mathbb{E}_{q_\phi(t_i,z|X)}\left[\log p_\theta(x_{t_i} \mid Z_{t_i})\right], \tag{6}$$

where the residual $r_{\theta,\phi}$ is defined by $g_\theta(t,z)r_{\theta,\phi}(t,z,X) = h_\theta(t,z) - f_\phi(t,z,X)$; for simplicity we assume throughout that $g_\theta(t,z)$ is square and invertible, so that $r_{\theta,\phi}$ is unique. The first two terms together are the Kullback–Leibler (KL) divergence between the path measures of the posterior and prior processes, given by Girsanov's theorem (Øksendal, 2003) and finite because the two processes share a diffusion coefficient. We call the second term the *path-space KL term*; it is the only term of $\mathcal{L}$ that depends on the posterior process through more than its marginals.

Simulation-Free Estimation If $q_\phi(t, z \mid X)$ can be sampled directly for any t, every term of eq. (6) is an expectation that can be estimated by sampling a time t uniformly on $[0,T]$ and a state $Z_t \sim q_\phi(t, z \mid X)$, without simulating the posterior process (Course & Nair, 2023a; Bartosh et al., 2025). SDE Matching (Bartosh et al., 2025) achieves this in three steps: (i) define the marginals through a map of noise, $Z_t = F_\phi(t, \varepsilon, X)$ with $\varepsilon \sim \mathcal{N}(0, I)$; (ii) read F_ϕ as a flow map with ε in the role of z_0, and obtain from eq. (1) the vector field $\bar{f}_\phi$ of the ODE that generates these marginals; (iii) convert the ODE into an SDE with the diffusion coefficient of the prior process, preserving the marginals (Song et al., 2021),

$$f_{\phi,\theta}(t,z,X) = \bar{f}_\phi(t,z,X) + \frac{1}{2}\Sigma_\theta(t,z)\nabla_z \log q_\phi(t,z \mid X) + \frac{1}{2}\nabla \cdot \Sigma_\theta(t,z), \tag{7}$$

where $\Sigma_\theta = g_\theta g_\theta^\top$; we write f_ϕ for short when the dependence on θ is not needed. Bartosh et al. (2025) take F_ϕ affine in ε, so that the posterior marginals are Gaussian. The marginals, however, fix the posterior process only up to a marginal-preserving vector field, so this construction implicitly selects one compatible drift among many; making the selection explicit is the subject of the next section.

3 Gaussian Flow Dynamics

Gaussian flow dynamics are deterministic and stochastic dynamics constructed directly from a time-dependent affine transformation of a standard Gaussian, so that their marginals are the Gaussian laws it defines; the flow-map perspective exposes the degrees of freedom that are invisible from the marginals alone. Throughout this section we write q for the marginal densities, anticipating their role as posterior marginals in section 4, and suppress observations and parameters.

Gaussian Flow We define the *Gaussian flow map* Ψ as an affine transformation of Gaussian noise,

$$\Psi(t,\varepsilon) = m(t) + S(t)\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \tag{8}$$

for continuously differentiable $m : [0,T] \to \mathbb{R}^d$ and invertible, continuously differentiable $S : [0,T] \to \mathbb{R}^{d\times d}$. The random variable $Z_t = \Psi(t,\varepsilon)$ has density

$$q(t,z) = \mathcal{N}(z \mid m(t), S(t)S(t)^\top), \tag{9}$$

and by eq. (1) its trajectories solve the ODE

$$\mathrm{d}Z_t = \bar{f}(t, Z_t)\mathrm{d}t, \qquad \bar{f}(t,z) = \dot{m}(t) + \dot{S}(t)S(t)^{-1}(z - m(t)), \tag{10}$$

with $Z_0 \sim \mathcal{N}(m(0), S(0)S(0)^\top)$, which we call the *Gaussian flow*. Its vector field is affine in z, and its marginals eq. (9) solve the continuity equation eq. (2) for this vector field (appendix A.1).

Gaussian Flow Diffusion We define the *Gaussian flow diffusion* as the SDE induced by Ψ and a diffusion coefficient $g(t,z)$ via the conversion eq. (7),

$$\mathrm{d}Z_t = f(t, Z_t)\mathrm{d}t + g(t, Z_t)\mathrm{d}W_t, \quad Z_0 \sim \mathcal{N}(m(0), S(0)S(0)^\top), \tag{11}$$

$$f(t,z) = \dot{m}(t) + \left(\dot{S}(t)S(t)^{-1} - \frac{1}{2}\Sigma(t,z)S(t)^{-\top}S(t)^{-1}\right)(z - m(t)) + \frac{1}{2}\nabla \cdot \Sigma(t,z),$$

with $\Sigma(t,z) = g(t,z)g(t,z)^\top$. Its marginals are again the Gaussians eq. (9), which solve the FPK equation eq. (3) for this drift and diffusion coefficient (appendix A.2); it is the posterior process that SDE Matching associates with the affine map $F = \Psi$. When $\Sigma(t,z) \equiv \Sigma(t)$ is state-independent, the drift coefficient is affine in z and the process is Gaussian; for a state-dependent diffusion coefficient its joint distributions are in general not Gaussian, although its marginals are.

Representation of Linear SDEs Conversely, every linear SDE with state-independent noise,

$$\mathrm{d}Z_t = (A(t)Z_t + u(t))\,\mathrm{d}t + g(t)\mathrm{d}W_t, \qquad Z_0 \sim \mathcal{N}(m_0, S_0 S_0^\top), \tag{12}$$

is a Gaussian flow diffusion. We write $\mathrm{GL}(d)$ for the group of invertible $d \times d$ matrices.

Theorem 1 (Representation of Linear SDEs)**.** *Let A, u, g in eq.* (12) *be continuous on $[0,T]$ and $S_0 \in \mathrm{GL}(d)$. Then there exist unique continuously differentiable $m : [0,T] \to \mathbb{R}^d$ and $S : [0,T] \to \mathrm{GL}(d)$ with $m(0) = m_0$ and $S(0) = S_0$ such that the Gaussian flow diffusion eq.* (11) *with flow map $m(t) + S(t)\varepsilon$ and diffusion coefficient g coincides with eq.* (12) *(proof in appendix A.3).*

Replacing S_0 by another square root $S_0 R$ of the initial covariance only relabels the trajectories, since $\Psi(t, R\varepsilon)$ defines the same process; up to this relabeling, the representation is unique.

Marginals and Gauge By Theorem 1, distinct linear SDEs with the same marginal covariance path are represented by distinct square roots $S(t)$: the square root carries more information than the covariance. To make this explicit, we factorize

$$S(t) = L(t)O(t), \tag{13}$$

where $L(t)$ is the Cholesky factor of the covariance, lower triangular with positive diagonal, and $O(t)$ is orthogonal; both inherit continuous differentiability from $S(t)$. The pair (m, L) determines the marginals, while O leaves them invariant. To see how O enters the dynamics, define $\Omega(t) = \dot{O}(t)O(t)^\top$, which is skew-symmetric because $OO^\top \equiv I$; we call it the *gauge generator*. Then $\dot{S}S^{-1} = \dot{L}L^{-1} + L\Omega L^{-1}$, and since the drift coefficients in eqs. (10) and (11) depend on S only through $\dot{S}S^{-1}$ and $SS^\top = LL^\top$, the orthogonal factor enters the dynamics only through the affine vector field

$$v(t,z) = L(t)\Omega(t)L(t)^{-1}(z - m(t)), \tag{14}$$

added to the drift coefficient of the process with the same marginals and $O \equiv I$. Since the FPK equation eq. (3) depends on the drift coefficient only through $\nabla \cdot [f(t,z)q(t,z)]$, adding a field v to the drift coefficient leaves the marginals unchanged if and only if v is *q-divergence-free*, $\nabla \cdot [v(t,z)q(t,z)] = 0$ (Smith et al., 2026); the fields eq. (14) are exactly the affine ones.

Proposition 2 (Affine Gauge Fields)**.** *Let $q(t,z) = \mathcal{N}(z \mid m(t), L(t)L(t)^\top)$ with $L(t)$ invertible. An affine vector field $v(t,z) = M(t)z + c(t)$ is q-divergence-free if and only if $v(t,z) = L(t)\Omega(t)L(t)^{-1}(z - m(t))$ for a skew-symmetric $\Omega(t)$, in which case $\Omega(t) = L(t)^{-1}M(t)L(t)$ (proof in appendix A.4; the Lyapunov form of this statement is Proposition 3 of Smith et al., 2026).*

Consequently, for fixed marginals eq. (9) and diffusion coefficient g, letting O range over all continuously differentiable orthogonal functions of time produces exactly the SDEs with these marginals and this g whose drift differs from that of $O \equiv I$ by an affine field; for state-independent g, these are all the Gaussian processes with these marginals and this g (Theorem 1). Following Smith et al. (2026), we call Ω the *gauge*. Existing methods commit to a particular Ω; we call their choices *fixed gauges*: the *square root gauge* of SDE Matching (Bartosh et al., 2025), the *symmetric gauge* of Course & Nair (2023a), and the simplest, $O \equiv I$, that is, $\Omega \equiv 0$, which we call the *Cholesky gauge* (appendix B).

Parameterization The dynamics depend on O only through Ω, and sampling from the marginals requires only m and L, so the orthogonal factor never has to be formed: we parameterize Gaussian flow dynamics directly by $m(t)$, $L(t)$ and $\Omega(t)$, with neural networks that output $d^2 + d$ numbers, the entries of m, the nonzero entries of L with a positive transform on its diagonal, and the entries of Ω above its diagonal. In the whitened state $\eta = L(t)^{-1}(z - m(t))$, the drift coefficient eq. (11) reads

$$f(t,z) = \dot{m}(t) + \dot{L}(t)\eta + L(t)\Omega(t)\eta - \frac{1}{2}\Sigma(t,z)L(t)^{-\top}\eta + \frac{1}{2}\nabla \cdot \Sigma(t,z). \tag{15}$$

Its evaluation requires $\dot{m}$ and $\dot{L}$ (from forward-mode differentiation), two triangular solves and a few matrix–vector products, $\mathcal{O}(d^2)$ in total, also for state-dependent diffusion coefficients whenever $\Sigma(t,z)w$ and $\nabla \cdot \Sigma(t,z)$ cost $\mathcal{O}(d^2)$, as for the diagonal structures common in latent SDE models (Li et al., 2020; Bartosh et al., 2025). Alternatively, $O(t)$ itself can be parameterized, for instance as a Householder flow (Tomczak & Welling, 2016); see appendix B.

4 GAUGE MATCHING

Gauge Matching is SDE Matching with a Gaussian flow diffusion as posterior process, so that its marginals and its gauge are both learned, while the objective eq. (6) and its simulation-free estimator are unchanged. After specifying the method, we characterize the gauge that the objective selects and relate it to the correction of Smith et al. (2026).

Posterior Process and Training Given observations X, the posterior process is the Gaussian flow diffusion eq. (11) with the diffusion coefficient g_θ of the prior process, parameterized as in section 3 by a mean $m_\phi(t, X)$, a Cholesky factor $L_\phi(t, X)$ and a gauge generator $\Omega_\phi(t, X)$, neural networks of t and a representation of X such as the output of a sequence model (Bartosh et al., 2025); its marginals are $q_\phi(t, z \mid X) = \mathcal{N}(z \mid m_\phi, L_\phi L_\phi^\top)$ and its drift coefficient is eq. (15) with $\Sigma_\theta = g_\theta g_\theta^\top$. SDE Matching is the member of this family with diagonal L_ϕ, for which its square root gauge has $\Omega_\phi \equiv 0$; the fixed gauges of appendix B prescribe Ω_ϕ as a function of L_ϕ, $\dot{L}_\phi$ and Σ_θ; Gauge Matching leaves Ω_ϕ free. Training minimizes eq. (6) jointly over θ and ϕ with the simulation-free estimator of section 2.2: the path-space KL term is estimated at a single time $t \sim \mathcal{U}[0, T]$ and the reconstruction term at a single index $i \sim \mathcal{U}\{1, \dots, N\}$ (algorithm 1 in appendix D). Sampling $Z_t = m_\phi + L_\phi \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$ makes the whitened state in eq. (15) simply $\eta = \varepsilon$, and an iteration costs the network evaluations, one forward-mode derivative of m_ϕ and L_ϕ in t, and $\mathcal{O}(d^2)$ linear algebra, whatever the number of observations and the length of the time interval.

The Optimal Gauge The gauge enters eq. (6) only through the path-space KL term; this term alone therefore selects the posterior process among those compatible with the learned marginals. To see which, fix the prior process and the marginals. The integrand at time t depends on the gauge only through $\Omega_\phi(t, X)$, so the minimization decouples across time and we study it at a fixed t, suppressing observations and parameters as in section 3. Let f_0 denote the drift coefficient of the Cholesky gauge; by eq. (14), the posterior process with gauge Ω has drift coefficient $f_0 + L\Omega L^{-1}(z - m)$, and with $\|v\|^2_{\Sigma^{-1}} = v^\top \Sigma^{-1} v$ the squared residual in eq. (6) is $\|r\|_2^2 = \|h - f\|^2_{\Sigma^{-1}}$, a quadratic function of Ω whose minimizer can be found explicitly.

Proposition 3 (Optimal Gauge)**.** *Fix $t \in [0, T]$ and m, L, $\dot{m}$, $\dot{L}$ at t. Let $h(t, \cdot)$ and $g(t, \cdot)$ be continuously differentiable, with h, its Jacobian $\nabla_z h$ (entries $\partial h_i / \partial z_j$), $\Sigma = gg^\top$ and $\nabla \cdot \Sigma$ of at most polynomial growth and Σ uniformly positive definite. The integrand of the path-space KL term,*

$$\mathcal{J}(\Omega) = \frac{1}{2} \mathbb{E}_{q(t,z)} \left[\left\| h(t, Z_t) - f_0(t, Z_t) - L\Omega L^{-1}(Z_t - m) \right\|^2_{\Sigma(t, Z_t)^{-1}} \right], \tag{16}$$

is a strictly convex quadratic function of the skew-symmetric matrix Ω and has a unique minimizer $\Omega^\star$, the optimal gauge*. If Σ is state-independent, $\Omega^\star$ is the unique solution of*

$$G\Omega^\star + \Omega^\star G = L^\top \left(\Sigma^{-1} \bar{J} - \bar{J}^\top \Sigma^{-1} \right) L, \quad G = L^\top \Sigma^{-1} L, \quad \bar{J} = \mathbb{E}_{q(t,z)} \left[\nabla_z h(t, Z_t) \right] - \dot{L} L^{-1}. \tag{17}$$

The proof is in appendix C.1. To interpret eq. (17), note that $\bar{J}$ is, up to a term that cancels on the right-hand side, the Jacobian of the residual $h - f_0$ averaged over the posterior marginal, and that the right-hand side is the skew-symmetric part of $\Sigma^{-1}\bar{J}$ transformed by L: the objective selects the gauge that reproduces the rotational component of the prior dynamics as seen from the posterior marginals, which the marginals themselves cannot represent. For the Ornstein–Uhlenbeck spiral of section 5.1, the stationary linear SDE with drift $(-\alpha I + \omega J)z$ and $J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, eq. (17) gives $\Omega^\star = \omega J$, for which the posterior process coincides with the prior process. The same computation tells us when a fixed gauge is already optimal (appendix C.2): the Cholesky gauge if and only if $\Sigma^{-1}\bar{J}$ is symmetric, and, for isotropic Σ, the symmetric gauge if and only if $\mathbb{E}_{q(t,z)}[\nabla_z h(t, Z_t)]$ is symmetric, as for a gradient prior drift; neither holds for prior dynamics with a rotational component.

Relation to Helmholtz-SDE Helmholtz-SDE (Smith et al., 2026) starts from a fixed-gauge posterior process with a state-independent diffusion coefficient and adds to its drift coefficient the q-divergence-free component of the residual $h - f$, which minimizes the path-space KL term over all marginal-preserving fields; to make it computable, the residual is replaced by its Taylor expansion about the mean $m(t)$. At order $\ell = 1$, used in all their experiments, the correction is affine, hence a gauge field by Proposition 2, and the corrected gauge is, irrespective of the fixed gauge it starts

Table 1: Posterior processes with Gaussian marginals in simulation-free latent SDE learning. Cost is that of evaluating the drift coefficient at one (t, z) beyond the network evaluations, for a full $d \times d$ covariance (SDE Matching: diagonal or triangular S). The Cholesky gauge ($\Omega \equiv 0$) is Gauge Matching without the gauge.

Method	Gauge	Cost	State-dependent g
Course & Nair (2023a;b)	symmetric, fixed	$\mathcal{O}(d^3)$	no
SDE Matching (Bartosh et al., 2025)	square root, fixed	$\mathcal{O}(d^2)$	yes
Cholesky gauge (ablation)	$\Omega \equiv 0$, fixed	$\mathcal{O}(d^2)$	yes
Helmholtz-SDE (Smith et al., 2026)	order-one correction, computed	$\mathcal{O}(d^3)$	no
Gauge Matching (ours)	learned	$\mathcal{O}(d^2)$	yes

from, the solution of eq. (17) with the Jacobian at the mean, $\nabla_z h(t, m)$, in place of its expectation (appendix C.2). The two gauges coincide whenever $\nabla_z h$ is affine in z, as for the quadratic Lorenz drift, and differ otherwise: Helmholtz-SDE finds the optimal gauge of the linearized prior process, whereas Gauge Matching minimizes the exact objective over the same family; only corrections of order $\ell \geq 2$ leave this family. The larger difference is computational. Solving eq. (17) at every evaluation requires the Jacobian of the prior drift network and $\mathcal{O}(d^3)$ linear algebra, and for a state-dependent diffusion coefficient it becomes a linear system whose coefficients must be estimated numerically (appendix C.2). Gauge Matching learns Ω_ϕ by stochastic gradient descent, at $\mathcal{O}(d^2)$, without derivatives of the prior drift coefficient and for state-dependent diffusion coefficients (table 1); since $\mathcal{J}$ is strictly convex in Ω, this adds no local optima at fixed marginals.

Exactness in the Conjugate Setting When the prior process is linear and the observation model is linear-Gaussian, the true posterior process is a Gaussian process, the Kalman–Bucy smoother (Särkkä & Solin, 2019), which the square root and symmetric gauges cannot represent (Smith et al., 2026). The family with a learned gauge can.

Corollary 4 (Conjugate Setting)**.** *Let the prior process be the linear SDE eq.* (12) *with $S_0 \in \mathrm{GL}(d)$ and $\Sigma(t)$ positive definite, and let $p_\theta(x \mid z) = \mathcal{N}(x \mid Cz, R)$ with R positive definite. Then (i) between consecutive observation times, the true posterior process is a Gaussian flow diffusion with diffusion coefficient g, continuously differentiable m, L and continuous Ω; and (ii) the infimum of the KL divergence from Gauge Matching posterior processes to the true one is zero (proof in appendix C.3).*

At each observation time, m and L of the true posterior process have kinks and Ω jumps, whereas a neural network in t is continuously differentiable; the infimum is therefore approached rather than attained, as for Helmholtz-SDE (Smith et al., 2026, Theorem 1), with the discrepancy confined to neighborhoods of the observation times.

5 Experiments

We ask whether gradient descent on the path-space KL term finds the optimal gauge of Proposition 3; whether the learned gauge closes the gap of the fixed gauges in the conjugate setting and how it compares with the exact gauge of Helmholtz-SDE there; whether it matches Helmholtz-SDE on nonlinear systems and extends beyond its scope; and how it scales with d. All posterior processes share the SDE Matching architecture (Bartosh et al., 2025) and differ only in L_ϕ and the gauge: the square root gauge with a diagonal L_ϕ, which is SDE Matching; the Cholesky ($\Omega_\phi \equiv 0$) and symmetric gauges with a full L_ϕ; Helmholtz-SDE at order one, which we reimplement; and the learned gauge of Gauge Matching (appendix D). We call each independent observation sequence X a *trial*.

5.1 Linear Dynamics

Learning the Gauge We take as prior process the two-dimensional linear SDE eq. (12) with $A = -\alpha I + \omega J$, $\Sigma = I$ and the stationary initial distribution $\mathcal{N}(0, I/(2\alpha))$, the Ornstein–Uhlenbeck spiral of Smith et al. (2026). Its drift is a contracting gradient field plus a rotation that preserves the

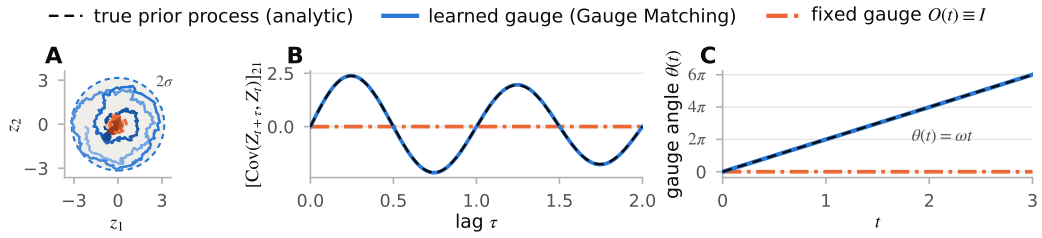

Figure 1: Learning the gauge of the stationary Ornstein–Uhlenbeck spiral ($\alpha = 0.2, \omega = 2\pi$) without observations. *A:* one sample path of each fitted model, learned gauge (blue) and fixed gauge $O(t) \equiv I$ (orange), over the true 2σ disk (shaded) and each model's own 2σ circle (dashed): the fixed gauge cannot circulate and shrinks the marginals instead. *B:* lagged cross-covariance $[\mathrm{Cov}(Z_{t+\tau}, Z_t)]_{21}$ of the true and the two fitted processes. *C:* rotation angle $\theta(t)$ of the learned $O(t)$ against ωt.

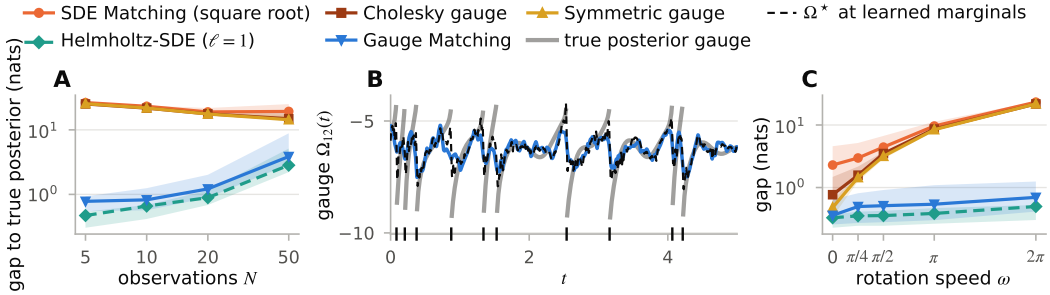

Figure 2: Gauge Matching in the conjugate setting ($\sigma = 0.3$; median and interquartile range over 16 trials). *A:* KL divergence to the true posterior process against the number of observations N at $\omega = 2\pi$; the learned gauge is 4–34 times below the fixed gauges and within a nat of Helmholtz-SDE, whose gauge is exact here. *B:* learned gauge $\Omega_{12}(t)$ for one trial with $N = 10$, against the optimal gauge eq. (17) at the learned marginals and the gauge of the true posterior process, which jumps at the observation times (ticks). *C:* KL divergence against the rotation speed ω at $N = 10$.

stationary law, so the marginals carry no information about ω: the temporal structure lies in the gauge alone. Without observations, the posterior process should recover the prior process. We fit a Gaussian flow diffusion by minimizing the first two terms of eq. (6), with $O(t)$ a Householder flow whose rotation angle $\theta(t)$ can be read off. The learned gauge recovers the prior process: the marginals converge to the stationary law, $\theta(t)$ to ωt, and the path-space KL term drops to 0.007 nats over $[0, 3]$ (fig. 1). The fixed gauge $O(t) \equiv I$, with which all three fixed gauges coincide here (appendix B), cannot represent the rotation and distorts the marginals to compensate, shrinking the variance from the true 2.5 to 0.080 while still incurring about 18 nats.

Conjugate Setting We next add observations $\mathcal{N}(x \mid Cz + c, \sigma^2 I)$ with a random matrix C and offset c at N random times in $[0, T]$, with $T = 5$ and $\alpha = 0.2$, and fix the generative model at the truth. Since all drift coefficients are affine, the KL divergence of each fitted posterior process to the true posterior process, which we call the *gap*, is available in closed form (appendix D). For $\omega = 2\pi$ and $\sigma = 0.3$ (fig. 2A), the three fixed gauges incur a gap of 14–26 nats, within about 2 nats of the best their variational families can do; the Cholesky gauge improves little on the square root gauge, so it is the gauge, not the covariance, that is missing. Gauge Matching reduces the gap to 0.8–3.9 nats. Helmholtz-SDE, whose gauge is exact here and costs nothing to compute (the prior Jacobian is the constant A), reaches 0.5–2.8 nats with the same network and budget: knowing the optimal gauge is worth 0.2–1 nat over learning it, and what remains for both is the error of a smooth posterior at the kinks of the true one (Corollary 4). The learned $\Omega_\phi(t)$ tracks the optimal gauge eq. (17) at the learned marginals and, up to the jumps, the gauge of the true posterior process (fig. 2B); as ω grows

Table 2: Noisy Lorenz attractor (Smith et al., 2026), $d = 4$; mean ± 2 standard errors over 5 seeds on held-out trials; arrows give the better direction. *Negative ELBO*: per trial, nats. *Latent LL*: log-density of the true state under the posterior predictive. *Autocorr. error*: distance between the lagged correlation matrices of the learned and the true prior process. *Gauge excess*: $\mathcal{J}(\Omega) - \mathcal{J}(\Omega^\star)$ of eq. (16) at the method's own gauge and marginals, nats per unit time (appendix D).

Posterior process	Negative ELBO ↓	Latent LL ↑	Autocorr. error ↓	Gauge excess ↓
SDE Matching (square root)	77.9 ± 0.6	-0.22 ± 0.01	0.134 ± 0.009	2.84 ± 0.08
Cholesky gauge	75.5 ± 0.8	-0.20 ± 0.01	0.111 ± 0.007	2.52 ± 0.09
Symmetric gauge	76.9 ± 0.8	-0.22 ± 0.01	0.111 ± 0.006	2.61 ± 0.11
Helmholtz-SDE ($\ell = 1$)	53.0 ± 0.6	-0.06 ± 0.01	0.052 ± 0.006	0.001 ± 0.000
Gauge Matching	53.2 ± 0.7	-0.06 ± 0.01	0.056 ± 0.009	0.044 ± 0.003
Gauge Matching, diagonal L_ϕ	55.3 ± 1.9	-0.10 ± 0.02	0.064 ± 0.006	0.022 ± 0.001

(fig. 2C), the gap of the fixed gauges rises from 0.6–2.3 to 22–23 nats, linearly in ωT (Smith et al., 2026, Theorem 4), while that of the learned gauge stays below 0.8.

5.2 Nonlinear Dynamics

Noisy Lorenz Attractor We follow Smith et al. (2026): the stochastic Lorenz system on $[0, 5]$, with process noise large enough that the posterior does not collapse onto the true trajectory, is observed every $\Delta t = 0.25$ through a noisy linear map, for 1024 training trials; the latent SDE has $d = 4$, a neural prior drift and a fixed diffusion coefficient, and the prior and observation model are learned jointly with the posterior process. In table 2, the fixed gauges reach a negative ELBO of 75–78 nats per trial, whether with diagonal or with full covariance. Gauge Matching and Helmholtz-SDE reach 53 nats; the difference lies in the path-space KL term, and their marginals assign a higher likelihood to the true latent state. The learned dynamics improve too: the autocorrelation error of the prior process drops from 0.11–0.13 to 0.05–0.06, and only these two models reproduce the outward spiral within each lobe of the attractor (fig. 8). The gauge excess confirms that gradient descent finds the optimal gauge, to within 0.05 nats per unit time against 2.5–2.8 for the fixed gauges, and the last row shows that the gain is the gauge, not the covariance: a diagonal L_ϕ with a learned gauge reaches 55.3 nats, a full L_ϕ without one stays at 75.5. The picture holds across observation spacings and noise levels (appendix D), and the agreement of Gauge Matching and Helmholtz-SDE is what section 4 predicts: the fitted drift networks are nearly affine over the posterior support.

Known Nonlinear Drift Gauge Matching and Helmholtz-SDE differ when the prior drift is known and strongly nonlinear: on a Stuart–Landau oscillator with shear and the generative model fixed at the truth (table 4 in appendix D), the gauge excess of the learned gauge is 10 times smaller than that of the order-one correction, and its negative ELBO is lower by 1.6 ± 0.3 nats, on every seed.

State-Dependent Noise The stochastic Lotka–Volterra system (Vadillo, 2019; Bartosh et al., 2025) has multiplicative noise, $g(z) = \sigma \operatorname{diag}(z)$, which Helmholtz-SDE does not support; in log coordinates the noise becomes additive and the drift acquires exponentials. We learn a latent SDE from sparse, noisy observations in each coordinate system, with the true diffusion coefficient in log coordinates and a learned state-dependent one in the original ones, and evaluate the learned prior process by the period of its cycle, the predator's lag behind the prey and the autocorrelation error (table 5 in appendix D). In log coordinates the posterior is narrow relative to the curvature of the drift, so the order-one correction is already optimal: Helmholtz-SDE and Gauge Matching agree, as section 4 predicts, with gauge excess 0.000 and 0.004 against 0.215 for the fixed gauge. In the original coordinates, where Helmholtz-SDE does not apply, Gauge Matching reduces the period error from 22% to 13% and improves all three metrics on every seed (fig. 10). The same holds in $d = 8$: on a rotating Ornstein–Uhlenbeck process with multiplicative noise $g(z) = \sigma \operatorname{diag}(\sqrt{1 + z_i^2})$, Gauge Matching lowers the negative ELBO of the fixed gauges by 6.6 ± 0.8 nats with the generative model known and 4.8 ± 0.3 with a learned prior, at an 18–29 times smaller gauge excess (table 6).

Motion Capture On the motion capture benchmark of Gan et al. (2015), 50-dimensional recordings with 300 dense, nearly noise-free observations each, Gauge Matching with $d = 6$ reaches a test

mean squared error of 4.52 ± 0.29, on par with SDE Matching (4.50 ± 0.32) and, like it, behind the simulation-based latent SDE of Li et al. (2020) (4.03 ± 0.20): the posterior concentrates on the dense observations, so the gauge has little to add, and nothing is lost.

Scaling in the Latent Dimension The learned gauge adds $d(d-1)/2$ network outputs, 2016 at $d = 64$; Helmholtz-SDE pays d forward-mode passes through the prior drift network and an $\mathcal{O}(d^3)$ solve of eq. (17). On a rotating Ornstein–Uhlenbeck process in $d \in \{2, 4, \ldots, 64\}$ dimensions with the generative model known, the gap of the learned gauge stays within a factor of two of that of Helmholtz-SDE, which is exact here, and 4–27 times below the fixed gauges at every d: gradient descent finds the optimal gauge in 2016 dimensions (fig. 5 in appendix D). With a learned generative model and the state fully observed, the fixed gauges are 3–9 nats worse than Helmholtz-SDE at $d = 8$ and, for a full L_ϕ, at every $d \leq 32$ on every seed, while Gauge Matching recovers most of the difference. Per training iteration on a GPU, Gauge Matching costs 1.05–1.5 times as much as SDE Matching for $d \leq 128$ and Helmholtz-SDE 1.3–3.2 times, the Jacobian and the cubic solve dominating at large d; end to end, Helmholtz-SDE trained 1.2 times longer than Gauge Matching on the Lorenz system and 1.6 times on the Stuart–Landau oscillator.

6 Related Work

Variational Inference for Latent SDEs Simulation-based VI parameterizes the posterior process by its drift coefficient and optimizes the gauge implicitly, at the cost of one numerical integration per gradient step: with an affine drift the marginals follow from moment ODEs (Archambeau et al., 2007a;b; Duncker et al., 2019; Wildner & Koeppl, 2021; Köhs et al., 2021; Verma et al., 2024; Hu et al., 2025; Park et al., 2025); with a neural drift the process is simulated and differentiated through (Movellan et al., 2002; Ryder et al., 2018; Tzen & Raginsky, 2019; Hodgkinson et al., 2020; Li et al., 2020; Kidger et al., 2021), as in latent ODEs (Chen et al., 2018; Rubanova et al., 2019).

Simulation-Free Methods and the Gauge Diffusions with prescribed marginals go back to Brigo (2000; 2002) and were used for variational path estimation by Sutter et al. (2016). Simulation-free VI with Gaussian marginals (Course & Nair, 2023a;b; Bartosh et al., 2025) fixes the gauge (appendix B); Kiyohara et al. (2025) use a mean-field family, and Neklyudov et al. (2023); Zhang et al. (2024) learn simulation-free dynamics in observation space. Smith et al. (2026) identified a post-hoc gauge correction (section 4) and briefly mention learning Ω in their Appendix A.5 without developing or evaluating it; they suggest this sacrifices SDE Matching's closed-form structure, whereas our $O(t)$ construction retains its explicit closed-form flow map.

Marginal-Preserving Fields The same freedom appears in generative modeling, where the probability-flow ODE (Song et al., 2021) is one of many transports between the marginals of a diffusion model: Horvat & Pfister (2024) analyze this freedom, Petrović et al. (2025) exploit it in flow matching (Lipman et al., 2023) to learn non-gradient dynamics, and Hua et al. (2025) parameterize a density and a dynamics jointly so that the continuity equation holds by construction. Helmholtz–Hodge decompositions also structure Gaussian-process models of vector fields (Berlinghieri et al., 2023; Xu & Pan, 2024). In these works the divergence-free component is a modeling choice; here it is a coordinate of the variational family that the marginals leave free. Diffusion models with learned noising processes (Singhal et al., 2023; Bartosh et al., 2024a; Nielsen et al., 2024; Sahoo et al., 2024; Bartosh et al., 2024b; Du et al., 2024) are latent SDEs with one observation and a learned posterior.

7 Conclusion

We introduced Gaussian flow dynamics, which turn the marginal-preserving freedom of a process with Gaussian marginals into an explicit skew-symmetric generator, at quadratic cost and for state-dependent diffusion coefficients, and Gauge Matching, which lets the SDE Matching objective select this generator. The learned gauge removes most of the gap of the fixed gauges in the conjugate setting, matches Helmholtz-SDE on nonlinear systems at lower cost and up to $d = 64$, improves on it when the prior drift is known and strongly nonlinear, and extends to state-dependent noise, where Helmholtz-SDE does not apply. Nonlinear marginal-preserving fields (Hua et al., 2025) and non-Gaussian marginals are left for future work.

AI USE STATEMENT

Large language models (Claude, Anthropic) were used in the preparation of this paper as a writing and coding assistant under the direction of the authors: for drafting and editing prose, for writing analysis and plotting code, and for checking derivations numerically. All mathematical results, experimental designs and conclusions were specified and verified by the authors, who take full responsibility for the content of the paper.

ETHICS STATEMENT

This is methodological work on variational inference for latent SDEs. Its experiments use synthetic data generated from known dynamical systems and one public benchmark, the CMU motion capture recordings of Gan et al. (2015) with the preprocessing of Wang et al. (2008); no data were collected from human subjects for this work. We foresee no direct negative societal impact beyond the general dual use of time-series models.

REPRODUCIBILITY STATEMENT

All theoretical results are proved in the appendix: appendix A for section 3, appendix B for the gauges of existing methods, and appendix C for section 4. Appendix D specifies the architectures, hyperparameters, data generation and metric definitions of every experiment. Code for all experiments, including the baselines and our implementation of Helmholtz-SDE, together with the data generation scripts, will be published upon acceptance.

The remaining items of the required disclosure, such as the use of human subjects, personal data or proprietary datasets, do not apply to this work.

REFERENCES

Cédric Archambeau, Dan Cornford, Manfred Opper, and John Shawe-Taylor. Gaussian process approximations of stochastic differential equations. In *Gaussian Processes in Practice*, pp. 1–16. PMLR, 2007a.

Cédric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, and John Shawe-Taylor. Variational inference for diffusion processes. In *Advances in Neural Information Processing Systems*, volume 20, 2007b.

Grigory Bartosh, Dmitry Vetrov, and Christian A. Naesseth. Neural diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3073–3095. PMLR, 2024a.

Grigory Bartosh, Dmitry Vetrov, and Christian A Naesseth. Neural flow diffusion models: Learnable forward process for improved diffusion modelling. *Advances in Neural Information Processing Systems*, 37:73952–73985, 2024b.

Grigory Bartosh, Dmitry Vetrov, and Christian A. Naesseth. SDE Matching: Scalable and simulation-free training of latent stochastic differential equations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3054–3070. PMLR, 2025.

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.

Damiano Brigo. On SDEs with marginal laws evolving in finite-dimensional exponential families. *Statistics & Probability Letters*, 49(2):127–134, 2000.

Damiano Brigo. The general mixture-diffusion SDE and its relationship with an uncertain-volatility option model with volatility-asset decorrelation. *arXiv preprint arXiv:0812.4052*, 2002. Working paper dated September 2002.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Kevin Course and Prasanth B Nair. Amortized reparametrization: efficient and scalable variational inference for latent SDEs. *Advances in Neural Information Processing Systems*, 36:78296–78318, 2023a.

Kevin Course and Prasanth B Nair. State estimation of a physical system with unknown governing equations. *Nature*, 622(7982):261–267, 2023b.

Yuanqi Du, Michael Plainer, Rob Brekelmans, Chenru Duan, Frank Noé, Carla P. Gomes, Alán Aspuru-Guzik, and Kirill Neklyudov. Doob's Lagrangian: A sample-efficient variational approach to transition path sampling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 65791–65822, 2024.

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.

Zhe Gan, Chunyuan Li, Ricardo Henao, David E Carlson, and Lawrence Carin. Deep temporal sigmoid belief networks for sequence modeling. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Larry C Grove. *Classical groups and geometric algebra*, volume 39 of *Graduate Studies in Mathematics*. American Mathematical Society, 2002.

Liam Hodgkinson, Chris van der Heide, Fred Roosta, and Michael W Mahoney. Stochastic normalizing flows. *arXiv preprint arXiv:2002.09547*, 2020.

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

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

Mengjian Hua, Eric Vanden-Eijnden, and Ricky T. Q. Chen. Simulation-free differential dynamics through neural conservation laws. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1730–1744. PMLR, 2025.

Patrick Kidger, James Foster, Xuechen Li, and Terry Lyons. Efficient and accurate gradients for neural SDEs. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18747–18761, 2021.

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

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

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

John M Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, 2 edition, 2012.

Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3870–3882. PMLR, 2020.

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

Javier R Movellan, Paul Mineiro, and Ruth J Williams. A Monte Carlo EM approach for partially observable diffusion processes: Theory and applications to neural networks. *Neural Computation*, 14(7):1507–1544, 2002.

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

Beatrix M. G. Nielsen, Anders Christensen, Andrea Dittadi, and Ole Winther. DiffEnc: Variational diffusion with a learned encoder. In *International Conference on Learning Representations*, 2024.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, Berlin, 6 edition, 2003.

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.

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. In *Advances in Neural Information Processing Systems*, 2025.

Yulia Rubanova, Ricky T. Q. Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tom Ryder, Andrew Golightly, A Stephen McGough, and Dennis Prangle. Black-box variational inference for stochastic differential equations. In *International Conference on Machine Learning*, pp. 4423–4432. PMLR, 2018.

Subham Sekhar Sahoo, Aaron Gokaslan, Christopher De Sa, and Volodymyr Kuleshov. Diffusion models with learned adaptive noise. In *Advances in Neural Information Processing Systems*, volume 37, pp. 105730–105779, 2024.

Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, Cambridge, 2019.

R. Singhal, M. Goldstein, and R. Ranganath. Where to diffuse, how to diffuse, and how to get back: Automated learning for multivariate diffusions. In *International Conference on Learning Representations*, 2023.

Henry D Smith, Brian L Trippe, and Scott W Linderman. Closing the approximation gap in simulation-free latent SDEs. *arXiv preprint arXiv:2606.16138*, 2026.

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.

Tobias Sutter, Arnab Ganguly, and Heinz Koeppl. A variational approach to path estimation and parameter inference of hidden diffusion processes. *Journal of Machine Learning Research*, 17 (190):1–37, 2016.

Jakub M Tomczak and Max Welling. Improving variational auto-encoders using Householder flow. *arXiv preprint arXiv:1611.09630*, 2016.

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

Fernando Vadillo. Comparing stochastic Lotka–Volterra predator-prey models. *Applied Mathematics and Computation*, 360:181–189, 2019.

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

Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298, 2008.

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

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

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

A Proofs for Section 3

This appendix proves that the Gaussian flow and the Gaussian flow diffusion have the marginals eq. (9) (Propositions 5 and 6), and then proves Theorem 1 and Proposition 2.

A.1 Gaussian Flow and the Continuity Equation

Proposition 5. *The Gaussian flow eq.* (10) *satisfies the continuity equation* $\partial_t q(t,z) = -\nabla \cdot [\bar{f}(t,z)q(t,z)]$ *for* $q(t,z) = \mathcal{N}(z \mid m(t), S(t)S(t)^\top)$.

Proof. Since $q > 0$, the continuity equation eq. (2) for q is equivalent to the equation obtained by dividing it by $q(t,z)$. Using the product rule for the divergence, $\nabla \cdot [\bar{f}q] = (\nabla \cdot \bar{f})q + \bar{f}^\top \nabla q$, it becomes

$$\partial_t \log q(t,z) + \nabla \cdot \bar{f}(t,z) + \bar{f}(t,z)^\top \nabla \log q(t,z) = 0. \tag{18}$$

We compute the three terms in turn and show that they sum to zero. To lighten the notation we drop the argument t from m, S and their derivatives, and write $P = SS^\top$ for the covariance, so that $P^{-1} = S^{-\top}S^{-1}$, and $\bar{z} = z - m$.

First term. Since $\log q(t,z) = -\frac{1}{2}\log\det P - \frac{1}{2}\bar{z}^\top P^{-1}\bar{z} + \text{const}$,

$$\partial_t \log q(t,z) = -\frac{1}{2}\operatorname{tr}\left(P^{-1}\dot{P}\right) + \dot{m}^\top P^{-1}\bar{z} - \frac{1}{2}\bar{z}^\top \partial_t\left(P^{-1}\right)\bar{z},$$

where we used the derivative of the log-determinant, $\partial_t \log\det P = \operatorname{tr}(P^{-1}\dot{P})$, the product rule for the quadratic form, and the symmetry of P^{-1}. The two remaining derivatives, $\operatorname{tr}(P^{-1}\dot{P})$ and $\partial_t(P^{-1})$, are, with $\dot{P} = \dot{S}S^\top + S\dot{S}^\top$,

$$\begin{aligned}\operatorname{tr}\left(P^{-1}\dot{P}\right) &= \operatorname{tr}\left(S^{-\top}S^{-1}\dot{S}S^\top\right) + \operatorname{tr}\left(S^{-\top}S^{-1}S\dot{S}^\top\right) = 2\operatorname{tr}\left(S^{-1}\dot{S}\right),\\ \partial_t\left(P^{-1}\right) &= -P^{-1}\dot{P}P^{-1} = -S^{-\top}S^{-1}\dot{S}S^{-1} - S^{-\top}\dot{S}^\top S^{-\top}S^{-1},\end{aligned}$$

where the first line uses the cyclic property of the trace and the second the derivative of a matrix inverse. Substituting, and using that the two terms of $-\frac{1}{2}\bar{z}^\top\partial_t(P^{-1})\bar{z}$ are transposes of each other and hence equal,

$$\partial_t \log q(t,z) = -\operatorname{tr}\left(S^{-1}\dot{S}\right) + \dot{m}^\top S^{-\top}S^{-1}\bar{z} + \bar{z}^\top S^{-\top}\dot{S}^\top S^{-\top}S^{-1}\bar{z}.$$

Second term. The vector field $\bar{f}(t,z) = \dot{m} + \dot{S}S^{-1}\bar{z}$ is affine in z, so its divergence is the trace of its linear part,

$$\nabla \cdot \bar{f}(t,z) = \operatorname{tr}\left(\dot{S}S^{-1}\right) = \operatorname{tr}\left(S^{-1}\dot{S}\right).$$

Third term. With $\nabla \log q(t,z) = -S^{-\top}S^{-1}\bar{z}$,

$$\bar{f}(t,z)^\top \nabla \log q(t,z) = -\dot{m}^\top S^{-\top}S^{-1}\bar{z} - \bar{z}^\top S^{-\top}\dot{S}^\top S^{-\top}S^{-1}\bar{z}.$$

Summing the three terms, the traces cancel, the terms linear in $\bar{z}$ cancel, and the quadratic forms cancel, so eq. (18) holds. □

A.2 Gaussian Flow Diffusion and the FPK Equation

Proposition 6. *The Gaussian flow diffusion eq.* (11) *satisfies the FPK equation* $\partial_t q(t,z) = -\nabla \cdot [f(t,z)q(t,z)] + \frac{1}{2}\nabla \cdot \nabla \cdot [g(t,z)g(t,z)^\top q(t,z)]$ *for* $q(t,z) = \mathcal{N}(z \mid m(t), S(t)S(t)^\top)$.

Proof. The idea is that the two terms by which the drift coefficient eq. (11) differs from the Gaussian flow eq. (10) are designed to cancel the diffusion term of the FPK equation, so that the equation reduces to the continuity equation of Proposition 5. Throughout, $\Sigma = gg^\top$.

Step 1: rewrite the drift coefficient. By the conversion eq. (7) that defines the Gaussian flow diffusion,

$$f(t,z) = \bar{f}(t,z) + \frac{1}{2}\Sigma(t,z)\nabla_z \log q(t,z) + \frac{1}{2}\nabla \cdot \Sigma(t,z),$$

and, since $\nabla_z \log q = \nabla_z q / q$, the last two terms combine into

$$\begin{aligned}\frac{1}{2}\Sigma(t,z)\frac{\nabla_z q(t,z)}{q(t,z)} + \frac{1}{2}\nabla \cdot \Sigma(t,z) &= \frac{1}{2q(t,z)}\left(\Sigma(t,z)\nabla_z q(t,z) + q(t,z)\nabla \cdot \Sigma(t,z)\right) \\ &= \frac{1}{2q(t,z)}\nabla \cdot [\Sigma(t,z)q(t,z)],\end{aligned}$$

where the last equality is the product rule for the divergence applied to each column of the symmetric matrix Σq. Hence

$$f(t,z)q(t,z) = \bar{f}(t,z)q(t,z) + \frac{1}{2}\nabla \cdot [\Sigma(t,z)q(t,z)].$$

Step 2: substitute into the FPK equation. Using the linearity of the divergence,

$$\begin{aligned}&-\nabla \cdot [f(t,z)q(t,z)] + \frac{1}{2}\nabla \cdot \nabla \cdot [\Sigma(t,z)q(t,z)] \\ &\quad = -\nabla \cdot [\bar{f}(t,z)q(t,z)] - \frac{1}{2}\nabla \cdot \nabla \cdot [\Sigma(t,z)q(t,z)] + \frac{1}{2}\nabla \cdot \nabla \cdot [\Sigma(t,z)q(t,z)] \\ &\quad = -\nabla \cdot [\bar{f}(t,z)q(t,z)].\end{aligned}$$

The FPK equation for f and g is therefore equivalent to the continuity equation for $\bar{f}$, which $q(t,z) = \mathcal{N}(z \mid m(t), S(t)S(t)^\top)$ satisfies by Proposition 5. □

A.3 Proof of Theorem 1

We show that for continuous A, u, g and $S_0 \in \mathrm{GL}(d)$ there are unique continuously differentiable m and S with $m(0) = m_0$, $S(0) = S_0$ such that the Gaussian flow diffusion with flow map $m + S\varepsilon$ and diffusion coefficient g is the linear SDE eq. (12).

Proof. For a state-independent diffusion coefficient $g(t)$, with $\Sigma(t) = g(t)g(t)^\top$, the drift coefficient eq. (11) of the Gaussian flow diffusion is

$$f(t,z) = \dot{m}(t) + \left(\dot{S}(t)S(t)^{-1} - \frac{1}{2}\Sigma(t)S(t)^{-\top}S(t)^{-1}\right)(z - m(t)).$$

It coincides with the drift coefficient $A(t)z + u(t)$ of eq. (12) for all z if and only if the linear parts and the constant parts agree,

$$\dot{S}(t)S(t)^{-1} - \frac{1}{2}\Sigma(t)S(t)^{-\top}S(t)^{-1} = A(t), \tag{19}$$

$$\dot{m}(t) - \left(\dot{S}(t)S(t)^{-1} - \frac{1}{2}\Sigma(t)S(t)^{-\top}S(t)^{-1}\right)m(t) = u(t). \tag{20}$$

The initial distributions agree by the choice $m(0) = m_0$, $S(0) = S_0$. It therefore remains to show that eqs. (19) and (20) have unique continuously differentiable solutions $S : [0,T] \to \mathrm{GL}(d)$ and $m : [0,T] \to \mathbb{R}^d$ with these initial values. We proceed in three steps: we construct S from the covariance of eq. (12) and an orthogonal factor, show that this S is the unique solution, and then solve for m.

Step 1: construction of S. Multiplied by $S(t)$ from the right, eq. (19) is equivalent to the matrix ODE

$$\dot{S}(t) = A(t)S(t) + \frac{1}{2}\Sigma(t)S(t)^{-\top}. \tag{21}$$

Let $P_0 = S_0 S_0^\top \succ 0$ and let $P(t)$ solve the differential Lyapunov equation

$$\dot{P}(t) = A(t)P(t) + P(t)A(t)^\top + \Sigma(t), \quad P(0) = P_0,$$

which is the covariance equation of the linear SDE eq. (12). For continuous A and $\Sigma \succeq 0$ it has the unique solution

$$P(t) = \Phi_A(t,0)P_0\Phi_A(t,0)^\top + \int_0^t \Phi_A(t,s)\Sigma(s)\Phi_A(t,s)^\top \mathrm{d}s,$$

where Φ_A is the state-transition matrix of A, the solution of $\partial_t\Phi_A(t,s) = A(t)\Phi_A(t,s)$ with $\Phi_A(s,s) = I$ (Särkkä & Solin, 2019). Since $P_0 \succ 0$ and $\Phi_A(t,0)$ is invertible, $P(t) \succ 0$ for all t. Let $P(t) = L(t)L(t)^\top$ be its Cholesky factorization; since the Cholesky factor is a smooth function of a positive definite matrix, L is continuously differentiable and invertible. Define

$$\Omega(t) = L(t)^{-1}\left[A(t)L(t) + \frac{1}{2}\Sigma(t)L(t)^{-\top} - \dot{L}(t)\right].$$

To see that Ω is skew-symmetric, multiply by L on the left and by $L^\top$ on the right:

$$L\Omega L^\top + \left(L\Omega L^\top\right)^\top = \left(AP + PA^\top + \Sigma\right) - \left(\dot{L}L^\top + L\dot{L}^\top\right) = \dot{P} - \dot{P} = 0,$$

using the Lyapunov equation for $\dot{P}$ and $\dot{P} = \dot{L}L^\top + L\dot{L}^\top$; since L is invertible, $\Omega + \Omega^\top = 0$. Now let $O(t)$ solve

$$\dot{O}(t) = \Omega(t)O(t), \quad O(0) = L(0)^{-1}S_0.$$

The initial condition is orthogonal, since $O(0)O(0)^\top = L(0)^{-1}P_0L(0)^{-\top} = I$, and skew-symmetry of Ω implies $\frac{\mathrm{d}}{\mathrm{d}t}(OO^\top) = 0$, so $O(t)$ is orthogonal for all t. Hence $S(t) = L(t)O(t)$ is invertible with $S(0) = S_0$, and

$$\dot{S} = \dot{L}O + L\Omega O = \dot{L}O + \left(AL + \tfrac{1}{2}\Sigma L^{-\top} - \dot{L}\right)O = AS + \tfrac{1}{2}\Sigma S^{-\top},$$

since $L^{-\top}O = S^{-\top}$ for orthogonal O; so S satisfies eq. (21). Here $\Omega = \dot{O}O^\top$ is exactly the gauge generator of section 3 for the factorization $S = LO$; appendix C.3 uses this identification.

Step 2: uniqueness of S. The right-hand side of eq. (21) is locally Lipschitz in S on the open set $\mathrm{GL}(d)$, so the solution with $S(0) = S_0$ is unique as long as it exists; the solution constructed in Step 1 exists on all of $[0,T]$, so it is the unique one.

Step 3: the mean. Substituting eq. (19) into eq. (20) gives the linear ODE

$$\dot{m}(t) = A(t)m(t) + u(t), \quad m(0) = m_0,$$

which has a unique continuously differentiable solution since A and u are continuous (Särkkä & Solin, 2019). □

A.4 Proof of Proposition 2

We show that an affine vector field is q-divergence-free for a Gaussian q with Cholesky factor L if and only if it is a gauge field $L\Omega L^{-1}(z-m)$ with skew-symmetric Ω.

Proof. We write the q-divergence-free condition in whitened coordinates, where it becomes a polynomial identity in the whitened state, and read off the constraints on M and c degree by degree. Fix t and drop it from the notation. Let $q(z) = \mathcal{N}(z \mid m, LL^\top)$ with L invertible, and let $v(z) = Mz + c$ be an affine vector field. By the product rule for the divergence,

$$\nabla \cdot [v(z)q(z)] = \left(\nabla \cdot v(z) + v(z)^\top \nabla \log q(z)\right) q(z),$$

so v is q-divergence-free if and only if $\nabla \cdot v(z) + v(z)^\top \nabla \log q(z) = 0$ for all z. Since $\nabla \cdot v(z) = \mathrm{tr}(M)$ and $\nabla \log q(z) = -L^{-\top}L^{-1}(z-m)$, in the whitened coordinate $\eta = L^{-1}(z-m)$ this condition reads

$$\mathrm{tr}(M) - \left(L^{-1}ML\,\eta + L^{-1}(Mm+c)\right)^\top \eta = 0 \quad \text{for all } \eta \in \mathbb{R}^d. \tag{22}$$

The left-hand side is a polynomial of degree two in η, and it vanishes identically if and only if each of its homogeneous parts vanishes. The quadratic part, $-\eta^\top L^{-1}ML\,\eta$, vanishes for all η if and only

if $L^{-1}ML$ is skew-symmetric. The linear part, $-(Mm+c)^\top L^{-\top}\eta$, vanishes for all η if and only if $Mm+c=0$. The constant part requires $\mathrm{tr}(M)=\mathrm{tr}(L^{-1}ML)=0$, which skew-symmetry already implies.

Hence v is q-divergence-free if and only if $c=-Mm$ and $\Omega=L^{-1}ML$ is skew-symmetric, that is, if and only if $v(z)=L\Omega L^{-1}(z-m)$ for a skew-symmetric Ω. Both directions of Proposition 2 follow, and $\Omega=L^{-1}ML$ is determined by v. □

The condition eq. (22) does not involve the diffusion coefficient. Since the continuity and FPK equations depend on the drift coefficient only through $\nabla\cdot[fq]$, adding a q-divergence-free v to the drift coefficient of eq. (10) or eq. (11) leaves both equations, and hence the marginal laws, unchanged, for any g, state-dependent or not.

B Fixed Gauges of Existing Methods

This appendix identifies the gauges that existing simulation-free methods use, so that they can be placed in the family of section 3. Existing methods with Gaussian marginals $q(t,z)=\mathcal{N}(z\mid m(t),P(t))$ and a state-independent diffusion coefficient use a posterior process with an affine drift coefficient $f(t,z)=A(t)z+u(t)$; we call $A(t)$ its *state matrix*. Comparing with eq. (11), the state matrix has the form

$$A(t)=K(t)-\frac{1}{2}\Sigma(t)P(t)^{-1},\qquad \dot P(t)=K(t)P(t)+P(t)K(t)^\top, \tag{23}$$

where $K(t)=\dot S(t)S(t)^{-1}$, and the Lyapunov equation on the right follows by differentiating $P=SS^\top$. Conversely, every solution $K(t)$ of this Lyapunov equation is realized by some square root $S(t)$ (Smith et al., 2026). In terms of the factorization $S=LO$ of section 3, $K(t)=\dot L(t)L(t)^{-1}+L(t)\Omega(t)L(t)^{-1}$ with $L(t)$ the Cholesky factor of $P(t)$. Each method therefore corresponds to a particular choice of the skew-symmetric $\Omega(t)$ as a function of $(P,\dot P,\Sigma)$. The choice $O\equiv I$, that is, $\Omega\equiv 0$ and $K=\dot L L^{-1}$, is the Cholesky gauge of section 3.

Square Root Gauge SDE Matching (Bartosh et al., 2025) defines the posterior process through the affine map $F(t,\varepsilon)=m(t)+S(t)\varepsilon$ for a learned square root $S(t)$, so that $K=\dot S S^{-1}$ directly. Which gauge this selects depends on how S is parameterized. For the symmetric square root $S=P^{1/2}$, Smith et al. (2026) call the resulting choice the *square root gauge*. In the coordinates $S=LO$ it corresponds to $\Omega=L^{-1}DL$ with $D=\frac{\mathrm{d}}{\mathrm{d}t}P^{1/2}\,P^{-1/2}-\dot L L^{-1}$. Both terms of D solve the Lyapunov equation in eq. (23), so their difference D satisfies $DP+PD^\top=0$; hence $D(z-m)$ is q-divergence-free, and Ω is skew-symmetric by Proposition 2. The implementation of Bartosh et al. (2025) uses a diagonal $S(t)$. The marginal covariance is then diagonal, the symmetric square root and the Cholesky factor coincide, and the square root and Cholesky gauges agree. Parameterizing a general $S(t)$ would in principle span every gauge, but it conflates the marginals with the gauge; the factorization $S=LO$ separates them.

Symmetric Gauge Course & Nair (2023a;b) choose $A(t)$ to be the unique symmetric solution of $\dot P-\Sigma=AP+PA^\top$; we call this the *symmetric gauge*. To express it in our coordinates, write $A=A_0+L\Omega L^{-1}$, where $A_0=\dot L L^{-1}-\frac{1}{2}\Sigma P^{-1}$ is the state matrix of the Cholesky gauge. Symmetry of A is then equivalent to the Sylvester equation

$$G_0\Omega+\Omega G_0=-L^\top(A_0-A_0^\top)L,\qquad G_0=L^\top L,$$

which has a unique skew-symmetric solution because G_0 is positive definite and the right-hand side is skew-symmetric; it has the form of eq. (17) with $\Sigma=I$. Unlike the square root and Cholesky gauges, the symmetric gauge depends on Σ, and it does not extend to state-dependent diffusion coefficients: when Σ depends on z, the drift coefficient of the Cholesky gauge is no longer affine in z, and no state-independent Ω can make its Jacobian symmetric for all z (Smith et al., 2026).

Learned Gauge For all three fixed gauges, Cholesky, square root and symmetric, the gauge is a deterministic function of the marginals and Σ, chosen independently of the objective. In Gauge

Matching, $\Omega(t)$ is instead a free function, and Proposition 2 shows that no other affine choice exists: every fixed gauge is a point in the family we optimize over. One special case matters for section 5.1. When the marginals are stationary, $\dot{P} = 0$, the Cholesky and square root gauges give $K = 0$, and the symmetric gauge gives $K = 0$ as well whenever Σ and P commute, as in the isotropic setting there. The fixed gauge $O \equiv I$ used in section 5.1 therefore represents all three fixed gauges at once. Table 1 summarizes how the fixed gauges, the order-one Helmholtz correction of section 4 and the learned gauge compare in cost and scope.

Householder Parameterization of the Gauge Instead of parameterizing the generator Ω directly, the orthogonal factor $O(t)$ of eq. (13) can be parameterized itself, for instance as a Householder flow (Tomczak & Welling, 2016) of $n = 2\lfloor d/2 \rfloor$ reflections, $O(t) = \prod_{i=1}^{n} H(v_i(t))$ with $H(v) = I - 2vv^\top/\|v\|^2$, which represents every rotation (Grove, 2002, Theorem 6.6); the gauge is then $\Omega = \dot{O}O^\top$, obtained by forward-mode differentiation of the v_i. This recovers an explicit flow map $\Psi(t, \varepsilon) = m(t) + L(t)O(t)\varepsilon$ at the same asymptotic cost, at the price of a redundant parameterization, since O and OR for constant orthogonal R give the same process. We use it in section 5.1, where the gauge angle of $O(t)$ can be read off directly.

C Proofs for Section 4

C.1 Proof of Proposition 3

The proof has two parts. We first show that $\mathcal{J}$ is a strictly convex quadratic on the skew-symmetric matrices, which gives existence and uniqueness of the minimizer for any diffusion coefficient. We then derive the stationarity condition, which for a state-independent diffusion coefficient reduces to eq. (17) by Stein's lemma. After the proof, appendix C.2 draws the consequences used in section 4: the case of a state-dependent diffusion coefficient, the optimality criterion for fixed gauges, the stationary linear example, and the identification of the order-one Helmholtz correction.

Proof. Fix t and drop it from the notation, so that $q(z) = \mathcal{N}(z \mid m, LL^\top)$, and write $Z = m + L\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$. In these coordinates the gauge field eq. (14) is $L\Omega L^{-1}(Z - m) = L\Omega\varepsilon$, and with the residual of the Cholesky gauge $r_0(z) = h(z) - f_0(z)$ the objective reads

$$\mathcal{J}(\Omega) = \frac{1}{2}\mathbb{E}_\varepsilon \left[\|r_0(Z) - L\Omega\varepsilon\|^2_{\Sigma(Z)^{-1}} \right].$$

Under the hypotheses of Proposition 3, the divergence term of f_0 is defined, r_0 has at most polynomial growth, $\Sigma(z)^{-1}$ is bounded because Σ is uniformly positive definite, and $\mathcal{J}(\Omega)$ is finite for every Ω.

Strict Convexity Expanding the square,

$$\mathcal{J}(\Omega) = \mathcal{J}(0) - \mathbb{E}_\varepsilon \left[r_0(Z)^\top \Sigma(Z)^{-1} L\Omega\varepsilon \right] + \frac{1}{2}\mathbb{E}_\varepsilon \left[\varepsilon^\top \Omega^\top L^\top \Sigma(Z)^{-1} L\Omega\varepsilon \right],$$

a quadratic function of Ω. Its quadratic part is nonnegative because $\Sigma(z)^{-1}$ is positive definite for every z, and it vanishes only if $L\Omega\varepsilon = 0$ almost surely, that is, only if $\Omega = 0$, since ε has full support and L is invertible. The quadratic part is therefore positive definite on the linear space $\mathfrak{so}(d)$ of skew-symmetric matrices. Hence $\mathcal{J}$ is strictly convex and coercive on $\mathfrak{so}(d)$ and has a unique minimizer $\Omega^\star$. This part of the argument uses no structure of Σ beyond positive definiteness; in particular, it holds for state-dependent diffusion coefficients.

State-Independent Diffusion Coefficient Let Σ be constant. Using $\mathbb{E}_\varepsilon[\varepsilon^\top X\varepsilon] = \operatorname{tr}(X)$ for any matrix X and writing $\Xi = \mathbb{E}_\varepsilon[r_0(Z)\varepsilon^\top]$,

$$\mathcal{J}(\Omega) = \mathcal{J}(0) - \operatorname{tr}\left(\Omega\Xi^\top\Sigma^{-1}L\right) + \frac{1}{2}\operatorname{tr}\left(\Omega^\top G\Omega\right), \qquad G = L^\top\Sigma^{-1}L.$$

The directional derivative of $\mathcal{J}$ in a skew-symmetric direction E is $\operatorname{tr}(E^\top[G\Omega - L^\top\Sigma^{-1}\Xi])$, where we used that G is symmetric. A matrix X satisfies $\operatorname{tr}(E^\top X) = 0$ for all $E \in \mathfrak{so}(d)$ if and only if X is symmetric, since the symmetric and skew-symmetric matrices are orthogonal complements under the trace inner product. The derivative therefore vanishes in every skew-symmetric direction if and

only if $G\Omega - L^\top\Sigma^{-1}\Xi$ is symmetric, that is, if and only if it equals its transpose $-\Omega G - \Xi^\top\Sigma^{-1}L$, which rearranges to

$$G\Omega + \Omega G = L^\top\Sigma^{-1}\Xi - \Xi^\top\Sigma^{-1}L. \tag{24}$$

The right-hand side is skew-symmetric. Since G is positive definite, the Sylvester operator $\Omega \mapsto G\Omega + \Omega G$ has eigenvalues $\lambda_i + \lambda_j > 0$, where λ_i are the eigenvalues of G, and is therefore invertible on $\mathbb{R}^{d\times d}$; so eq. (24) has exactly one solution. Transposing eq. (24) and using that its right-hand side is skew-symmetric shows that $-\Omega^\top$ solves it whenever Ω does, so by uniqueness the solution is skew-symmetric.

It remains to compute Ξ. Stein's lemma states that $\mathbb{E}_\varepsilon[\varphi(\varepsilon)\varepsilon^\top] = \mathbb{E}_\varepsilon[\nabla_\varepsilon\varphi(\varepsilon)]$ for continuously differentiable φ of at most polynomial growth, where $\nabla_\varepsilon\varphi$ is the Jacobian with entries $\partial\varphi_i/\partial\varepsilon_j$, the convention used for $\nabla_z h$ throughout. Applied to $\varphi(\varepsilon) = r_0(m + L\varepsilon)$, it gives

$$\Xi = \mathbb{E}_\varepsilon\left[\nabla_z r_0(Z)\right] L = \left(\mathbb{E}_{q(z)}\left[\nabla_z h(Z)\right] - A_0\right) L, \qquad A_0 = \dot{L}L^{-1} - \frac{1}{2}\Sigma L^{-\top}L^{-1},$$

where A_0 is the state matrix of f_0 in eq. (11), whose divergence term vanishes for constant Σ. Substituting into eq. (24), the term $\frac{1}{2}\Sigma L^{-\top}L^{-1}$ of A_0 drops out, since $\Sigma^{-1}\cdot\Sigma L^{-\top}L^{-1} = L^{-\top}L^{-1}$ is symmetric, and eq. (17) follows with $\bar{J} = \mathbb{E}_{q(z)}[\nabla_z h(Z)] - \dot{L}L^{-1}$. □

C.2 Consequences of Proposition 3

State-Dependent Diffusion Coefficient For a general $\Sigma(z)$ the minimizer is still the solution of a linear system, since $\mathcal{J}$ is a strictly convex quadratic. Writing $\Omega = \sum_k \gamma_k E_k$ in a basis $E_1, \ldots, E_{d(d-1)/2}$ of $\mathfrak{so}(d)$, the stationarity conditions $\partial\mathcal{J}/\partial\gamma_j = 0$ read $\sum_k Q_{jk}\gamma_k = b_j$ with

$$Q_{jk} = \mathbb{E}_\varepsilon\left[\varepsilon^\top E_j^\top L^\top\Sigma(Z)^{-1}LE_k\varepsilon\right], \qquad b_j = \mathbb{E}_\varepsilon\left[r_0(Z)^\top\Sigma(Z)^{-1}LE_j\varepsilon\right],$$

and Q is positive definite by the strict convexity argument above. Unlike in the state-independent case, these coefficients do not reduce to a Jacobian: the factor $\Sigma(Z)^{-1}$ inside the expectations depends on Z, so Stein's lemma no longer applies. Q and b therefore have to be estimated at every evaluation, by Monte Carlo sampling or, in low dimension, by numerical quadrature. Gauge Matching sidesteps this by learning Ω_ϕ directly.

Optimality of Fixed Gauges The stationarity condition also tells us when a given gauge is optimal. Let Ω be any gauge, with state matrix $A_\Omega = A_0 + L\Omega L^{-1}$. By the computation above, Ω is optimal if and only if $G\Omega - L^\top\Sigma^{-1}\Xi$ is symmetric. Substituting $G = L^\top\Sigma^{-1}L$ and $\Xi = (\mathbb{E}_{q(z)}[\nabla_z h(Z)] - A_0)L$,

$$\begin{aligned} G\Omega - L^\top\Sigma^{-1}\Xi &= L^\top\Sigma^{-1}\left(L\Omega L^{-1} + A_0 - \mathbb{E}_{q(z)}[\nabla_z h(Z)]\right) L \\ &= L^\top\Sigma^{-1}\left(A_\Omega - \mathbb{E}_{q(z)}[\nabla_z h(Z)]\right) L, \end{aligned}$$

and since congruence by the invertible L preserves symmetry, Ω is optimal if and only if

$$\Sigma^{-1}\left(A_\Omega - \mathbb{E}_{q(z)}\left[\nabla_z h(Z)\right]\right) \text{ is symmetric.} \tag{25}$$

Two cases are used in section 4. For the Cholesky gauge, $A_\Omega = A_0$, and eq. (25) holds if and only if $\Sigma^{-1}\bar{J}$ is symmetric, since the term $\frac{1}{2}\Sigma L^{-\top}L^{-1}$ of A_0 drops out as before. For the symmetric gauge, A_Ω is symmetric by definition; if in addition $\Sigma = \sigma^2 I$ is isotropic, then eq. (25) holds if and only if $\mathbb{E}_{q(z)}[\nabla_z h(Z)]$ is symmetric. This is the case for a gradient prior drift coefficient $h = -\nabla U$, whose Jacobian $-\nabla^2 U$ is symmetric at every z.

The Stationary Linear SDE of section 5.1 Let $h(z) = Az$ with $A = -\alpha I + \omega J$, $J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, $\Sigma = I$, and let the marginals be the stationary law $m = 0$, $LL^\top = I/(2\alpha)$, so $\dot{L} = 0$ and $L = I/\sqrt{2\alpha}$. Then $G = I/(2\alpha)$, $\bar{J} = A$, and the right-hand side of eq. (17) is $(A - A^\top)/(2\alpha) = \omega J/\alpha$, so $\Omega^\star = \omega J$. Here $A_0 = \dot{L}L^{-1} - \frac{1}{2}\Sigma L^{-\top}L^{-1} = -\alpha I$, so the drift coefficient of the optimal gauge is $A_0 z + L\Omega^\star L^{-1}z = -\alpha z + \omega J z = h(z)$: the optimal gauge recovers the prior process, and the path-space KL term vanishes.

The Order-One Helmholtz Correction We now show that the order-one correction of Helmholtz-SDE (Smith et al., 2026) is the optimal gauge of the linearized prior process, as claimed in section 4. Helmholtz-SDE starts from a reference posterior process with a fixed gauge Ω_{ref}, whose drift coefficient $f_{\mathrm{ref}} = f_0 + L\Omega_{\mathrm{ref}}L^{-1}(z-m)$ is affine, and adds to it the q-divergence-free component of the residual $h - f_{\mathrm{ref}}$. This component is defined by the q-weighted Helmholtz decomposition of Smith et al. (2026): it is the field v that minimizes $\mathbb{E}_{q(z)}[\|h(Z) - f_{\mathrm{ref}}(Z) - v(Z)\|^2_{\Sigma^{-1}}]$ over all q-divergence-free fields. At order $\ell = 1$, the residual is replaced by its linearization about m; since f_{ref} is already affine, this amounts to replacing h by its linearization $\tilde{h}(z) = h(m) + \nabla_z h(m)(z-m)$.

Three observations identify the resulting gauge.

1. The Helmholtz decomposition of an affine field with respect to a Gaussian q has affine components (Smith et al., 2026), so the correction is an affine q-divergence-free field, that is, by Proposition 2, a gauge field $L\Omega_{\mathrm{H}}L^{-1}(z-m)$.
2. The correction minimizes over all q-divergence-free fields and, by the first observation, lies in the subset of gauge fields; it is therefore also the minimizer over gauge fields. Since $f_{\mathrm{ref}} + L\Omega_{\mathrm{H}}L^{-1}(z-m) = f_0 + L(\Omega_{\mathrm{ref}} + \Omega_{\mathrm{H}})L^{-1}(z-m)$, the gauge $\Omega_{\mathrm{ref}} + \Omega_{\mathrm{H}}$ is the unique minimizer of $\mathcal{J}$ with h replaced by $\tilde{h}$.
3. $\nabla_z \tilde{h} \equiv \nabla_z h(m)$, so by Proposition 3 this minimizer is the solution of eq. (17) with $\bar{J}$ replaced by $\nabla_z h(m) - \dot{L}L^{-1}$.

The corrected gauge therefore does not depend on Ω_{ref}, and it coincides with $\Omega^\star$ whenever $\mathbb{E}_{q(z)}[\nabla_z h(Z)] = \nabla_z h(m)$, in particular whenever $\nabla_z h$ is affine in z. The order-$\ell \geq 2$ corrections of Smith et al. (2026) are marginal-preserving fields of higher polynomial degree, which lie outside the Gauge Matching family.

C.3 Proof of Corollary 4

We show that for a linear prior process and a linear-Gaussian observation model the true posterior process is a Gaussian flow diffusion between observation times, and that Gauge Matching posterior processes approach it in KL divergence. Let Z follow the linear SDE eq. (12) with $Z_0 \sim \mathcal{N}(m_0, S_0 S_0^\top)$, $S_0 \in \mathrm{GL}(d)$, and $\Sigma(t) = g(t)g(t)^\top$ positive definite, and let $x_{t_i} \sim \mathcal{N}(CZ_{t_i}, R)$ with R positive definite. We take the observation times to be interior, $0 < t_1 < \cdots < t_N < T$, and write $t_0 = 0$ and $t_{N+1} = T$; an observation at an endpoint introduces no kink and only simplifies the argument. The Gauge Matching posterior processes over which the infimum is taken are those with m_ϕ, L_ϕ continuously differentiable and Ω_ϕ continuous on $[0, T]$, the regularity that neural networks in t provide.

The proof has three steps. We first describe the true posterior process and show that its covariance is positive definite. We then apply Theorem 1 on each interval between observations, which gives the first claim of the corollary and an explicit formula for the jump of the gauge at an observation time. Finally, we smooth the resulting $m^\dagger$, $L^\dagger$, $\Omega^\dagger$ in small neighborhoods of the observation times and show that the KL divergence of the smoothed process to the true posterior process can be made arbitrarily small. Throughout, a dagger marks quantities of the true posterior process, to distinguish them from the optimal gauge $\Omega^\star$ of Proposition 3.

Proof. Step 1: the true posterior process. Conditioned on X, the latent path is a Markov process with initial law $q^\dagger(0, z) \propto p_\theta(0, z)\, p_\theta(x_{t_1}, \ldots, x_{t_N} \mid Z_0 = z)$ and a drift coefficient given by Doob's h-transform (Bartosh et al., 2025, Appendix B): on the open interval (t_i, t_{i+1}),

$$f^\dagger(t, z) = A(t)z + u(t) + \Sigma(t)\nabla_z \log p_\theta(x_{t_{i+1}}, \ldots, x_{t_N} \mid Z_t = z).$$

The likelihood of the future observations given the state is an unnormalized Gaussian function of z, so the last term is affine in z. Its coefficients are continuous functions of the transition matrix and the process noise covariance of eq. (12) (Särkkä & Solin, 2019), hence continuous in t on $[t_i, t_{i+1}]$. We conclude that $f^\dagger(t, z) = A^\dagger(t)z + u^\dagger(t)$ with $A^\dagger$, $u^\dagger$ continuous on each closed interval $[t_i, t_{i+1}]$. Across each t_i, however, the set of future observations changes, and with it $A^\dagger$ and $u^\dagger$ change discontinuously. The marginals of the true posterior process are the smoothing distributions $q^\dagger(t, z) = \mathcal{N}(z \mid m^\dagger(t), P^\dagger(t))$, which are continuous in t on $[0, T]$.

We also need that $P^\dagger(t) \succ 0$ for every t. Conditioning on more information can only decrease a Gaussian covariance, so it suffices to exhibit a finer conditioning under which the covariance of

Z_t is still positive definite. Condition, in addition to X, on the latent states Z_{t_j} at all observation times $t_j \neq t$; given these states, the observations x_{t_j} carry no further information about Z_t. The covariance of Z_t is then that of a Gaussian bridge of eq. (12) over an interval of positive length, or of the unconditioned process before t_1 or after t_N; both are positive definite since $\Sigma \succ 0$ and $S_0 S_0^\top \succ 0$. If t is itself an observation time, conditioning on x_t adds the finite precision $C^\top R^{-1} C$, which keeps the covariance positive definite. Hence $P^\dagger(t) \succ 0$.

Step 2: representation between observations. On $[t_i, t_{i+1}]$ the true posterior process is the linear SDE with coefficients $A^\dagger$, $u^\dagger$, g started from $\mathcal{N}(m^\dagger(t_i), P^\dagger(t_i))$ with $P^\dagger(t_i) \succ 0$. By Theorem 1 it is a Gaussian flow diffusion on this interval, whose marginal parameters are necessarily $m = m^\dagger$ and $L = L^\dagger$, the Cholesky factor of $P^\dagger$. Both are continuously differentiable on $[t_i, t_{i+1}]$: $m^\dagger$ and $P^\dagger$ solve $\dot{m} = A^\dagger m + u^\dagger$ and $\dot{P} = A^\dagger P + P A^{\dagger\top} + \Sigma$ with continuous coefficients, and the Cholesky factorization is a smooth map on positive definite matrices. By the proof in appendix A.3, the gauge generator of this representation is

$$\Omega^\dagger(t) = L^\dagger(t)^{-1} \left[A^\dagger(t) L^\dagger(t) + \frac{1}{2} \Sigma(t) L^\dagger(t)^{-\top} - \dot{L}^\dagger(t) \right],$$

which is continuous on $[t_i, t_{i+1}]$. This proves the first claim of Corollary 4.

Across an observation time, $m^\dagger$ and $L^\dagger$ are continuous, but in general their derivatives jump, and so does $\Omega^\dagger$. The jump is explicit. The last term of $f^\dagger$ is Σ times the gradient of the log-likelihood of the future observations, so $A^\dagger = A - \Sigma\Lambda$, where Λ is the negative Hessian of this log-likelihood in z, the precision of the future observations about Z_t. This precision includes $C^\top R^{-1} C$ for $t < t_i$, when x_{t_i} is still in the future, and not for $t > t_i$. Writing Δ for the jump from t_i^- to t_i^+, Λ drops by $C^\top R^{-1} C$ and the state matrix therefore jumps by $\Delta A^\dagger = \Sigma C^\top R^{-1} C$, and the derivative $\dot{P}^\dagger = A^\dagger P^\dagger + P^\dagger A^{\dagger\top} + \Sigma$ jumps by $\Delta A^\dagger P^\dagger + P^\dagger \Delta A^{\dagger\top}$.

The jump of $\dot{L}^\dagger$. Let $\mathrm{ltri}(X)$ denote the strictly lower triangular part of X plus half its diagonal; differentiating $P = LL^\top$ and solving for the lower triangular $\dot{L}$ gives $\dot{L} = L\ \mathrm{ltri}(L^{-1}\dot{P}L^{-\top})$. Since $L^\dagger$ is continuous across t_i, only $\dot{L}^\dagger$ jumps. With $B = L^{\dagger -1} \Delta A^\dagger L^\dagger = L^{\dagger -1} \Sigma C^\top R^{-1} C L^\dagger$ we have $L^{\dagger -1} \Delta \dot{P}^\dagger L^{\dagger -\top} = B + B^\top$, so the jump of $\dot{L}^\dagger$ is $L^\dagger\ \mathrm{ltri}(B + B^\top)$. Substituting into the formula for $\Omega^\dagger$,

$$\Delta\Omega^\dagger = L^{\dagger -1} \Delta A^\dagger L^\dagger - L^{\dagger -1} \Delta \dot{L}^\dagger = B - \mathrm{ltri}(B + B^\top).$$

This matrix is skew-symmetric, because $\mathrm{ltri}(X) + \mathrm{ltri}(X)^\top = X$ for symmetric X. It vanishes if and only if $B = \mathrm{ltri}(B + B^\top)$, that is, if and only if B is lower triangular. For $\Sigma = I$ and an isotropic added precision $C^\top R^{-1} C \propto I$, as under the observation model of Smith et al. (2026), B is diagonal and the gauge does not jump (appendix D).

Step 3: the infimum is zero. Let $Q^\dagger$ be the law of the true posterior process and Q_ϕ that of a Gauge Matching posterior process with the same diffusion coefficient. By Girsanov's theorem,

$$\begin{aligned} D_{\mathrm{KL}}\left(Q_\phi \,\middle\|\, Q^\dagger\right) &= D_{\mathrm{KL}}\left(q_\phi(0, z \mid X) \,\middle\|\, q^\dagger(0, z)\right) \\ &\quad + \frac{1}{2} \int_0^T \mathbb{E}_{q_\phi(t,z|X)} \left[\left\| f_\phi(t, Z_t, X) - f^\dagger(t, Z_t) \right\|^2_{\Sigma(t)^{-1}} \right] \mathrm{d}t, \end{aligned}$$

and this divergence equals the slack $\mathcal{L} + \log p_\theta(X)$ of the bound eq. (6), so its infimum is zero if and only if the bound is tight. The integral on the second line has the form of the path-space KL term of eq. (6) with the prior drift coefficient replaced by $f^\dagger$; we refer to it as the path-space term of this divergence. We construct a family of Gauge Matching posterior processes Q_δ, indexed by $\delta > 0$, with $D_{\mathrm{KL}}\left(Q_\delta \,\middle\|\, Q^\dagger\right) \to 0$.

Smoothing. Fix $\delta > 0$ smaller than half the smallest spacing between consecutive elements of $\{t_0, \ldots, t_{N+1}\}$, and let $U_\delta = \bigcup_{i=1}^N (t_i - \delta, t_i + \delta)$ be the union of δ-neighborhoods of the observation times. Define m_δ and L_δ to agree with $m^\dagger$ and $L^\dagger$ outside U_δ and, on each component of U_δ, to be the cubic Hermite interpolants that match the values and one-sided derivatives of $m^\dagger$, $L^\dagger$ at the endpoints of the component. Define Ω_δ analogously by linear interpolation of $\Omega^\dagger$. Then m_δ, L_δ are continuously differentiable and Ω_δ is continuous on $[0, T]$, so Q_δ, the Gauge Matching posterior process with these functions, is admissible.

Uniform bounds. Since $m^\dagger$, $L^\dagger$ are Lipschitz and $\dot{m}^\dagger$, $\dot{L}^\dagger$, $\Omega^\dagger$ are bounded on $[0,T]$, the interpolants have derivatives $\dot{m}_\delta$, $\dot{L}_\delta$ and values Ω_δ bounded by a constant independent of δ. Moreover, $\sup_t \|L_\delta(t) - L^\dagger(t)\| = \mathcal{O}(\delta)$, and $\lambda_{\min}(P^\dagger(t))$ is bounded away from zero on $[0,T]$, since $P^\dagger$ is positive definite by Step 1 and continuous on the compact interval $[0,T]$. For small δ, therefore, $L_\delta(t)$ is lower triangular with positive diagonal and $\|L_\delta(t)^{-1}\|$ is bounded uniformly in t and δ.

The divergence. The drift coefficient of Q_δ, which is eq. (15) evaluated at $(m_\delta, L_\delta, \Omega_\delta)$, agrees with $f^\dagger$ outside U_δ, so the integrand of the path-space term vanishes there. On U_δ, both drift coefficients are affine in z with coefficients bounded uniformly in δ, and the second moment $\mathbb{E}_{q_\delta(t,z|X)}[\|Z_t\|^2] = \|m_\delta(t)\|^2 + \mathrm{tr}(L_\delta(t)L_\delta(t)^\top)$ is bounded uniformly as well. The integrand is therefore bounded on U_δ by a constant c independent of δ, and since U_δ has length at most $2N\delta$, the path-space term, with its factor $\frac{1}{2}$, is at most $cN\delta$. The initial term vanishes since $0 \notin U_\delta$. Hence $D_{\mathrm{KL}}\left(Q_\delta \,\middle\|\, Q^\dagger\right) \le cN\delta \to 0$ as $\delta \to 0$, which proves the second claim of Corollary 4. The argument is that of Smith et al. (2026, Theorem 1), with the gauge smoothed alongside the marginals. □

D Experimental Details

Algorithm 1 summarizes one training iteration of Gauge Matching, the simulation-free estimator of section 2.2 applied to the posterior process of section 4.

Algorithm 1 Gauge Matching (one training iteration)

Require: observations X; generative model $p_\theta(0,z)$, h_θ, g_θ, $p_\theta(x \mid z)$; posterior process m_ϕ, L_ϕ, Ω_ϕ

$\hat{\mathcal{L}} \leftarrow D_{\mathrm{KL}}\left(q_\phi(0,z \mid X) \,\|\, p_\theta(0,z)\right)$ ▷ initial KL term

$t \sim \mathcal{U}[0,T]$, $\varepsilon \sim \mathcal{N}(0,I)$, $Z_t \leftarrow m_\phi(t,X) + L_\phi(t,X)\varepsilon$

$f \leftarrow \dot{m}_\phi + \dot{L}_\phi\varepsilon + L_\phi\Omega_\phi\varepsilon - \frac{1}{2}\Sigma_\theta(t,Z_t)L_\phi^{-\top}\varepsilon + \frac{1}{2}\nabla\cdot\Sigma_\theta(t,Z_t)$ ▷ eq. (15) with $\eta = \varepsilon$

solve $g_\theta(t,Z_t)\, r = h_\theta(t,Z_t) - f$, $\hat{\mathcal{L}} \leftarrow \hat{\mathcal{L}} + \frac{T}{2}\|r\|_2^2$ ▷ path-space KL term

$i \sim \mathcal{U}\{1,\dots,N\}$, $\varepsilon' \sim \mathcal{N}(0,I)$, $Z_{t_i} \leftarrow m_\phi(t_i,X) + L_\phi(t_i,X)\varepsilon'$

$\hat{\mathcal{L}} \leftarrow \hat{\mathcal{L}} - N\log p_\theta(x_{t_i} \mid Z_{t_i})$ ▷ reconstruction

gradient step on (θ,ϕ) with respect to $\hat{\mathcal{L}}$

All methods are implemented in a common code base (see the reproducibility statement). This appendix lists the choices shared by all experiments and then, for each experiment, the data, the model, deviations of the posterior process from the shared architecture, the training schedule, and the definition of every reported metric.

Posterior architecture. All posterior processes share the amortized architecture of Bartosh et al. (2025): a GRU processes the observations of a trial into per-observation embeddings (when the observation times are irregular, it also receives the times and their spacings); the embedding at a query time t is a softmax-weighted average of these, with weights centered at t over the observation times; and an MLP with two hidden layers of width 100 and SiLU activations maps the embedding and t to the posterior parameters. For Gauge Matching these are the d entries of m_ϕ, the $d(d+1)/2$ entries of L_ϕ, with a softplus transform on the diagonal, and the $d(d-1)/2$ entries of Ω_ϕ above the diagonal, the last from a separate head of the same MLP.

Baselines. The three fixed gauges use the same network without the gauge head: SDE Matching, that is, the square root gauge, with a diagonal L_ϕ, and the Cholesky and symmetric gauges with a full L_ϕ. Helmholtz-SDE uses the network of the Cholesky gauge and, at every evaluation, sets Ω to the solution of eq. (17) with $\nabla_z h_\theta(t, m_\phi)$ in place of the expected Jacobian. By appendix C.2, the resulting posterior process is the one that the order-one correction produces from any fixed gauge with these marginals, so no separate reference gauge is needed. The Jacobian is computed by d forward-mode passes through the prior drift network, the Sylvester equation eq. (17) is solved by the Kronecker-sum formula for $d \le 4$ and by the eigendecomposition described in appendix D.1 for larger d, and gradients flow through the solve.

Prior process. The prior drift coefficient is an MLP with one hidden layer of width 100 and softplus activations, as in Li et al. (2020) and Smith et al. (2026). The diffusion coefficient is state-independent and diagonal unless stated otherwise.

Optimization. All models are trained with Adam (Kingma & Ba, 2015) and gradient clipping at norm 5, using one sample each of t, ε and i per trial and iteration as in algorithm 1; in the gauge-learning, conjugate and Protocol 1 experiments the batch consists of samples rather than trials. Learning rates, batch sizes and iteration counts are listed per experiment. Unless stated otherwise, every experiment is repeated over 5 seeds, and we report the mean and twice the standard error over seeds.

Gap and Gauge Excess Two quantities recur in the experiments. The *gap* of a fitted posterior process is its KL divergence to the true posterior process (section 5.1). For two processes with a common diffusion coefficient this divergence is, by Girsanov's theorem, an initial KL term plus a path-space term (the display in Step 3 of appendix C.3). When both drift coefficients are affine, as for a linear prior process and every posterior process considered here, the integrand of the path-space term is a Gaussian expectation of a quadratic form, and we evaluate the gap in closed form on a time grid of step $5 \cdot 10^{-4}$.

The *gauge excess* of table 2 is $\mathcal{J}(\Omega) - \mathcal{J}(\Omega^\star)$ in nats per unit time, where $\mathcal{J}$ is the integrand of Proposition 3 at the marginals of the trained model, Ω is the gauge the method uses at each time t, and $\Omega^\star$ is the solution of eq. (17) with the expected Jacobian estimated from 10^3 Monte Carlo samples; $\mathcal{J}$ itself is estimated from the same samples. The result is averaged over t on a grid of 100 points and over 64 held-out trials; every gauge excess reported below uses this estimator. Since $\mathcal{J}$ is a strictly convex quadratic function of Ω, the excess is nonnegative, and it is zero exactly when the method's gauge is the optimal one for its own marginals.

D.1 Linear Dynamics

Learning the Gauge Data. The prior process is eq. (12) with $A = -\alpha I + \omega J$, $u = 0$, $g = I$, $\alpha = 0.2$, $\omega = 2\pi$ and the stationary initial law $\mathcal{N}(0, I/(2\alpha))$ on $[0, T]$ with $T = 3$. There are no observations and hence no encoder.

Posterior. m_ϕ, L_ϕ and the Householder vectors $v_i(t)$ of $O(t)$ are MLPs of t with two hidden layers of width 64, and the gauge $\Omega = \dot{O}O^\top$ is obtained by forward-mode differentiation.

Training. We minimize the initial and path-space KL terms of eq. (6) with the estimator of section 2.2 for 10^4 iterations, with 2^{10} samples of (t, ε) per iteration and learning rate 10^{-3}. The initial term is needed to identify the initial law: without observations, any initial law is compatible with a zero path-space KL term once the process has relaxed to stationarity.

Metrics. The gauge angle in fig. 1 is $\theta(t) = \operatorname{atan2}(O_{21}(t), O_{11}(t))$, unwrapped, with $\theta(0)$ subtracted since the initial rotation is not identifiable. We call the path-space KL term per unit time the *rate*. The variance 0.080 quoted in section 5.1 is that of the trained fixed-gauge process averaged over $t \in [T/6, 5T/6]$; we exclude the ends because the process has boundary layers there, where the initial KL term at $t = 0$ and the absence of a future cost at $t = T$ let the variance rise to about 0.2. Its rate averaged over $[0, T]$ is 6.06 nats per unit time, which gives the 18 nats over $[0, 3]$ quoted there. Both values agree with the closed-form optimum derived next (0.080 and 6.09).

Marginal distortion of the fixed gauge. For stationary isotropic marginals $\mathcal{N}(0, vI)$ and $\Omega = 0$, eq. (15) gives the posterior drift coefficient $f_0(z) = -z/(2v)$, and with $Z = \sqrt{v}\,\varepsilon$ the rate is

$$\begin{aligned}\frac{1}{2}\mathbb{E}\left[\|AZ - f_0(Z)\|^2\right] &= \frac{v}{2}\mathbb{E}\left[\left\|\left(A + \tfrac{1}{2v}I\right)\varepsilon\right\|^2\right] = \frac{v}{2}\operatorname{tr}\left(A^\top A + \tfrac{1}{v}A + \tfrac{1}{4v^2}I\right)\\ &= (\alpha^2 + \omega^2)v - \alpha + \frac{1}{4v},\end{aligned}$$

using $A^\top A = (\alpha^2 + \omega^2)I$ and $\operatorname{tr} A = -2\alpha$. The minimizer is $v = 1/(2\sqrt{\alpha^2 + \omega^2})$ with minimal rate $\sqrt{\alpha^2 + \omega^2} - \alpha$, whereas at the stationary variance $v = 1/(2\alpha)$ the rate is $\omega^2/(2\alpha)$. For $\alpha = 0.2$ and $\omega = 2\pi$ the minimizing variance is $v = 0.080$ against the stationary 2.5, and the corresponding rates are 6.09 against 98.7 nats per unit time; the learned gauge attains rate zero at the stationary variance.

The restriction to isotropic marginals is for the closed form only: the trained model is free to learn any m_ϕ, L_ϕ, and a numerical minimization over lower-triangular $L(t)$ confirms that the isotropic optimum is the global one.

Conjugate Setting Data. We follow Smith et al. (2026, Appendix C.2): the prior process of section 5.1 with $T = 5$, $\alpha = 0.2$ and $\omega \in \{0, \pi/4, \pi/2, \pi, 2\pi\}$; observations $x_{t_i} \sim \mathcal{N}(Cz_{t_i} + c, \sigma^2 I)$ at $N \in \{5, 10, 20, 50\}$ uniformly random times with $\sigma \in \{0.05, 0.3, 1.0\}$, where C has standard normal entries normalized to unit determinant and c is standard normal; 16 independent trials per configuration. The generative model is fixed at its true value, so only the posterior process is trained.

Deviation from Smith et al. (2026). Their observation noise is $\sigma^2 CC^\top$ rather than $\sigma^2 I$. Under their choice the precision $C^\top R^{-1} C = I/\sigma^2$ that an observation adds is isotropic, so the covariance of the true posterior process stays isotropic, its gauge is the constant ωJ, and the square root gauge fails for the reason of section 5.1 alone. Under isotropic noise the added precision $C^\top C/\sigma^2$ is anisotropic and the gauge of the true posterior process is time-dependent: by appendix C.3 it jumps by $B - \mathrm{ltri}(B + B^\top)$ with $B = L^{\dagger -1} C^\top C L^\dagger / \sigma^2$ at every observation time, as fig. 2B illustrates.

True posterior process. It is the Kalman–Bucy smoother (Särkkä & Solin, 2019). We obtain its drift coefficient by Doob's h-transform as in appendix C.3 on a grid of step $5 \cdot 10^{-3}$, and its gauge, shown in fig. 2B, by Proposition 2 from its state matrix and the Cholesky factor of its covariance.

Posterior. Since the posterior process is fitted per trial, the encoder of the shared architecture is replaced by a learned per-trial embedding of dimension 100. The smoother has kinks at every observation time, which an MLP of t alone resolves poorly (every posterior process stalled at a gap of about 100 nats), so we add Fourier features of time, $(\sin(2\pi k t/T), \cos(2\pi k t/T))_{k \le 160}$, to the input t; with $k \le 8$ instead, the gap of Gauge Matching at $N = 10$, $\sigma = 0.3$, $\omega = 2\pi$ is 9.2 nats rather than 1.5 at the shorter budget below. The time derivatives $\dot{m}_\phi$, $\dot{L}_\phi$ pass through these features by forward-mode differentiation.

Training. 2^8 samples of t and of i per trial and iteration; Adam with learning rate 10^{-3} decayed exponentially to 10^{-5} over $5 \cdot 10^4$ iterations for Helmholtz-SDE and Gauge Matching, and to 10^{-4} over $2 \cdot 10^4$ iterations for the three fixed gauges. The longer budget lowers the gap of Gauge Matching from 1.5–5.2 to 0.8–3.9 nats (fig. 2A) and that of Helmholtz-SDE from 1.0–4.4 to 0.5–2.8 nats, so the two are compared at equal budget. We did not extend the fixed gauges: at $\sigma \ge 0.3$ they already lie within 0.04–2.2 nats of the optima of their variational families (below), so their 14–26 nats are a property of the family, not of the optimizer. Since the prior Jacobian is the constant A, the order-one correction of Helmholtz-SDE is exact and costs nothing beyond the Sylvester solve; the per-iteration cost comparison of section 5 does not apply here.

Metrics. The gap is computed in closed form as above. As a reference for what training leaves on the table, we also minimize the closed-form gap directly over each variational family by L-BFGS on the grid, with a Gaussian kernel of width 0.02 applied to the optimized functions standing in for the smoothness of a network; fig. 4 shows these optima next to the trained models of fig. 3. At $\sigma = 0.3$ and $\omega = 2\pi$ the optima are 0.3–2.2 nats for the two gauge-corrected families and 12–26 nats for the fixed gauges, which places Helmholtz-SDE 0.2–0.6 nats, Gauge Matching 0.3–1.7 nats and the fixed gauges 0.6–2.1 nats above the optimum of their family.

Remarks. At $\sigma = 0.05$ the trained models lie 3–41 nats (Helmholtz-SDE and Gauge Matching alike) and 4–54 nats (fixed gauges) above their optima, so that row of fig. 3 measures the difficulty of fitting a sharply kinked smoother with a smooth network, not the effect of the gauge; the optima themselves (fig. 4) put the gauge-corrected families 2–4 times below the fixed ones there as well. At $\omega = 0$ the fixed gauges retain a small gap in fig. 2C, which comes from the covariance structure that the observations induce: the square root gauge, with its diagonal L_ϕ, can represent the true posterior process only if its covariance is isotropic, which it is not for a generic C, while the Cholesky and symmetric gauges are close to the gauge-corrected models there (fig. 3).

Scaling in the Latent Dimension Data. The prior process is the d-dimensional rotating Ornstein–Uhlenbeck process $\mathrm{d}Z = (-\alpha I + \omega K) Z \mathrm{d}t + \mathrm{d}W$ with $\alpha = 0.2$, $\omega = 2\pi$ and $T = 5$ as in the conjugate setting, where $K = (G - G^\top)/\|G - G^\top\|_2$ for a standard normal $d \times d$ matrix G: a random skew-symmetric matrix of unit spectral norm, which rotates $\lfloor d/2 \rfloor$ random planes at angular velocities of at most ω. Since $A + A^\top = -2\alpha I$, the stationary law is $\mathcal{N}(0, I/(2\alpha))$ for every K, and

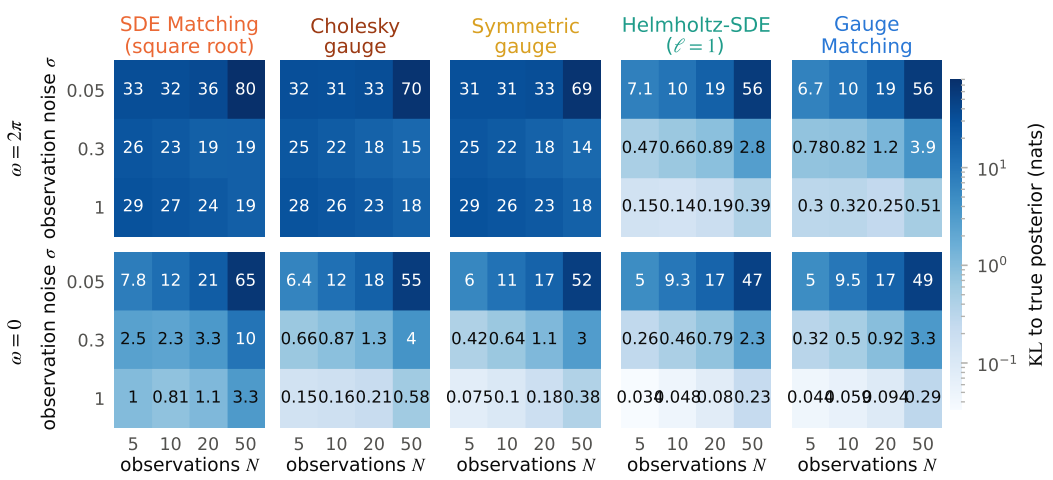

Figure 3: KL divergence to the true posterior process in the conjugate setting for every combination of observation noise σ and number of observations N, for the five posterior processes of section 5, at $\omega = 2\pi$ (top) and $\omega = 0$ (bottom). Median over 16 trials. At $\omega = 0$ the prior drift is a gradient field, so the symmetric gauge is optimal and Helmholtz-SDE coincides with it; the learned gauge lies within a factor of 1.5 of both, and the square root gauge retains a gap because the posterior covariance is not isotropic under a generic C. The residual error grows with N and shrinks with σ: it is the error of a continuously differentiable posterior at the observation times. At $\sigma = 0.05$ every method is instead limited by optimization (fig. 4 and the text).

the transition $e^{As} = e^{-\alpha s} e^{\omega K s}$ with covariance $(1 - e^{-2\alpha s})/(2\alpha)\, I$ lets us sample the latent state exactly at the observation times.

Observations are $x = Cz + c + \mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.3$, C a $\lceil d/2 \rceil \times d$ matrix with orthonormal rows (the transpose of the Q factor of a standard normal $d \times \lceil d/2 \rceil$ matrix: a random half of the state is observed, the other half is inferred through the dynamics) and c standard normal, at $N = 10$ distinct uniformly random times per trial on a grid of step $5 \cdot 10^{-3}$. One system (K, C, c) is drawn per seed and shared by all trials and all five posterior processes; $d \in \{2, 4, 8, 16, 32, 64\}$, 3 seeds, the seed of the system being also that of the model.

Protocol 1 (known generative model). As in the conjugate setting, the prior process, the initial law and the observation model are fixed at their true values, and the posterior process is fitted to 16 trials at once with the architecture of the conjugate setting (per-trial embedding, Fourier time features) and heads for m_ϕ, L_ϕ and, for Gauge Matching, the $d(d-1)/2$ entries of Ω_ϕ. Training: 2^8 samples of t and of i per trial and iteration, Adam with learning rate 10^{-3} decayed exponentially to 10^{-5} over $3 \cdot 10^4$ iterations, the same budget for all five gauges. Metric: the closed-form gap to the Kalman–Bucy smoother in d dimensions, as median and interquartile range over the 3×16 trials.

Protocol 2 (learned generative model). The protocol of the Lorenz system with a d-dimensional latent state: a Gaussian initial law with learned mean and diagonal covariance, the shared prior drift network, the fixed diffusion coefficient $g = I$, a linear-Gaussian observation model with learned C and c and the fixed noise σ, and the amortized GRU posterior of width 100; 1024 training and 2048 test trials; $3 \cdot 10^4$ iterations, batches of 256 trials, learning rate 10^{-3}. Metrics on the test trials as in table 2: the negative ELBO on a grid of 500 times with 4 noise samples, the latent log-likelihood of the noise-free $Cz + c$ under the posterior predictive at the observation times, and the gauge excess.

Sylvester solve. The symmetric gauge, Helmholtz-SDE and the gauge excess solve eq. (17). For $d \leq 4$ we solve the $d^2 \times d^2$ Kronecker-sum system directly; for $d \geq 8$ we use the eigendecomposition $G = V \Lambda V^\top$ of the symmetric positive definite G, $\Omega = V\big[(V^\top R V)_{ij}/(\lambda_i + \lambda_j)\big]V^\top$ with R the right-hand side of eq. (17), at cost $\mathcal{O}(d^3)$ per sample, and obtain its gradient from the adjoint Sylvester equation with the same eigendecomposition rather than by differentiating through the

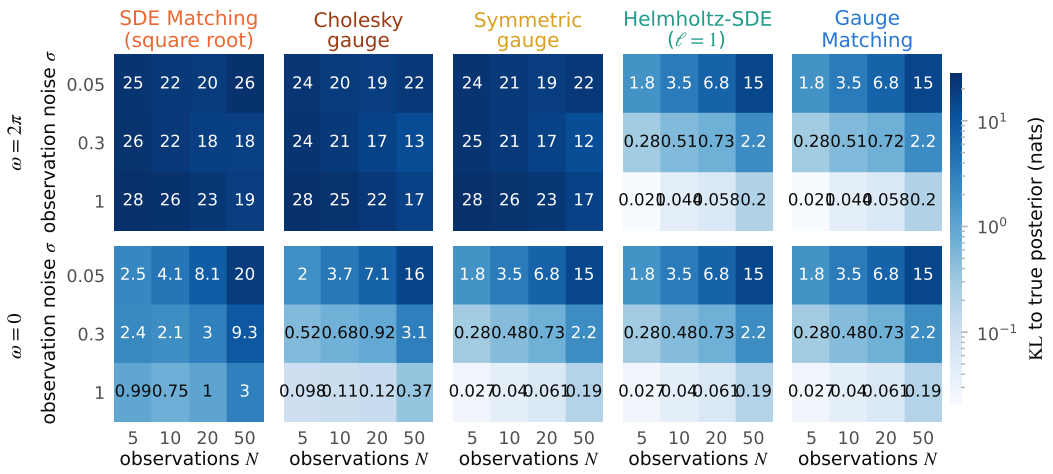

Figure 4: What the variational family admits versus what training reaches. Same grid as fig. 3, but each entry is the minimum of the closed-form KL divergence over the variational family of that gauge, found by L-BFGS directly on the time grid (functions smoothed by a Gaussian kernel of width 0.02), rather than the divergence of a trained network. Median over 16 trials. The gap of the fixed gauges is a property of the family: their optima are 12–28 nats at $\omega = 2\pi$, and the trained models of fig. 3 lie within 0.04–2.2 nats of them at $\sigma \geq 0.3$. The gauge-corrected families admit 0.02–2.2 nats there, and at $\sigma = 0.05$ they admit 1.8–15 nats against the 5–56 nats that training reaches. Helmholtz-SDE at order one and the learned gauge have the same optimum, since the order-one correction is exact for a linear prior drift; at $\omega = 0$ the symmetric gauge coincides with both.

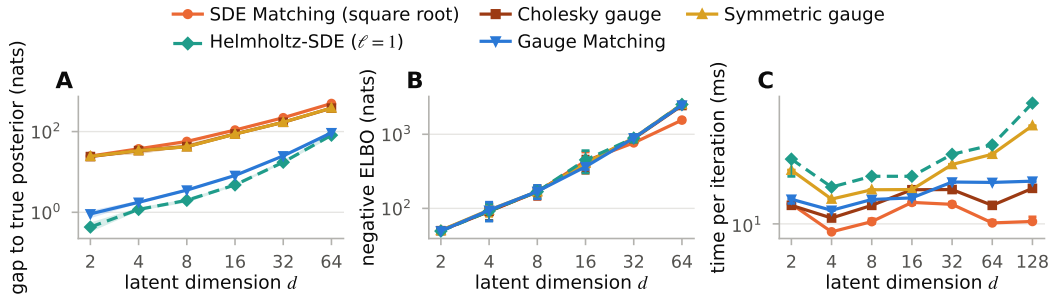

Figure 5: Scaling in the latent dimension on the rotating Ornstein–Uhlenbeck system. *A:* gap to the true posterior process, generative model known (median and interquartile range over 16 trials and 3 seeds). *B:* negative ELBO per trial, generative model learned (mean ± 2 standard errors over 3 seeds). *C:* GPU time per training iteration, models as in B up to $d = 128$ (batch of 256 trials; median and interquartile range over 50 iterations).

eigendecomposition, which is singular for repeated eigenvalues. The two solvers agree to machine precision.

Timing. Figure 5C times one training iteration of the models of Protocol 2 (forward and backward pass, gradient clipping and Adam step, batch of 256 trials) for $d \in \{2, 4, \ldots, 128\}$ on one otherwise idle GPU, as the median and interquartile range over 50 iterations after 5 warm-up iterations with device synchronization, and on a CPU for $d \leq 64$ (10 iterations).

Results. With the generative model known (fig. 5A), the gap of Helmholtz-SDE, whose gauge is exact here, grows from 0.4 nats at $d = 2$ to 81 at $d = 64$ (0.4/1.2/2.0/4.8/17/81 at $d = 2/4/8/16/32/64$) and that of Gauge Matching from 0.9 to 94 (0.9/1.8/3.6/8.2/25/94), while the fixed gauges lie at 24–25 and 381–496 (the Cholesky and symmetric gauges agree at every d, since the symmetric gauge is

nearly zero for a drift that is an isotropic contraction plus a rotation). The part of the gap due to the mean is the same for the two gauge-corrected processes at every d (0.06 at $d = 2$, 63 at $d = 64$; it is the training error of the fit to 16 trials). The part due to the covariance, which the gauge controls, is 0.4–17 for Helmholtz-SDE and 0.9–32 for Gauge Matching against 24–393 for the fixed gauges: the learned gauge removes 89–97% of the fixed gauges' covariance gap at every d, and its norm is within 2% of that of the exact gauge.

With a learned generative model (fig. 5B) the five posterior processes reach the same negative ELBO within seed noise for $d \leq 16$. The fitted prior drift recovers little of the rotation from ten partial observations per trial (the imaginary parts of the eigenvalues of its mean Jacobian are 0.02–1.1 against the true 2π), so it is close to a gradient field, for which every gauge is near-optimal, and the gauge excess of the fixed gauges is only 0.02–0.4 nats per unit time. Pilot runs with $N = 20$ or 40 observations, full observation of the state, or slower rotation ($\omega = \pi/2$, π) at $d = 8$ did not change this.

For $d \geq 32$ the diagonal posterior of SDE Matching reaches a lower negative ELBO than the four full-covariance posteriors, which agree with each other (1554 ± 78 against 2444–2538 at $d = 64$). The difference lies entirely in the reconstruction term: the learned observation matrix of every full-covariance posterior is rank deficient (5–8 of its 32 singular values below 0.05 at $d = 64$, 1–3 of 16 at $d = 32$, none for the diagonal posterior), so these models settle in a local optimum in which a subset of latent directions is not observed and reverts to the prior, which lowers the path-space KL term and raises the reconstruction term. This is common to all four full-covariance gauges and not an effect of the gauge; training three times longer at $d = 64$ (one seed) halves the difference in training loss, and the same collapse appears in the identifiable setting below when the observation model is learned.

Per iteration (fig. 5C), the GPU step is dominated by kernel launches up to $d \approx 32$, so the Sylvester solve costs Helmholtz-SDE and the symmetric gauge a factor of 1.3–1.6 over SDE Matching there and 2.2 (3.2) at $d = 64$ (128), while Gauge Matching costs 1.05–1.5; on the CPU the $\mathcal{O}(d^3)$ solve dominates from $d = 32$ on, with Helmholtz-SDE at 6.4–7.3 times SDE Matching against 1.4–1.7 for Gauge Matching.

Scaling in the Latent Dimension: Identifiable Setting Protocol 2 above measures the five posterior processes on a prior that the data barely identify. To measure them when the rotation is identified, we repeat it at the setting of table 6, with additive noise so that Helmholtz-SDE applies.

Data. The rotating Ornstein–Uhlenbeck process above with $\alpha = 0.5$, $\omega = \pi$ and $g = I$ (angular velocities of at most π, eight observations per period of the fastest plane), $d \in \{8, 16, 32\}$, observed in full: C is a random $d \times d$ orthogonal matrix, c standard normal, $N = 21$ distinct uniformly random observation times per trial on $[0, 5]$, noise $\sigma = 0.3$; 1024 training and 2048 test trials, one system per seed, 3 seeds.

Model. As in Protocol 2, except that the observation model (C, c, σ) is fixed at its true value, as for the Lotka–Volterra system and table 6: the latent basis is then pinned by the data, and the learned prior drift, initial law and posterior process must explain every latent direction. Five posterior processes: SDE Matching, the Cholesky gauge, Helmholtz-SDE, Gauge Matching, and Gauge Matching with a diagonal L_ϕ (appendix D.2); the symmetric gauge is omitted, since it coincides with the Cholesky gauge on this prior (Protocol 1).

Training. $3 \cdot 10^4$ iterations, batches of 256 trials, learning rate 10^{-3}, as in Protocol 2; at $d = 16$ and 32 all five posterior processes were also trained for $6 \cdot 10^4$ iterations.

Metrics. As in Protocol 2. Since the seed-to-seed spread of the negative ELBO is that of the random system (457–489 nats at $d = 16$ for one gauge), we report paired differences to Helmholtz-SDE per seed as the median over the 2048 held-out trials of the per-trial difference (mean and 2 standard errors over seeds). The median is used because a few held-out trials, for which the amortized posterior is far off, have a negative ELBO above 10^4 nats (at most 15 of 2048 per seed, plus one non-finite trial for Gauge Matching): at $3 \cdot 10^4$ iterations this occurs only for the gauge-corrected models at $d = 32$, at $6 \cdot 10^4$ iterations for every full-covariance posterior and on single trials for the two diagonal ones. We also report the rotation rates of the learned drift, the imaginary parts of the eigenvalues of the skew part of its mean Jacobian over posterior samples, against the true $\omega|\lambda_j(K)|$.

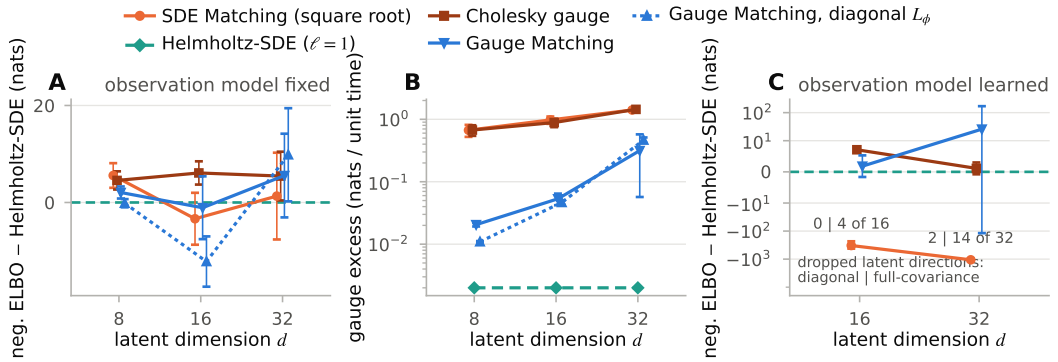

Figure 6: Dimension sweep at an identifiable setting (learned prior drift, full observation, $N = 21$, $\omega = \pi$; see the text). A: negative ELBO minus that of Helmholtz-SDE, paired per seed as the median over held-out trials of the per-trial difference (mean and 2 standard errors over 3 seeds), with the observation model fixed. B: gauge excess (Helmholtz-SDE is zero by construction and drawn at the floor of the axis). C: as A, with the observation model learned: the full-covariance posteriors drop latent directions (numbers: mean over seeds for the diagonal | full-covariance posteriors) and lose by hundreds of nats to the diagonal one.

Results. Figure 6. At $d = 8$ the negative ELBO is 218–219 nats for the two fixed gauges against 212–214 for the three gauge-corrected posterior processes; paired per trial and per seed, SDE Matching is 3–9 nats and the Cholesky gauge 4–7 nats above Helmholtz-SDE on every seed, Gauge Matching is 0.5–3 above it and Gauge Matching with a diagonal L_ϕ within 1.5 of it. The difference is the path-space KL term (112–113 for the fixed gauges against 104–107), of which Gauge Matching gives 3–5 nats back in the reconstruction term. The gauge excess of the fixed gauges is 0.54–0.80 nats per unit time against 0.01–0.02 for the two learned gauges, whose $\|\Omega_\phi\|_F$ matches $\|\Omega^\star\|_F$ within 2%. Unlike in Protocol 2, the fitted prior rotates, at 83–86% of the true rates for the gauge-corrected posteriors and 69–75% for the fixed gauges, which as in table 6 absorb part of the rotation into the drift they are trained with.

At $d = 16$ and 32 the comparison at equal L_ϕ still favors the gauge, and more so at the longer budget: the Cholesky gauge is 1.4–11.6 nats above Helmholtz-SDE at $3 \cdot 10^4$ iterations, and 5–21 above Helmholtz-SDE and 5–30 above Gauge Matching at $6 \cdot 10^4$, on every seed at both d. Gauge Matching with a diagonal L_ϕ is 3–14 nats below SDE Matching at $d = 16$ (on two of three seeds at $3 \cdot 10^4$ iterations, on every seed at $6 \cdot 10^4$; gauge excess 0.04–0.05 against 0.9–1.1), but not consistently at $d = 32$: at $6 \cdot 10^4$ iterations it is 26 nats above on one seed and 4–13 below on the other two, its $\|\Omega_\phi\|_F$ exceeds $\|\Omega^\star\|_F$ by 7–9%, and its gauge head diverges near the end of the interval on one seed.

Across parameterizations, however, the diagonal L_ϕ is fitted faster than a full one. At $3 \cdot 10^4$ iterations SDE Matching is within 8 nats of Helmholtz-SDE at $d = 16$, the sign depending on the seed, Gauge Matching with a diagonal L_ϕ has the lowest seed mean there, and the five models agree within 12 nats at $d = 32$. At $6 \cdot 10^4$ iterations Gauge Matching with a diagonal L_ϕ is the best model at $d = 16$ on every seed, 8–9 nats below Helmholtz-SDE, which SDE Matching and Gauge Matching tie within 5 and 2 nats; at $d = 32$ SDE Matching is 24–27 nats below Helmholtz-SDE and 9–26 below Gauge Matching on every seed, in test and training loss alike, although its gauge excess is 1.7 nats per unit time, 8 nats over the trajectory, against 0.15 for the learned gauge.

Remarks. With the observation model learned instead of fixed, the same sweep gives a different picture (fig. 6C): all four full-covariance posteriors converge to a rank-deficient observation matrix, with 3–5 of 16 ($d = 16$) and 13–15 of 32 ($d = 32$) singular values below 0.05 against 0–2 for the diagonal one, and their reconstruction term is 650 and 1950 nats against 218 and 660; the ordering of the gauges within the full-covariance family is unchanged (Cholesky 7.1 and 1.1 nats above Helmholtz-SDE, paired medians). This is the mechanism behind the lag of the full-covariance posteriors at $d \geq 32$ in Protocol 2, and the reason why the identifiable setting fixes C.

D.2 NONLINEAR DYNAMICS

Noisy Lorenz Attractor **Data.** As in Smith et al. (2026, Appendix D.1): the stochastic Lorenz system with parameters $(10, 28, 8/3)$, additive noise of scale 5 and initial state $\mathcal{N}(0, I)$ is simulated on $[0, 5]$ by Euler–Maruyama with step 10^{-3}; the standardized state is observed every $\Delta t = 0.25$ with Gaussian noise of standard deviation 0.3; 1024 training and 2048 test trials.

Model. The latent SDE has $d = 4$; a Gaussian initial distribution with learned mean and diagonal covariance; the shared prior drift network; the fixed diffusion coefficient $g = I$ for all methods, including SDE Matching, which in Bartosh et al. (2025) uses a state-dependent one; and a linear-Gaussian observation model $\mathcal{N}(x \mid Cz + c, R)$ with learned C and c and fixed R.

Training. $3 \cdot 10^4$ iterations, batches of 256 trials, learning rate 10^{-3}, for all methods.

Metrics. The *negative ELBO* is eq. (6) per test trial. Its initial KL term is 0.5–0.6 nats for every method; its path-space KL term is 57.3–59.6 for the three fixed gauges against 33.5 and 33.7 for Helmholtz-SDE and Gauge Matching, which pay about one nat more in the reconstruction term (19.0 and 18.9 against 17.6–17.7) for a 24–26 nat lower path-space term.

The *latent log-likelihood* follows Smith et al. (2026): writing $y^{(k)}(t_i)$ for the noise-free standardized true state of test trial k at observation time t_i, and m_ϕ, $P_\phi = L_\phi L_\phi^\top$ for the posterior marginal of that trial at t_i, it is $\frac{1}{NK} \sum_{k=1}^{K} \sum_{i=1}^{N} \log \mathcal{N}\big(y^{(k)}(t_i) \mid Cm_\phi(t_i) + c,\, CP_\phi(t_i)C^\top + R\big)$ over the N observation times and $K = 2048$ test trials, with the learned observation model (C, c, R): the log-density of the true state under the posterior predictive.

The *autocorrelation error* is computed from 2048 sample paths of the learned prior process started from the posterior at $t = 0$. For lags τ on a grid of step 0.05 in $[0, 2]$, we form the 3×3 correlation matrix of the observed coordinates between times t and $t + \tau$, average the squared Frobenius distance to the corresponding matrix of the true prior process over t and τ, and report the root mean square over entries. The Monte Carlo floor of this estimator at this sample size is 0.024, and the true dynamics started from the fitted posterior at $t = 0$ score 0.030; the gauge-corrected models, at 0.052–0.056, are therefore within 0.03 of what the estimator can resolve.

The *gauge excess* is defined above. The *training time* quoted in section 5.2 is the wall-clock time to $3 \cdot 10^4$ iterations, averaged over seeds; the five methods shared one GPU, so only ratios are meaningful: relative to SDE Matching, the Cholesky gauge, the symmetric gauge, Helmholtz-SDE and Gauge Matching took 1.06, 1.27, 1.43 and 1.17 times as long. On the Stuart–Landau oscillator at $\kappa = 3$, measured the same way, Helmholtz-SDE took 1.57 times as long as Gauge Matching and Gauge Matching 1.29 times as long as SDE Matching.

Comparison with published numbers. Our negative ELBO values are 6–10 nats above those reported by Smith et al. (2026) for the same methods, uniformly across methods, so the differences between SDE Matching and Helmholtz-SDE agree with theirs. The offset is not overfitting, since training and test values agree to within 1 nat; it is consistent with a difference in batch size or data draw and, for SDE Matching, with our use of the fixed diffusion coefficient.

Robustness to observation spacing and noise. We repeat the experiment for the observation spacings $\Delta t \in \{0.125, 0.25, 0.5\}$, that is $N = 41$, 21 or 11 observations per trial on $[0, 5]$, and the observation noise levels $\sigma \in \{0.15, 0.3, 0.6\}$; $(0.25, 0.3)$ is the protocol of table 2. The same 1024 training and 2048 test trajectories underlie every protocol, so only the sampling of the observations changes; the model, the training and the metrics are those of table 2, with the fixed observation noise $R = \sigma^2 I$ set to the true value, for SDE Matching, the Cholesky gauge, Helmholtz-SDE and Gauge Matching (the symmetric gauge, which tracks the Cholesky gauge in table 2, is omitted), three seeds per protocol.

The negative ELBO is a trimmed mean over the test trials that excludes trials more than ten interquartile ranges above the median; this excludes no trial in most runs and never changes a mean by more than 0.5 nats except where the plain mean is infinite (see the remarks). Table 3 and fig. 7 report the results.

For $\Delta t \leq 0.25$ and $\sigma \leq 0.3$ the fixed gauges trail Helmholtz-SDE by 20–25 nats in every protocol (their *deficit*, the difference of the negative ELBOs), all of it in the path-space KL term ($+25$ to $+29$ nats against -1 to -6 in the reconstruction term), while Gauge Matching is within -0.7 to $+1.8$ nats

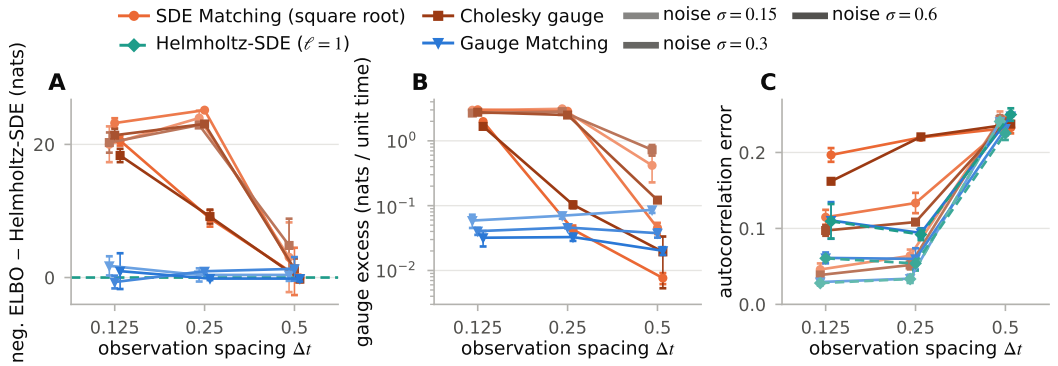

Figure 7: Noisy Lorenz attractor under nine observation protocols: spacing $\Delta t \in \{0.125, 0.25, 0.5\}$ ($N = 41, 21, 11$ observations per trial) and noise $\sigma \in \{0.15, 0.3, 0.6\}$ (light to dark), $d = 4$, the model and training of table 2, three seeds per protocol. A: negative ELBO relative to Helmholtz-SDE, paired over seeds. B: gauge excess (log scale). C: autocorrelation error of the learned prior process. Mean ± 2 standard errors over seeds.

of Helmholtz-SDE in every protocol; the gauge excess of the fixed gauges is 2.5–3.1 nats per unit time and that of Gauge Matching 0.04–0.07, and the norm of the learned gauge matches that of the optimal gauge at its own marginals to within 1%.

The deficit of the fixed gauges does not grow with sparser or noisier observations: it disappears. At $\sigma = 0.6$ it is 18–21 nats for $\Delta t = 0.125$ but 9 nats for $\Delta t = 0.25$, and at $\Delta t = 0.5$ it is -0.5 to $+4.8$ nats, within two standard errors of zero in five of the six cases. The autocorrelation error explains why: it is 0.03–0.06 for the gauge-corrected models and 0.04–0.13 for the fixed gauges where the deficit is large, but 0.23–0.25 for all four posterior processes at $\Delta t = 0.5$, so with eleven observations per trial the fitted prior drift no longer recovers the rotation within the lobes; the optimal gauge at the fixed-gauge marginals shrinks from $\|\Omega^\star\| \approx 12$ to 0.3–4 and their gauge excess to below 0.8 nats per unit time.

Remarks. On rare test trials (1–6 of 2048 in twelve of the 96 new runs) the amortized posterior returns a marginal covariance with one eigenvalue of order 10^{-9}. For Helmholtz-SDE the Sylvester system of eq. (17) is then nearly singular, the gauge it returns has norm 10^6–10^{12}, and the path-space KL term of that trial is 10^8–10^{12} nats, which is why the plain mean of two runs is 10^8 or infinite; the same trials are unremarkable for Gauge Matching, whose Ω_ϕ is a bounded network output. Two training runs diverged, Helmholtz-SDE at $(0.5, 0.3)$ and SDE Matching at $(0.5, 0.6)$, in both cases a posterior standard deviation collapsing within a few iterations, reproducibly for the seed; they were replaced by a further seed. Nothing of this kind occurred for $\Delta t \leq 0.25$.

Known Nonlinear Drift: Stuart–Landau Oscillator To isolate the effect of the expected Jacobian in eq. (17) from the learning of the prior, we take a two-dimensional oscillator with a cubic drift coefficient and fix the generative model at the truth, as in the conjugate setting, so that only the posterior process is learned.

Why the drift must be known. With a learned prior drift, Gauge Matching and Helmholtz-SDE coincide on every system we tried, including van der Pol and Stuart–Landau oscillators: a drift network fitted through posteriors as wide as those encountered here is close to affine over the posterior support, its expected Jacobian equals its Jacobian at the mean to 10^{-3}, and the order-one correction is exact for it. This, and not the degree of the true drift, is why the Lorenz and Lotka–Volterra systems do not separate the two methods.

Which oscillator. With the true drift fixed, the separation depends on the oscillator. For $d = 2$ and isotropic $\Sigma = s^2 I$ the gauge excess of the order-one correction has the closed form $\delta^2 \det P / (2 s^2 \operatorname{tr} P)$, where P is the posterior covariance and $\delta = \Delta J_{21} - \Delta J_{12}$ is twice the off-diagonal entry of the skew-symmetric part of $\Delta J = \mathbb{E}_{q(z)}[\nabla_z h(Z)] - \nabla_z h(m)$, the difference

Table 3: Noisy Lorenz attractor under nine observation protocols (spacing Δt, N observations per trial, noise σ); the row $(0.25, 21, 0.3)$ is the protocol of table 2, here over its first three seeds. Negative ELBO per test trial and gauge excess in nats per unit time, mean ± 2 standard errors over three seeds; the gauge excess of the Cholesky gauge is 10% below the square root's where the deficit is large, and that of Helmholtz-SDE is 0.000–0.005.

			Negative ELBO ↓				Gauge excess ↓	
Δt	N	σ	Sq. root	Cholesky	Helmholtz	Gauge M.	Sq. root	Gauge M.
0.125	41	0.15	45.8 ± 2.1	46.0 ± 1.1	25.7 ± 0.7	27.5 ± 2.1	2.97 ± 0.02	0.059 ± 0.014
0.125	41	0.3	98.2 ± 0.4	96.4 ± 0.8	75.0 ± 0.3	74.3 ± 0.7	3.02 ± 0.09	0.040 ± 0.005
0.125	41	0.6	164.6 ± 0.1	162.4 ± 0.2	144.1 ± 0.8	145.0 ± 2.1	1.98 ± 0.06	0.032 ± 0.009
0.25	21	0.15	60.8 ± 1.1	59.7 ± 0.5	36.8 ± 0.5	37.2 ± 1.4	3.12 ± 0.04	0.070 ± 0.007
0.25	21	0.3	77.8 ± 0.5	75.7 ± 0.3	52.6 ± 0.4	53.5 ± 0.8	2.86 ± 0.10	0.046 ± 0.001
0.25	21	0.6	92.7 ± 0.6	93.0 ± 1.7	83.9 ± 0.7	83.7 ± 0.5	0.04 ± 0.01	0.033 ± 0.004
0.5	11	0.15	49.3 ± 5.2	51.1 ± 3.2	46.3 ± 1.0	46.7 ± 0.9	0.42 ± 0.19	0.086 ± 0.009
0.5	11	0.3	41.6 ± 1.2	42.4 ± 0.2	40.6 ± 1.4	42.4 ± 0.0	0.05 ± 0.01	0.038 ± 0.005
0.5	11	0.6	47.6 ± 0.5	47.9 ± 0.2	48.1 ± 0.5	48.0 ± 0.0	0.01 ± 0.00	0.020 ± 0.003

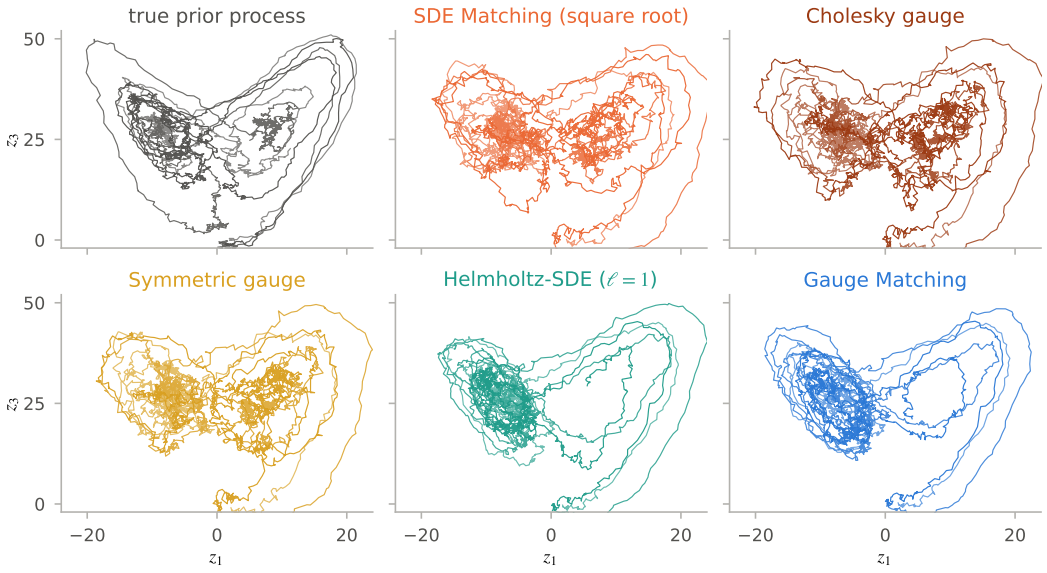

Figure 8: Three samples from the true prior process of the noisy Lorenz attractor and from the prior processes learned with each posterior process. The fixed gauges miss the outward spiral within each lobe of the attractor; the learned gauge and Helmholtz-SDE recover it.

between the expected Jacobian and the Jacobian at the mean; the symmetric part of ΔJ does not enter.

For the van der Pol oscillator $h(z) = (z_2,\ \mu(1 - z_1^2)z_2 - z_1)$ one finds $\delta = -2\mu P_{12}$, which vanishes for the nearly isotropic posteriors of a well-observed limit cycle (in pilot runs with $\mu \in \{2, 3\}$ the excess of Helmholtz-SDE stayed below 0.001 nats), and the Duffing oscillator behaves alike. We therefore use the Stuart–Landau oscillator with shear, $h(z) = (a_0 - |z|^2)z + (\omega + \kappa|z|^2)Jz$, for which $\delta = 4\kappa \operatorname{tr} P$ and the excess is $8\kappa^2 \operatorname{tr} P \det P / s^2$.

Data. $a_0 = \omega = 1$, $\kappa \in \{2, 3\}$, additive noise $s = 0.5$, $z(0) \sim \mathcal{N}(0, I)$, on $[0, 10]$ by Euler–Maruyama with step 10^{-3}; observations of both coordinates with Gaussian noise of standard deviation 0.5 at $t = 0, 1, \ldots, 10$ ($N = 11$); 1024 training and 2048 test trials.

Model. The prior drift, the diffusion coefficient sI, the initial law $\mathcal{N}(0, I)$ and the observation model $C = I$, $c = 0$ are fixed at their true values.

Table 4: Stuart–Landau oscillator with shear $\kappa = 3$, $d = 2$, generative model known, posterior process learned. Columns as in table 2 without the time, with the autocorrelation error over lags in $[0, 4]$; the negative ELBO excludes held-out trials whose path-space KL term exceeds 10^3 nats (one of 5×2048 for Helmholtz-SDE, none otherwise). Mean ± 2 standard errors over 5 seeds; the seed-paired ELBO difference between Gauge Matching and Helmholtz-SDE is 1.6 ± 0.3 nats.

Posterior process	Negative ELBO ↓	Latent LL ↑	Autocorr. error ↓	Gauge excess ↓
SDE Matching (square root)	64.5 ± 0.1	-1.89 ± 0.01	0.054 ± 0.001	0.685 ± 0.020
Cholesky gauge	64.3 ± 0.1	-1.89 ± 0.02	0.053 ± 0.002	0.658 ± 0.015
Symmetric gauge	64.5 ± 0.2	-1.89 ± 0.01	0.054 ± 0.001	0.686 ± 0.017
Helmholtz-SDE ($\ell = 1$)	54.1 ± 0.1	-1.67 ± 0.01	0.043 ± 0.002	0.135 ± 0.006
Gauge Matching	52.5 ± 0.3	-1.69 ± 0.01	0.044 ± 0.004	0.014 ± 0.003
Gauge Matching, diagonal L_ϕ	52.9 ± 0.2	-1.71 ± 0.01	0.045 ± 0.003	0.012 ± 0.003

Posterior. The shared architecture, with a single gauge entry Ω_{12} for Gauge Matching.

Training. $3 \cdot 10^4$ iterations, batches of 256 trials, learning rate 10^{-3}, 5 seeds.

Metrics. Those of table 2 on the 2048 test trials. The negative ELBO is evaluated on a grid of 500 times with 4 noise samples each and excludes the rare test trials whose path-space KL term exceeds 10^3 nats: the cubic drift makes this term explode for a trial when the amortized posterior is badly off for it (one of 5×2048 trials for Helmholtz-SDE at $\kappa = 3$; three in total at $\kappa = 2$). The latent log-likelihood is evaluated at the observation times. The autocorrelation error uses lags in $[0, 4]$ and 2048 prior sample paths started from the posterior at $t = 0$; its Monte Carlo floor is 0.022, and the true dynamics started from the fitted posterior score 0.040–0.044 at $\kappa = 3$. The gauge excess uses the expected Jacobian of the true drift.

Results. At $\kappa = 3$ the mean posterior variance is $P_{11} \approx P_{22} \approx 0.06$, for which the closed form predicts a gauge excess of the order-one correction of 0.12–0.14 nats per unit time; we measure 0.135 ± 0.006, against 0.014 ± 0.003 for Gauge Matching and 0.66–0.69 for the fixed gauges (table 4, fig. 9). The negative ELBO of Gauge Matching is lower than that of Helmholtz-SDE by 1.6 ± 0.3 nats, paired over seeds and in the same direction for all five: it gains 3.0 nats in the path-space KL term and 0.1 in the initial KL term, and gives 1.5 back in the reconstruction term through a slightly wider posterior, whose latent log-likelihood is accordingly 0.025 lower. At $\kappa = 2$ the gauge excess of the order-one correction drops to 0.052 ± 0.004 (0.028 ± 0.003 for Gauge Matching), roughly the factor $(2/3)^2$ that the closed form predicts, and the ELBO difference, 0.1 ± 0.8 nats, is within seed noise.

State-Dependent Noise: Stochastic Lotka–Volterra **Data.** The system is $\mathrm{d}x = (\alpha x - \beta xy)\mathrm{d}t + \sigma x \mathrm{d}W^1$, $\mathrm{d}y = (\delta xy - \gamma y)\mathrm{d}t + \sigma y \mathrm{d}W^2$ with $(\alpha, \beta, \delta, \gamma) = (2/3, 4/3, 1, 1)$, $\sigma = 0.15$ and initial state $(0.9, 0.2)$ perturbed by $\mathcal{N}(0, 0.05^2 I)$, as in Course & Nair (2023a) and Bartosh et al. (2025). Trajectories are simulated by Euler–Maruyama in log coordinates with step 10^{-3} on $[0, 40]$, about five periods of the deterministic cycle through the initial state, which has period 8.0. Observations are the two populations with Gaussian noise of standard deviation 0.1 at $N = 30$ uniformly random times per trial; 2048 training and 512 test trials. In *log coordinates* the same trajectories are transformed to $(\ell_1, \ell_2) = (\log x, \log y)$; by Itô's lemma the diffusion coefficient is then the constant σI and the drift coefficient is $(\alpha - \beta e^{\ell_2} - \sigma^2/2,\ \delta e^{\ell_1} - \gamma - \sigma^2/2)$, and the observation noise is added to the log-populations.

Model. The latent SDE has $d = 2$ and the shared prior drift network. The observation model is fixed at the identity, $C = I$ and $c = 0$, with the true noise level, so that the latent state is the (log-)population vector. In log coordinates the diffusion coefficient is fixed at its true value σI, as $g = I$ is for the Lorenz system. In the original coordinates it is learned and diagonal, $g_{ii}(z) = \sqrt{z_i^2 + \epsilon^2}\, s_i(z_i)$ with $\epsilon = 0.1$ and s_i the sigmoid-bounded network of Bartosh et al. (2025) (one hidden layer of width 100, softplus activations, output in $(0, 1)$, initialized at 0.3). The noise thus scales with the population as the true σz_i does, and the floor ϵ, of the size of the observation noise, keeps the path-space KL term bounded where the Gaussian posterior puts mass at non-positive populations; without it training diverges. Helmholtz-SDE and the symmetric gauge are not applicable in the original coordinates.

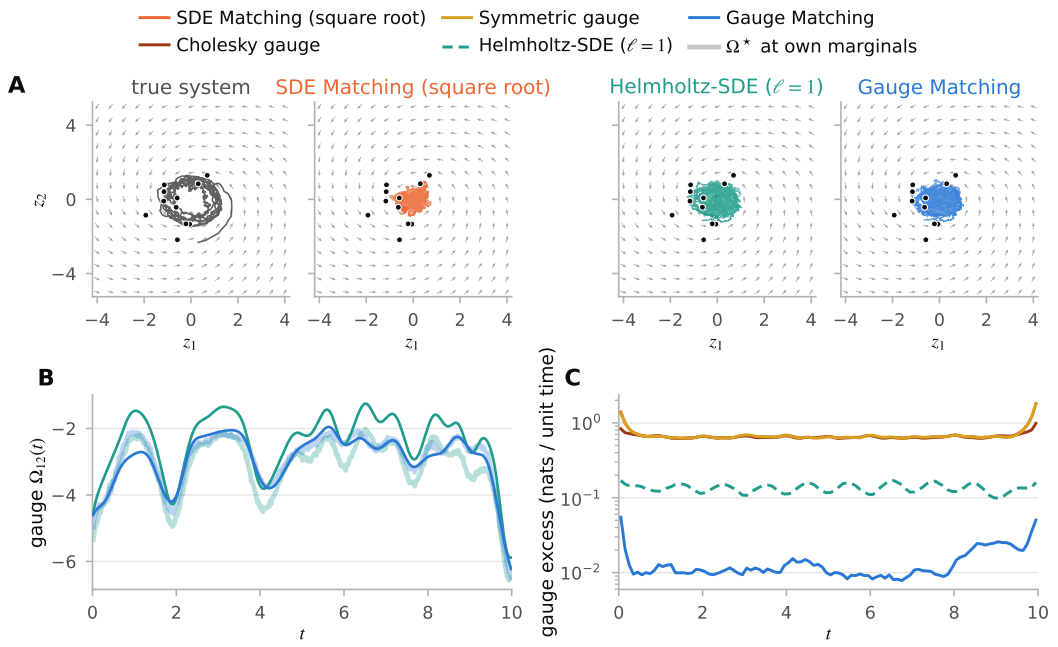

Figure 9: Stuart–Landau oscillator with shear $\kappa = 3$, known generative model. *A:* the true drift field with the trajectory and observations (dots) of one test trial, and three sample paths of the true prior process started from the initial marginal $q_\phi(0, z \mid X)$ fitted by SDE Matching, Helmholtz-SDE and Gauge Matching; since the generative model is fixed at the truth, the panels differ only through the fitted initial marginal. *B:* the gauge $\Omega_{12}(t)$ that each of the two gauge-corrected methods uses for this trial, against the optimal gauge eq. (17) at its own marginals (thick, light). *C:* gauge excess $\mathcal{J}(\Omega) - \mathcal{J}(\Omega^\star)$ over time, averaged over 64 test trials and all seeds, for the five posterior processes (log scale): the order-one correction is off by the closed-form amount $8\kappa^2 \operatorname{tr} P \det P/s^2 \approx 0.13$ nats per unit time, the learned gauge by a tenth of it.

Posterior. The shared architecture; since $T = 40$, the posterior MLP receives t/T alongside t.

Training. $3 \cdot 10^4$ iterations, batches of 128 trials, learning rate 10^{-3}, 10 seeds.

Metrics. All three metrics are computed from 1024 sample paths of the learned prior process, started from the learned initial distribution and simulated with step $2 \cdot 10^{-3}$ over a horizon of 80 time units, about eight cycles. In the original coordinates the learned diffusion coefficient keeps the populations positive only approximately, and sample paths that leave the box $[-1, 20]^2$ are excluded (7–8% on average, up to 31% for one seed, most of them beyond the training horizon $T = 40$; clipping the populations at zero instead changes no reported value).

The *period* is the location of the first nonzero maximum of the autocorrelation of the prey coordinate, averaged over paths. The *lag* is the location, within one period, of the maximum of the cross-correlation between predator and prey. Both are reported as relative errors against the values for the true system computed by the same procedure, which are period 10.2 and lag 2.9 in log coordinates and 9.8 and 1.5 in the original coordinates; these periods exceed the deterministic 8.0 because the process noise drives trajectories onto wider orbits of the conserved-quantity family, whose periods are longer. The *autocorrelation error* is computed as for the Lorenz system, with lags in $[0, 10]$.

Results. Table 5 reports the metrics. In log coordinates the gauge excess is 0.215 ± 0.006 for the fixed gauge, 0.000 for Helmholtz-SDE and 0.004 ± 0.001 for Gauge Matching: the posterior covariance is small relative to the curvature of the exponential drift terms, so the Jacobian at the mean already gives the optimal gauge, and the two gauge-corrected methods agree within seed noise. In the original coordinates Gauge Matching improves all three metrics over the fixed gauge on every seed, with a

Table 5: Stochastic Lotka–Volterra system with sparse noisy observations, $d = 2$, in log coordinates, where the diffusion coefficient is state-independent and fixed at its true value, and in the original coordinates, where it is multiplicative and learned, and Helmholtz-SDE does not apply. Columns: relative error of the period of the cycle; relative error of the lag of the predator behind the prey; autocorrelation error as in table 2, lags in $[0, 10]$, all for the learned prior process against the true system; gauge excess in nats per unit time as in table 2, in the original coordinates from the state-dependent linear system of appendix C.2 estimated by Monte Carlo (10^3 samples at 100 times on 64 test trials). Mean ± 2 standard errors over 10 seeds.

Coord.	Posterior process	Period err. ↓	Lag err. ↓	Autocorr. err. ↓	Gauge excess ↓
log	SDE Matching (square root)	0.12 ± 0.04	0.09 ± 0.05	0.119 ± 0.019	0.215 ± 0.006
	Helmholtz-SDE ($\ell = 1$)	0.09 ± 0.03	0.07 ± 0.03	0.100 ± 0.015	0.000 ± 0.000
	Gauge Matching	0.07 ± 0.02	0.05 ± 0.03	0.093 ± 0.018	0.004 ± 0.001
original	SDE Matching (square root)	0.22 ± 0.05	0.42 ± 0.03	0.191 ± 0.022	0.081 ± 0.002
	Cholesky gauge	0.22 ± 0.05	0.44 ± 0.04	0.192 ± 0.024	0.082 ± 0.004
	Gauge Matching, diagonal L_ϕ	0.13 ± 0.05	0.35 ± 0.03	0.139 ± 0.022	0.002 ± 0.001
	Gauge Matching	0.13 ± 0.03	0.34 ± 0.02	0.140 ± 0.015	0.003 ± 0.001

gauge excess of 0.003 against 0.081 nats per unit time; the Cholesky gauge and Gauge Matching with a diagonal L_ϕ (appendix D.2) confirm that the gain is the gauge.

Remarks.

- An independent draw of the true system scores 0.01–0.02 on all three metrics, which is the noise floor of the estimators.
- The negative ELBO on the training trials, evaluated with the test estimator, lies about 30 nats below the test value for every method (about 45 with 512 training trials), so the held-out metrics are not limited by overfitting.
- In the original coordinates the learned relative noise level s_i is 0.25–0.30 at population 1, against the true $\sigma = 0.15$, which widens the sample paths relative to the true system (fig. 10).

State-Dependent Noise in Higher Dimension To test the state-dependent case beyond $d = 2$, where the optimal gauge is not a Jacobian but the solution of the $d(d-1)/2$-dimensional linear system of appendix C.2, we use the rotating Ornstein–Uhlenbeck process of the dimension sweep with a multiplicative diffusion coefficient.

Data. $\mathrm{d}Z = (-\alpha I + \omega K)Z\mathrm{d}t + \sigma \operatorname{diag}\big(\sqrt{1 + Z_i^2}\big)\mathrm{d}W$ in $d = 8$ with $\alpha = 0.5$, $\omega = \pi$, $\sigma = 0.6$ and $K = (G - G^\top)/\|G - G^\top\|_2$ for a standard normal G, which for the system used rotates four planes at angular velocities $\pi|\lambda_j(K)| = 3.14, 1.70, 0.69, 0.02$. Since $Z^\top K Z = 0$, the process is stationary for $\sigma^2 < 2\alpha$ with second moment $s_0^2 = \sigma^2/(2\alpha - \sigma^2) = 0.5625$ per coordinate, and $Z_0 \sim \mathcal{N}(0, s_0^2 I)$. Trajectories are simulated on $[0, 5]$ by Euler–Maruyama with step 10^{-3}; all d coordinates are observed through the identity every $\Delta t = 0.25$ ($N = 21$) with Gaussian noise of standard deviation 0.3; 1024 training, 2048 test trials and 4096 independent reference trajectories, one system K shared by all methods and seeds.

Model. Two protocols. *Known generative model:* drift Az, diffusion coefficient $\sigma \operatorname{diag}(\sqrt{1 + z_i^2})$, initial law $\mathcal{N}(0, s_0^2 I)$ and observation model $C = I$, $c = 0$ fixed at their true values, so that only the posterior process is learned, as for the Stuart–Landau oscillator. *Learned generative model:* the shared prior drift network, the state-dependent diagonal diffusion network of Bartosh et al. (2025), $g_{ii}(z) = 2\,\mathrm{sigmoid}(\mathrm{MLP}_i(z_i))$ with one hidden layer of width 100 and softplus activations, a Gaussian initial law with learned mean and diagonal covariance, and the observation model fixed at the identity with the true noise level, as for the Lotka–Volterra system. Helmholtz-SDE and the symmetric gauge do not apply in either protocol.

Posterior. The shared architecture; four posterior processes: SDE Matching (diagonal L_ϕ), the Cholesky gauge (full L_ϕ), Gauge Matching (full L_ϕ and the 28 entries of Ω_ϕ) and Gauge Matching with a diagonal L_ϕ (appendix D.2).

Training. $3 \cdot 10^4$ iterations, batches of 256 trials, learning rate 10^{-3}, 5 seeds.

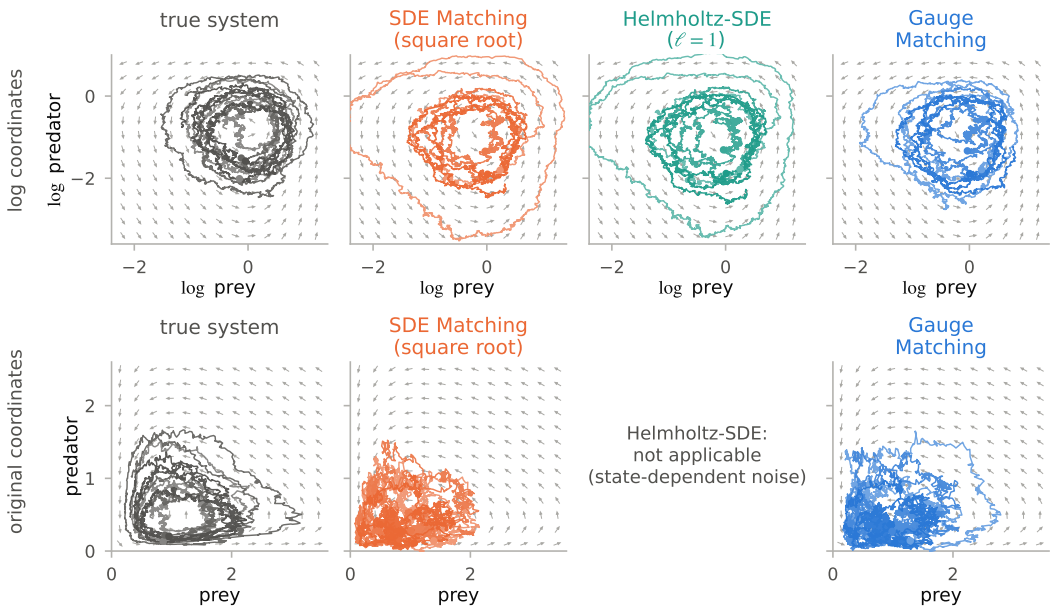

Figure 10: Learned prior dynamics of the stochastic Lotka–Volterra system. Drift field and three sample paths of the true system and of the models learned with each posterior process, in log coordinates with the diffusion coefficient fixed at its true value (top) and in the original coordinates with a learned state-dependent diffusion coefficient (bottom). In log coordinates all three posterior processes learn a drift field that sustains the cycle. In the original coordinates the learned multiplicative noise is larger than the true one, which widens the sample paths, and the cycle of the fixed gauge is smaller and less regular than that of the learned gauge.

Metrics. Negative ELBO and latent log-likelihood as in table 2; the gauge excess with $\Omega^\star$ from the linear system $Q\gamma = b$ of appendix C.2, whose coefficients and $\mathcal{J}$ are estimated from the same Monte Carlo samples as above, the residual r_0 including the divergence term $\frac{1}{2}\nabla \cdot \Sigma$ of eq. (15). For the learned generative model also the autocorrelation error of 2048 prior sample paths over lags in $[0, 2]$ against the reference trajectories, the relative root-mean-square error of the learned diffusion coefficient against $\sigma\sqrt{1 + z_i^2}$ over 10^4 stationary states, and the rotation rates of the learned drift, the imaginary parts of the eigenvalues of its Jacobian averaged over stationary states, against $\pi|\lambda_j(K)|$.

Results. Table 6. With the generative model known, Gauge Matching gains 6.6 ± 0.8 nats of negative ELBO over SDE Matching and 6.9 ± 0.6 over the Cholesky gauge, paired over seeds and on every seed. The gain is in the path-space KL term (94–95 for the fixed gauges against 84–85 for the learned ones), of which the gauge-corrected models give 1–2 nats back in the reconstruction term (53–54 against 55) through a slightly wider posterior, whose latent log-likelihood is accordingly 0.05 lower. The gauge excess of the fixed gauges is 1.29–1.40 nats per unit time against 0.041–0.049 for the learned ones, whose $\|\Omega_\phi\|_F = 5.2$–5.3 matches $\|\Omega^\star\|_F = 5.3$ at their own marginals; the optimal gauge at the fixed gauges' marginals has the same norm, so the fixed gauges miss a rotation of that size.

With the generative model learned, the picture is the same at a smaller scale: a paired gain of 4.8 ± 0.3 nats over SDE Matching and 5.1 ± 0.5 over the Cholesky gauge (8 in the path-space term, 3 given back in the reconstruction term), and a gauge excess of 0.85–0.88 against 0.040–0.046. The learned dynamics improve with the gauge: the autocorrelation error of the prior process drops from 0.072 to 0.040 (Monte Carlo floor 0.029), and the rotation rates of the learned drift are $2.94, 1.60, 0.61$ against the true $3.14, 1.70, 0.69$ for the learned gauges but 2.55–$2.58, 1.41, 0.54$ for the fixed ones, which absorb part of the rotation into a more contractive drift, as in the dimension sweep. The learned diffusion coefficient is the same for all four posterior processes and only approximately right (relative root-mean-square error 0.15): the sigmoid network captures the level of the noise at $z_i = 0$ but

Table 6: Rotating Ornstein–Uhlenbeck process in $d = 8$ with multiplicative noise, where Helmholtz-SDE and the symmetric gauge do not apply. Top: generative model fixed at the truth, posterior process learned. Bottom: prior drift, diffusion coefficient and initial law learned jointly with the posterior process. Columns as in table 2 without the time, with the gauge excess evaluated with the state-dependent optimal gauge of appendix C.2 and, for the learned model, the autocorrelation error over lags in $[0, 2]$; the negative ELBO excludes held-out trials above 10^3 nats (0–3 of 2048 per seed). Mean ± 2 standard errors over 5 seeds.

Model	Posterior process	Neg. ELBO ↓	Latent LL ↑	Autocorr. err. ↓	Gauge excess ↓
known	SDE Matching (square root)	154.8 ± 0.8	-1.03 ± 0.04	—	1.40 ± 0.02
	Cholesky gauge	155.1 ± 0.8	-1.04 ± 0.04	—	1.29 ± 0.02
	Gauge Matching	148.2 ± 0.8	-1.09 ± 0.03	—	0.049 ± 0.002
	Gauge Matching, diagonal L_ϕ	147.3 ± 0.5	-1.08 ± 0.03	—	0.041 ± 0.001
learned	SDE Matching (square root)	150.5 ± 0.5	-1.05 ± 0.02	0.072 ± 0.003	0.88 ± 0.02
	Cholesky gauge	150.8 ± 0.2	-1.06 ± 0.01	0.071 ± 0.003	0.85 ± 0.02
	Gauge Matching	145.7 ± 0.3	-1.06 ± 0.01	0.040 ± 0.002	0.046 ± 0.001
	Gauge Matching, diagonal L_ϕ	145.4 ± 0.2	-1.06 ± 0.02	0.040 ± 0.002	0.040 ± 0.002

underestimates its growth with $|z_i|$, equally for every gauge. Diagonal and full L_ϕ agree within a nat in both protocols, as for the Stuart–Landau and Lotka–Volterra systems (appendix D.2).

Remarks.

- The negative ELBO excludes test trials whose total exceeds 10^3 nats (the criterion of table 4, applied to the total instead of the path-space KL term): with $d = 8$ and a wide posterior, the amortized posterior is far off for 0–3 of the 2048 test trials per seed (one trial of the Cholesky gauge with a learned model at $9 \cdot 10^4$ nats), which dominates the plain mean without changing the ranking; the plain means are 158.5, 156.3, 149.0, 148.4 (known) and 150.5, 159.7, 145.7, 146.1 (learned) in the row order of table 6. The path-space and reconstruction terms quoted above are plain means and sum to the plain negative ELBO.
- Since the rotation is fully observed ($C = I$, eight observations per period of the fastest plane), the learned prior recovers most of it, unlike in the partially observed dimension sweep; the gain over the fixed gauges is smaller than with the known model (4.8 against 6.6 nats) because the learned drift co-adapts to the gauge it is trained with.
- The state-dependent optimal gauge is never computed during training, only for the gauge-excess metric.

Ablation: Gauge versus Covariance Gauge Matching differs from SDE Matching in two respects: the covariance factor L_ϕ is a full lower-triangular matrix rather than a diagonal one, and the gauge Ω_ϕ is learned rather than fixed at zero. To attribute the improvement, we add the missing corner of the 2×2 design, *Gauge Matching with a diagonal* L_ϕ: the marginals of SDE Matching (d means and d positive standard deviations) together with the $d(d-1)/2$ entries of a learned skew-symmetric Ω_ϕ, all read off the same posterior network; with L_ϕ diagonal, the triangular solves in eq. (15) are divisions. The Cholesky gauge of the main tables is the other corner, a full L_ϕ with $\Omega \equiv 0$.

Data, Model, Training. Identical to the rows the variant joins: the Lorenz protocol of table 2, the Stuart–Landau protocol of table 4 ($\kappa = 3$), the Lotka–Volterra protocol of table 5 in the original coordinates, the two protocols of table 6, the identifiable dimension sweep of fig. 6, and the conjugate setting at $\sigma = 0.3$, $\omega = 2\pi$, $N \in \{5, 10, 20, 50\}$ with the $5 \cdot 10^4$-iteration budget of the two gauge-corrected models. The variant has the parameter count of the Cholesky gauge (54 141 at $d = 4$ against 53 535 for SDE Matching and 54 747 for Gauge Matching).

Results. On the Lorenz system the variant reaches a negative ELBO of 55.3 ± 1.9 nats (one seed at 58.8, the others 53.5–55.6), against 77.9 for SDE Matching, 75.5 for the Cholesky gauge and 53.2 for Gauge Matching: of the 24.7 nats between SDE Matching and Gauge Matching, the gauge accounts for 22.5 and the covariance shape for 2.2 ± 1.8, paired over seeds and entirely in the path-space KL term. Its gauge excess is 0.022 ± 0.001. On the Stuart–Landau oscillator the variant is within 0.4 ± 0.2 nats of Gauge Matching, the difference lying in the reconstruction term, with a gauge excess of 0.012 ± 0.003, while the Cholesky gauge stays with the fixed gauges at 64.3. In

the original coordinates of the Lotka–Volterra system, where the diffusion coefficient is learned and state-dependent, the Cholesky gauge scores like SDE Matching and the variant like Gauge Matching on all three metrics (paired differences within 0.02 over 10 seeds; table 5), the variant beating the Cholesky gauge on every seed.

In the conjugate setting the picture differs: the gap of the variant is 2.4, 2.6, 2.9 and 8.6 nats at $N = 5, 10, 20, 50$, against 0.8, 0.8, 1.2, 3.9 for Gauge Matching and 19–26 for SDE Matching; the mean part of the gap is the same for the two learned-gauge models, and the whole difference is in the covariance part (2.3–7.9 against 0.7–3.0).

Remarks. The smoother covariance of the rotating Ornstein–Uhlenbeck process is anisotropic in the fixed frame, and a diagonal L_ϕ can rotate its samples but not its covariance, which is why the gap is largest in the conjugate setting and at large N, where the posterior is narrowest and most anisotropic; on the nonlinear systems the learned prior and observation model absorb most of the anisotropy. The gauge head is insensitive to its initialization and learning rate: on the Lorenz system, initializing it at the library default instead of zero, or training it with a learning rate of 0.3 or 3 times that of the rest of the network, changes the negative ELBO of Gauge Matching by $+0.3$ to $+0.6$ nats (3 seeds, paired, within 2 standard errors of zero) and leaves the gauge excess at 0.045–0.049 and $\|\Omega_\phi\|_F$ within 0.5% of $\|\Omega^\star\|_F$.

Motion Capture Data. We follow Li et al. (2020) and Bartosh et al. (2025) exactly: the 50-dimensional recordings of subject 35 of Gan et al. (2015) with the preprocessing of Wang et al. (2008); 16 training, 3 validation and 4 test sequences of 300 observations each; $d = 6$; and the architectures and hyperparameters of Li et al. (2020).

Training. 10 random seeds instead of the default 5.

Metrics. The test mean squared error is defined as in Bartosh et al. (2025, Table 2); Gauge Matching reaches 4.52 ± 0.29 (mean $\pm$ standard deviation over the 10 seeds), while the values quoted for SDE Matching (4.50 ± 0.32) and the latent SDE of Li et al. (2020) (4.03 ± 0.20) are those reported by Bartosh et al. (2025) under the same protocol.

Remarks. With 300 nearly noise-free observations per sequence the posterior concentrates on the trajectory and the path-space KL term is small for every gauge, so the gauge has little to contribute; the experiment checks that the additional $d(d-1)/2 = 15$ gauge outputs cost nothing in accuracy.