VEST: VARIATIONAL DRIFT ESTIMATION FROM NOISY OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Nonparametric drift estimation for stochastic differential equations from multiple independent trajectories with additive observation noise remains challenging in statistics and machine learning. However, direct regression on noisy increments amplifies measurement error, whereas regression on smoothed trajectories treated as exact discards uncertainty and dependence between consecutive latent states. Thus, we propose Variational Drift Estimation from Noisy Observations (VEST), a method that alternates between Gaussian Markov smoothing and neural drift learning. An iterated extended Kalman smoother proposes updates to the Gaussian approximation, while the drift is updated by regressing on conditional targets derived from its adjacent state moments. An acceptance rule based on a shared variational objective determines whether each proposed update is accepted. Moreover, we give theoretical guarantees on the discretization error of the variational objective and the convergence of the algorithm. Experiments show that VEST outperforms existing related methods.

1 INTRODUCTION

The drift function of a stochastic differential equation (SDE) describes the local mean evolution of the underlying process. Estimating this function from observations helps characterize unknown dynamics, with applications in finance (Takamizawa, 2008), cell biology (Yeo et al., 2021), and climate science (Ditlevsen & Ditlevsen, 2023). Parametric approaches assume a prescribed functional form for the drift and estimate its parameters (Aït-Sahalia, 2002). When a suitable functional form is difficult to specify, nonparametric approaches allow the drift to be learned more flexibly from observed trajectories (Ruttor et al., 2013).

Existing nonparametric approaches include kernel-based (Dalalyan, 2005) and spectral (Gobet et al., 2004) estimators for a single ergodic trajectory observed over a long horizon. For multiple independent trajectories observed over a fixed horizon, Comte & Genon-Catalot (2020) developed projection estimators. Neural network drift estimators have also been studied for discretely observed long ergodic trajectories (Oga & Koike, 2024) and for high-frequency observations of independent paths (Zhao et al., 2026). In practice, however, measurements may be corrupted by observation noise.

Approaches to drift estimation from noisy observations include local averaging or deconvolution, particle methods, and variational inference, among others. Local averaging reduces noise by averaging observations over short time intervals (Schmisser, 2011), while deconvolution uses the noise distribution to correct for measurement error in drift estimation (El Kolei & Navarro, 2022). Particle methods approximate latent trajectory distributions using weighted samples and use these samples to update the drift. For example, Ganguly et al. (2025) combine sequential Monte Carlo with expectation–maximization for drift estimation. Variational methods approximate the latent trajectory posterior with a tractable distribution and optimize this approximation together with the drift model. SING (Hu et al., 2025) combines natural gradient variational inference with drift learning in latent SDE models and its formulation assumes a constant diffusion coefficient.

In this work, we study nonparametric drift estimation from multiple independent trajectories with additive Gaussian noise. We allow the known diffusion coefficient to depend on the state and focus on incorporating posterior uncertainty into neural drift regression. To this end, we propose VEST (Figure 1), a variational method for neural drift estimation from noisy trajectories. The method

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.

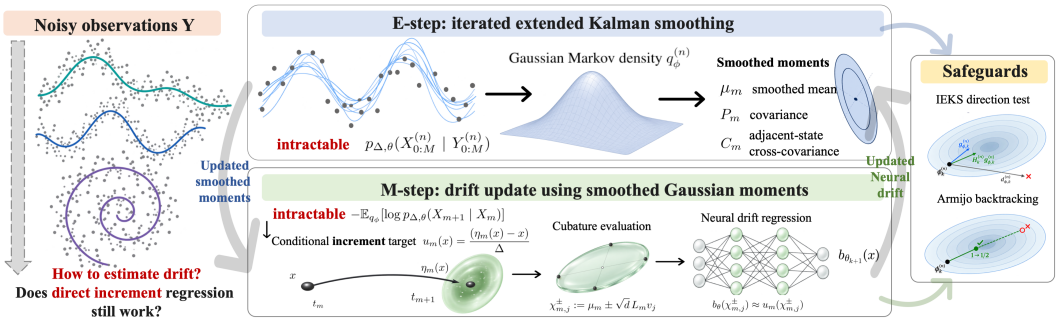


Figure 1: Overview of VEST. The method alternates Gaussian smoothing with neural drift regression, using pairwise moments to define conditional increment targets and cubature points.

incorporates pairwise Gaussian smoothing uncertainty into drift regression and evaluates Gaussian and neural updates using a common cubature variational objective. Our contributions are threefold:

- **A variational formulation of posterior-increment regression.** We derive conditional posterior-increment targets from pairwise Gaussian smoothing marginals, retaining the dependence between adjacent latent states. For a fixed smoothing distribution, minimizing the resulting cubature regression loss is equivalent to maximizing the common objective over the drift parameters.
- **A particle-free computational scheme combining Gaussian smoothing and deterministic cubature.** We use IEKS to construct Gaussian proposals and spherical-radial cubature to integrate the regression loss over current-state uncertainty. The drift update incorporates marginal and cross-covariance information without sampling latent posterior paths.
- **Convergence guarantees for safeguarded updates.** We apply acceptance tests based on the common objective to Gaussian proposals and neural gradient steps. For the ideal algorithm under the stated smoothness, interior compactness, and exact-gradient assumptions, we establish monotone objective ascent, vanishing gradients, stationary accumulation points, and an $O(1/K)$ bound on the minimum squared gradient norm over the first K iterations. Full-sequence convergence holds under the Kurdyka–Łojasiewicz property.

Notation. For vectors, $\|\cdot\| = \|\cdot\|_2$ denotes the Euclidean norm. For symmetric matrices, $A \succeq B$ means that $A - B$ is positive semidefinite, and $A \succ 0$ means that A is positive definite. For $W \succ 0$, write $\|v\|_W^2 = v^\top W v$. We write $z = (\phi, \theta)$, with $\theta \in \mathbb{R}^p$, $z_k = (\phi_k, \theta_k)$, and $\bar{z}_k = (\phi_{k+1}, \theta_k)$, the intermediate iterate after the Gaussian update. For probability measures P, Q on $C([0,T]; \mathbb{R}^d)$, let $W_{2,\infty}(P,Q) := \inf_{\gamma \in \Pi(P,Q)} \left(\mathbb{E}_\gamma \sup_{t \le T} \|X_t - Z_t\|^2\right)^{1/2}$, where $\Pi(P,Q)$ is the set of couplings.

2 Problem Formulation

Consider a continuous SDE over a fixed time horizon $[0,T]$. Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a complete filtered probability space satisfying the usual conditions on which a standard d-dimensional Brownian motion W is defined. We consider the following classical time-homogeneous SDE:

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t,\ X_0 \sim \mathcal{N}(m_0, P_0), \tag{1}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ is the unknown drift function, $\sigma : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ is the known diffusion coefficient, $m_0 \in \mathbb{R}^d$ and $P_0 \in \mathbb{S}^d \succ 0$ denote the known mean and covariance matrix of X_0, respectively. We assume that b and σ are locally Lipschitz continuous and satisfy a linear growth condition. Then the SDE admits a unique strong solution on $[0,T]$. Define the diffusion covariance matrix by $a(x) := \sigma(x)\sigma(x)^\top$, and assume that there exist constants $0 < \underline{a} \le \overline{a} < \infty$ such that $\underline{a} I_d \preceq a(x) \preceq \overline{a} I_d$ for all $x \in \mathbb{R}^d$.

A classical setting for drift estimation assumes discrete observations of the state process X (Gobet et al., 2004). In practical applications, however, X is not directly observable. Instead, on the time

grid $t_m = m\Delta$, $\Delta = T/M$, $m = 0, \dots, M$, we observe

$$Y_{t_m} = X_{t_m} + \varepsilon_m,\ \varepsilon_m \sim \mathcal{N}(0, R),$$

where R is a known positive definite covariance matrix. Suppose that N independent noisy trajectories are available and the observation errors are independent across times and trajectories and are independent of all latent processes. For the n-th trajectory, denote $Y_m^{(n)} := Y_{t_m}^{(n)}$ and $Y_{0:M}^{(n)} := (Y_0^{(n)}, \dots, Y_M^{(n)})$. The observed dataset is

$$\mathcal{D}_N := \left\{ Y_{0:M}^{(1)},\ Y_{0:M}^{(2)},\ \cdots,\ Y_{0:M}^{(N)} \right\}.$$

Our objective is to estimate the unknown drift function b from $\mathcal{D}_N$.

When the states are directly observed, a standard approach regresses normalized state increments on their preceding states (Zhao et al., 2026). For the n-th trajectory, we have the exact identity

$$\mathbb{E}\left[\frac{X_{t_{m+1}}^{(n)} - X_{t_m}^{(n)}}{\Delta} \,\middle|\, X_{t_m}^{(n)} = x \right] = \frac{1}{\Delta} \int_{t_m}^{t_{m+1}} \mathbb{E}\left[b(X_s^{(n)}) \,\middle|\, X_{t_m}^{(n)} = x \right] \mathrm{d}s.$$

This conditional mean converges to $b(x)$ as $\Delta \to 0$. Applying the same regression to the noisy observations,

$$\mathbb{E}\left[\frac{Y_{m+1}^{(n)} - Y_m^{(n)}}{\Delta} \,\middle|\, Y_m^{(n)} = y \right] = \frac{1}{\Delta} \int_{t_m}^{t_{m+1}} \mathbb{E}\left[b(X_s^{(n)}) \,\middle|\, Y_m^{(n)} = y \right] \mathrm{d}s - \frac{1}{\Delta} \mathbb{E}\left[\varepsilon_m^{(n)} \,\middle|\, Y_m^{(n)} = y \right].$$

The first term averages the drift over latent states conditional on the noisy observation $Y_m^{(n)} = y$. The second term reflects the conditional mean of the current observation noise, which need not vanish and is scaled by Δ^{-1}. Thus, for fixed observation-noise covariance R, the regression target does not generally converge to $b(y)$ as $\Delta \to 0$. This motivates incorporating latent-state uncertainty into the drift update.

To this end, we formulate a marginal-likelihood objective based on a discretized latent-state model. We estimate the drift by a neural network $b_\theta : \mathbb{R}^d \to \mathbb{R}^d$. We use the Euler–Maruyama approximation (Särkkä & Svensson, 2023, Section 4.3) to obtain explicit Gaussian transition densities for constructing the observation likelihood. Let X_m denote the latent state at observation time t_m. Its one-step transition is

$$X_{m+1} = X_m + \Delta b_\theta(X_m) + w_m,\ w_m \mid X_m \sim \mathcal{N}(0, \Delta a(X_m)). \tag{2}$$

Let $p_{\Delta,\theta}(x_{m+1} \mid x_m)$ denote the Gaussian transition density specified by (2), and let $p_R(y_m \mid x_m) = \mathcal{N}(y_m; x_m, R)$ denote the observation density. Under this discretized model, the marginal likelihood of an observed trajectory $y_{0:M}$ is obtained by integrating out the latent states:

$$p_{\Delta,\theta}(y_{0:M}) = \int_{\mathbb{R}^{d(M+1)}} \mu_0(x_0) \prod_{m=0}^{M-1} p_{\Delta,\theta}(x_{m+1} \mid x_m) \times \prod_{m=0}^{M} p_R(y_m \mid x_m)\, \mathrm{d}x_{0:M}, \tag{3}$$

where $\mu_0(x_0) = \mathcal{N}(x_0; m_0, P_0)$ is the initial density. For N independent observed trajectories, we consider the regularized marginal-likelihood loss function

$$\mathcal{J}_\Delta(\theta) := -\frac{1}{N} \sum_{n=1}^{N} \log p_{\Delta,\theta}\left(Y_{0:M}^{(n)} \right) + \lambda \mathcal{R}(\theta),$$

where $\lambda \geq 0$ and $\mathcal{R}(\theta)$ regularizes the learned drift. However, evaluating (3) exactly is generally intractable, as it requires integration over the latent path. In the next section, we develop a variational approximation to this objective, alternating Gaussian state inference and neural drift updates.

3 Variational EM for Drift Estimation

The latent state posterior depends on the unknown drift, while estimating the drift requires information about the latent states. We address this coupling through a generalized EM algorithm that alternates between iterated Kalman smoothing and neural drift updates. Two variants of this scheme, VEST-NM and VEST-LR, are described in Appendix C.

3.1 GAUSSIAN VARIATIONAL OBJECTIVE

For each trajectory n, we use a Gaussian Markov density $q_\phi^{(n)}$ to approximate the latent-path posterior $p_{\Delta,\theta}(X_{0:M}^{(n)} \mid Y_{0:M}^{(n)})$. Let $\phi^{(n)}$ denote its variational parameters and set $\phi = (\phi^{(1)}, \dots, \phi^{(N)})$. By Jensen's inequality,

$$\mathcal{F}_\Delta(\phi, \theta) := \frac{1}{N} \sum_{n=1}^{N} \left\{ \mathbb{E}_{q_\phi^{(n)}} \left[\log \mu_0(X_0^{(n)}) + \sum_{m=0}^{M-1} \log p_{\Delta,\theta}(X_{m+1}^{(n)} \mid X_m^{(n)}) + \sum_{m=0}^{M} \log p_R(Y_m^{(n)} \mid X_m^{(n)}) \right] + \mathcal{H}(q_\phi^{(n)}) \right\} - \lambda \mathcal{R}(\theta) \le -\mathcal{J}_\Delta(\theta), \tag{4}$$

where $\mathcal{H}(q) := -\mathbb{E}_q[\log q(X_{0:M})]$ is the differential entropy. We seek to maximize this regularized evidence lower bound (ELBO) over ϕ and θ. To optimize the variational objective in (4), we alternate between updating the Gaussian Markov distributions $q_\phi^{(n)}$ and the drift parameter θ. For fixed θ, each $q_\phi^{(n)}$ approximates the latent-path posterior $p_{\Delta,\theta}(X_{0:M}^{(n)} \mid Y_{0:M}^{(n)})$. For fixed q_ϕ, we update θ using the inferred path distributions. The initial-state, observation, and entropy terms can be evaluated in closed form. The expected transition log density generally requires numerical approximation. To evaluate this expectation, we first condition on the current state and then apply cubature to the remaining Gaussian expectation. Section 3.2 describes the Gaussian state update for fixed θ, and Section 3.3 describes the drift update for fixed Gaussian distributions. Section 3.4 specifies an acceptance rule for both updates based on a common computable objective.

3.2 E-STEP: ITERATED EXTENDED KALMAN SMOOTHING

At iteration k of the alternating updates, we fix $\theta = \theta_k$ and update the Gaussian Markov law $q_\phi^{(n)}$ to approximate the latent-path posterior $p_{\Delta,\theta_k}(X_{0:M}^{(n)} \mid Y_{0:M}^{(n)})$ for each trajectory n. Since the nonlinear drift makes this posterior intractable, we use IEKS to construct a Gaussian candidate. IEKS repeatedly linearizes the Euler transitions around a reference trajectory and smooths the resulting linear Gaussian model. The Euler transition under θ_k is

$$X_{m+1} = X_m + \Delta b_{\theta_k}(X_m) + w_m,\ w_m \mid X_m \sim \mathcal{N}(0, \Delta a(X_m)). \tag{5}$$

At the start of iteration k, we initialize the reference path for each trajectory n with the marginal means of the current Gaussian smoothing distribution:

$$\bar{x}_m^{(n)} = \mathbb{E}_{q_{\phi_k}^{(n)}}[X_m^{(n)}], \qquad m = 0, \dots, M.$$

At $k = 0$, $q_{\phi_0}^{(n)}$ is obtained by Kalman smoothing under the zero-drift auxiliary model with diffusion frozen at m_0, as detailed in Appendix B. We suppress the trajectory index n below.

For the current reference trajectory $\bar{x}_{0:M}$, we linearize the drift as $b_{\theta_k}(x) \approx b_{\theta_k}(\bar{x}_m) + J_{b_{\theta_k}}(\bar{x}_m)(x - \bar{x}_m)$, where $J_{b_{\theta_k}}$ denotes the drift Jacobian. Substituting this approximation into (5) gives the affine transition mean $A_m x + c_m$, where

$$A_m := I_d + \Delta J_{b_{\theta_k}}(\bar{x}_m),\ c_m := \Delta \big[b_{\theta_k}(\bar{x}_m) - J_{b_{\theta_k}}(\bar{x}_m)\bar{x}_m \big],\ Q_m := \Delta a(\bar{x}_m). \tag{6}$$

We also freeze the diffusion covariance at the reference state $\bar{x}_m$. Thus, for the current reference trajectory, the linearized state-space model is

$$\begin{cases} X_{m+1} = A_m X_m + c_m + w_m,\ w_m \sim \mathcal{N}(0, Q_m), \\ Y_m = X_m + \varepsilon_m,\ \varepsilon_m \sim \mathcal{N}(0, R), \end{cases}$$

with $X_0 \sim \mathcal{N}(m_0, P_0)$. Kalman filtering and RTS smoothing can then be applied.

At each IEKS iteration, we recompute A_m, c_m, Q_m from the current reference trajectory and apply Kalman filtering and RTS smoothing to obtain the smoothed means μ_m^s, covariances P_m^s, and adjacent-state cross-covariances C_m^s. We use the smoothed means to update the reference trajectory with damping and repeat until it stabilizes or the prescribed iteration limit is reached. The final smoothing moments define the candidate Gaussian distribution for trajectory n; detailed recursions are given in Appendix B.

3.3 M-STEP: DRIFT UPDATE USING SMOOTHED GAUSSIAN MOMENTS

With the Gaussian laws $q_\phi^{(n)}$ fixed, the M-step updates the drift parameter θ. In (4), only the transition log density and the regularizer depend on θ. Write $r_\theta(x, x') = x' - x - \Delta b_\theta(x)$. Up to terms independent of θ, minimizing the negative variational objective amounts to minimizing

$$\mathcal{Q}_\phi(\theta) = \frac{1}{2N\Delta} \sum_{n=1}^{N} \sum_{m=0}^{M-1} \mathbb{E}_{q_\phi^{(n)}} \left[\left\| r_\theta(X_m^{(n)}, X_{m+1}^{(n)}) \right\|^2_{a(X_m^{(n)})^{-1}} \right] + \lambda \mathcal{R}(\theta).$$

For the Gaussian law $q_\phi^{(n)}$ fixed in the M-step, let $\mu_m = \mathbb{E}_{q_\phi^{(n)}}[X_m^{(n)}]$, $P_m = \mathrm{Cov}_{q_\phi^{(n)}}(X_m^{(n)})$, and $C_m = \mathrm{Cov}_{q_\phi^{(n)}}(X_m^{(n)}, X_{m+1}^{(n)})$. Suppressing the trajectory index in these moments, its pairwise marginal is

$$q_\phi^{(n)}(X_m, X_{m+1}) = \mathcal{N}\left(\begin{bmatrix} X_m \\ X_{m+1} \end{bmatrix}; \begin{bmatrix} \mu_m \\ \mu_{m+1} \end{bmatrix}, \begin{bmatrix} P_m & C_m \\ C_m^\top & P_{m+1} \end{bmatrix} \right).$$

Moreover,

$$\eta_m(x) := \mathbb{E}_{q_\phi}[X_{m+1} \mid X_m = x] = \mu_{m+1} + C_m^\top P_m^{-1}(x - \mu_m), \tag{7}$$

$$V_m := \mathrm{Cov}_{q_\phi}(X_{m+1} \mid X_m) = P_{m+1} - C_m^\top P_m^{-1} C_m. \tag{8}$$

Define $u_m(x) := (\eta_m(x) - x)/\Delta$, the conditional expected normalized increment under the Gaussian smoothing distribution. Then,

$$\mathbb{E}_{q_\phi}\left[\|r_\theta(X_m, X_{m+1})\|^2_{a(x)^{-1}} \mid X_m = x \right] = \Delta^2 \|u_m(x) - b_\theta(x)\|^2_{a(x)^{-1}} + \mathrm{tr}\big(a(x)^{-1} V_m\big).$$

With the smoothing distribution fixed, the trace term remains unchanged as we update the drift parameters. To form the M-step objective, we average this squared discrepancy under the marginal $q_\phi(X_m) = \mathcal{N}(\mu_m, P_m)$. We approximate this expectation using $2d$ equally weighted cubature points (Arasaratnam & Haykin, 2009):

$$\chi_{m,j}^{\pm} := \mu_m \pm \sqrt{d}\, L_m v_j, \qquad j = 1, \ldots, d, \tag{9}$$

where $L_m L_m^\top = P_m \succ 0$ is the Cholesky factorization, v_j is the j-th standard basis vector, and each point has weight $1/(2d)$. After the Gaussian update, we hold the resulting distributions $q_{\phi_{k+1}}^{(n)}$ fixed. The neural M-step then approximately minimizes

$$\widehat{\mathcal{L}}_k(\theta) = \frac{\Delta}{4dN} \sum_{n=1}^{N} \sum_{m=0}^{M-1} \sum_{j=1}^{d} \sum_{s \in \{-1,+1\}} \left\| u_m^{(n)}(\chi_{m,j}^{(n),s}) - b_\theta(\chi_{m,j}^{(n),s}) \right\|^2_{a(\chi_{m,j}^{(n),s})^{-1}} + \lambda \mathcal{R}(\theta). \tag{10}$$

Thus, the M-step uses the cubature loss $\widehat{\mathcal{L}}_k(\theta)$ to approximately minimize $\mathcal{Q}_{\phi_{k+1}}(\theta)$. The covariance P_m determines where the loss is evaluated, while C_m enters the conditional target $u_m(x)$.

3.4 SAFEGUARDED GAUSSIAN AND NEURAL UPDATES

IEKS constructs a Gaussian proposal by linearizing the drift and freezing the diffusion covariance along a reference path. When the proposal moves into regions where this local model is inaccurate, the computable variational objective can decrease. A finite neural step can also overshoot and increase the cubature loss. Motivated by line search for nonlinear Kalman smoothing (Särkkä & Svensson, 2020), we check both updates against the same objective before acceptance.

We first make the acceptance criterion explicit by writing the objective in terms of the Gaussian moments and the drift parameters. Based on the EM steps above, we evaluate both updates using the following deterministic cubature approximation of (4):

$$\widetilde{\mathcal{F}}_\Delta(\phi, \theta) := \frac{1}{N} \sum_{n=1}^{N} \left[A_n(\phi) - \sum_{m=0}^{M-1} \frac{1}{2d} \sum_{j=1}^{d} \sum_{s \in \{-,+\}} \ell_m^{(n)}\big(\chi_{m,j}^{(n),s}; \phi, \theta\big) \right] - \lambda \mathcal{R}(\theta), \tag{11}$$

where

$$A_n(\phi) := \mathbb{E}_{q_\phi^{(n)}}[\log \mu_0(X_0^{(n)})] + \sum_{m=0}^{M} \mathbb{E}_{q_\phi^{(n)}}[\log p_R(Y_m^{(n)} \mid X_m^{(n)})] + \mathcal{H}(q_\phi^{(n)}),$$

$$\ell_m^{(n)}(x;\phi,\theta) = \frac{d}{2}\log(2\pi\Delta) + \frac{1}{2}\log\det a(x) + \frac{\Delta}{2}\|u_m^{(n)}(x) - b_\theta(x)\|^2_{a(x)^{-1}} + \frac{1}{2\Delta}\operatorname{tr}\Big(a(x)^{-1}V_m^{(n)}\Big).$$

Each outer iteration k first updates the Gaussian parameters for all trajectories with θ_k fixed, yielding ϕ_{k+1}. It then performs J neural inner steps, indexed by $j = 0, \dots, J-1$, with ϕ_{k+1} fixed throughout, yielding θ_{k+1}. We first describe the Gaussian update. For fixed θ, we check directions and select step sizes separately for each trajectory. Let $\widetilde{\mathcal{F}}_{\Delta,n}(\phi^{(n)}, \theta)$ denote the nth bracketed term in (11). The vector $\phi^{(n)}$ collects the marginal means $\mu_{0:M}^{(n)}$, the independent entries of the marginal covariances $P_{0:M}^{(n)}$, and the adjacent-state cross-covariances $C_{0:M-1}^{(n)}$.

For each trajectory, we first assess the proposed direction using the gradient of its objective; the step size is chosen afterward. At iteration k, IEKS produces candidate moments $\widehat{\phi}_{k+1}^{(n)}$ for each trajectory n. Define the IEKS direction and the gradient at the current Gaussian law:

$$d_{\phi,k}^{(n)} := \widehat{\phi}_{k+1}^{(n)} - \phi_k^{(n)}, \qquad g_{\phi,k}^{(n)} := \nabla_{\phi^{(n)}} \widetilde{\mathcal{F}}_{\Delta,n}(\phi_k^{(n)}, \theta_k).$$

We use the IEKS direction if it is sufficiently aligned with the gradient and has controlled length:

$$\big(g_{\phi,k}^{(n)}\big)^\top d_{\phi,k}^{(n)} \geq \gamma_E \|g_{\phi,k}^{(n)}\|_2^2, \qquad \|d_{\phi,k}^{(n)}\|_2 \leq \Gamma_E \|g_{\phi,k}^{(n)}\|_2, \tag{12}$$

where $\gamma_E, \Gamma_E > 0$. If either condition fails, we set $d_{\phi,k}^{(n)} = H_k^{(n)} g_{\phi,k}^{(n)}$, where $0 < \underline{b} I \preceq H_k^{(n)} \preceq \overline{b} I$ and $\gamma_E \leq \underline{b} \leq \overline{b} \leq \Gamma_E$.

To preserve a valid Gaussian law, we require positive-definite marginal covariances and adjacent-state joint covariances

$$B_m^{(n)} := \begin{bmatrix} P_m^{(n)} & C_m^{(n)} \\ (C_m^{(n)})^\top & P_{m+1}^{(n)} \end{bmatrix}, \ m = 0, \dots, M-1.$$

The line search checks the covariance constraints and the amount of improvement in the objective. Starting with $\alpha = 1$, we halve α until the trial moments $\phi_k^{(n)} + \alpha d_{\phi,k}^{(n)}$ satisfy $P_m^{(n)} \succ 0$ for $m = 0, \dots, M$ and $B_m^{(n)} \succ 0$ for $m = 0, \dots, M-1$, and the Armijo condition holds:

$$\widetilde{\mathcal{F}}_{\Delta,n}(\phi_k^{(n)} + \alpha d_{\phi,k}^{(n)}, \theta_k) \geq \widetilde{\mathcal{F}}_{\Delta,n}(\phi_k^{(n)}, \theta_k) + c_E \alpha \big(g_{\phi,k}^{(n)}\big)^\top d_{\phi,k}^{(n)}, \qquad c_E \in (0,1). \tag{13}$$

At each feasible Gaussian trial, we keep $\theta = \theta_k$ fixed and recompute all moment-dependent quantities before evaluating the objective. With the accepted step size $\alpha_{n,k}$, we set $\phi_{k+1}^{(n)} = \phi_k^{(n)} + \alpha_{n,k} d_{\phi,k}^{(n)}$.

We then update the drift parameter with ϕ_{k+1} fixed. Maximizing $\widetilde{\mathcal{F}}_\Delta(\phi_{k+1}, \theta)$ is equivalent to minimizing $\widehat{\mathcal{L}}_k(\theta)$ in (10). Starting from $\theta_{k,0} = \theta_k$, we perform J neural updates. For $j = 0, \dots, J-1$, we use the direction $g_{\theta,k,j} = -\nabla_\theta \widehat{\mathcal{L}}_k(\theta_{k,j})$ and choose $\beta_{k,j}$ by Armijo backtracking on $\widehat{\mathcal{L}}_k$, with parameter $c_M \in (0,1)$. Each accepted step gives $\theta_{k,j+1} = \theta_{k,j} + \beta_{k,j} g_{\theta,k,j}$; we then set $\theta_{k+1} = \theta_{k,J}$. The complete procedure is summarized in Algorithm 1 in Appendix F.1.

4 Theoretical Guarantees

We first relate the discrete variational objective to a continuous path model. We consider one trajectory and suppress its index. Let P_θ denote the path law of (1) with drift b_θ, and let $p_{\Delta,\theta}$ denote the path density induced by (2). Define the continuous Euler interpolation by

$$\mathrm{d}\bar{X}_t = b_\theta(\bar{X}_{t_m})\,\mathrm{d}t + a(\bar{X}_{t_m})^{1/2}\,\mathrm{d}W_t, \quad t \in [t_m, t_{m+1}), \qquad \bar{X}_0 \sim \mu_0, \tag{14}$$

where the initial state is independent of W. Its path law is denoted by $\bar{P}_{\Delta,\theta}$. On the observation grid, its density is $p_{\Delta,\theta}$. Extend q_ϕ to a continuous path law $\bar{Q}_{\Delta,\phi}$ by using the conditional bridges of $\bar{P}_{\Delta,\theta}$ between adjacent grid states. These bridges depend only on their endpoints and the known diffusion covariance; their explicit form is given in (23).

For a continuous path law Q, define the unregularized ELBOs

$$\bar{\mathcal{L}}_\Delta(Q,\theta) := \sum_{m=0}^{M} \mathbb{E}_Q[\log p_R(Y_m \mid X_{t_m})] - D_{\mathrm{KL}}(Q\|\bar{P}_{\Delta,\theta}),$$

$$\mathcal{L}_\Delta(\phi,\theta) := \sum_{m=0}^{M} \mathbb{E}_{q_\phi}[\log p_R(Y_m \mid X_m)] - D_{\mathrm{KL}}(q_\phi\|p_{\Delta,\theta}).$$

Theorem 1. *Under* (A1) *in Appendix D, the following statements hold in any dimension $d \geq 1$.*

(i) *For every Gaussian Markov density q_ϕ in Section 3.1, the extension $\bar{Q}_{\Delta,\phi}$ has grid density q_ϕ, is independent of θ for fixed ϕ, and satisfies*

$$D_{\mathrm{KL}}(\bar{Q}_{\Delta,\phi}\|\bar{P}_{\Delta,\theta}) = D_{\mathrm{KL}}(q_\phi\|p_{\Delta,\theta}), \tag{15}$$

$$\bar{\mathcal{L}}_\Delta(\bar{Q}_{\Delta,\phi},\theta) = \mathcal{L}_\Delta(\phi,\theta). \tag{16}$$

(ii) *There is a constant $C > 0$, independent of Δ, ϕ, and the observation values, such that*

$$W_{2,\infty}(\bar{P}_{\Delta,\theta}, P_\theta) \leq C\sqrt{\Delta}. \tag{17}$$

Theorem 1 identifies the discrete ELBO with the ELBO of a continuous Euler path model. As the observation grid is refined, the Euler prior converges to the original SDE prior in path-space Wasserstein distance at rate $O(\sqrt{\Delta})$. Together, these results connect the discrete variational formulation to the underlying continuous-time model.

We next establish monotonicity of $\widetilde{\mathcal{F}}_\Delta$ and vanishing gradients for the VEST iterations.

Theorem 2. *Under assumptions* (A2)–(A4) *in Appendix E.1, it holds*

$$\widetilde{\mathcal{F}}_\Delta(z_{k+1}) \geq \widetilde{\mathcal{F}}_\Delta(z_k), \ \lim_{k\to\infty} \|\nabla\widetilde{\mathcal{F}}_\Delta(z_k)\|_2 = 0. \tag{18}$$

Moreover, for a constant $C > 0$ independent of K,

$$\min_{0\leq k<K} \|\nabla\widetilde{\mathcal{F}}_\Delta(z_k)\|_2^2 \leq C/K. \tag{19}$$

If the objective additionally satisfies the Kurdyka–Łojasiewicz property, the full sequence (z_k) converges to a single stationary point.

The Gaussian update uses an IEKS proposal based on local linearization, while the neural update takes a finite number of gradient steps. The safeguards assess both types of update against the common objective. Under the stated assumptions, the direction conditions and acceptance rules ensure sufficient improvement in the common objective. Theorem 2 shows that this progress drives the alternating procedure toward stationarity.

5 Experiments

5.1 Drift estimation accuracy

We evaluate VEST on seven stochastic dynamical systems across ten experimental settings with state dimension $d \in \{1, 2, 5, 10\}$. The one-dimensional systems comprise a linear Ornstein-Uhlenbeck (OU) process, a system with nonlinear saturating drift, and a double-well system with two stable states. The two-dimensional systems are a linear spiral with rotational dynamics and a Van der Pol oscillator whose deterministic dynamics exhibit a stable limit cycle. To assess drift estimation with coupled coordinates, we include a five-dimensional ring of bistable units and a ten-dimensional Kuramoto model of coupled phase oscillators. We also consider state-dependent diffusion in the OU, double-well, and Van der Pol systems, giving ten experimental settings in total.

The baselines are naive increment regression (Naive), the adapted Schmisser estimator (Penalized LS) (Schmisser, 2011), EM–SMC–RKHS (Ganguly et al., 2025), SING (Hu et al., 2025), and SDE Matching (Bartosh et al., 2025). SING assumes constant diffusion, and Penalized LS is applicable to one-dimensional systems. Table 1 reports the mean drift root mean squared error (RMSE) and standard deviation across five data seeds. A dash indicates a setting outside the scope of the corresponding baseline implementation.

Table 1: Comparison of drift estimation RMSE across methods. Boldface marks the lowest mean RMSE in each row, and underlining marks the second-lowest.

System	d	VEST (ours)	EM-SMC-RKHS (Ganguly et al., 2025)	SDE Matching (Bartosh et al., 2025)	SING (Hu et al., 2025)	Penalized LS (Schmisser, 2011)	Naive
Constant diffusion							
OU	1	**0.084** $\pm$ 0.030	0.237 $\pm$ 0.090	<u>0.116</u> $\pm$ 0.035	<u>0.116</u> $\pm$ 0.015	0.638 $\pm$ 0.099	10.499 $\pm$ 0.824
Saturating	1	**0.096** $\pm$ 0.042	0.601 $\pm$ 0.178	<u>0.227</u> $\pm$ 0.087	0.662 $\pm$ 0.035	0.259 $\pm$ 0.078	9.169 $\pm$ 0.718
Double well	1	**0.198** $\pm$ 0.076	<u>0.202</u> $\pm$ 0.061	0.404 $\pm$ 0.052	0.229 $\pm$ 0.067	0.273 $\pm$ 0.132	4.086 $\pm$ 0.664
Linear spiral	2	**0.159** $\pm$ 0.012	1.451 $\pm$ 0.412	<u>0.254</u> $\pm$ 0.046	0.325 $\pm$ 0.037	—	6.906 $\pm$ 0.519
Van der Pol	2	**0.684** $\pm$ 0.157	1.640 $\pm$ 0.445	0.917 $\pm$ 0.115	<u>0.744</u> $\pm$ 0.124	—	7.504 $\pm$ 1.235
Bistable ring	5	<u>0.956</u> $\pm$ 0.155	2.695 $\pm$ 0.091	**0.945** $\pm$ 0.129	1.084 $\pm$ 0.133	—	13.065 $\pm$ 0.494
Kuramoto	10	**0.654** $\pm$ 0.033	6.602 $\pm$ 0.471	<u>0.757</u> $\pm$ 0.083	1.173 $\pm$ 0.151	—	26.435 $\pm$ 0.820
State-dependent diffusion							
OU	1	**0.077** $\pm$ 0.014	0.327 $\pm$ 0.048	<u>0.110</u> $\pm$ 0.024	—	0.539 $\pm$ 0.113	8.520 $\pm$ 0.559
Double well	1	**0.191** $\pm$ 0.051	<u>0.234</u> $\pm$ 0.092	0.406 $\pm$ 0.062	—	0.277 $\pm$ 0.121	4.089 $\pm$ 0.513
Van der Pol	2	**0.791** $\pm$ 0.200	1.317 $\pm$ 0.438	<u>0.903</u> $\pm$ 0.115	—	—	7.472 $\pm$ 1.306

5.2 Effect of observation noise and sampling frequency

To assess sensitivity to measurement error and sparse observations, we conduct two sensitivity sweeps on the constant-diffusion versions of the double-well, linear spiral, and bistable ring systems. We denote the observation-noise standard deviation and sampling interval used in the preceding drift accuracy experiment by τ_0 and Δ_0, respectively. In this section, we vary one at a time by a factor $r \in \{0.25, 0.5, 1, 2, 4\}$: the noise sweep uses $\tau = r\tau_0$ with $\Delta = \Delta_0$, and the interval sweep uses $\Delta = r\Delta_0$ with $\tau = \tau_0$.

For each seed and setting, all methods are trained on the same observations and evaluated on common independent test states. Figure 2 reports the mean drift RMSE and standard deviation over five seeds and the numerical results are given in Appendix F.3. As observation noise increases, naive increment regression deteriorates sharply, whereas VEST is much less affected. On the bistable ring, SDE Matching remains close to VEST. As the sampling interval increases, naive regression improves markedly, while VEST remains comparatively stable and achieves the lowest mean error at every interval setting across all three systems.

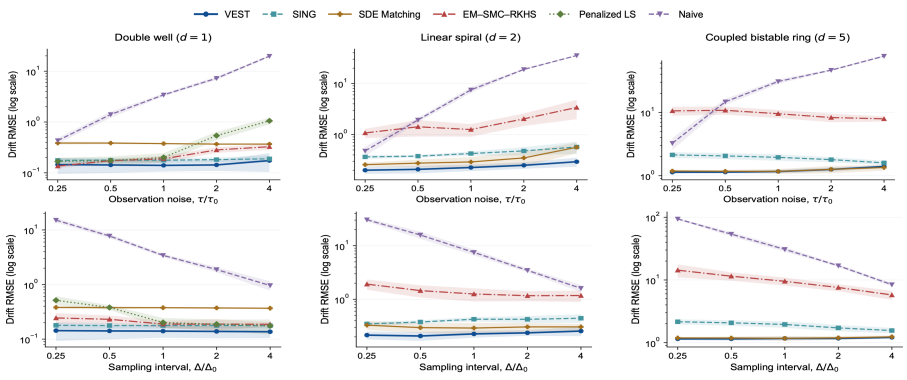


Figure 2: Sensitivity of drift RMSE to observation noise and sampling interval. From left to right, the columns show the double-well system ($d = 1$), linear spiral ($d = 2$), and coupled bistable ring ($d = 5$). The upper row varies τ/τ_0 at fixed $\Delta = \Delta_0$; the lower row varies Δ/Δ_0 at fixed $\tau = \tau_0$.

5.3 LATENT TRAJECTORY RECOVERY

Moreover, we evaluate latent-state recovery on test trajectories from the constant diffusion double well and linear spiral systems (Figure 3). Alongside state and drift RMSE, we compare learned and true terminal distributions using the 1-Wasserstein distance (W_1) for the double well and the sliced 1-Wasserstein distance (sliced W_1) for the linear spiral. VEST has lower drift RMSE and terminal distance than SING in both systems.

System	Metric	SING	VEST
Double well	Drift RMSE	0.2264	**0.1981**
	State RMSE	0.0906	**0.0904**
	W_1	0.1443	**0.1160**
Linear spiral	Drift RMSE	0.2913	**0.1522**
	State RMSE	0.0674	**0.0672**
	Sliced W_1	0.0750	**0.0593**

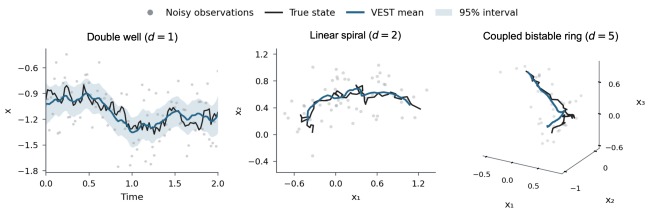


Figure 3: Latent-state recovery. Left: Five-seed mean errors and terminal distribution distances for VEST and SING. Right: Noisy observations (gray), true states (black), and VEST posterior means (blue) on test trajectories. Shading shows the pointwise 95% posterior interval.

5.4 REAL DATA: POTENTIAL RECOVERY IN OPTICAL FEEDBACK TRAPS

In a bistable optical feedback trap, the barrier between two potential wells governs transitions between stable states. Estimating its height is difficult when position measurements are noisy and each short trajectory samples only one well. Bryan et al. (2022) addressed this problem with SKIPPER, which infers the potential from noisy trajectories. We use their Eb10 dataset to test whether VEST can recover the barrier from its estimated drift. We compare VEST with direct neural increment regression (NN) and SKIPPER. Further details are given in Appendix F.5.

The estimated drift and potential satisfy $\mathrm{d}(\widehat{U}/k_BT)/\mathrm{d}x = -\widehat{b}(x)/D$. We measure the barrier from the deeper inferred well to the maximum between the wells and compare it with the nominal design value of $10\,k_BT$ (Table 2). VEST gives the closest estimate.

Table 2: Eb10 barrier estimates.

Method	Barrier (k_BT)	Deviation(%)
VEST	**8.71**	**12.9**
Direct NN	6.71	32.9
SKIPPER	8.12	18.8

We also analyze three two-dimensional harmonic-trap recordings (n1, n4, and n7) from Bryan et al. (2022), acquired at different measurement-noise levels. Using the noisier recordings, VEST consistently captures the reference well shape across all three noise levels (Figure 4).

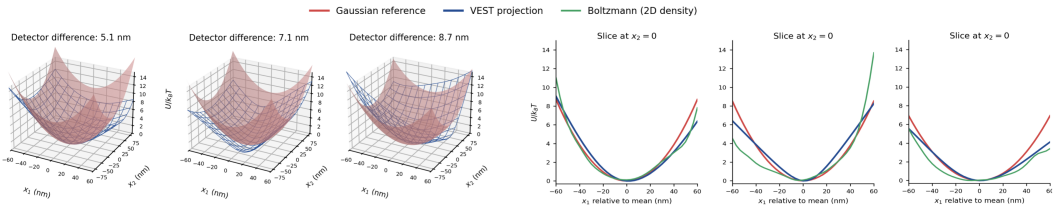


Figure 4: Potential estimates for harmonic-trap recordings n1, n4, and n7 (Bryan et al., 2022). First three panels: references from paired-detector covariance (red) and conservative projections of VEST drift (blue). Last three: $x_2 = 0$ slices and Boltzmann estimates from noisy positions (green).

6 CONCLUSION

We introduced VEST, combining Gaussian smoothing, conditional posterior-increment targets, and deterministic cubature for neural drift estimation. Under the stated assumptions, safeguarding both updates with a common objective yields stationary accumulation points. Synthetic and optical trap experiments illustrate drift and potential recovery. Future work could address unknown diffusion.

AI USE STATEMENT

We used OpenAI Codex to assist with literature search and language polishing. The authors reviewed the AI-assisted material and take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The appendix provides detailed experimental settings, evaluation protocols, and sources for the real data and external baseline implementations to support reproduction of the reported results.

REFERENCES

Yacine Aït-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica*, 70(1):223–262, 2002. doi: 10.1111/1468-0262.00274. URL https://www.princeton.edu/~yacine/mle.pdf.

Ienkaran Arasaratnam and Simon Haykin. Cubature Kalman filters. *IEEE Transactions on Automatic Control*, 54(6):1254–1269, 2009. doi: 10.1109/TAC.2009.2019800.

Cédric Archambeau, Dan Cornford, Manfred Opper, and John Shawe-Taylor. Gaussian process approximations of stochastic differential equations. In *Gaussian Processes in Practice*, volume 1 of *Proceedings of Machine Learning Research*, pp. 1–16. PMLR, 2007. URL https://proceedings.mlr.press/v1/archambeau07a.html.

Hédy Attouch, Jérôme Bolte, and Benar F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013. doi: 10.1007/s10107-011-0484-9.

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. URL https://proceedings.mlr.press/v267/bartosh25a.html.

Bradley M. Bell. The iterated Kalman smoother as a Gauss–Newton method. *SIAM Journal on Optimization*, 4(3):626–636, 1994. doi: 10.1137/0804035.

J. Shepard Bryan, IV, Prithviraj Basak, John Bechhoefer, and Steve Pressé. Inferring potential landscapes from noisy trajectories of particles within an optical feedback trap. *iScience*, 25(9):104731, 2022. doi: 10.1016/j.isci.2022.104731.

Fabienne Comte and Valentine Genon-Catalot. Nonparametric drift estimation for i.i.d. paths of stochastic differential equations. *Annals of Statistics*, 48(6):3336–3365, 2020. doi: 10.1214/19-AOS1933.

Kevin Course and Prasanth B. Nair. State estimation of a physical system with unknown governing equations. *Nature*, 622(7982):261–267, 2023. doi: 10.1038/s41586-023-06574-8.

Arnak Dalalyan. Sharp adaptive estimation of the drift function for ergodic diffusions. *Annals of Statistics*, 33(6):2507–2528, 2005. doi: 10.1214/009053605000000615.

Peter Ditlevsen and Susanne Ditlevsen. Warning of a forthcoming collapse of the Atlantic meridional overturning circulation. *Nature Communications*, 14:4254, 2023. doi: 10.1038/s41467-023-39810-w.

Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable continuous-time models of latent stochastic dynamical systems. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1726–1734. PMLR, 2019. URL https://proceedings.mlr.press/v97/duncker19a.html.

Salima El Kolei and Fabien Navarro. Contrast estimation for noisy observations of diffusion processes via closed-form density expansions. *Statistical Inference for Stochastic Processes*, 25(2): 303–336, 2022. doi: 10.1007/s11203-021-09256-2.

Benjamin Favetto. Parameter estimation by contrast minimization for noisy observations of a diffusion process. *Statistics*, 48(6):1344–1370, 2014. doi: 10.1080/02331888.2013.828058.

Benjamin Favetto. Estimating functions for noisy observations of ergodic diffusions. *Statistical Inference for Stochastic Processes*, 19(1):1–28, 2016. doi: 10.1007/s11203-015-9121-1.

Arnab Ganguly, Riten Mitra, and Jinpu Zhou. Nonparametric learning of stochastic differential equations from sparse and noisy data. *arXiv preprint arXiv:2508.11597*, 2025.

Zoubin Ghahramani and Sam T. Roweis. Learning nonlinear dynamical systems using an EM algorithm. In *Advances in Neural Information Processing Systems*, volume 11, pp. 431–437, 1998. URL https://proceedings.neurips.cc/paper/1998/hash/0ebcc77dc72360d0eb8e9504c78d38bd-Abstract.html.

Emmanuel Gobet, Marc Hoffmann, and Markus Reiß. Nonparametric estimation of scalar diffusions based on low frequency data. *Annals of Statistics*, 32(5):2223–2253, 2004. doi: 10.1214/009053604000000797.

Georg A. Gottwald. Finite-size effects in a stochastic Kuramoto model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(10):101103, 2017. doi: 10.1063/1.5004618.

Desmond J. Higham, Xuerong Mao, and Andrew M. Stuart. Strong convergence of Euler-type methods for nonlinear stochastic differential equations. *SIAM Journal on Numerical Analysis*, 40(3): 1041–1063, 2002. doi: 10.1137/S0036142901389530. URL https://webhomes.maths.ed.ac.uk/~dhigham/Publications/P44.pdf.

Amber Hu, Henry Smith, and Scott Linderman. SING: SDE inference via natural gradients. In *Advances in Neural Information Processing Systems*, volume 38, pp. 170822–170875, 2025. doi: 10.52202/085713-5146.

Wooyong Lee, Priscilla E. Greenwood, Nancy Heckman, and Wolfgang Wefelmeyer. Pre-averaged kernel estimators for the drift function of a diffusion process in the presence of microstructure noise. *Statistical Inference for Stochastic Processes*, 20(2):237–252, 2017. doi: 10.1007/s11203-016-9141-5.

Jakob Lindqvist, Simo Särkkä, Ángel F. García-Fernández, Matti Raitoharju, and Lennart Svensson. Posterior linearisation smoothing with robust iterations. *arXiv preprint arXiv:2112.03969*, 2021.

Neda Mohammadi, Leonardo V. Santoro, and Victor M. Panaretos. Nonparametric estimation for SDE with sparsely sampled paths: An FDA perspective. *Stochastic Processes and their Applications*, 167:104239, 2024. doi: 10.1016/j.spa.2023.104239.

Akihiro Oga and Yuta Koike. Drift estimation for a multi-dimensional diffusion process using deep neural networks. *Stochastic Processes and their Applications*, 170:104240, 2024. doi: 10.1016/j.spa.2023.104240.

Herbert E. Rauch, F. Tung, and Charlotte T. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, 1965. doi: 10.2514/3.3166.

Andreas Ruttor, Philipp Batz, and Manfred Opper. Approximate gaussian process inference for the drift function in stochastic differential equations. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2040–2048, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/021bbc7ee20b71134d53e20206bd6feb-Abstract.html.

Simo Särkkä and Lennart Svensson. Levenberg–Marquardt and line-search extended Kalman smoothers. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5875–5879, 2020. doi: 10.1109/ICASSP40776.2020.9054686.

Simo Särkkä and Lennart Svensson. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2nd edition, 2023. doi: 10.1017/9781108917407.

Emeline Schmisser. Non-parametric drift estimation for diffusions from noisy data. *Statistics & Decisions*, 28(2):119–150, 2011. doi: 10.1524/stnd.2011.1063.

Hideyuki Takamizawa. Is nonlinear drift implied by the short end of the term structure? *The Review of Financial Studies*, 21(1):311–346, 2008. doi: 10.1093/rfs/hhm072.

Prakhar Verma, Vincent Adam, and Arno Solin. Variational Gaussian process diffusion processes. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1909–1917. PMLR, 2024. URL `https://proceedings.mlr.press/v238/verma24a.html`.

Yu Wang and Arnab Ganguly. Variational smoothing and inference for SDEs from sparse data with dynamic neural flows. *arXiv preprint arXiv:2605.05606*, 2026.

C. F. Jeff Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1): 95–103, 1983. doi: 10.1214/aos/1176346060.

Grace Hui Ting Yeo, Sachit D. Saksena, and David K. Gifford. Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. *Nature Communications*, 12:3222, 2021. doi: 10.1038/s41467-021-23518-w.

Yuzhen Zhao, Yating Liu, and Marc Hoffmann. Drift estimation for diffusion processes using neural networks based on discretely observed independent paths. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(34):28778–28785, 2026. doi: 10.1609/aaai.v40i34.40111.

APPENDIX CONTENTS

A Related Work **14**

B Zero-drift initialization and IEKS recursions **15**

C Variational and Drift Extensions **16**

C.1 VEST-NM: q_ϕ without Markov property 16

C.2 VEST-LR: b_θ with linear model 17

D Proof of Theorem 1 **17**

E Proof of Theorem 2 **19**

E.1 Standing assumptions 20

E.2 Preliminary lemmas 20

E.3 Monotone ascent 22

E.4 Sufficient ascent 22

E.5 Convergence of the gradient at z_k 23

E.6 Stationarity of accumulation points 23

E.7 Iteration complexity 23

E.8 Full-sequence convergence under the Kurdyka–Łojasiewicz property 24

F Experimental Settings **25**

F.1 VEST algorithm 25

F.2 Drift estimation settings 26

F.3 Sensitivity results 28

F.4 Latent-state recovery 29

F.5 Real data experiments 30

F.6 Additional experiment: effect of the trajectory number N 31

F.7 Additional experiment: VEST-NM and VEST-LR 32

F.8 Additional experiment: ablation results 33

Notation. For matrices, $\|\cdot\| = \|\cdot\|_{\mathrm{op}}$ denotes the operator norm, and $\|\cdot\|_F$ denotes the Frobenius norm. For a symmetric matrix A, $\lambda_{\min}(A)$ is its smallest eigenvalue. Write $B(z, r) := \{v : \|v - z\| < r\}$ for the open Euclidean ball. The domain $\mathcal{D}$ consists of the Gaussian moment parameters ϕ for which every marginal covariance $P_m^{(n)}$ and adjacent pair covariance $B_m^{(n)}$ is positive definite. Constants denoted by C may change between displays.

A Related Work

Drift Estimation from Noisy Observations. Several estimators study a single long noisy trajectory through local averaging or deconvolution. Schmisser (2011) combines pre-averaging with penalized least squares, while Lee et al. (2017) use pre-averaged observations in a Nadaraya–Watson estimator. For parametric diffusions, Favetto (2014; 2016) develops contrasts and estimating functions based on local means. El Kolei & Navarro (2022) combine transition-density expansions with deconvolution for parameter estimation.

For multiple independent noisy trajectories, Mohammadi et al. (2024) use pooled moments to estimate time-varying drift and diffusion under prescribed state dependence. However, none of these methods uses a neural network to estimate the drift. VEST learns a state-dependent neural drift from multiple independent noisy trajectories while inferring their latent states.

Latent-State Inference Coupled with Drift Learning. Approximate EM methods couple latent-path reconstruction with drift estimation. Classical nonlinear state-space EM methods combine extended Kalman smoothing with parameter updates that average over latent-state uncertainty; for example, Ghahramani & Roweis (1998) obtain tractable updates using radial-basis-function dynamics. Ruttor et al. (2013) combine a piecewise linearized posterior process with sparse Gaussian-process drift regression. Duncker et al. (2019) jointly infer continuous latent paths and a sparse Gaussian-process drift conditioned on learned fixed points and local Jacobians, targeting interpretable nonlinear dynamics from noisy, irregular observations. Ganguly et al. (2025) use sequential Monte Carlo to approximate the E-step and penalized RKHS regression in the M-step for sparse, noisy observations.

Gaussian variational path approximations were developed by Archambeau et al. (2007). Building on this continuous-time formulation, Verma et al. (2024) use an exponential-family site parameterization to replace fixed-point posterior updates with natural-gradient-like optimization and improve generative-model parameter learning. SVISE (Course & Nair, 2023) uses stochastic variational inference with Markov Gaussian processes to jointly estimate states and unknown dynamics. SING (Hu et al., 2025) optimizes a Gaussian Markov posterior by natural gradients and demonstrates variational EM with Gaussian-process, neural-network, and polynomial drifts. Its constant-diffusion formulation supports both Monte Carlo and deterministic quadrature.

VEST accommodates known state-dependent diffusion covariance, a setting outside SING's original constant-diffusion formulation. VEST formulates neural drift learning as conditional posterior-increment regression within a common deterministic cubature objective. Adjacent-state cross-covariances enter the regression targets, while marginal means and covariances determine the cubature points used to approximate the expected loss. This yields a deterministic neural regression update that incorporates uncertainty under the Gaussian smoothing approximation. For a fixed smoothing distribution, minimizing this cubature regression loss is equivalent to maximizing the common cubature objective over the drift parameters.

Complementary approaches learn alternative representations of the smoothing distribution. SDE Matching (Bartosh et al., 2025) parameterizes a variational posterior with directly sampleable time marginals, enabling a simulation-free training objective without integrating posterior trajectories. While SDE Matching uses directly sampled time marginals, VEST uses a Gaussian Markov smoother that retains covariance between consecutive latent states for neural drift regression. Wang & Ganguly (2026) instead learn a conditional backward score that defines a posterior SDE and use samples from the induced process in a variational EM procedure. VEST updates the neural drift by deterministic cubature of smoothed moments, without sampling posterior trajectories.

Globalized Gaussian Smoothing and Safeguarded Alternating Updates. Bell (1994) establishes the connection between iterated Kalman smoothing and Gauss-Newton optimization. Särkkä

& Svensson (2020) developed line-search and Levenberg-Marquardt variants of IEKS, and Lindqvist et al. (2021) extend these techniques to posterior linearisation smoothing. These methods stabilize latent-state estimation for a fixed dynamical model.

VEST applies this globalization principle to alternating Gaussian and neural drift updates. Direction safeguards and backtracking check IEKS-based Gaussian proposals and neural gradient steps against a shared deterministic cubature objective. Under the assumptions of Theorem 2, the ideal safeguarded scheme achieves monotone ascent of the common cubature objective, with the joint gradient in the Gaussian and drift parameters tending to zero.

B ZERO-DRIFT INITIALIZATION AND IEKS RECURSIONS

Zero-drift initialization. For each observed trajectory n, we initialize $q_{\phi_0}^{(n)}$ by smoothing an auxiliary model with zero drift and process covariance frozen at $a(m_0)$. This model is

$$\begin{aligned} X_0^{(n)} &\sim \mathcal{N}(m_0, P_0), \\ X_{m+1}^{(n)} \mid X_m^{(n)} = x &\sim \mathcal{N}(x, \Delta a(m_0)), \qquad m = 0, \dots, M-1, \\ Y_m^{(n)} \mid X_m^{(n)} = x &\sim \mathcal{N}(x, R), \qquad m = 0, \dots, M. \end{aligned}$$

It is linear Gaussian, with $A_m = I_d$, $c_m = 0$, and $Q_m = \Delta a(m_0)$. Starting from $(\mu_0^-, P_0^-) = (m_0, P_0)$ and suppressing the trajectory index in the recursions, we apply the measurement update below and, for $m = 0, \dots, M-1$, predict with

$$\mu_{m+1}^- = \mu_m^f, \qquad P_{m+1}^- = P_m^f + \Delta a(m_0).$$

The backward pass starts at $(\mu_M^s, P_M^s) = (\mu_M^f, P_M^f)$ and uses the RTS recursions below with $G_m = P_m^f (P_{m+1}^-)^{-1}$. Its marginal means μ_m^s, covariances P_m^s, and adjacent cross-covariances C_m^s specify the Gaussian Markov initialization $q_{\phi_0}^{(n)}$. We set the initial reference path to $\bar{x}_m = \mu_m^s$ for $m = 0, \dots, M$. For $k > 0$, the E-step instead starts from the smoothed means of the previously accepted $q_{\phi_k}^{(n)}$.

IEKS filtering and smoothing. At each IEKS iteration, form A_m, c_m, Q_m from the current reference path using (6). These matrices define a linear Gaussian model for that iteration. Starting with $(\mu_0^-, P_0^-) = (m_0, P_0)$, the forward measurement update for $m = 0, \dots, M$ and prediction for $m = 0, \dots, M-1$ are

$$\begin{aligned} S_m &= P_m^- + R, & K_m &= P_m^- S_m^{-1}, \\ \mu_m^f &= \mu_m^- + K_m(Y_m - \mu_m^-), & P_m^f &= (I_d - K_m) P_m^- (I_d - K_m)^\top + K_m R K_m^\top, \\ \mu_{m+1}^- &= A_m \mu_m^f + c_m, & P_{m+1}^- &= A_m P_m^f A_m^\top + Q_m. \end{aligned}$$

Here the superscripts $-$ and f denote predicted and filtered moments. Starting from $(\mu_M^s, P_M^s) = (\mu_M^f, P_M^f)$, the backward Rauch–Tung–Striebel recursion for $m = M-1, \dots, 0$ (Rauch et al., 1965; Särkkä & Svensson, 2023) is

$$\begin{aligned} G_m &= P_m^f A_m^\top (P_{m+1}^-)^{-1}, \\ \mu_m^s &= \mu_m^f + G_m(\mu_{m+1}^s - \mu_{m+1}^-), \\ P_m^s &= P_m^f + G_m(P_{m+1}^s - P_{m+1}^-) G_m^\top, \\ C_m^s &:= \mathrm{Cov}(X_m, X_{m+1} \mid Y_{0:M}) = G_m P_{m+1}^s. \end{aligned}$$

All covariances in these recursions are under the current linear Gaussian model, and the superscript s denotes smoothed moments. After a backward pass, we update the reference path using

$$\bar{x}_m \leftarrow (1-\rho)\bar{x}_m + \rho \mu_m^s, \qquad \rho \in (0, 1],$$

and repeat the linearization, filtering, and smoothing steps until the reference path stabilizes or the prescribed IEKS iteration limit is reached. A final filtering and smoothing pass at the updated reference path supplies consistent moments (μ_m^s, P_m^s, C_m^s) for the Gaussian candidate.

C VARIATIONAL AND DRIFT EXTENSIONS

In this section, we consider two extensions of Section 3: a general Gaussian variational family for q_ϕ and a drift model linear in its parameters.

C.1 VEST-NM: q_ϕ WITHOUT MARKOV PROPERTY

In Section 3.1, q_ϕ is modeled as a Gaussian Markov law. We consider a kernel-based Gaussian family with non-Markov dependence as an alternative variational approximation. For a fixed observed trajectory $Y_{0:M} = y_{0:M}$, we approximate the smoothing posterior $p_{\Delta,\theta}(X_{0:M} \mid Y_{0:M} = y_{0:M})$ by the Gaussian variational law

$$q_\phi(X_{0:M}) := \mathcal{N}(\mu, K_\eta),\ \phi := (\mu, \eta),\ K_\eta \succ 0,$$

where $\mu \in \mathbb{R}^{(M+1)d}$ and $K_\eta \in \mathbb{R}^{(M+1)d\times(M+1)d}$. For a Gaussian law, the Markov property is characterized by its inverse covariance:

$$[K_\eta^{-1}]_{ij} = 0 \text{ for } |i-j| > 1.$$

We consider two common covariance kernels that generally induce non-Markov Gaussian laws: the radial basis function (RBF) kernel and the Matérn-$3/2$ kernel. For either kernel, we set

$$[K_\eta]_{ij} = s_\eta(t_i)s_\eta(t_j)\kappa_\eta(t_i,t_j)I_d + \epsilon\delta_{ij}I_d, \qquad \epsilon > 0,$$

where $s_\eta(t) > 0$ allows the posterior marginal variance to vary over time. The two kernels are

$$\kappa_{\mathrm{RBF}}(t,t') := \exp\left(-\frac{(t-t')^2}{2\ell^2}\right),$$
$$\kappa_{\mathrm{M32}}(t,t') := \left(1 + \frac{\sqrt{3}|t-t'|}{\ell}\right)\exp\left(-\frac{\sqrt{3}|t-t'|}{\ell}\right), \qquad \ell > 0.$$

As for the E-step, the method in Section 3.2 updates the Gaussian Markov parameters (μ, P, C). We instead optimize (μ, η), where η determines the full covariance K_η through the RBF or Matérn-$3/2$ kernel. The E-step updates (μ, η) by ascending the variational objective using the full Gaussian entropy

$$\mathcal{H}(q_\phi) = \frac{1}{2}\log\det K_\eta + \frac{(M+1)d}{2}\log(2\pi e).$$

Then, we extract $P_m = [K_\eta]_{m,m}$ and $C_m = [K_\eta]_{m,m+1}$. The drift M-step then proceeds as in Section 3.3.

Notice that $\log p_{\Delta,\theta}(X_{0:M}, Y_{0:M})$ depends only on individual states and adjacent pairs (X_m, X_{m+1}). Then, we construct a Gaussian Markov law from these distributions:

$$q_\phi^\star(x_{0:M}) := q_\phi(x_0)\prod_{m=0}^{M-1} q_\phi(x_{m+1} \mid x_m). \tag{20}$$

By construction, $q_\phi^\star$ and q_ϕ have the same distributions for each X_m and each adjacent pair (X_m, X_{m+1}). Thus, $\mathbb{E}_{q_\phi}[\log p_{\Delta,\theta}(X_{0:M}, Y_{0:M})] = \mathbb{E}_{q_\phi^\star}[\log p_{\Delta,\theta}(X_{0:M}, Y_{0:M})]$ and the variational objectives differ only in entropy.

The entropy of the Gaussian Markov law $q_\phi^\star$ is

$$\mathcal{H}(q_\phi^\star) = \frac{(M+1)d}{2}\log(2\pi e) + \frac{1}{2}\left(\log\det P_0 + \sum_{m=0}^{M-1}\log\det V_m\right),$$

where V_m is defined in (8). Applying the objective in (11) to q_ϕ and $q_\phi^\star$, the terms other than entropy coincide. Hence,

$$\widetilde{\mathcal{F}}_\Delta(q_\phi,\theta) - \widetilde{\mathcal{F}}_\Delta(q_\phi^\star,\theta) = \mathcal{H}(q_\phi) - \mathcal{H}(q_\phi^\star) = -D_{\mathrm{KL}}(q_\phi \| q_\phi^\star).$$

Proposition 1. *For any Gaussian law $q_\phi = \mathcal{N}(\mu, K_\eta)$ with $K_\eta \succ 0$, let $q_\phi^\star$ be its Gaussian Markovization defined in* (20)*. Then, for any fixed θ,*

$$\widetilde{\mathcal{F}}_\Delta(q_\phi, \theta) \le \widetilde{\mathcal{F}}_\Delta(q_\phi^\star, \theta),$$

with equality if and only if $q_\phi = q_\phi^\star$.

The proposition follows from the identity above and $D_{\mathrm{KL}}(q_\phi \| q_\phi^\star) \ge 0$. This result shows that every Gaussian variational law has a Gaussian Markov counterpart with an equal or higher objective at fixed θ. Thus, over the unrestricted Gaussian family, it suffices to consider Gaussian Markov laws.

C.2 VEST-LR: b_θ WITH LINEAR MODEL

Let $\varphi : \mathbb{R}^d \to \mathbb{R}^C$ be a fixed, continuously differentiable feature function. We set $b_\theta(x) = W\varphi(x)$, where $W \in \mathbb{R}^{d\times C}$, and use $\mathcal{R}(W) = \|W\|_F^2 = \sum_{r,c} W_{rc}^2$. For RBF features, $\varphi_c(x) = \exp(-\|x - z_c\|^2/(2\ell^2))$, where $z_c \in \mathbb{R}^d$ is a fixed center and $\ell > 0$ is the lengthscale. At fixed q_ϕ, the cubature M-step objective in (10) is quadratic in $w = \mathrm{vec}(W)$. Index its $2dMN$ cubature points by i, with inputs x_i and targets u_i. Let Φ stack $\Phi_i = \varphi(x_i)^\top \otimes I_d$, let U stack the targets, and set

$$\Omega = \mathrm{blkdiag}\left(\frac{1}{2d} a(x_i)^{-1}\right)_i.$$

For $\lambda > 0$, the unique minimizer is

$$w^\star = \left(\Phi^\top \Omega \Phi + \frac{2N\lambda}{\Delta} I_{dC}\right)^{-1} \Phi^\top \Omega U, \qquad W^\star = \mathrm{unvec}_{d\times C}(w^\star).$$

Here $\mathrm{blkdiag}$ forms a block-diagonal matrix, vec stacks the columns of W, and $\mathrm{unvec}_{d\times C}$ reshapes that vector back into a $d \times C$ matrix. The E-step evaluates b_θ and $\nabla b_\theta(x) = W\nabla\varphi(x)$. For the RBF features,

$$\nabla\varphi_c(x) = \frac{z_c - x}{\ell^2}\varphi_c(x).$$

D PROOF OF THEOREM 1

In addition to the model conditions in Section 2, we assume:

(A1): The functions b_θ and a are globally Lipschitz on $\mathbb{R}^d$.

The constants below may depend on T, d, m_0, P_0, the Lipschitz constants, $\|b_\theta(0)\|$, and the diffusion bounds in Section 2, but not on Δ, ϕ, or the observation values.

Lemma 1. *For symmetric matrices $A, B \succeq \underline{a} I_d$,*

$$\|A^{1/2} - B^{1/2}\|_F \le \frac{1}{2\sqrt{\underline{a}}}\|A - B\|_F.$$

Proof. Set $H = A^{1/2} - B^{1/2}$. Since $A^{1/2}H + HB^{1/2} = A - B$, taking the Frobenius inner product with H gives

$$2\sqrt{\underline{a}}\|H\|_F^2 \le \mathrm{tr}(H^\top A^{1/2} H) + \mathrm{tr}(H^\top H B^{1/2}) = \mathrm{tr}(H^\top (A - B)) \le \|H\|_F \|A - B\|_F.$$

Division by $\|H\|_F$ proves the claim when $H \neq 0$; the case $H = 0$ is immediate. □

Let X solve

$$\mathrm{d}X_t = b_\theta(X_t)\,\mathrm{d}t + a(X_t)^{1/2}\,\mathrm{d}W_t, \qquad X_0 \sim \mu_0,$$

and construct $\bar{X}$ in (14) using the same initial state and Brownian motion. Write $\kappa_\Delta(s) = t_m$ for $s \in [t_m, t_{m+1})$, and $\kappa_\Delta(T) = T$.

Lemma 2. *Under* (A1)*,*

$$\mathbb{E}\sup_{t\le T}\|X_t\|^2 + \mathbb{E}\sup_{t\le T}\|\bar{X}_t\|^2 \le C, \tag{21}$$

$$\mathbb{E}\sup_{t\le T}\|X_t - \bar{X}_t\|^2 \le C\Delta. \tag{22}$$

Proof. We use the classical Euler strong error argument (Higham et al., 2002). Lemma 1 and (A1) imply that $a^{1/2}$ is globally Lipschitz. Also, $\|b_\theta(x)\| \le C(1+\|x\|)$ and $\|a(x)^{1/2}\|_F^2 = \operatorname{tr} a(x) \le d\bar{a}$. Thus both processes are well defined on $[0,T]$. The integral form of (14) is

$$\bar{X}_t = X_0 + \int_0^t b_\theta(\bar{X}_{\kappa_\Delta(s)})\,\mathrm{d}s + \int_0^t a(\bar{X}_{\kappa_\Delta(s)})^{1/2}\,\mathrm{d}W_s.$$

Using $\kappa_\Delta(s) \le s$, the Cauchy–Schwarz and Doob inequalities give

$$\begin{aligned}\mathbb{E}\sup_{r\le t}\|\bar{X}_r\|^2 &\le 3\mathbb{E}\|X_0\|^2 + 3T\int_0^t \mathbb{E}\|b_\theta(\bar{X}_{\kappa_\Delta(s)})\|^2\,\mathrm{d}s + 12d\bar{a}t \\ &\le C + C\int_0^t \mathbb{E}\sup_{r\le s}\|\bar{X}_r\|^2\,\mathrm{d}s.\end{aligned}$$

The same estimate holds for X. These estimates can first be applied up to the exit time of each process from a ball. Gronwall's inequality and Fatou's lemma then give (21), since $\mathbb{E}\|X_0\|^2 = \|m_0\|^2 + \operatorname{tr} P_0 < \infty$.

For $s \in [t_m, t_{m+1}]$, the Euler interpolation satisfies

$$\bar{X}_s - \bar{X}_{t_m} = (s-t_m)b_\theta(\bar{X}_{t_m}) + a(\bar{X}_{t_m})^{1/2}(W_s - W_{t_m}).$$

Conditioning on the information at t_m and using (21), we obtain

$$\mathbb{E}\|\bar{X}_s - \bar{X}_{t_m}\|^2 = (s-t_m)^2\mathbb{E}\|b_\theta(\bar{X}_{t_m})\|^2 + (s-t_m)\mathbb{E}\operatorname{tr} a(\bar{X}_{t_m}) \le C\Delta,$$

where $\Delta \le T$. Let $e_t = X_t - \bar{X}_t$. Then

$$e_t = \int_0^t \big[b_\theta(X_s) - b_\theta(\bar{X}_{\kappa_\Delta(s)})\big]\,\mathrm{d}s + \int_0^t \big[a(X_s)^{1/2} - a(\bar{X}_{\kappa_\Delta(s)})^{1/2}\big]\,\mathrm{d}W_s.$$

Applying the same inequalities and the Lipschitz bounds gives

$$\begin{aligned}\mathbb{E}\sup_{r\le t}\|e_r\|^2 &\le C\int_0^t \mathbb{E}\|X_s - \bar{X}_{\kappa_\Delta(s)}\|^2\,\mathrm{d}s \\ &\le C\int_0^t \mathbb{E}\sup_{r\le s}\|e_r\|^2\,\mathrm{d}s + C\int_0^t \mathbb{E}\|\bar{X}_s - \bar{X}_{\kappa_\Delta(s)}\|^2\,\mathrm{d}s \\ &\le C\int_0^t \mathbb{E}\sup_{r\le s}\|e_r\|^2\,\mathrm{d}s + C\Delta.\end{aligned}$$

Gronwall's inequality proves (22). □

Proof of Theorem 1:

(i) On each interval, (14) has transition

$$\bar{X}_{t_{m+1}} \mid \bar{X}_{t_m} = x \sim \mathcal{N}\big(x + \Delta b_\theta(x), \Delta a(x)\big).$$

Independent Brownian increments therefore give the grid density $p_{\Delta,\theta}$. Conditional on the grid states, the paths on distinct intervals are independent. On $[t_m, t_{m+1}]$, the conditional mean is the linear interpolation of the endpoints: conditioning on the endpoint replaces $W_{t_m+s} - W_{t_m}$ by its conditional mean $s(W_{t_{m+1}} - W_{t_m})/\Delta$, and the drift term cancels. The remaining covariance is that of a Brownian bridge with diffusion covariance $a(X_m)$.

Consequently, sample $X_{0:M} \sim q_\phi$ and independent standard d-dimensional Brownian bridges β_m on $[0,1]$, independent of $X_{0:M}$, and set

$$Z_{t_m+s} = \Big(1 - \frac{s}{\Delta}\Big)X_m + \frac{s}{\Delta}X_{m+1} + \sqrt{\Delta}\,a(X_m)^{1/2}\beta_m(s/\Delta), \quad 0 \le s \le \Delta. \tag{23}$$

This defines $\bar{Q}_{\Delta,\phi}$ and proves that its grid law is q_ϕ. The formula contains no θ, so this path law is independent of θ for fixed ϕ.

For a continuous path z, write $\pi_\Delta z = (z_{t_0}, \dots, z_{t_M})$. Let $K_\Delta(x, \mathrm{d}z)$ be the conditional bridge law specified by (23). The two path measures have the disintegrations

$$\bar{P}_{\Delta,\theta}(\mathrm{d}z) = \int p_{\Delta,\theta}(x) K_\Delta(x, \mathrm{d}z)\, \mathrm{d}x, \qquad \bar{Q}_{\Delta,\phi}(\mathrm{d}z) = \int q_\phi(x) K_\Delta(x, \mathrm{d}z)\, \mathrm{d}x.$$

The kernel is supported on paths with $\pi_\Delta z = x$, and $p_{\Delta,\theta}$ is strictly positive. Hence, for any bounded measurable path functional f,

$$\begin{aligned}\int f(z) \frac{q_\phi(\pi_\Delta z)}{p_{\Delta,\theta}(\pi_\Delta z)}\, \bar{P}_{\Delta,\theta}(\mathrm{d}z) &= \int q_\phi(x) \int f(z) K_\Delta(x, \mathrm{d}z)\, \mathrm{d}x \\ &= \int f(z)\, \bar{Q}_{\Delta,\phi}(\mathrm{d}z).\end{aligned}$$

It follows that

$$\frac{\mathrm{d}\bar{Q}_{\Delta,\phi}}{\mathrm{d}\bar{P}_{\Delta,\theta}}(z) = \frac{q_\phi(\pi_\Delta z)}{p_{\Delta,\theta}(\pi_\Delta z)}. \tag{24}$$

For each fixed grid, the linear growth of b_θ and the diffusion bounds imply

$$|\log p_{\Delta,\theta}(x_{0:M})| \le C_{\Delta,\theta} \left(1 + \sum_{m=0}^{M} \|x_m\|^2\right).$$

The nondegenerate Gaussian density q_ϕ also has a quadratic log density and finite second moments. Thus the log ratio in (24) is integrable under $\bar{Q}_{\Delta,\phi}$. Integrating it proves (15). The observation log densities are quadratic and depend only on grid states, so their expectations are finite and agree under $\bar{Q}_{\Delta,\phi}$ and q_ϕ. This proves (16).

(ii) First identify the law of X in Lemma 2 with P_θ. For a solution of the original SDE with diffusion σ, the process

$$\widetilde{W}_t = \int_0^t a(X_s)^{-1/2} \sigma(X_s)\, \mathrm{d}W_s$$

is a continuous local martingale with quadratic variation tI_d, and hence is Brownian motion by Lévy's characterization. The same state process solves the SDE with diffusion $a^{1/2}$ driven by $\widetilde{W}$. Global Lipschitz continuity of b_θ and $a^{1/2}$ gives uniqueness in law, so its path law is P_θ. The coupling in Lemma 2 therefore has marginals P_θ and $\bar{P}_{\Delta,\theta}$. By the definition of $W_{2,\infty}$,

$$W_{2,\infty}(\bar{P}_{\Delta,\theta}, P_\theta)^2 \le \mathbb{E} \sup_{t \le T} \|X_t - \bar{X}_t\|^2 \le C\Delta.$$

Taking square roots proves (17). □

For N trajectories, applying (16) to each trajectory gives

$$\frac{1}{N} \sum_{n=1}^{N} \bar{\mathcal{L}}_\Delta(\bar{Q}^{(n)}_{\Delta,\phi}, \theta) - \lambda \mathcal{R}(\theta) = \mathcal{F}_\Delta(\phi, \theta).$$

This identity concerns the exact objective in (4); the cubature approximation is the one defined in Section 3.3.

E Proof of Theorem 2

Proof sketch. This overview records the sufficient-ascent inequalities used in the proof. We use the notation and assumptions stated below. The blockwise ascent lemma gives, for either block and a direction d, whenever $\alpha \ge 0$ and the trial segment $[z, z + \alpha d]$ lies in $\mathcal{U}$,

$$\widetilde{\mathcal{F}}_\Delta(z + \alpha d) \ge \widetilde{\mathcal{F}}_\Delta(z) + \alpha \nabla \widetilde{\mathcal{F}}_\Delta(z)^\top d - \frac{L\alpha^2}{2} \|d\|_2^2.$$

The two gradient-related inequalities imply that sufficiently small steps satisfy the corresponding Armijo condition, so both backtracking procedures terminate at every trajectory and neural inner update. Compactness and uniform constants give a positive lower bound on accepted step sizes. For

each Gaussian trajectory, apply the ascent lemma to $\widetilde{\mathcal{F}}_{\Delta,n}$ with Lipschitz constant NL. Sum the resulting trajectory-wise Armijo inequalities and divide by N. Adding the sufficient ascent from the Gaussian and neural blocks, summed over all J neural inner steps, gives

$$\widetilde{\mathcal{F}}_{\Delta}(z_{k+1}) - \widetilde{\mathcal{F}}_{\Delta}(z_k) \geq c\big(\|g_{\phi,k}\|_2^2 + \sum_{j=0}^{J-1} \|g_{\theta,k,j}\|_2^2\big), \tag{25}$$

for some $c > 0$. The objective is bounded above on the compact set; summing (25) therefore makes the Gaussian and all J neural squared-gradient series finite. The block gradients vanish. Lipschitz continuity transfers this conclusion from $w_{k,0} = \bar{z}_k$, using $g_{\theta,k,0}$, to z_k, proving (18). Averaging (25) over the first K outer iterations, together with the cross-block gradient bound, gives (19). The final statement follows from the standard finite-length argument for objectives with the Kurdyka–Łojasiewicz property, using the multi-step displacement bound established below. This is a generalized-EM-style stationary point result (Wu, 1983); it does not imply a global optimum.

E.1 Standing assumptions

We analyze the trajectory-specific Gaussian updates of Algorithm 2 and the J neural updates of Algorithm 3 in each outer iteration, with N and the positive integer J fixed. Define the neural inner iterates by

$$w_{k,j} := (\phi_{k+1}, \theta_{k,j}), \quad j = 0, \ldots, J.$$

Thus, $w_{k,0} = \bar{z}_k$ and $w_{k,J} = z_{k+1}$. Define

$$\begin{aligned} g_{\phi,k} &:= \nabla_\phi \widetilde{\mathcal{F}}_{\Delta}(z_k) = N^{-1} \operatorname{col}\big(g_{\phi,k}^{(1)}, \ldots, g_{\phi,k}^{(N)}\big), \\ d_{\phi,k} &:= \operatorname{col}\big(d_{\phi,k}^{(1)}, \ldots, d_{\phi,k}^{(N)}\big). \end{aligned} \tag{26}$$

The local gradients $g_{\phi,k}^{(n)}$ are those of the unaveraged objectives $\widetilde{\mathcal{F}}_{\Delta,n}$; $g_{\phi,k}$ is the gradient of their average. For the neural updates, define

$$g_{\theta,k,j} := \nabla_\theta \widetilde{\mathcal{F}}_{\Delta}(w_{k,j}), \qquad j = 0, \ldots, J-1.$$

Let $d_{\theta,k,j}$ denote the direction used at inner step j. Constants below may depend on N and J, but are uniform in k, j, and n. We formalize the assumptions of Section 4 as follows.

(A2): **Interior compactness.** All z_k and $w_{k,j}$, for $k \geq 0$ and $j = 0, \ldots, J$, lie in a compact set $\mathcal{Z} \subset \mathcal{D} \times \mathbb{R}^p$. There is a convex open set $\mathcal{U}$ in this domain containing $\mathcal{Z}$, and $r > 0$ such that the Euclidean ball $B(z, r)$ lies in $\mathcal{U}$ for every $z \in \mathcal{Z}$.

(A3): **Smoothness.** $\widetilde{\mathcal{F}}_{\Delta}$ is continuously differentiable on $\mathcal{U}$, with a full-gradient Lipschitz constant $L > 0$ there. This includes both within-block and cross-block variations. Local Lipschitz regularity of b_0 alone is not this assumption.

(A4): **Exact gradients and gradient-related directions.** The two blocks use the exact gradients of their respective objectives, uncapped backtracking, and no finite-tolerance early stopping of these safeguarded updates. These objectives are $\widetilde{\mathcal{F}}_{\Delta,n}$ for trajectory n and $\widetilde{\mathcal{F}}_{\Delta}(\phi_{k+1}, \cdot)$ for every neural update within outer iteration k. There exist $\gamma_E, \Gamma_E, \gamma_M, \Gamma_M > 0$ such that every accepted E-step direction for each trajectory satisfies (12), and every accepted M-step direction $d_{\theta,k,j}$ satisfies the analogous inequalities with (γ_M, Γ_M) in place of (γ_E, Γ_E) at every inner step $j = 0, \ldots, J-1$. The Armijo constants satisfy $c_E, c_M \in (0, 1)$.

E.2 Preliminary lemmas

Lemma 3 (Boundedness on the compact set)**.** *There are finite constants $\widetilde{\mathcal{F}}_{\Delta}^*$ and G such that $\widetilde{\mathcal{F}}_{\Delta}(z) \leq \widetilde{\mathcal{F}}_{\Delta}^*$ and $\|\nabla \widetilde{\mathcal{F}}_{\Delta}(z)\| \leq G$ for every $z \in \mathcal{Z}$.*

Proof. $\widetilde{\mathcal{F}}_{\Delta}$ and $\nabla \widetilde{\mathcal{F}}_{\Delta}$ are continuous on $\mathcal{Z}$ by (A3), and $\mathcal{Z}$ is compact by (A2); a continuous function on a compact set attains its supremum. □

Lemma 4 (The fallback direction is gradient related)**.** *If $g \neq 0$ and $d = Hg$ for a symmetric matrix H with $\underline{b} I \preceq H \preceq \overline{b} I$, $\underline{b}, \overline{b} > 0$, then $g^\top d \geq \underline{b}\|g\|^2$ and $\|d\| \leq \overline{b}\|g\|$.*

Proof. $g^\top d = g^\top H g \geq \underline{b}\|g\|^2$ because $H - \underline{b}I \succeq 0$. For the second bound, $\|d\| = \|Hg\| \leq \|H\|_{\mathrm{op}}\|g\| \leq \bar{b}\|g\|$, since the operator norm of a symmetric matrix is its largest eigenvalue, at most $\bar{b}$. □

By Lemma 4, the fallback direction always satisfies the direction test with constants $(\underline{b}, \bar{b})$; taking $\gamma_E \leq \underline{b}$ and $\Gamma_E \geq \bar{b}$ makes the E-step direction conditions in (A4) hold uniformly whether the IEKS proposal or the fallback is selected. For the neural update, Algorithm 3 uses $d_{\theta,k,j} = g_{\theta,k,j}$ at every inner step, so the M-step direction conditions hold for any $\gamma_M \leq 1 \leq \Gamma_M$; the gradient identity follows from (27).

Lemma 5 (Two-sided norm control)**.** *Under (A4), the stacked Gaussian direction satisfies $N\gamma_E\|g_{\phi,k}\| \leq \|d_{\phi,k}\| \leq N\Gamma_E\|g_{\phi,k}\|$, and each neural direction satisfies*

$$\gamma_M\|g_{\theta,k,j}\| \leq \|d_{\theta,k,j}\| \leq \Gamma_M\|g_{\theta,k,j}\|, \qquad j = 0, \ldots, J-1,$$

with no factor N.

Proof. The upper bound follows by squaring and summing the trajectory-wise bounds in (A4), then using (26). The same scaling gives

$$g_{\phi,k}^\top d_{\phi,k} = N^{-1}\sum_{n=1}^{N}(g_{\phi,k}^{(n)})^\top d_{\phi,k}^{(n)} \geq N\gamma_E\|g_{\phi,k}\|^2.$$

If $g_{\phi,k} = 0$, then every local gradient is zero, and the trajectory-wise upper bounds in (A4) imply $d_{\phi,k} = 0$, so the claim is immediate. Hence suppose $g_{\phi,k} \neq 0$. For the lower bound, Cauchy–Schwarz gives $g_{\phi,k}^\top d_{\phi,k} \leq \|g_{\phi,k}\|\|d_{\phi,k}\|$; combined with $g_{\phi,k}^\top d_{\phi,k} \geq N\gamma_E\|g_{\phi,k}\|^2$ and dividing by $\|g_{\phi,k}\| > 0$ gives $\|d_{\phi,k}\| \geq N\gamma_E\|g_{\phi,k}\|$. The same argument applies to each neural inner step, without the trajectory scaling factor N. □

Lemma 6 (Ascent lemma)**.** *Under (A3), for every $z \in \mathcal{U}$, direction d, and $\alpha \geq 0$ whose entire segment $[z, z + \alpha d]$ lies in $\mathcal{U}$,*

$$\widetilde{\mathcal{F}}_\Delta(z + \alpha d) \geq \widetilde{\mathcal{F}}_\Delta(z) + \alpha\nabla\widetilde{\mathcal{F}}_\Delta(z)^\top d - \frac{L}{2}\alpha^2\|d\|^2.$$

Proof. By the fundamental theorem of calculus and Cauchy–Schwarz,

$$\begin{aligned}\widetilde{\mathcal{F}}_\Delta(z + \alpha d) - \widetilde{\mathcal{F}}_\Delta(z) - \alpha\nabla\widetilde{\mathcal{F}}_\Delta(z)^\top d &= \int_0^\alpha \left[\nabla\widetilde{\mathcal{F}}_\Delta(z + td) - \nabla\widetilde{\mathcal{F}}_\Delta(z)\right]^\top d\,\mathrm{d}t \\ &\geq -\int_0^\alpha Lt\|d\|^2\,\mathrm{d}t = -\frac{L}{2}\alpha^2\|d\|^2,\end{aligned}$$

using the Lipschitz bound of (A3). □

Lemma 7 (Uniform lower bound on accepted step sizes)**.** *For every nonzero block gradient, the corresponding backtracking accepts a step bounded below by a positive constant independent of k, the neural inner-step index j, and of the trajectory index n. The constants depend on L, r, G, N, the direction bounds, and the Armijo constants.*

Proof. Consider one trajectory update or one neural inner update, with constants γ, Γ, c_A. Let f be the objective of that update, x its current parameter vector, and g, d its gradient and direction. For trajectory n, f is $\widetilde{\mathcal{F}}_{\Delta,n}$ with θ_k fixed; for the neural update, $f = \widetilde{\mathcal{F}}_\Delta(\phi_{k+1}, \cdot)$. By separability, Gaussian trials can be embedded at z_k with the other coordinates fixed; neural inner step j starts at $w_{k,j}$. Both starting points lie in $\mathcal{Z}$ by (A2). By (A3) and (26), the local gradient at the current iterate is bounded by $G_f = NG$ and its Lipschitz constant is $L_f = NL$ for a Gaussian trajectory update. For the neural update, take $G_f = G$ and $L_f = L$. These Lipschitz bounds hold on trial segments embedded in $\mathcal{U}$ with the other parameters fixed. By (A4), $\|d\| \leq \Gamma\|g\| \leq \Gamma G_f$. When $G = 0$, all iterates are stationary and no search is needed. Otherwise every $0 \leq \alpha \leq r/(2\Gamma G_f)$ keeps the

whole trial segment of the full parameter vector inside $\mathcal{U}$, and hence inside the feasible domain. On such a segment the ascent lemma gives

$$f(x+\alpha d) - f(x) \geq \alpha g^\top d - \frac{L_f \alpha^2}{2}\|d\|^2.$$

The Armijo inequality holds whenever $\alpha \leq 2(1-c_A)\gamma/(L_f\Gamma^2)$, using the two direction bounds. Consequently, with

$$t_* = \min\left\{1, \frac{r}{2\Gamma G_f}, \frac{2(1-c_A)\gamma}{L_f\Gamma^2}\right\} > 0,$$

every $\alpha \in (0, t_*]$ is both feasible and acceptable. Halving from one therefore accepts a step at least $t_*/2$. Apply this argument to the Gaussian trajectory updates and each neural inner update separately to obtain $\bar{\alpha}$ and $\bar{\beta}$. The Gaussian constants are common to all trajectories, so $\alpha_{n,k} \geq \bar{\alpha}$ whenever $g^{(n)}_{\phi,k} \neq 0$. The neural bound is $\beta_{k,j} \geq \bar{\beta}$ whenever $g_{\theta,k,j} \neq 0$, uniformly in k and j. If a block gradient is zero, (A4) forces its direction to be zero and that block is left unchanged at that update. □

Lemma 7 shows in particular that both backtracking loops in Algorithm 1 terminate after finitely many halvings at every trajectory and neural inner update, so the algorithm is well defined at every iteration.

E.3 Monotone ascent

Lemma 8 (Monotone ascent)**.** *For every k, $\widetilde{\mathcal{F}}_\Delta(\bar{z}_k) \geq \widetilde{\mathcal{F}}_\Delta(z_k)$ and $\widetilde{\mathcal{F}}_\Delta(z_{k+1}) \geq \widetilde{\mathcal{F}}_\Delta(\bar{z}_k)$; hence $\widetilde{\mathcal{F}}_\Delta(z_{k+1}) \geq \widetilde{\mathcal{F}}_\Delta(z_k)$, as stated in* (18)*.*

Proof. By (13) and (A4), separability of the Gaussian objective gives

$$\begin{aligned}\widetilde{\mathcal{F}}_\Delta(\bar{z}_k) - \widetilde{\mathcal{F}}_\Delta(z_k) &= \frac{1}{N}\sum_{n=1}^{N}\left[\widetilde{\mathcal{F}}_{\Delta,n}(\phi^{(n)}_{k+1}, \theta_k) - \widetilde{\mathcal{F}}_{\Delta,n}(\phi^{(n)}_k, \theta_k)\right] \\ &\geq \frac{c_E}{N}\sum_{n=1}^{N}\alpha_{n,k}(g^{(n)}_{\phi,k})^\top d^{(n)}_{\phi,k} \geq 0.\end{aligned}$$

For the neural update, fixing ϕ_{k+1} in (11) and comparing with (10) gives

$$\widehat{\mathcal{L}}_k(\theta) = -\widetilde{\mathcal{F}}_\Delta(\phi_{k+1}, \theta) + K_{\phi_{k+1}}, \tag{27}$$

where $K_{\phi_{k+1}}$ is independent of θ. Together with the Armijo condition in Algorithm 3, this identity gives, upon summing over the J neural inner steps,

$$\begin{aligned}\widetilde{\mathcal{F}}_\Delta(z_{k+1}) - \widetilde{\mathcal{F}}_\Delta(\bar{z}_k) = \widehat{\mathcal{L}}_k(\theta_k) - \widehat{\mathcal{L}}_k(\theta_{k+1}) &= \sum_{j=0}^{J-1}\left[\widehat{\mathcal{L}}_k(\theta_{k,j}) - \widehat{\mathcal{L}}_k(\theta_{k,j+1})\right] \\ &\geq c_M\sum_{j=0}^{J-1}\beta_{k,j}\|g_{\theta,k,j}\|^2 \geq 0.\end{aligned}$$

□

E.4 Sufficient ascent

Define $R_k \geq 0$ by $R_k^2 := \|g_{\phi,k}\|^2 + \sum_{j=0}^{J-1}\|g_{\theta,k,j}\|^2$.

Lemma 9 (Sufficient ascent)**.** *There is $c > 0$, depending only on the constants of Lemma 7 and (A4), such that*

$$\widetilde{\mathcal{F}}_\Delta(z_{k+1}) - \widetilde{\mathcal{F}}_\Delta(z_k) \geq cR_k^2 \qquad \textit{for every } k.$$

This is (25)*.*

Proof. By (13), Lemma 7, and (A4), summing over the trajectories yields

$$\begin{aligned}\widetilde{\mathcal{F}}_\Delta(\bar{z}_k) - \widetilde{\mathcal{F}}_\Delta(z_k) &\geq \frac{c_E \bar{\alpha} \gamma_E}{N} \sum_{n=1}^{N} \|g_{\phi,k}^{(n)}\|^2 \\ &= c_E \bar{\alpha} \gamma_E N \|g_{\phi,k}\|^2.\end{aligned}$$

Zero local gradients contribute zero to this inequality. For the neural update, summing the Armijo bounds from Lemma 8 and using Lemma 7 gives

$$\widetilde{\mathcal{F}}_\Delta(z_{k+1}) - \widetilde{\mathcal{F}}_\Delta(\bar{z}_k) \geq c_M \bar{\beta} \sum_{j=0}^{J-1} \|g_{\theta,k,j}\|^2.$$

Zero neural gradients contribute zero, so their step sizes do not affect this bound. Summing and setting $c := \min(c_E \bar{\alpha} \gamma_E N,\ c_M \bar{\beta})$ gives the claim. □

E.5 Convergence of the gradient at z_k

Proof of (18). By Lemma 3, $\widetilde{\mathcal{F}}_\Delta \leq \widetilde{\mathcal{F}}_\Delta^*$ on $\mathcal{Z}$; by Lemma 8, $(\widetilde{\mathcal{F}}_\Delta(z_k))_k$ is nondecreasing, hence converges to some $\widetilde{\mathcal{F}}_\Delta^\infty \leq \widetilde{\mathcal{F}}_\Delta^*$. Summing the sufficient-ascent inequality of Lemma 9 from $k = 0$ to $K - 1$ and telescoping,

$$\begin{aligned}c \sum_{k=0}^{K-1} R_k^2 &\leq \widetilde{\mathcal{F}}_\Delta(z_K) - \widetilde{\mathcal{F}}_\Delta(z_0) \\ &\leq \widetilde{\mathcal{F}}_\Delta^* - \widetilde{\mathcal{F}}_\Delta(z_0)\end{aligned}$$

for every K; letting $K \to \infty$, the series on the left converges, so in particular

$$\|g_{\phi,k}\| \to 0, \qquad \|g_{\theta,k,j}\| \to 0 \quad \text{for each } j = 0, \ldots, J-1. \tag{28}$$

Since $g_{\phi,k} = \nabla_\phi \widetilde{\mathcal{F}}_\Delta(z_k)$ exactly, $\|\nabla_\phi \widetilde{\mathcal{F}}_\Delta(z_k)\| \to 0$. For the θ-block, $g_{\theta,k,0} = \nabla_\theta \widetilde{\mathcal{F}}_\Delta(\bar{z}_k)$ is evaluated at $\bar{z}_k$, not z_k; by (A3) and Lemma 5,

$$\begin{aligned}\|\nabla_\theta \widetilde{\mathcal{F}}_\Delta(z_k) - g_{\theta,k,0}\| &= \|\nabla_\theta \widetilde{\mathcal{F}}_\Delta(\phi_k, \theta_k) - \nabla_\theta \widetilde{\mathcal{F}}_\Delta(\phi_{k+1}, \theta_k)\| \\ &\leq L \|\phi_{k+1} - \phi_k\| \\ &= L \left(\sum_{n=1}^{N} \alpha_{n,k}^2 \|d_{\phi,k}^{(n)}\|^2 \right)^{1/2} \leq L N \Gamma_E \|g_{\phi,k}\|,\end{aligned}$$

using $\alpha_{n,k} \leq 1$ for every n and (26). Hence $\|\nabla_\theta \widetilde{\mathcal{F}}_\Delta(z_k)\| \leq \|g_{\theta,k,0}\| + L N \Gamma_E \|g_{\phi,k}\| \to 0$ by (28). Combining the two blocks,

$$\begin{aligned}\|\nabla \widetilde{\mathcal{F}}_\Delta(z_k)\|^2 &= \|\nabla_\phi \widetilde{\mathcal{F}}_\Delta(z_k)\|^2 + \|\nabla_\theta \widetilde{\mathcal{F}}_\Delta(z_k)\|^2 \\ &\leq \|g_{\phi,k}\|^2 + 2\|g_{\theta,k,0}\|^2 + 2L^2 N^2 \Gamma_E^2 \|g_{\phi,k}\|^2 \to 0,\end{aligned}$$

which is (18). □

E.6 Stationarity of accumulation points

Proof that accumulation points are stationary. Since $\mathcal{Z}$ is compact, (z_k) has at least one accumulation point. Let $z_{k_j} \to z^* \in \mathcal{Z}$ be any convergent subsequence. By continuity of $\nabla \widetilde{\mathcal{F}}_\Delta$ (A3), $\nabla \widetilde{\mathcal{F}}_\Delta(z_{k_j}) \to \nabla \widetilde{\mathcal{F}}_\Delta(z^*)$; by (18), $\nabla \widetilde{\mathcal{F}}_\Delta(z_k) \to 0$ along the full sequence and hence along this subsequence. Therefore $\nabla \widetilde{\mathcal{F}}_\Delta(z^*) = 0$. □

E.7 Iteration complexity

Proof of (19). From the proof of (18), for every $K \geq 1$,

$$\sum_{k=0}^{K-1} R_k^2 \leq \frac{\widetilde{\mathcal{F}}_\Delta^* - \widetilde{\mathcal{F}}_\Delta(z_0)}{c} =: \frac{C_0}{c}.$$

Also from that proof, $\|\nabla\widetilde{\mathcal{F}}_\Delta(z_k)\|^2 \le C_1 R_k^2$ with $C_1 := \max(1+2L^2N^2\Gamma_E^2,\, 2)$, since $\|g_{\theta,k,0}\|^2 \le \sum_{j=0}^{J-1}\|g_{\theta,k,j}\|^2$. Hence

$$\min_{0\le k<K}\|\nabla\widetilde{\mathcal{F}}_\Delta(z_k)\|^2 \le C_1\cdot\frac{1}{K}\sum_{k=0}^{K-1}R_k^2 \le \frac{C_1C_0}{cK}\le\frac{C}{K},$$

which is (19). Here $C := \max\{1, C_1C_0/c\} > 0$, K counts outer iterations, and C is independent of K for fixed N and J. □

E.8 Full-sequence convergence under the Kurdyka–Łojasiewicz property

We use the standard Kurdyka–Łojasiewicz (KL) property (Attouch et al., 2013), adapted to a function that is ascended rather than descended. Let $\mathcal{K}_\eta$ denote the set of concave functions $\varphi : [0,\eta) \to [0,\infty)$ with $\varphi(0) = 0$, continuously differentiable on $(0,\eta)$ with $\varphi' > 0$ there, and continuous at 0.

Definition 1 (KL property, ascent form). *$\widetilde{\mathcal{F}}_\Delta$ satisfies the KL property at $z^* \in \mathcal{Z}$ if there exist $\eta > 0$, a neighborhood U of z^*, and $\varphi \in \mathcal{K}_\eta$ such that*

$$\varphi'\big(\widetilde{\mathcal{F}}_\Delta(z^*) - \widetilde{\mathcal{F}}_\Delta(z)\big)\,\|\nabla\widetilde{\mathcal{F}}_\Delta(z)\| \ge 1 \qquad \textit{for all } z\in U \textit{ with } 0<\widetilde{\mathcal{F}}_\Delta(z^*)-\widetilde{\mathcal{F}}_\Delta(z)<\eta.$$

This is the usual KL inequality applied to $h := -\widetilde{\mathcal{F}}_\Delta$ at z^*; z^* need not be a local minimizer of h. Real-analytic objectives and smooth semialgebraic objectives satisfy this property locally. We retain the KL property as an explicit hypothesis for the general model (Attouch et al., 2013), in agreement with the hypothesis of Theorem 2.

Proof of full-sequence convergence. Let z^* be any accumulation point of (z_k) (one exists by compactness), with $z_{k_j} \to z^*$ along a subsequence. Because $\widetilde{\mathcal{F}}_\Delta(z_k) \to \widetilde{\mathcal{F}}_\Delta^\infty$ monotonically and $\widetilde{\mathcal{F}}_\Delta$ is continuous, $\widetilde{\mathcal{F}}_\Delta(z_{k_j}) \to \widetilde{\mathcal{F}}_\Delta(z^*)$, forcing $\widetilde{\mathcal{F}}_\Delta(z^*) = \widetilde{\mathcal{F}}_\Delta^\infty$.

If $\widetilde{\mathcal{F}}_\Delta(z_{k_0}) = \widetilde{\mathcal{F}}_\Delta^\infty$ for some finite k_0, then by monotonicity $\widetilde{\mathcal{F}}_\Delta(z_k) = \widetilde{\mathcal{F}}_\Delta^\infty$ for all $k \ge k_0$, so Lemma 9 forces $g_{\phi,k} = g_{\theta,k,j} = 0$ for every $j = 0,\dots,J-1$, hence $z_{k+1} = z_k$, for all $k \ge k_0$: the sequence is eventually constant at the stationary point z_{k_0}, and the claim holds trivially. Assume from now on that $\widetilde{\mathcal{F}}_\Delta(z_k) < \widetilde{\mathcal{F}}_\Delta^\infty$ for every k.

Fix KL data (η, U, φ) at z^*, and write $U = B(z^*,\epsilon_0)$ without loss of generality. By Lemma 5, $\beta_{k,j} \le 1$, and Cauchy–Schwarz,

$$\|\theta_{k+1}-\theta_k\| \le \sum_{j=0}^{J-1}\beta_{k,j}\|d_{\theta,k,j}\| \le \Gamma_M\sqrt{J}\left(\sum_{j=0}^{J-1}\|g_{\theta,k,j}\|^2\right)^{1/2}.$$

Together with $\|\phi_{k+1}-\phi_k\| \le N\Gamma_E\|g_{\phi,k}\|$, this gives

$$\|z_{k+1}-z_k\| \le \Gamma_{\max}\,R_k, \qquad \Gamma_{\max} := \max(N\Gamma_E, \sqrt{J}\,\Gamma_M),$$

and from the proof of (19), $\|\nabla\widetilde{\mathcal{F}}_\Delta(z_k)\| \le \sqrt{C_1}\,R_k$, and from Lemma 9, $\widetilde{\mathcal{F}}_\Delta(z_{k+1}) - \widetilde{\mathcal{F}}_\Delta(z_k) \ge cR_k^2$.

Choice of k_1. Since $z_{k_j} \to z^*$ and $\widetilde{\mathcal{F}}_\Delta(z_{k_j}) \to \widetilde{\mathcal{F}}_\Delta^\infty$, choose $\rho \in (0,\epsilon_0)$ and k_1 along the subsequence such that $z_{k_1} \in B(z^*,\rho/2)$, $\widetilde{\mathcal{F}}_\Delta^\infty - \widetilde{\mathcal{F}}_\Delta(z_{k_1}) < \eta$, and

$$\frac{\Gamma_{\max}\sqrt{C_1}}{c}\,\varphi\big(\widetilde{\mathcal{F}}_\Delta^\infty - \widetilde{\mathcal{F}}_\Delta(z_{k_1})\big) < \frac{\rho}{2}; \tag{29}$$

this is possible because φ is continuous at 0 with $\varphi(0) = 0$, and $\widetilde{\mathcal{F}}_\Delta(z_{k_j}) \to \widetilde{\mathcal{F}}_\Delta^\infty$ along the subsequence.

Induction. We claim $z_k \in B(z^*,\rho)$ for every $k \ge k_1$. This holds for $k = k_1$ by construction. Suppose it holds for $k_1,\dots,k$. Then $\widetilde{\mathcal{F}}_\Delta(z_{k_1}) \le \widetilde{\mathcal{F}}_\Delta(z_j) < \widetilde{\mathcal{F}}_\Delta^\infty$ for $j = k_1,\dots,k$ by monotonicity,

so $0 < \widetilde{\mathcal{F}}_\Delta^\infty - \widetilde{\mathcal{F}}_\Delta(z_j) \le \widetilde{\mathcal{F}}_\Delta^\infty - \widetilde{\mathcal{F}}_\Delta(z_{k_1}) < \eta$ and $z_j \in U$; the KL inequality therefore applies at each such z_j:

$$\varphi'\big(\widetilde{\mathcal{F}}_\Delta^\infty - \widetilde{\mathcal{F}}_\Delta(z_j)\big)\,\|\nabla\widetilde{\mathcal{F}}_\Delta(z_j)\| \ge 1, \qquad j = k_1, \dots, k.$$

Set $\Delta_j := \varphi(\widetilde{\mathcal{F}}_\Delta^\infty - \widetilde{\mathcal{F}}_\Delta(z_j)) \ge 0$, a nonincreasing sequence since φ is increasing and $\widetilde{\mathcal{F}}_\Delta(z_j)$ is nondecreasing. By concavity of φ,

$$\Delta_j - \Delta_{j+1} \ge \varphi'\big(\widetilde{\mathcal{F}}_\Delta^\infty - \widetilde{\mathcal{F}}_\Delta(z_j)\big)\big[\widetilde{\mathcal{F}}_\Delta(z_{j+1}) - \widetilde{\mathcal{F}}_\Delta(z_j)\big] \ge \frac{1}{\|\nabla\widetilde{\mathcal{F}}_\Delta(z_j)\|}\cdot cR_j^2 \ge \frac{c}{\sqrt{C_1}}R_j,$$

using the KL inequality and then $\|\nabla\widetilde{\mathcal{F}}_\Delta(z_j)\| \le \sqrt{C_1}R_j$. Hence $R_j \le (\sqrt{C_1}/c)(\Delta_j - \Delta_{j+1})$, and

$$\begin{aligned}\|z_{k+1} - z_{k_1}\| &\le \sum_{j=k_1}^{k}\|z_{j+1} - z_j\| \le \Gamma_{\max}\sum_{j=k_1}^{k} R_j \\ &\le \frac{\Gamma_{\max}\sqrt{C_1}}{c}\sum_{j=k_1}^{k}(\Delta_j - \Delta_{j+1}) \le \frac{\Gamma_{\max}\sqrt{C_1}}{c}\Delta_{k_1} < \frac{\rho}{2}\end{aligned}$$

by (29). Combined with $\|z_{k_1} - z^*\| < \rho/2$, the triangle inequality gives $\|z_{k+1} - z^*\| < \rho$, completing the induction.

Finite length. The bound above holds for every $k \ge k_1$ with the same right-hand side, so letting $k \to \infty$,

$$\sum_{j=k_1}^{\infty}\|z_{j+1} - z_j\| \le \frac{\Gamma_{\max}\sqrt{C_1}}{c}\Delta_{k_1} < \infty.$$

A sequence with finite total variation is Cauchy, hence convergent; since (z_k) already has z^* as an accumulation point, its limit must be z^*. Thus $z_k \to z^*$, and z^* is a stationary point by the preceding subsection. This proves the final claim of Theorem 2. □

This completes the proof of Theorem 2. □

F Experimental Settings

F.1 VEST algorithm

Algorithm 1 VEST

Require: Observations $\mathcal{D}_N$; Δ, K, L, J, ρ; known m_0, P_0, σ, R

1: Initialize ϕ_0 by zero-drift smoothing using $\mathcal{D}_N, \Delta, m_0, P_0, \sigma, R$.
2: Initialize the drift-network parameters θ_0.
3: **for** $k = 0, \dots, K-1$ **do**
4: $\quad\phi_{k+1} \leftarrow \textsc{GaussianUpdate}(\phi_k, \theta_k, \mathcal{D}_N, \Delta, m_0, P_0, \sigma, R, L, \rho)$
5: $\quad\theta_{k+1} \leftarrow \textsc{DriftUpdate}(\phi_{k+1}, \theta_k, \Delta, \sigma, J)$
6: **end for**
7: Set $\widehat{\theta}_{\Delta,K} \leftarrow \theta_K$.
8: **return** $\widehat{b}_{\Delta,K} := b_{\widehat{\theta}_{\Delta,K}}$.

Algorithm 2 Gaussian update in VEST

procedure GAUSSIANUPDATE($\phi_k, \theta_k, \mathcal{D}_N, \Delta, m_0, P_0, \sigma, R, L, \rho$)
  for $n = 1, \dots, N$ **do**
    Set $\bar{x}_m^{(n)} \leftarrow \mathbb{E}_{q_{\phi_k}^{(n)}}[X_m^{(n)}]$ for $m = 0, \dots, M$.
    for $\ell = 1, \dots, L$ **do**
      Compute A_m, c_m, Q_m at $\bar{x}_m^{(n)}$ using (6).
      Filter $Y_{0:M}^{(n)}$ using m_0, P_0, R and apply RTS smoothing to obtain μ^s, P^s, C^s.
      Set $\bar{x}_m^{(n)} \leftarrow (1-\rho)\bar{x}_m^{(n)} + \rho\mu_m^s$ for $m = 0, \dots, M$.
      if $\bar{x}^{(n)}$ has stabilized **then**
        break
      end if
    end for
    Recompute A_m, c_m, Q_m at the final $\bar{x}^{(n)}$ and smooth again.
    Set $\widehat{\phi}_{k+1}^{(n)} \leftarrow (\mu^s, P^s, C^s)$.
    Set $g_{\phi,k}^{(n)} \leftarrow \nabla_{\phi^{(n)}} \widetilde{\mathcal{F}}_{\Delta,n}(\phi_k^{(n)}, \theta_k)$ and $d_{\phi,k}^{(n)} \leftarrow \widehat{\phi}_{k+1}^{(n)} - \phi_k^{(n)}$.
    if $(g_{\phi,k}^{(n)})^\top d_{\phi,k}^{(n)} < \gamma_E \|g_{\phi,k}^{(n)}\|_2^2$ **or** $\|d_{\phi,k}^{(n)}\|_2 > \Gamma_E \|g_{\phi,k}^{(n)}\|_2$ **then**
      Set $d_{\phi,k}^{(n)} \leftarrow H_k^{(n)} g_{\phi,k}^{(n)}$.
    end if
    Choose the first $\alpha_{n,k} \in \{1, 1/2, \dots\}$ for which the trial $P_m^{(n)}, B_m^{(n)} \succ 0$ for every m and $\widetilde{\mathcal{F}}_{\Delta,n}(\phi_k^{(n)} + \alpha_{n,k} d_{\phi,k}^{(n)}, \theta_k) \geq \widetilde{\mathcal{F}}_{\Delta,n}(\phi_k^{(n)}, \theta_k) + c_E \alpha_{n,k} (g_{\phi,k}^{(n)})^\top d_{\phi,k}^{(n)}$.
    Set $\phi_{k+1}^{(n)} \leftarrow \phi_k^{(n)} + \alpha_{n,k} d_{\phi,k}^{(n)}$.
  end for
  return $\phi_{k+1} := (\phi_{k+1}^{(1)}, \dots, \phi_{k+1}^{(N)})$.
end procedure

Algorithm 3 Drift update in VEST

procedure DRIFTUPDATE($\phi_{k+1}, \theta_k, \Delta, \sigma, J$)
  Compute $u_m^{(n)}$ and $\chi_{m,j}^{(n),\pm}$ from the accepted Gaussian moments.
  Form $\widehat{\mathcal{L}}_k$ using (10). Set $\theta_{k,0} \leftarrow \theta_k$.
  for $j = 0, \dots, J-1$ **do**
    Set $g_{\theta,k,j} \leftarrow -\nabla_\theta \widehat{\mathcal{L}}_k(\theta_{k,j})$.
    Choose the first $\beta_{k,j} \in \{1, 1/2, \dots\}$ satisfying $\widehat{\mathcal{L}}_k(\theta_{k,j} + \beta_{k,j} g_{\theta,k,j}) \leq \widehat{\mathcal{L}}_k(\theta_{k,j}) - c_M \beta_{k,j} \|g_{\theta,k,j}\|_2^2$.
    Set $\theta_{k,j+1} \leftarrow \theta_{k,j} + \beta_{k,j} g_{\theta,k,j}$.
  end for
  return $\theta_{k,J}$.
end procedure

F.2 DRIFT ESTIMATION SETTINGS

Systems Setup. For each system, we simulate N independent trajectories from $dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t$ by Euler–Maruyama at the inference step size Δ. The one-dimensional systems use $N = 40$ trajectories, and the two-dimensional systems use $N = 25$. The five- and ten-dimensional systems use $N = 15$. All systems use $\Delta = 0.02$. The one- and ten-dimensional systems use $M = 100$ steps, the two-dimensional systems use $M = 60$ steps, and the five-dimensional ring uses $M = 40$ steps. Observations satisfy $Y_m = X_{m\Delta} + \varepsilon_m$, with independent $\varepsilon_m \sim \mathcal{N}(0, R)$. The initial state is drawn from $\mathcal{N}(m_0, P_0)$. Within each system and data seed, the compared methods use the same simulated paths, observations, and drift-evaluation points. Table 1 reports results over data seeds 0–4. VEST's network initialization uses seed 42. The following are the system settings.

Ornstein–Uhlenbeck (OU) with constant diffusion:

$$dX_t = -0.7X_t\,dt + 0.4\,dW_t,\ (m_0, P_0, R) = (1, 0.3, 0.35^2).$$

Saturating drift:

$$dX_t = -2\tanh(2X_t)\,dt + 0.5\,dW_t,\ (m_0, P_0, R) = (0, 0.5, 0.25^2).$$

Double well with constant diffusion:

$$dX_t = (X_t - X_t^3)\,dt + 0.45\,dW_t,\ (m_0, P_0, R) = (0, 1, 0.25^2).$$

Linear spiral:

$$dX_t = \begin{pmatrix} -0.5 & -1.5 \\ 1.5 & -0.5 \end{pmatrix} X_t\,dt + 0.3\,dW_t,\ (m_0, P_0, R) = ((1,0)^\top, 0.3I_2, 0.2^2 I_2).$$

Van der Pol oscillator with constant diffusion:

$$dX_t = \begin{pmatrix} X_{2,t} \\ (1 - X_{1,t}^2)X_{2,t} - X_{1,t} \end{pmatrix} dt + 0.3\,dW_t,\ (m_0, P_0, R) = ((0.5, 0.5)^\top, 0.4I_2, 0.2^2 I_2).$$

Bistable ring. For $i = 1, \ldots, 5$, indices are interpreted modulo five:

$$dX_{i,t} = \big[X_{i,t} - X_{i,t}^3 + 0.5(X_{i-1,t} - 2X_{i,t} + X_{i+1,t})\big]dt + 0.35\,dW_{i,t},$$

with $(m_0, P_0, R) = (0, 0.6I_5, 0.2^2 I_5)$.

Stochastic Kuramoto system. For $i = 1, \ldots, 10$:

$$dX_{i,t} = \left[\omega_i + \frac{1.5}{10}\sum_{j=1}^{10}\sin(X_{j,t} - X_{i,t})\right] dt + 0.35\,dW_{i,t},$$

with $m_{0,i} = -0.9 + 0.2(i-1)$, $(P_0, R) = (0.15I_{10}, 0.2^2 I_{10})$. The frequencies ω_i are evenly spaced from -0.5 to 0.5 (Gottwald, 2017).

Ornstein–Uhlenbeck (OU) with state-dependent diffusion:

$$dX_t = -0.7X_t\,dt + 0.3\sqrt{1 + 1.2X_t^2}\,dW_t,\ (m_0, P_0, R) = (1, 0.3, 0.3^2).$$

Double well with state-dependent diffusion:

$$dX_t = (X_t - X_t^3)\,dt + 0.35\sqrt{1 + 0.8X_t^2}\,dW_t,\ (m_0, P_0, R) = (0, 1, 0.25^2).$$

Van der Pol oscillator with state-dependent diffusion:

$$dX_t = \begin{pmatrix} X_{2,t} \\ (1 - X_{1,t}^2)X_{2,t} - X_{1,t} \end{pmatrix} dt + 0.22\sqrt{1 + 0.2\|X_t\|_2^2}\,dW_t,$$

with $(m_0, P_0, R) = ((0.5, 0.5)^\top, 0.4I_2, 0.2^2 I_2)$.

Network settings. VEST represents the drift with a $\tanh$ multilayer perceptron whose input and output dimensions equal the state dimension. The network has one hidden layer of width 24 for the one-dimensional systems, 32 for the two-dimensional systems, and 48 for the five- and ten-dimensional systems. The constant-diffusion double-well setting instead uses two hidden layers, each of width 48.

Training settings. The IEKS iteration cap is 20 for one-dimensional systems, 15 for two-dimensional systems, and 12 for five- and ten-dimensional systems. Under constant diffusion, OU, saturating drift, and double well each use 16 outer iterations with 10 neural updates per outer iteration. Linear spiral uses 10 outer iterations with 8 neural updates per outer iteration. Van der Pol uses 24 outer iterations with 20 neural updates per outer iteration. Coupled bistable ring and Kuramoto each use 6 outer iterations with 6 neural updates per outer iteration. Under state-dependent diffusion, OU and double well each use 14 outer iterations with 8 neural updates per outer iteration. Van der Pol uses 24 outer iterations with 20 neural updates per outer iteration. During smoothing, covariance eigenvalues below 10^{-4} are set to 10^{-4}, and the initial adjacent-state covariances are adjusted if needed. For each Gaussian or neural update, we try at most 25 successively halved step sizes and skip the update if none is accepted. Training stops before K rounds if the relative drift change on fixed probe points stays below 10^{-4} for two rounds with no Gaussian search failures, or if all Gaussian searches fail and no neural step is accepted.

Evaluation. Let $E = \{x_i\}_{i=1}^{N_e}$ be the evaluation points. We report

$$\mathrm{RMSE}_E = \left[\frac{1}{N_e} \sum_{i=1}^{N_e} \|\widehat{b}(x_i) - b_0(x_i)\|_2^2 \right]^{1/2}.$$

In one dimension, E consists of the points on a fixed 200-point grid that fall between the empirical 2nd and 98th percentiles of the latent states. In two dimensions, E is a 25×25 grid over the corresponding coordinatewise percentile box. In five dimensions, E consists of 400 states sampled from training trajectories. In ten dimensions, it consists of 400 states from independently simulated trajectories.

Baselines. The baselines cover five approaches to drift estimation from noisy trajectories. Naive increment regression (Naive) uses observed increments directly, whereas the adapted Schmisser estimator (Penalized LS) combines pre-averaging with penalized least squares to estimate drift in one-dimensional systems (Schmisser, 2011). The adapted EM-SMC-RKHS estimator combines particle-based latent-state inference with kernel drift estimation (Ganguly et al., 2025). SING learns a neural drift through variational inference under constant diffusion (Hu et al., 2025), and SDE Matching matches posterior and prior SDE dynamics (Bartosh et al., 2025).

For each system and seed, all methods are fitted to the same noisy trajectories and evaluated at the same points. We use the authors' released implementations of SING[1] and SDE Matching[2]. Our wrappers set the observation map to the identity and the noise covariance to the known value R. For SING, we use five-point Gauss–Hermite quadrature in one and two dimensions and one-sample Monte Carlo integration in five and ten dimensions. For SDE Matching, we supply the known diagonal diffusion, rescale time to $[0, 1]$ during training, and convert the estimated drift back to physical time for evaluation. We include the reconstruction factor specified in Eq. (24) of Bartosh et al. (2025) and train for 4000 Adam steps with learning rate 10^{-3} and hidden width 100.

We implement Naive, Penalized LS, and EM–SMC–RKHS in our benchmark code. Naive regresses $(Y_{m+1} - Y_m)/\Delta$ on Y_m using a one-hidden-layer network. Its hidden width is 24 in one dimension, 32 in two dimensions, and 48 in five or ten dimensions. We train it for 3000 steps in one and two dimensions and 2000 steps otherwise, with learning rate 0.02 and regularization coefficient 10^{-4}. For Penalized LS, we average observations in non-overlapping blocks of five within each trajectory, pool the resulting data, and select among up to 21 trigonometric basis functions with penalty coefficient 2. Our adapted EM–SMC–RKHS uses six particles, retains the three highest-weight particles, and performs eight EM iterations with 96 kernel centers and ridge coefficient 10^{-2}.

F.3 Sensitivity results

Simulation and evaluation protocol. We simulate latent trajectories with Euler-Maruyama at a step size of 0.0005. For the double well, linear spiral, and coupled bistable ring, respectively, we retain the main experiment trajectory counts and horizons $(N, T) = (40, 2.0)$, $(25, 1.2)$, and

[1] https://github.com/lindermanlab/sing.

[2] https://github.com/GrigoryBartosh/sde_matching.

$(15, 0.8)$, as well as the system diffusion coefficients and VEST architectures. The reference interval is $\Delta_0 = 0.02$; the reference noise standard deviation is $\tau_0 = 0.25$ for the double well and 0.20 for the other two systems. For the sampling interval sweep, we observe each simulated trajectory at different time intervals while keeping its total duration T fixed. For the observation noise sweep, we generate one set of standard Gaussian noise values and scale it to each noise level. The following table provides the numerical values shown in Figure 2, reported as mean $\pm$ standard deviation over five seeds.

Table F.1: Drift RMSE in the observation noise and sampling interval sweeps. In the upper block, $r = \tau/\tau_0$ at $\Delta = \Delta_0$; in the lower block, $r = \Delta/\Delta_0$ at $\tau = \tau_0$.

System	r	VEST	SING	SDE Matching	EM–SMC–RKHS	Penalized LS	Naive
Observation noise							
Double well	0.25	0.145 ± 0.050	0.176 ± 0.021	0.386 ± 0.024	$\mathbf{0.136} \pm 0.013$	0.168 ± 0.030	0.432 ± 0.054
	0.5	$\mathbf{0.143} \pm 0.044$	0.178 ± 0.021	0.386 ± 0.021	0.174 ± 0.012	0.175 ± 0.035	1.407 ± 0.190
	1	$\mathbf{0.140} \pm 0.034$	0.177 ± 0.021	0.376 ± 0.010	0.189 ± 0.027	0.202 ± 0.043	3.419 ± 0.220
	2	$\mathbf{0.143} \pm 0.034$	0.182 ± 0.019	0.368 ± 0.009	0.282 ± 0.035	0.541 ± 0.100	7.256 ± 0.534
	4	$\mathbf{0.174} \pm 0.070$	0.189 ± 0.030	0.368 ± 0.009	0.327 ± 0.045	1.054 ± 0.156	19.749 ± 1.021
Linear spiral	0.25	$\mathbf{0.201} \pm 0.031$	0.366 ± 0.031	0.258 ± 0.024	1.078 ± 0.220	—	0.479 ± 0.101
	0.5	$\mathbf{0.209} \pm 0.035$	0.378 ± 0.021	0.276 ± 0.048	1.424 ± 0.504	—	1.938 ± 0.197
	1	$\mathbf{0.226} \pm 0.034$	0.425 ± 0.040	0.291 ± 0.072	1.256 ± 0.345	—	7.513 ± 0.906
	2	$\mathbf{0.250} \pm 0.036$	0.479 ± 0.090	0.349 ± 0.109	2.035 ± 0.557	—	18.887 ± 0.819
	4	$\mathbf{0.294} \pm 0.059$	0.573 ± 0.155	0.564 ± 0.134	3.396 ± 1.391	—	35.365 ± 0.801
Bistable ring	0.25	$\mathbf{1.134} \pm 0.039$	2.132 ± 0.219	1.186 ± 0.053	10.564 ± 1.709	—	3.262 ± 0.697
	0.5	$\mathbf{1.139} \pm 0.049$	2.054 ± 0.224	1.176 ± 0.061	10.784 ± 1.582	—	14.699 ± 1.610
	1	$\mathbf{1.167} \pm 0.074$	1.951 ± 0.230	1.173 ± 0.104	9.525 ± 1.381	—	30.752 ± 2.462
	2	$\mathbf{1.256} \pm 0.094$	1.796 ± 0.164	1.260 ± 0.165	8.294 ± 1.287	—	46.406 ± 1.522
	4	1.399 ± 0.066	1.597 ± 0.128	$\mathbf{1.346} \pm 0.144$	7.914 ± 0.756	—	77.358 ± 3.074
Sampling interval							
Double well	0.25	$\mathbf{0.143} \pm 0.050$	0.179 ± 0.025	0.381 ± 0.017	0.244 ± 0.055	0.514 ± 0.097	15.105 ± 1.403
	0.5	$\mathbf{0.141} \pm 0.042$	0.178 ± 0.022	0.377 ± 0.018	0.231 ± 0.040	0.380 ± 0.038	7.749 ± 0.662
	1	$\mathbf{0.140} \pm 0.034$	0.177 ± 0.021	0.376 ± 0.010	0.189 ± 0.027	0.202 ± 0.043	3.419 ± 0.220
	2	$\mathbf{0.138} \pm 0.035$	0.178 ± 0.019	0.372 ± 0.012	0.188 ± 0.043	0.188 ± 0.041	1.882 ± 0.206
	4	$\mathbf{0.137} \pm 0.032$	0.176 ± 0.016	0.368 ± 0.008	0.187 ± 0.018	0.176 ± 0.054	0.961 ± 0.188
Linear spiral	0.25	$\mathbf{0.215} \pm 0.036$	0.347 ± 0.034	0.330 ± 0.062	1.923 ± 0.418	—	30.478 ± 2.303
	0.5	$\mathbf{0.207} \pm 0.039$	0.378 ± 0.029	0.297 ± 0.059	1.451 ± 0.374	—	15.844 ± 2.362
	1	$\mathbf{0.226} \pm 0.034$	0.425 ± 0.040	0.291 ± 0.072	1.256 ± 0.345	—	7.513 ± 0.906
	2	$\mathbf{0.237} \pm 0.029$	0.424 ± 0.064	0.306 ± 0.055	1.168 ± 0.247	—	3.476 ± 0.282
	4	$\mathbf{0.256} \pm 0.051$	0.444 ± 0.080	0.306 ± 0.041	1.175 ± 0.282	—	1.606 ± 0.244
Bistable ring	0.25	$\mathbf{1.149} \pm 0.054$	2.153 ± 0.200	1.189 ± 0.035	14.338 ± 3.335	—	93.904 ± 3.171
	0.5	$\mathbf{1.150} \pm 0.075$	2.069 ± 0.200	1.185 ± 0.066	11.482 ± 1.732	—	53.759 ± 4.159
	1	$\mathbf{1.167} \pm 0.074$	1.951 ± 0.230	1.173 ± 0.104	9.525 ± 1.381	—	30.752 ± 2.462
	2	$\mathbf{1.170} \pm 0.059$	1.718 ± 0.194	1.194 ± 0.082	7.570 ± 0.986	—	16.869 ± 0.783
	4	$\mathbf{1.213} \pm 0.051$	1.560 ± 0.155	1.239 ± 0.069	5.758 ± 0.928	—	8.423 ± 0.531

F.4 Latent-state recovery

The quantitative comparison uses the constant-diffusion double-well and linear-spiral systems, with the training settings described in Appendix F.2. For each of five paired seeds, we simulate 100 independent test trajectories using Euler–Maruyama with step size 0.001 and record observations at $\Delta = 0.02$. To assess the learned dynamics, we then simulate 512 paths from each of 12 fixed initial states. We compare their terminal distributions with those from independently simulated true trajectories, using W_1 for the double well and sliced W_1 with 128 fixed projections for the linear spiral. Drift RMSE is evaluated on a grid spanning the 2nd to 98th percentiles of the training states along each coordinate. Table F.3 reports the five-seed means and standard deviations for the three

metrics summarized in the left panel of Figure 3, while Table F.2 extends the distribution comparison to three earlier prediction horizons.

Table F.2: Time-marginal distribution distances for VEST and SING (Hu et al., 2025). Distances are W_1 for the double well ($T = 2.0$) and sliced W_1 for the linear spiral ($T = 1.2$).

System	t/T	SING	VEST
Double well	0.10	0.0427 ± 0.0117	$\mathbf{0.0330 \pm 0.0108}$
	0.25	0.0733 ± 0.0241	$\mathbf{0.0593 \pm 0.0200}$
	0.50	0.1056 ± 0.0336	$\mathbf{0.0888 \pm 0.0307}$
	1.00	0.1443 ± 0.0544	$\mathbf{0.1160 \pm 0.0439}$
Linear spiral	0.10	0.0214 ± 0.0023	$\mathbf{0.0136 \pm 0.0010}$
	0.25	0.0404 ± 0.0036	$\mathbf{0.0270 \pm 0.0022}$
	0.50	0.0555 ± 0.0058	$\mathbf{0.0425 \pm 0.0041}$
	1.00	0.0750 ± 0.0055	$\mathbf{0.0593 \pm 0.0134}$

Table F.3: Comparison of latent-state recovery and learned dynamics between VEST and SING (Hu et al., 2025).

System	Method	Drift RMSE	State RMSE	Terminal distance
Double well	SING	0.2264 ± 0.0643	0.0906 ± 0.0015	0.1443 ± 0.0544
	VEST	$\mathbf{0.1981 \pm 0.0759}$	$\mathbf{0.0904 \pm 0.0015}$	$\mathbf{0.1160 \pm 0.0439}$
Linear spiral	SING	0.2913 ± 0.0291	0.0674 ± 0.0004	0.0750 ± 0.0055
	VEST	$\mathbf{0.1522 \pm 0.0103}$	$\mathbf{0.0672 \pm 0.0004}$	$\mathbf{0.0593 \pm 0.0134}$

F.5 Real data experiments

Data and noise calibration. We use `multitrack_Eb10.h5` from the public data of Bryan et al. (2022).[3] The file contains 100 release trajectories, each with 501 position measurements at the native interval $\delta = 10\,\mu$s. We split the complete trajectories into 60 for training, 20 for validation, and 20 for testing. All scaling and noise calibration use only the training records. For each training trajectory, we form position increments $Z_{n,m} = Y_{n,m+1} - Y_{n,m}$ at the native interval $\delta = 10\,\mu$s. We subtract the mean of all training increments, then compute their mean squared value v and the mean product c of consecutive increments within each trajectory. We estimate $q = 2D\delta$ and R by nonnegative least squares using

$$v \approx q + 2R, \qquad c \approx -R. \tag{30}$$

Here $v = 6.21668 \times 10^{-18}\,\mathrm{m}^2$ and $c = 2.76044 \times 10^{-19}\,\mathrm{m}^2$. Because c is positive, the constrained fit sets $R = 0$ and $q = v$. We retain $D = q/(2\delta) = 3.10834 \times 10^{-13}\,\mathrm{m}^2\mathrm{s}^{-1}$ and replace the zero noise variance with the numerical floor $R = 10^{-4}v = 6.21668 \times 10^{-22}\,\mathrm{m}^2$. We use these same calibration values for all methods.

Model settings. VEST and direct NN use a one-hidden-layer $\tanh$ network with a squared-parameter penalty of 10^{-5}. Mean validation CRPS for future noisy-position forecasts selects widths 16 and 32, respectively, from $\{16, 32\}$. Both networks use initialization seed 0. VEST runs at most 12 outer iterations, with up to three IEKS passes and 12 neural updates per iteration. Direct NN takes 1600 steps to regress increments between measured positions. We adapt the authors' public SKIPPER implementation by fixing D and R to the training estimates while retaining its potential and latent-trajectory updates, including Hamiltonian Monte Carlo (HMC) sampling. We use 100 inducing points and four chains (seeds 0–3), each run for 2000 iterations. Discarding the first 1000 iterations per chain leaves 4000 pooled draws.

Potential reconstruction. For VEST and direct NN, we reconstruct the potential from the fitted drift using $\mathrm{d}(\widehat{U}/k_BT)/\mathrm{d}x = -\widehat{b}/D$, up to an additive constant. This relation assumes a conservative

[3] Data: `https://doi.org/10.5281/zenodo.6680673`.

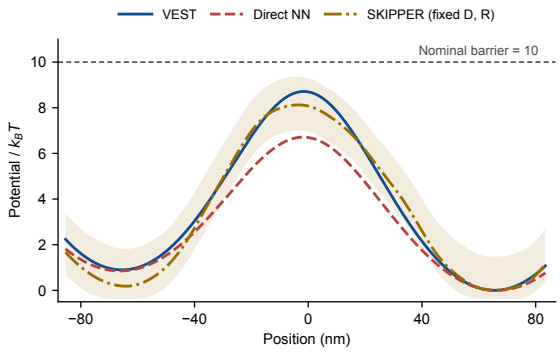


Figure F.1: Potential estimates on the Eb10 release data. Curves are shifted to zero at their minima; the SKIPPER curve shows its posterior mean potential.

one-dimensional overdamped model and the Einstein relation. We evaluate the potentials on a 401-point grid spanning the 2nd–98th percentiles of training positions. If a potential has two local minima x_L, x_R separated by a maximum x_b, we report the larger well-to-barrier height:

$$\widehat{H} = \frac{\widehat{U}(x_b) - \min\{\widehat{U}(x_L), \widehat{U}(x_R)\}}{k_B T}.$$

SKIPPER estimates the potential directly. We apply the same barrier definition to its posterior mean potential. Figure F.1 shows the estimated potentials. The designed potential has a barrier height of $10\,k_BT$ (Bryan et al., 2022), which we use as a reference for comparison.

Two-dimensional harmonic-trap protocol. For Figure 4, we use the experimental `harmwell2D_n1.h5`, `harmwell2D_n4.h5`, and `harmwell2D_n7.h5` records of Bryan et al. (2022) and their paired lower-noise detector recordings. Each pair contains 49,999 simultaneous two-dimensional positions sampled every $10\,\mu$s. We use the first 35,000 positions for calibration and fitting. We center positions at the training mean of the noisier detector, scale them by $20\,\mathrm{nm}$, and retain every fifth position. This gives a $50\,\mu$s sampling interval and a normalized time step of 0.02. From the subsampled training data, we take 22 evenly spaced windows of 41 positions.

Because the two detectors observe the same particle, their symmetrized cross-covariance estimates the stationary position covariance C_0 while removing independent detector noise in expectation. We estimate a linear transition matrix A from their lag-five cross-covariance and compute $Q = C_0 - AC_0A^\top$. The mean Q across conditions sets a common isotropic diffusion for VEST. For each condition, we use the covariance of the detector difference as the working observation covariance. This difference contains noise from both detectors; its pooled standard deviation is 5.1, 7.1, and $8.7\,\mathrm{nm}$ for n1, n4, and n7, respectively. VEST uses a two-input, two-output $\tanh$ network with one hidden layer of width 16, initialization seed 42, and a parameter penalty of 10^{-5}. We run eight outer iterations, each with at most ten IEKS passes and six neural updates.

The red Gaussian reference in Figure 4 is $U_{\mathrm{ref}}(z)/k_BT = z^\top C_0^{-1} z/2$, based on the paired-detector covariance. We evaluate VEST's fitted drift on a 37×37 grid spanning $[-60, 60]\,\mathrm{nm}$ and $[-85, 85]\,\mathrm{nm}$ and construct the blue potential by least-squares projection. The green Boltzmann estimate is minus the log of a two-dimensional kernel density fitted to the noisy training positions. The last three panels show $x_2 = 0$ slices.

F.6 Additional experiment: effect of the trajectory number N

Using the settings in Appendix F.2, we vary the number of training trajectories N: $\{10, 20, 40, 80\}$ for one-dimensional systems, $\{6, 12, 25, 50\}$ for the two-dimensional system, and $\{4, 8, 15, 30\}$ for five- and ten-dimensional systems. At each N, methods use the same observations, and evaluation points remain fixed within each system and seed. Figure F.2 shows six settings. Drift RMSE generally decreases as N increases, and VEST performs well across most systems.

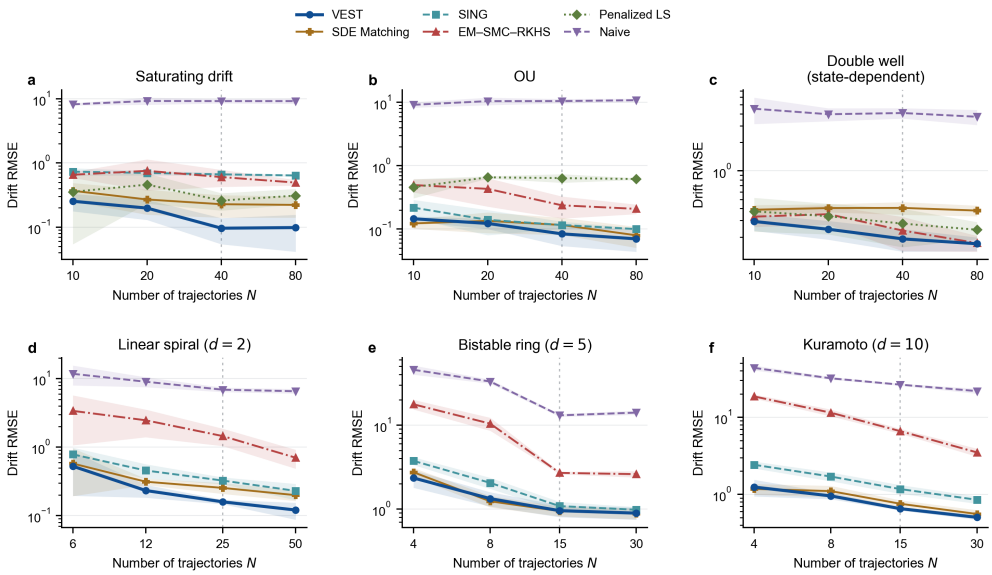


Figure F.2: Drift RMSE versus training trajectory count in six settings. Lines show the mean over five seeds and shaded bands indicate mean ± standard deviation.

F.7 Additional experiment: VEST-NM and VEST-LR

We evaluate VEST-NM and VEST-LR, introduced in Appendix C, on the ten settings in Table 1. Both variants use the same noisy trajectories for seeds 0–4 and the same drift-evaluation points. VEST-NM uses an RBF or Matérn-3/2 time kernel, with a positive scale at each time point and a learned lengthscale for each path. Its E-step optimizes the full Gaussian variational objective for at most ten iterations using a bound-constrained quasi-Newton method. The neural M-step uses VEST's architecture and training budget. VEST-LR uses 32 fixed RBF centers selected from the training observations by farthest-point traversal, with lengthscale set to the median pairwise center distance. Its E-step uses VEST's IEKS and safeguard settings. The M-step computes the closed-form candidate in Appendix C.2 and tests it by line search, using the same regularization coefficient as VEST. Overall, VEST maintains competitive accuracy across the ten settings. VEST-LR is particularly accurate for the double-well and Van der Pol systems.

Table F.4: Drift RMSE for VEST, VEST-NM, and VEST-LR on the same data and evaluation points as Table 1.

System	VEST	VEST-NM RBF	VEST-NM Matérn-3/2	VEST-LR RBF
Constant diffusion				
OU	0.084 ± 0.030	0.109 ± 0.046	0.101 ± 0.045	0.117 ± 0.022
Saturating	0.096 ± 0.042	0.198 ± 0.066	0.140 ± 0.024	0.166 ± 0.034
Double well	0.198 ± 0.076	0.234 ± 0.075	0.216 ± 0.073	0.156 ± 0.075
Linear spiral	0.159 ± 0.012	0.167 ± 0.015	0.163 ± 0.014	0.387 ± 0.091
Van der Pol	0.684 ± 0.157	0.674 ± 0.146	0.682 ± 0.150	0.501 ± 0.094
Bistable ring	0.956 ± 0.155	0.963 ± 0.145	0.957 ± 0.153	1.257 ± 0.063
Kuramoto	0.654 ± 0.033	0.670 ± 0.045	0.652 ± 0.041	2.016 ± 0.154
State-dependent diffusion				
OU	0.077 ± 0.014	0.110 ± 0.048	0.094 ± 0.033	0.127 ± 0.029
Double well	0.191 ± 0.051	0.203 ± 0.055	0.201 ± 0.053	0.144 ± 0.063
Van der Pol	0.791 ± 0.200	0.772 ± 0.171	0.787 ± 0.185	0.465 ± 0.105

F.8 Additional experiment: ablation results

We compare full VEST with three variants in a paired five-seed experiment. Within each seed and setting, all four versions use the same observations and drift-evaluation points. We score the saved models using the evaluation-point construction of Table 1 for the corresponding system and seed. In *Without* C_m, we omit the estimated covariance between consecutive latent states when computing the drift regression target. In *Posterior mean*, we calculate the drift loss only at the smoothed state mean, rather than averaging it over possible state values. In *Without safeguards*, we accept smoothing and drift updates without checking whether they improve the training objective; numerically invalid runs and completed runs with catastrophically divergent drift estimates are reported as unusable.

Table F.5: Ablation results on the evaluation points of Table 1. Drift RMSE is the mean $\pm$ population standard deviation over five seeds.

System	VEST	Without C_m	Posterior mean	Without safeguards
Constant diffusion				
OU	$\mathbf{0.084 \pm 0.030}$	0.332 ± 0.119	0.209 ± 0.050	$5/5$
Saturating	$\mathbf{0.096 \pm 0.042}$	0.181 ± 0.024	0.661 ± 0.057	$5/5$
Double well	$\mathbf{0.198 \pm 0.076}$	0.222 ± 0.075	0.226 ± 0.081	$5/5$
Linear spiral	$\mathbf{0.159 \pm 0.012}$	0.454 ± 0.056	0.226 ± 0.051	$5/5$
Van der Pol	$\mathbf{0.684 \pm 0.157}$	0.768 ± 0.158	0.720 ± 0.183	$5/5$
Coupled bistable ring	$\mathbf{0.956 \pm 0.155}$	0.983 ± 0.137	1.020 ± 0.172	$5/5$
Kuramoto	$\mathbf{0.654 \pm 0.033}$	1.159 ± 0.071	0.869 ± 0.056	$5/5$
State-dependent diffusion				
OU	$\mathbf{0.077 \pm 0.014}$	0.584 ± 0.193	0.178 ± 0.056	$5/5$
Double well	$\mathbf{0.191 \pm 0.051}$	0.213 ± 0.053	0.274 ± 0.056	$5/5$
Van der Pol	$\mathbf{0.791 \pm 0.200}$	0.974 ± 0.082	0.853 ± 0.198	$5/5$

Implementation details. Within each system and seed, all variants share the network initialization and training budget. Drift RMSE uses the evaluation points of Table 1 and Euclidean vector errors without normalization by dimension. In *Without* C_m, the target is $(\mu_{m+1}-x)/\Delta$. In *Posterior mean*, it is $(\mu_{m+1} - \mu_m)/\Delta$ and the loss is evaluated at $x = \mu_m$. Both changes apply to the Gaussian and drift updates. In *Without safeguards*, we remove the objective-based acceptance checks from both updates. Of the 50 runs, 43 terminate with numerical failures; the other seven finish but produce extremely large drift RMSEs. Thus, all five runs in each setting are reported as unusable in Table F.5.