GEOMETRIC CONSTRAINTS IMPROVE INFERENCE OF SPARSELY OBSERVED STOCHASTIC DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

The dynamics of systems of many degrees of freedom evolving on multiple scales are often modeled in terms of stochastic differential equations. Usually the structural form of these equations is unknown and the only manifestation of the system's dynamics are observations at discrete points in time. Despite their widespread use, accurately inferring these systems from sparse-in-time observations remains challenging. Conventional inference methods either focus on the temporal structure of observations, neglecting the geometry of the system's invariant density, or use geometric approximations of the invariant density, which are limited to conservative driving forces. To address these limitations, here, we introduce a novel approach that reconciles these two perspectives. We propose a path augmentation scheme that employs data-driven control to account for the geometry of the invariant system's density. Non-parametric inference on the augmented paths, enables efficient identification of the underlying deterministic forces of systems observed at low sampling rates.

1 INTRODUCTION

Unraveling a system's governing equations from time-series observations is often crucial for understanding unexplained natural phenomena. The goal is to find a mathematical representation that aligns with observational data and provides a comprehensive phenomenological understanding of the underlying mechanisms. This requires employing proper representations that capture key system properties while effectively simplify extraneous degrees of freedom. Stochastic differential equations (SDE) provide such a flexible representation (Arnold, 2014; Lande, 1976; Chandrasekhar, 1943; Nelson, 2004), by representing the dominant system forces in the deterministic part of the equation, (*drift function*) $f(\cdot) : R^d \to R^d$, and summarising the unresolved or irrelevant degrees as stochastic forces acting on the system (*diffusion*). The resulting evolution equation has the form

$$\mathrm{d}X_t = f(X_t)\mathrm{d}t + \sigma\,\mathrm{d}W_t, \qquad X_0 = x_0, \tag{1}$$

where $\sigma \in \mathcal{R}^{d\times d}$ stands for the noise amplitude, and W_t for the d-dimensional vector of independent Wiener processes. The equation should be interpreted according to the Ito formalism. We observe the system state at distinct points in time through $\mathcal{O}_k = \psi(X_{k\tau})$, where $X_{k\tau} \doteq X_t|_{t=\tau k}$, with $k = 1, 2, \ldots, K$ observations collected at **inter-observation intervals** τ, and want approximate the drift function $f(\cdot)$ from the observations $\{\mathcal{O}_k\}_{k=0}^{K}$. Here, we will consider $\psi(x) = x$, but the method generalises for monotonic functions $\psi(\cdot)$.

2 IDENTIFYING STOCHASTIC SYSTEMS FROM SPARSE-IN-TIME OBSERVATIONS

High-frequency observations

For small inter-observation intervals τ, we consider that we observe the continuous path of the system state $X_{0:T}$. In that case, we can identify the drift function $f(\cdot)$ by the first order Kramers-Moyal coefficient ((Kramers, 1940; Moyal, 1949; Tabar, 2019)) by empirically estimating conditional expectations of state increments ((Friedrich et al., 2000; Ragwitz & Kantz, 2001; Boninsegna et al., 2018; Siegert et al., 1998)). For the Bayesian non-parametric counterpart of this approach, (Ruttor et al., 2013; Batz et al., 2018) consider that transition probabilities between observations are

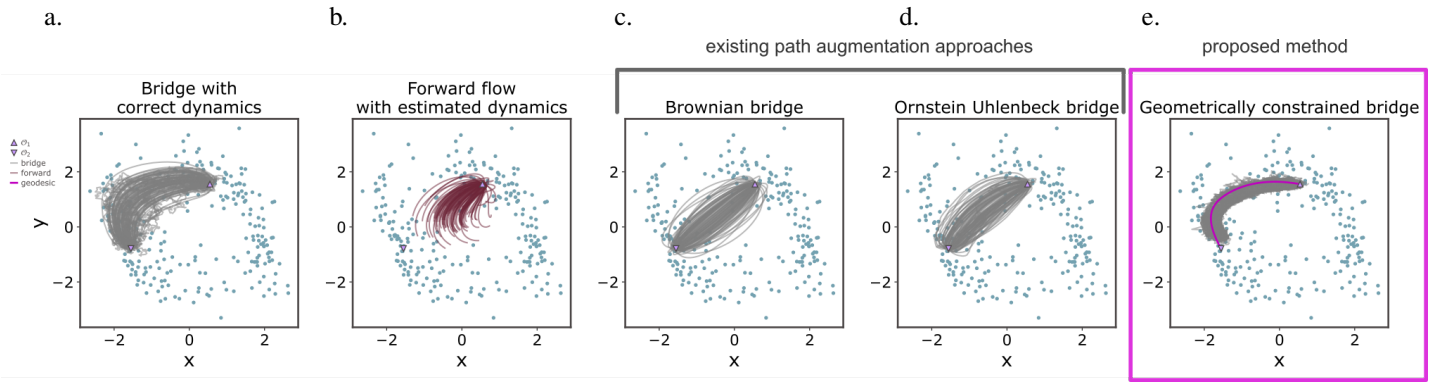

Figure 1: **Existing path augmentation strategies match poorly the underlying transition density between consecutive observations underestimating its curvature. a.)** Stochastic bridge marginal density (*grey*) between two successive observations $\mathcal{O}_1$ and $\mathcal{O}_2$ (*pink triangles*) following the ground truth dynamics. **b.)** Forward probability flow with estimated dynamics with a Gaussian likelihood (*maroon*) matches poorly the correct transition density and often fails to reach the second observation $\mathcal{O}_2$ (*downward pink triangle*). Common path augmentation strategies employ either: **c.)** Brownian bridges, or **d.)** Ornstein Uhlenbeck (linear) bridge marginals resulting from local linearisations of the estimated drift with Gaussian likelihood. Both approaches match poorly the correct transition density, because they underestimate its curvature. **e.)** The proposed geometrically constrained path augmentation provides a better approximation of the underlying transition density by forcing the bridge paths towards the geodesic curve that connects consecutive observations on the manifold induced by the observations.

Gaussian for $\mathrm{d}t \to 0$, resulting in a (Gaussian) likelihood for the drift (see Sec. A Eq. 7)

$$\mathcal{L}(X_{0:T} \mid f) = \exp\left[-\frac{1}{2}\int_0^T \|f(X_t)\|^2_{\sigma^2}\mathrm{d}t + \int_0^T \langle f(X_t), X_{t+\mathrm{d}t} - X_t\rangle \mathrm{d}t\right]. \tag{2}$$

To identify the drift Ruttor et al. (2013) impose a Gaussian process prior on the function values f (Eq. 12). In Eq. 2 we introduced the weighted inner product $\langle u, v\rangle \doteq u^\top \cdot \sigma^{-2} v$ and weighted norm $\|u\|_{\sigma^2} \doteq u^\top \cdot \sigma^{-2} u$.

Low-frequency observations

However, as the inter-observation interval τ increases, the transition probabilities between consecutive observations cannot be considered Gaussian, and thus the likelihood (Eq. 2) assumed between two successive observations is no longer valid if Eq. 1 is non-linear. The likelihood for the drift $\mathrm{P}(\{\mathcal{O}_k\}_{k=1}^K | f)$ for such settings takes the form of a *path integral*

$$\mathrm{P}(\{\mathcal{O}_k\}_{k=1}^K \mid f) = \int \mathrm{P}(\{\mathcal{O}_k\}_{k=1}^K, X_{0:T} \mid f)\mathcal{D}(X_{0:T}) = \int \mathrm{P}(\{\mathcal{O}_k\}_{k=1}^K \mid X_{0:T})\mathrm{P}(X_{0:T}|f)\mathcal{D}(X_{0:T}), \tag{3}$$

where $\{\mathcal{O}_k\}_{k=1}^K$ denotes the set of K discrete time observations, $\mathrm{P}(X_{0:T}|f)$ the prior path probability assuming the dynamics of Eq. 1, $\mathcal{D}(X_{0:T})$ identifies the formal volume element on the path space, while $\mathrm{P}(\{\mathcal{O}_k\}_{k=1}^K | X_{0:T})$ stands for the likelihood of observations given the latent path $X_{0:T}$.

From a geometric perspective, we consider that the nonlinear system dynamics induce an invariant density that may be approximated by a (possibly low dimensional) manifold. The sparse-in-time observations are samples of that manifold. For low-frequency observations, Euclidean distances employed for computing the state increments $X_{t+\tau} - X_t$ do not consider the geometry induced by the nonlinear dynamics, and thereby underestimate the curvature of the transition density between consecutive observations (Figure 4).

3 Geometric Path Augmentation

Since the likelihood of Eq. 3 is intractable, we consider the unobserved continuous path between observations as latent random variables $X_{0:T}$, and obtain a maximum a posteriori estimate for the drift through Expectation Maximisation (EM) (Dempster et al., 1977). Similar parametric (Elerian et al., 2001; Sermaidis et al., 2013) and non-parametric (Batz et al., 2018; Ruttor et al., 2013) methods have addressed the drift inference in the past, primarily in high-frequency observation settings.

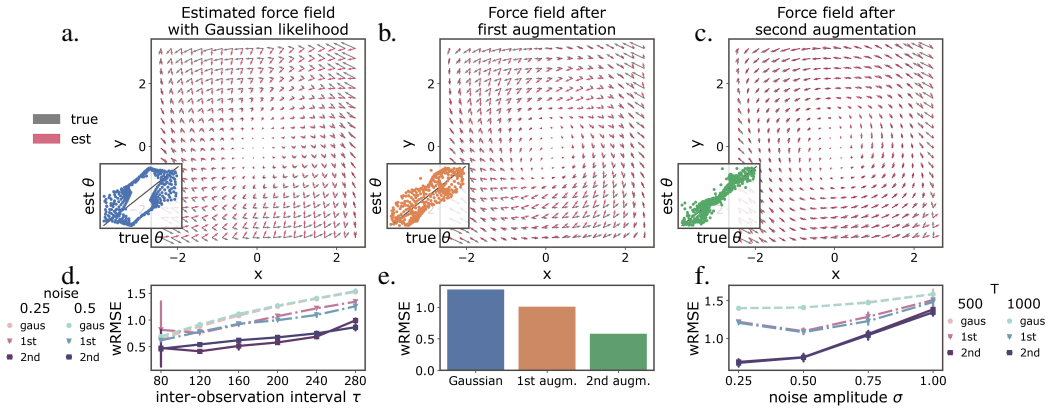

Figure 2: **Proposed path augmentation after two iterations already provides a good approximation of underlying drift.** Estimated (*red*) and true (*grey*) force field with **a.)** Gaussian likelihood **b.)** after one and **c.)** after second iteration of augmentations. (**insets**) Ground truth against estimated angles for each point on the two dimensional grid. **e.)** Weighted root mean square error (wRMSE) for estimated drifts after each iteration for the presented example. The weights for averaging the error at each grid point are obtained from a kernel density estimation on the observations $\{\mathcal{O}_k\}_{k=1}^K$. **d.)** wRMSE against inter-observation interval τ for different noise conditions $\sigma = \{0.25, 0.5\}$ for drift estimated with a Gaussian likelihood (*gaus*-circles), after first augmentation (*1st*-triangles), and after second augmentation (*2nd*-squares) for $T = 500$. **f.)** wRMSE against noise amplitude σ in the system for different trajectory durations $T = \{500, 1000\}$ time units for inter-observation interval $\tau = 240$. Markers follow the same coding as in d.). Errorbars indicate one standard deviation over 5 independent realisations.

Our approach is based on the non-parametric method discussed in (Batz et al., 2018; Ruttor et al., 2013), with two significant advancements:

- **(i)** We employ a path augmentation scheme following the **estimated nonlinear dynamics** resulting from inference with the Gaussian likelihood of Eq. 2 (as opposed to local linear approximations of these dynamics proposed in (Batz et al., 2018)).
- **(ii)** Importantly, we further **constrain the augmented paths to align with the geometry of the invariant density** between consecutive observations (Fig. 4 b.).

We follow an iterative algorithm, where at each iteration n we perform the two following steps:
(1.) An **E(xpectation) step**, where given a drift estimate $\hat{f}^n$ we construct an approximate posterior over the latent variables $Q(X_{0:T}) \approx \mathrm{P}(X_{0:T}|\{\mathcal{O}\}_{k=1}^K, \hat{f}^n(x))$.
(2.) A **M(aximisation) step**, where we update the drift estimation.

• **Approximate posterior over paths. (E-step)** We approximate the continuous path trajectory $X_{0:T}$ between observations by a posterior path measure defined as the minimiser of the free energy

$$\mathcal{F}[Q] = \frac{1}{2}\int_0^T \int \Big[\|g(x,t) - \hat{f}(x)\|^2_{\sigma^2} + U_{\mathcal{O}}(x,t) + U_{\mathcal{G}}(x,t)\Big]\, q_t(x)\, \mathrm{d}x\, \mathrm{d}t. \tag{4}$$

The term $U_{\mathcal{O}}(x,t) \doteq -\sum_{t_k} \ln \mathrm{P}(\mathcal{O}_k|x)\delta(t-t_k)$ forces the latent path to pass through the observations (or close to them depending on the observation process), while $U_{\mathcal{G}}(x,t) \doteq \|\Gamma_t - x\|^2$ guides the latent path towards the geodesic curves $\gamma^k_{t'}$ that connect consecutive observations on the manifold $\mathcal{M}$ induced by the system's invariant density (Sec. A.1.2). Here we denote $\Gamma_t \doteq \{\gamma^k_{t'}\}_{t=(k-1)\tau+t'\tau}$, where $\gamma^k_{t'}$ is the geodesic connecting $\mathcal{O}_k$ and $\mathcal{O}_{k+1}$, and $t' \in [0,1]$. We identify the geodesic $\gamma^k_{t'}$ for each interval by learning the local metric of the manifold $\mathcal{M}$ (see Sec. A.1.2 and Arvanitidis et al. (2019)).

Following Opper (2019), for each inter-observation interval $[\mathcal{O}_k, \mathcal{O}_{k+1}]$ we identify the posterior path measure (minimiser of Eq. equation 4) by the solution of a stochastic optimal control problem Maoutsa & Opper (2022; 2021); Maoutsa (2022) with the objective to obtain a time-dependent drift adjustment $u(x,t) := g(x,t) - \hat{f}(x)$ for the system with drift $\hat{f}(x)$ with **initial and terminal constraints** determined by $U_{\mathcal{O}}(x,t)$, and additional **path constraints** $U_{\mathcal{G}}(x,t)$.

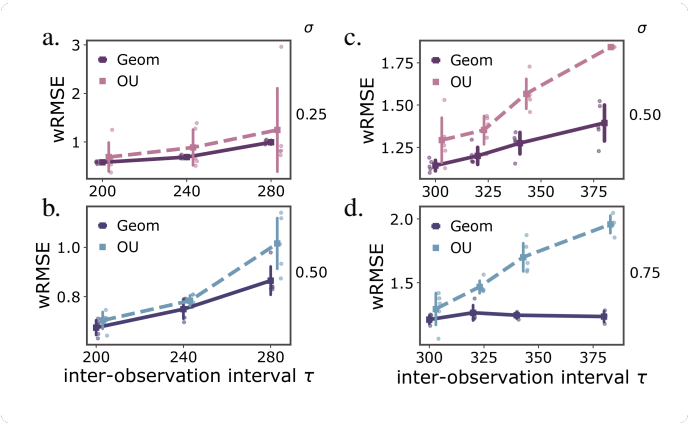

Figure 3: **Comparison of proposed path augmentation with Ornstein-Uhlenbeck augmentation.** Weighted root mean square error (wRMSE) against noise amplitude σ for different inter-observation intervals for noise amplitude for moderate inter-observation intervals with **a.)** $\sigma = 0.25$, and **b.)** $\sigma = 0.50$, and for large inter-observation intervals with **c.)** $\sigma = 0.50$, and **d.)** $\sigma = 0.75$ and known direction of movement. In **c., d.** the inter-observation interval results in only one observation per oscillation period.

• **Drift estimation. (M-step)** To estimate the drift from a sampled latent path, we assume a Gaussian process prior over function values and employ a sparse kernel approximation similar to Batz et al. (2018) (see Sec. A.2 for details).

4 Numerical Experiments

To demonstrate the performance of the proposed method we performed systematic estimations for a two-dimensional Van der Pol oscillator under different noise conditions σ, observed at different inter-observation intervals τ for different lengths of trajectories T (see Sec. B). For the examined noise amplitudes (Fig. 2 f.), the proposed path augmentation algorithm improves the naive estimation with Gaussian assumptions within two iterations for most noise amplitudes (Fig. 2). For increasing noise the improvement contributed by our approach decreases (Fig. 2f.), but is nevertheless not negligible. For low noise conditions, geodesics approximate the manifold structure better, however the path integral control is limited by the control costs proportional to inverse noise covariance. Our framework had comparable accuracy for all inter-observation lengths, but improvement was small for small lengths since in that setting the estimation with Gaussian likelihood already provides a good approximation of the ground truth drift. The proposed approach was compared to the augmentation method proposed by Batz et al. (2018) (augmentation with Ornstein-Uhlenbeck bridges) and delivered more accurate estimates for larger inter-observation intervals. For inter-observation intervals with only one observation per oscillation period (Figure 3c.,d.), the proposed approach delivered better results by considering additionally that we know the direction of movement in the state space (c.f. Sec. B). The variance of estimates from our method was smaller compared to Batz et al. due to conditioning on the invariant geometry of the system.

5 Conclusion and Discussion

We introduced a new method for identifying stochastic systems from sparse-in-time observations of the system's state that reconciles approaches that rely purely on the temporal structure of the observations with those that approximate the geometry of the invariant density. Our method employs a path augmentation strategy that uses the nonlinear dynamics of a coarse drift estimate and further constrains the augmented paths to follow the local geometry of the system's invariant density. We found that the proposed approach provides efficient recovery of the underlying drift function for periodic or quasi-periodic systems under several noise conditions.

Only a limited number of inference methods have attempted to merge geometric and temporal perspectives for the identification of stochastic systems, such as the Langevin regression (Callaham et al., 2021), TrajectoryNet (Tong et al., 2020), and the diffusion map method of Shnitzer et al. (2020; 2016). However, our method differs from these approaches by employing the geodesic approximation of the underlying data geometry.

REFERENCES

Ludwig Arnold. Stochastic differential equations as dynamical systems. In *Realization and Modelling in System Theory*, pp. 489–495. Springer, 1990. *(cited on page: 12)*

Stevan J Arnold. Phenotypic evolution: The ongoing synthesis: (american society of naturalists address), 2014. *(cited on page: 1)*

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017. doi: https://doi.org/10.48550/arXiv.1710.11379. *(cited on page: 13)*

Georgios Arvanitidis, Soren Hauberg, Philipp Hennig, and Michael Schober. Fast and robust shortest paths on manifolds learned from data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1506–1515. PMLR, 2019. *(cited on pages: 3 and 13)*

Philipp Batz, Andreas Ruttor, and Manfred Opper. Approximate bayes learning of stochastic differential equations. *Physical Review E*, 98(2):022109, 2018. *(cited on pages: 1, 2, 3, 4, and 9)*

Alexandros Beskos, Omiros Papaspiliopoulos, and Gareth O Roberts. Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, 12(6):1077–1098, 2006. *(cited on page: 9)*

Lorenzo Boninsegna, Feliks Nüske, and Cecilia Clementi. Sparse learning of stochastic dynamical equations. *The Journal of chemical physics*, 148(24):241723, 2018. doi: 10.1063/1.5018409. *(cited on page: 1)*

Jared L Callaham, J-C Loiseau, Georgios Rigas, and Steven L Brunton. Nonlinear stochastic modelling with Langevin regression. *Proceedings of the Royal Society A*, 477(2250):20210092, 2021. *(cited on page: 4)*

Andrew Carverhill. Flows of stochastic dynamical systems: ergodic theory. *Stochastics: An International Journal of Probability and Stochastic Processes*, 14(4):273–317, 1985. *(cited on page: 12)*

Subrahmanyan Chandrasekhar. Stochastic problems in physics and astronomy. *Reviews of modern physics*, 15(1):1, 1943. *(cited on page: 1)*

Raphaël Chetrite and Hugo Touchette. Variational and optimal control representations of conditioned and driven processes. *Journal of Statistical Mechanics: Theory and Experiment*, 2015 (12):P12001, 2015. *(cited on page: 10)*

Siddhartha Chib, Michael K Pitt, and Neil Shephard. Likelihood based inference for diffusion driven state space models. *Por Clasificar*, pp. 1–33, 2006. *(cited on page: 9)*

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977. *(cited on page: 2)*

Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*, volume 6. Springer, 1992. *(cited on pages: 12 and 13)*

Ola Elerian, Siddhartha Chib, and Neil Shephard. Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69(4):959–993, 2001. *(cited on page: 2)*

Neil Fenichel and JK Moser. Persistence and smoothness of invariant manifolds for flows. *Indiana University Mathematics Journal*, 21(3):193–226, 1971. *(cited on page: 12)*

Rudolf Friedrich and Joachim Peinke. Description of a turbulent cascade by a Fokker-Planck equation. *Physical Review Letters*, 78(5):863, 1997. doi: https://doi.org/10.1103/PhysRevLett.78.863. *(cited on page: 9)*

Rudolf Friedrich, Silke Siegert, Joachim Peinke, Marcus Siefert, Michael Lindemann, Jan Raethjen, Güntner Deuschl, Gerhard Pfister, et al. Extracting model equations from experimental data. *Physics Letters A*, 271(3):217–222, 2000. *(cited on page: 1)*

Christian Fröhlich, Alexandra Gessner, Philipp Hennig, Bernhard Schölkopf, and Georgios Arvanitidis. Bayesian Quadrature on Riemannian Data Manifolds. In *International Conference on Machine Learning*, pp. 3459–3468. PMLR, 2021. doi: https://doi.org/10.48550/arXiv.2102.06645. *(cited on page: 13)*

TV Girya and Igor Dmitrievich Chueshov. Inertial manifolds and stationary measures for stochastically perturbed dissipative dynamical systems. *Sbornik: Mathematics*, 186(1):29–45, 1995. *(cited on page: 12)*

Andrew Golightly and Darren J Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52(3):1674–1693, 2008. *(cited on page: 9)*

Christoph Honisch and Rudolf Friedrich. Estimation of kramers-moyal coefficients at low sampling rates. *Physical Review E*, 83(6):066701, 2011. *(cited on page: 9)*

Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940. *(cited on page: 1)*

Steven J Lade. Finite sampling interval effects in kramers–moyal analysis. *Physics Letters A*, 373 (41):3705–3709, 2009. *(cited on page: 9)*

Russell Lande. Natural selection and random genetic drift in phenotypic evolution. *Evolution*, pp. 314–334, 1976. *(cited on page: 1)*

John M Lee. *Introduction to Riemannian manifolds*, volume 176. Springer, 2018. *(cited on page: 12)*

Robert S Liptser and Albert N Shiryaev. *Statistics of random processes II: Applications*, volume 6. Springer Science & Business Media, 2013. doi: 10.1007/978-3-662-10028-8. *(cited on page: 8)*

Satya N Majumdar and Henri Orland. Effective langevin equations for constrained stochastic processes. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(6):P06039, 2015. *(cited on page: 10)*

Dimitra Maoutsa. Revealing latent stochastic dynamics from single-trial spike train observations. Bernstein Conference for Computational Neuroscience, Berlin 2022, 09 2022. *(cited on page: 3)*

Dimitra Maoutsa and Manfred Opper. Deterministic particle flows for constraining SDEs. *Machine Learning and the Physical Sciences, Workshop at the 35th Conference on Neural Information Processing Systems (NeurIPS), arXiv preprint arXiv:2110.13020*, 2021. doi: https://doi.org/10.48550/arXiv.2110.13020. *(cited on pages: 3 and 11)*

Dimitra Maoutsa and Manfred Opper. Deterministic particle flows for constraining stochastic nonlinear systems. *Phys. Rev. Research*, 4:043035, Oct 2022. doi: 10.1103/PhysRevResearch.4.043035. URL https://link.aps.org/doi/10.1103/PhysRevResearch.4.043035. *(cited on pages: 3, 11, 14, and 15)*

Dimitra Maoutsa, Sebastian Reich, and Manfred Opper. Interacting particle solutions of Fokker–Planck equations through gradient–log–density estimation. *Entropy*, 22(8):802, 2020. *(cited on page: 15)*

Salah-Eldin A Mohammed and Michael KR Scheutzow. The stable manifold theorem for stochastic differential equations. *Annals of probability*, pp. 615–652, 1999. doi: https://doi.org/10.1214/aop/1022677380. *(cited on page: 12)*

Jose Moyal. Stochastic processes and statistical physics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):150–210, 1949. *(cited on page: 1)*

Philip Nelson. *Biological physics*. WH Freeman New York, 2004. *(cited on page: 1)*

Manfred Opper. Variational inference for stochastic differential equations. *Annalen der Physik*, 531 (3):1800233, 2019. *(cited on pages: 3, 10, 11, and 14)*

Omiros Papaspiliopoulos, Yvo Pokern, Gareth O Roberts, and Andrew M Stuart. Nonparametric estimation of diffusions: a differential equations approach. *Biometrika*, 99(3):511–531, 2012. *(cited on page: 9)*

Mario Ragwitz and Holger Kantz. Indispensable finite time corrections for Fokker-Planck equations from time series data. *Physical Review Letters*, 87(25):254501, 2001. doi: 10.1103/physrevlett.87.254501. *(cited on page: 1)*

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pp. 63–71. Springer-Verlag, 2003. *(cited on page: 9)*

Andreas Ruttor, Philipp Batz, and Manfred Opper. Approximate gaussian process inference for the drift function in stochastic differential equations. *Advances in Neural Information Processing Systems*, 26, 2013. *(cited on pages: 1, 2, 3, 8, 9, and 14)*

Giorgos Sermaidis, Omiros Papaspiliopoulos, Gareth O Roberts, Alexandros Beskos, and Paul Fearnhead. Markov chain monte carlo for exact inference for diffusions. *Scandinavian Journal of Statistics*, 40(2):294–321, 2013. *(cited on pages: 2 and 9)*

Tal Shnitzer, Ronen Talmon, and Jean-Jacques Slotine. Manifold learning with contracting observers for data-driven time-series analysis. *IEEE Transactions on Signal Processing*, 65(4):904–918, 2016. doi: https://doi.org/10.1109/TSP.2016.2616334. *(cited on page: 4)*

Tal Shnitzer, Ronen Talmon, and Jean-Jacques Slotine. Manifold learning for data-driven dynamical system analysis. In *The Koopman Operator in Systems and Control*, pp. 359–382. Springer, 2020. *(cited on page: 4)*

Silke Siegert, R Friedrich, and J Peinke. Analysis of data sets of stochastic systems. *Physics Letters A*, 243(5-6):275–280, 1998. doi: 10.1016/S0375-9601(98)00283-7. *(cited on page: 1)*

M Tabar. Kramers–Moyal expansion and Fokker–Planck equation. In *Analysis and Data-Based Reconstruction of Complex Nonlinear Dynamical Systems*, pp. 19–29. Springer, 2019. *(cited on page: 1)*

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pp. 9526–9536. PMLR, 2020. *(cited on page: 4)*

Stephen Wiggins. *Normally hyperbolic invariant manifolds in dynamical systems*, volume 105. Springer Science & Business Media, 1994. *(cited on page: 12)*

Appendices

A Drift inference for high and low frequency observations **8**

A.1 Approximate posterior over paths. 10

A.1.1 Approximate posterior over paths without geometric constraints. 10

A.1.2 Approximate posterior over paths with geometric constraints. 12

A.2 Approximate posterior over drift functions. 14

B Details on numerical experiments **15**

C Additional figures **15**

A Drift inference for high and low frequency observations

We consider systems whose evolution is captured by the stochastic differential equation Eq. 1.

High frequency observations. When the system path $X_{0:T}$ is observed in continuous time, the infinitesimal transition probabilities of the diffusion process between consecutive observations are Gaussian, i.e.,

$$\mathrm{P}_f(X_{0:T} \mid f) \propto \exp\left(-\frac{1}{2\mathrm{d}t}\sum_t \|X_{t+\mathrm{d}t} - X_t - f(X_t)\mathrm{d}t\|^2_{\sigma^2}\right). \tag{5}$$

In turn, the transition probability of (discretised) Wiener paths $P_{\mathcal{W}}(X_{0:T})$ (i.e., paths from a driftless process) can be expressed as

$$\mathrm{P}_{\mathcal{W}}(X_{0:T}) = \exp\left(-\frac{1}{2\mathrm{d}t}\sum_t \|X_{t+\mathrm{d}t} - X_t\|^2_{\sigma^2}\right), \tag{6}$$

where $\|u\|_{\sigma^2} \doteq u^\top \cdot \sigma^{-2} u$ denotes the weighted norm with $D \doteq \sigma^2$ indicating the noise covariance. We can thus express the likelihood for the drift f by the Radon-Nykodym derivative between $P_f(X_{0:T}|f)$ and $\mathrm{P}_{\mathcal{W}}(X_{0:T})$ for paths $X_{0:T}$ within the time interval $[0, T]$ (Liptser & Shiryaev, 2013)

$$\mathcal{L}(X_{0:T} \mid f) = \exp\left[-\frac{1}{2}\sum_t \|f(X_t)\|^2_{\sigma^2}\mathrm{d}t + \sum_t \langle f(X_t), X_{t+\mathrm{d}t} - X_t\rangle_{\sigma^2}\right], \tag{7}$$

where for brevity we have introduced the notation $\langle u, v\rangle \doteq u^\top \cdot \sigma^{-2} v$ for the weighted inner product with respect to the inverse noise covariance σ^{-2}. This expression results from applying the Girsanov theorem on the path measures induced by a process with drift f and a Wiener process, with same diffusion σ, and employing an Euler-Maruyama discretisation on the continuous path $X_{0:T}$.

The likelihood given a continuously observed path of the SDE (Eq. 7) has a quadratic form in terms of the drift function. Therefore a Gaussian measure over function values (Gaussian process) is a natural conjugate prior for this likelihood. To identify the drift in a non-parametric form, we assume a Gaussian process prior for the function values $f \sim \mathrm{P}_0(f) = \mathrm{GP}(m^f, k^f)$, where m^f and k^f denote the mean and covariance function of the Gaussian process (Ruttor et al., 2013). The prior measure can be written as

$$\mathrm{P}_0(f) = \exp\left[-\frac{1}{2}\int\int f(x)\left(k^f(x,x')\right)^{-1} f(x')\mathrm{d}x\mathrm{d}x'\right], \tag{8}$$

if we consider a zero mean Gaussian process $m^f = 0$.

Bayesian inference for the drift function f requires the computation of a probability distribution in the function space, the posterior probability distribution $\mathrm{P}_f(f \mid X_{0:T})$. From the Bayes' rule the posterior can be expressed as

$$\mathrm{P}_f(f \mid X_{0:T}) = \frac{\mathrm{P}_0(f)\mathcal{L}(X_{0:T} \mid f)}{Z} \propto \mathrm{P}_0(f)\mathcal{L}(X_{0:T} \mid f), \tag{9}$$

where Z denotes a normalising factor defined as a path integral

$$Z = \int \mathrm{P}_0(f)\mathcal{L}(X_{0:T} \mid f)\mathcal{D}f, \tag{10}$$

where $\mathcal{D}f$ denotes integration over the Hilbert space $f : H_0[f] < \infty$. Here we have expressed the prior probability over functions as $\mathrm{P}_0(f) = e^{-H_0[f]}$. In (Ruttor et al., 2013) the authors show that in the continuous time limit, nonparametric estimation of drift functions becomes equivalent to Gaussian process regression, with the objective to identify the mapping from the system state X_t to state increments $\mathrm{d}X_t$ (Rasmussen, 2003). More precisely, we consider N observations of the system state X_t as the regressor, with associated response variables

$$Y_t = \frac{X_{t+\mathrm{d}t} - X_t}{\mathrm{d}t}, \tag{11}$$

and denote the kernel function of the Gaussian process by $k(x, x')$.

If we denote with $\mathcal{X} = \{X_t\}_{t=0}^{T-\mathrm{d}t}$ and $\mathcal{Y} = \{Y_t\}_{t=0}^{T-\mathrm{d}t}$ the set of state observations and observation increments, the mean of the posterior process over drift functions f can be expressed as

$$\bar{f}(x) = k^f(x, \mathcal{X})^\top \left(\mathcal{K} + \frac{\sigma^2}{\mathrm{d}t} I_N\right)^{-1} \mathcal{Y}, \tag{12}$$

where we abused the notation and denoted with $k^f(x, \mathcal{X})$ the vector resulting from evaluating the kernel k^f at points x and $\{\mathcal{O}_t\}_{k=1}^{K-1}$. Similarly $\mathcal{K} = k^f(\mathcal{X}, \mathcal{X})$ stands for the $(K-1) \times (K-1)$ matrix resulting from evaluation of the kernel on all observation pairs. In a similar vein, the posterior variance can be written as

$$\Sigma^2(x) = k^f(x, x) - k^f(x, \mathcal{X})^\top \left(\mathcal{K} + \frac{\sigma^2}{\mathrm{d}t}\right)^{-1} k^f(x, \mathcal{X}), \tag{13}$$

where the term $\sigma^2/\mathrm{d}t$ plays the role of observation noise.

Low frequency observations. When the inter-observation interval becomes large (*low frequency observations*), the Gaussian likelihood of Eq. 7 becomes invalid, since for large inter-observation intervals the transition density is no longer Gaussian. Thus, drift estimation with Gaussian assumptions (Friedrich & Peinke, 1997; Ruttor et al., 2013) becomes inaccurate. To mitigate this issue Lade (Lade, 2009) introduced a method to compute finite time corrections for the drift estimates, which has been applied (to the best of our knowledge) mostly to one dimensional problems (Honisch & Friedrich, 2011). On the other hand, the statistics community has proposed path augmentation schemes that augment the observed trajectory to a nearly continuous-time trajectory by sampling a simplified system's dynamics between observations (Golightly & Wilkinson, 2008; Papaspiliopoulos et al., 2012; Sermaidis et al., 2013; Beskos et al., 2006; Chib et al., 2006). However for large inter-observation intervals and for nonlinear systems the simplified dynamics employed for path augmentation match poorly the underlying path statistics, and these methods show poor convergence rates or fail to identify the correct dynamics (Figure 1 c. and d.). We point out here, that path augmentation with Ornstein Uhlenbeck bridges using as drift the local linearisation of the **correct** dynamics, provides a good approximation of the underlying transition density. However, during inference, the true underlying dynamics are unknown, and the proposed local linearisations on inaccurate drift estimates (Batz et al., 2018) perform poorly for low frequency observations.

Notice that as the inter-observation interval τ increases, the Gaussian likelihood assumed between two successive observations is no longer valid if the system is non-linear or when the noise is state dependent. The likelihood for the drift for such settings can be expressed in terms of a *path integral*

$$\mathrm{P}(\mathcal{O}_{1:K} \mid f) = \int \mathrm{P}(\mathcal{O}_{1:K} \mid X_{0:T})\mathrm{P}(X_{0:T} \mid f)\mathcal{D}(X_{0:T}), \tag{14}$$

where $\mathcal{O}_{1:K} \doteq \{\mathcal{O}_k\}_{k=1}^K$ denotes the set of K discrete time observations, $\mathrm{P}(X_{0:T} \mid f)$ the prior path probability resulting from a diffusion process with drift $f(x)$, $\mathcal{D}(X_{0:T})$ identifies the formal volume element on the path space, and $\mathrm{P}(\mathcal{O}_{1:K} \mid X_{0:T})$ stands for the likelihood of observations given the latent path $X_{0:T}$.

However, the path integral of Eq. 14 is intractable for nonlinear systems, thus we need to simultaneously estimate the drift and latent state of the diffusion process, i.e., to approximate the joint posterior measure of latent paths and drift functions $\mathrm{P}(X_{0:T}, f \mid \mathcal{O}_{1:K})$. Therefore we consider the unobserved continuous path $X_{0:T}$ as latent random variables and employ an Expectation Maximisation (EM) algorithm to identify a maximum a posteriori estimate for the drift function. More precisely, we follow an iterative algorithm, where at each iteration n we alternate between the two following steps:

An **Expectation** step, where given a drift estimate $\hat{f}^n(x)$ we construct an approximate posterior over the latent variables $Q(X_{0:T}) \approx \mathrm{P}(X_{0:T} \mid \mathcal{O}_{1:K}, \hat{f}^n(x))$, and compute the expected log-likelihood of the augmented path

$$\mathfrak{L}\big(\hat{f}^n(x), Q\big) = \mathbb{E}_Q\Big[\ln \mathscr{L}\big(X_{0:T} \mid \hat{f}^n(x)\big)\Big]. \tag{15}$$

A **Maximisation** step, where we update the drift estimation by maximising the expected log likelihood

$$f^{n+1}(x) = \arg\max_f \Big[\mathfrak{L}\big(f^n(x), Q\big) - \ln \mathrm{P}_0\big(f^n(x)\big)\Big]. \tag{16}$$

In Eq. 16 P_0 denotes the Gaussian process prior over function values.

A.1 Approximate posterior over paths.

Here we first formulate the approximate posterior over paths (conditional distribution for the path given the observations) by considering only individual observations as constraints (Section A.1.1). However, this approach results computationally taxing calculations during path augmentation, since the observations are atypical states of the initially estimated drift. To overcome this issue, we subsequently extend the formalism (Section A.1.2) to incorporate constraints that consider also the local geometry of the observations.

A.1.1 Approximate posterior over paths <u>without</u> geometric constraints.

Given a drift function (or a drift estimate) $\hat{f}(x)$ we can apply variational techniques to approximate the posterior measure over the latent path conditioned on the observations $\{\mathcal{O}_k\}_{k=1}^K$. We consider that the prior process (the process without considering the observations $\{\mathcal{O}_k\}_{k=1}^K$) is described by the equation

$$\mathrm{P}(X_{0:T} \mid \hat{f}): \qquad \mathrm{d}X_t = \hat{f}(X_t)\mathrm{d}t + \sigma \mathrm{d}W_t. \tag{17}$$

We will define an approximating (posterior) process that is conditioned on the observations. The conditioned process is also a diffusion process with the same diffusion as Eq. 17 but with a modified, time-dependent drift $g(x,t)$ that accounts for the observations (Chetrite & Touchette, 2015; Majumdar & Orland, 2015). We identify the approximate posterior measure Q with the posterior measure induced by an approximating process that is conditioned by the observations $\mathcal{O}_{1:K}$ (Opper, 2019), with governing equation

$$Q(X_{0:T}): \qquad \mathrm{d}X_t = g(X_t, t)\mathrm{d}t + \sigma \mathrm{d}W_t = \Big(\hat{f}(X_t) + \sigma^2 u(X_t, t)\Big)\,\mathrm{d}t + \sigma \mathrm{d}W_t. \tag{18}$$

The effective drift $g(X_t, t)$ of Eq. 18 may be obtained from the solution of the variational problem of minimising the free energy

$$\mathcal{F}[Q] = \mathcal{KL}\Big(Q(X_{0:T}) || \mathrm{P}(X_{0:T} \mid \hat{f})\Big) - \sum_{k=1}^{K} E_Q[\ln \mathrm{P}(\mathcal{O}_k \mid X_{t_k})]. \tag{19}$$

By applying the Cameron-Girsanov-Martin theorem we can express the Kullback-Leibler divergence between the two path measures induced by the diffusions with drift $\hat{f}(x)$ and $g(x,t)$ as

$$\mathcal{KL}\Big(Q(X_{0:T})||\mathrm{P}(X_{0:T}|\hat{f})\Big) = E_Q\left[\ln\left(\frac{dQ(X_{0:T})}{dP\left(X_{0:T}|\hat{f}\right)}\right)\right] \tag{20}$$

$$= E_Q\left[\left(-\frac{1}{2}\int_0^T \|\hat{f}(X_t) - g(X_t,t)\|^2_{\sigma^2}\mathrm{d}t + \int_0^T \frac{\hat{f}(X_t) - g(X_t,t)}{\sigma^2}\mathrm{d}W_t\right)\right] \tag{21}$$

$$= E_Q\left[\left(-\frac{1}{2}\int_0^T \|\hat{f}(X_t) - g(X_t,t)\|^2_{\sigma^2}\mathrm{d}t + V_T\right)\right] \tag{22}$$

$$= \frac{1}{2}\int_0^T\int \|g(x,t) - \hat{f}(x)\|^2_{\sigma^2}\, q_t(x)\,\mathrm{d}x\,\mathrm{d}t + \mathfrak{C}, \tag{23}$$

where $q_t(x)$ stands for the marginal density for X_t of the approximate process. In the third line we have introduced the random variable $V_T = \int_0^T \frac{\hat{f}(X_t)-g(X_t,t)}{\sigma^2}\mathrm{d}W_t$. Under the assumption that the function $\ell(X_t) = \hat{f}(X_t) - g(X_t,t)$ is bounded, piece-wise continuous, and in $L^2[0,\infty)$, V_T follows the distribution $\mathcal{N}\left(V_T \mid 0, \int_0^T \ell^2(s)\mathrm{d}s\right)$, which for a given T will result into a constant $\mathfrak{C}$. Thus the second term in Eq. 23 is not relevant for the minimisation of the free energy and will be omitted.

We can thus express the free energy of Eq. 19 as (Opper, 2019)

$$\mathcal{F}[Q] = \frac{1}{2}\int_0^T\int \Big[\|g(x,t) - \hat{f}(x)\|^2_{\sigma^2} + U(x,t)\Big]\, q_t(x)\,\mathrm{d}x\,\mathrm{d}t, \tag{24}$$

where the term $U(x,t)$ accounts for the observations $U(x,t) = -\sum_{t_k} \ln \mathrm{P}(\mathcal{O}_k \mid x)\delta(t-t_k)$.

The minimisation of the functional of the free energy can be construed as a stochastic control problem (Opper, 2019) with the objective to identify a time-dependent drift adjustment $u(x,t) := g(x,t) - \hat{f}(x)$ for the system with drift $\hat{f}(x)$ so that the controlled dynamics fulfil the constraints imposed by the observations.

For the case of exact observations, i.e., for an observation process $\psi(x) = x$, we can compute the drift adjustment for each of the $K-1$ inter-observation intervals independently. Thus for each interval between consecutive observations, we identify the optimal control $u(x,t)$ required to construct a stochastic bridge following the dynamics of Eq. 17 with initial and terminal states the respective observations $\mathcal{O}_k$ and $\mathcal{O}_{k+1}$.

The optimal drift adjustment for such a stochastic control problem for the inter-observation interval between $\mathcal{O}_k$ and $\mathcal{O}_{k+1}$ can be obtained from the solution of the backward equation (see (Maoutsa & Opper, 2022; 2021))

$$\frac{\partial \phi_t(x)}{\partial t} = -\mathcal{L}^{\dagger}_{\hat{f}}\phi_t(x) + U(x,t)\phi_t(x), \tag{25}$$

with terminal condition $\phi_T(x) = \chi(x) = \delta(x - \mathcal{O}_{k+1})$ and with $\mathcal{L}^{\dagger}_{\hat{f}}$ denoting the adjoint Fokker-Planck operator for the process of Eq. 17. As shown in Maoutsa et al. (Maoutsa & Opper, 2022; 2021) the optimal drift adjustment $u(x,t)$ can be expressed in terms of the difference of the logarithmic gradients of two probability flows

$$u^*(x,t) = D\Big(\nabla \ln q_{T-t}(x) - \nabla \ln \rho_t(x)\Big), \tag{26}$$

where ρ_t fulfils the forward (filtering) partial differential equation (PDE)

$$\frac{\partial \rho_t(x)}{\partial t} = \mathcal{L}_{\hat{f}}\rho_t(x) - U(x,t)\rho_t(x), \tag{27}$$

while q_t is the solution of a time-reversed PDE that depends on the logarithmic gradient of $\rho_t(x)$

$$\frac{\partial q_t(x)}{\partial t} = -\nabla \cdot \left[\Big(\sigma^2 \nabla \ln \rho_{T-t}(x) - f(x, T-t) \Big) q_t(x) \right] + \frac{\sigma^2}{2} \nabla^2 q_t(x), \tag{28}$$

with initial condition $q_0(x) \propto \rho_T(x)\chi(x)$.

A.1.2 Approximate posterior over paths <u>with</u> geometric constraints.

The previously described construction of the approximate measure in terms of stochastic bridges is relevant when the observations have non vanishing probability under the law of the prior diffusion process of Eq. 17. However, when the prior process (with the estimated drift $\hat{f}$) differs considerably from the process that generated the observations, such a construction might either provide a bad approximation of the underlying path measure, or show slow numerical convergence in the construction of the diffusion bridges. To overcome this issue, we consider here additional constraints for the posterior process that force the paths of the posterior measure to respect the local geometry of the observations. In the following we provide a brief introduction on the basics of Riemannian geometry and consequently continue with the geometric considerations of the proposed method.

Riemannian geometry. A d-dimensional **Riemannian manifold** (Do Carmo & Flaherty Francis, 1992; Lee, 2018) $(\mathcal{M}, h)$ embedded in a D-dimensional ambient space $\mathcal{X} = \mathcal{R}^D$ is a smooth curved d-dimensional surface endowed with a smoothly varying inner product (Riemannian) **metric** $h : x \to \langle \cdot | \cdot \rangle_x$ on $\mathcal{T}_x\mathcal{M}$. A tangent space $\mathcal{T}_x\mathcal{M}$ is defined at each point $x \in \mathcal{M}$. The Riemannian metric h defines a canonical volume measure on the manifold $\mathcal{M}$. Intuitively this characterises how to compute inner products locally between points on the tangent space of the manifold $\mathcal{M}$, and therefore determines also how to compute norms and thus distances between points on $\mathcal{M}$.

A **coordinate chart** (G, ϕ) provides the mapping from an open set G on $\mathcal{M}$ to an open set V in the Euclidean space. The dimensionality of the manifold is d if for each point $x \in \mathcal{M}$ there exists a local neighborhood $G \subset \mathcal{R}^d$. We can represent the metric h on the local chart (G, ϕ) by the positive definite matrix (**metric tensor**) $H(x) = (h_{i,j})_{x, 0 \leq i,j, \leq d} = \left(\langle \frac{\partial}{\partial x_i} | \frac{\partial}{\partial x_j} \rangle_x \right)_{0 \leq i,j, \leq d}$ at each point $x \in G$.

For $v, w \in \mathcal{T}_x\mathcal{M}$ and $x \in G$, their inner product can be expressed in terms of the matrix representation of the metric h on the tangent space $\mathcal{T}_x\mathcal{M}$ as $\langle v | w \rangle_x = v^\top H(x) w$, where $H(x) \in \mathcal{R}^{d \times d}$.

The **length of a curve** $\gamma : [0, 1] \to \mathcal{M}$ on the manifold is defined as the integral of the norm of the tangent vector

$$\ell(\gamma_{t'}) = \int_0^1 \|\dot{\gamma}_{t'}\|_h \mathrm{d}t' = \int_0^1 \sqrt{\dot{\gamma}_{t'}^\top H(\gamma_{t'}) \dot{\gamma}_{t'}} \mathrm{d}t', \tag{29}$$

where the dotted letter indicates the velocity of the curve $\dot{\gamma}_{t'} = \partial_{t'} \gamma_{t'}$. A **geodesic curve** is a locally length minimising smooth curve that connects two given points on the manifold.

Riemannian geometry of the observations. For approximating the posterior over paths we take into account the geometry of the invariant density as it is represented by the observations. To that end, we consider systems whose dynamics induce invariant (inertial) manifolds that contain the global attractor of the system and on which system trajectories concentrate (Wiggins, 1994; Mohammed & Scheutzow, 1999; Girya & Chueshov, 1995; Fenichel & Moser, 1971; Arnold, 1990; Carverhill, 1985). We assume thus that the continuous-time trajectories $X_{0:T} \in \mathcal{R}^d$ of the underlying system concentrates on an invariant manifold $\mathcal{M} \in \mathcal{R}^{m \leq d}$ of dimensionality m (possibly) smaller than d. The discrete-time observations $\mathcal{O}_k$ are thus samples of the manifold $\mathcal{M}$. The central premise of our approach is that **unobserved paths between successive observations will be lying either *on* or *in the vicinity* of the manifold** $\mathcal{M}$. In particular, we postulate that unobserved paths should lie **in the vicinity of geodesics that connect consecutive observations** on $\mathcal{M}$. To that end we propose a path augmentation framework that constraints the augmented paths to lie in the vicinity of identified geodesics between consecutive observations.

However, while this view of a lower dimensional manifold embedded in a higher dimensional ambient space helps to build our intuition for the proposed method, for computational purposes we

adopt a complementary view inspired by the discussion in (Fröhlich et al., 2021). According to this view, we consider the entire observation space $\mathcal{R}^d$ as a smooth Riemannian manifold, $\mathcal{M} \doteq \mathcal{R}^d$, characterised by a Riemannian metric h. The effect of the nonlinear geometry of the observations is then captured by the metric h. Thus to approximate the geometric structure of the system's invariant density, we learn the Riemannian metric tensor $H : \mathcal{R}^d \to \mathcal{R}^{d \times d}$ and compute the geodesics between consecutive observations according to the learned metric. Intuitively according to this view the observations $\{\mathcal{O}_k\}_{k=1}^K$ introduce distortions in the way we compute distances on the state space.

In effect this approach does not reduce the dimensionality of the space we operate, but changes the way we compute inner products and thus distances, lengths, and geodesic curves on $\mathcal{M}$. The alternative perspective of working on a lower dimensional manifold would strongly depend on the correct assessment of the dimensionality of said manifold. For example, one could use a Variational Autoencoder to approximate the observation manifold and subsequently obtain the Riemannian metric from the embedding of the manifold mediated by the decoder. However, our preliminary results of such an approach revealed that such a method requires considerable fine tuning to adapt to the characteristics of each dynamical system and is sensitive to the estimation of the dimensionality of the approximated manifold.

To learn the Riemannian metric and compute the geodesics we follow the framework proposed by Arvanitidis et al. (2019). In particular, we approximate the local metric induced by the observations at location $\mathbf{x}$ of the state space, in a non-parametric form by the inverse of the weighted local diagonal covariance computed on the observations as (Arvanitidis et al., 2019)

$$H_{dd}(\mathbf{x}) = \left(\sum_{i=1}^{K} w_i(\mathbf{x}) \left(x_i^{(d)} - x^{(d)} \right)^2 + \epsilon \right)^{-1}, \tag{30}$$

with weights $w_i(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}\|_2^2}{2\sigma_{\mathcal{M}}^2}\right)$, and $x^{(d)}$ denoting the d-th dimensional component of the vector $\mathbf{x}$. The parameter $\epsilon > 0$ ensures non-zero diagonals of the weighted covariance matrix, while $\sigma_{\mathcal{M}}$ characterises the curvature of the manifold.

Between consecutive observations for each interval $[\mathcal{O}_k, \mathcal{O}_{k+1}]$, we identify the geodesic $\gamma_{t'}^k$ as the energy minimising curve, i.e., as the minimiser of the kinetic energy functional $\mathcal{E}(\gamma_{t'}^k) = \int_0^1 L_{\mathcal{M}}(\gamma_{t'}^k, \dot{\gamma}_{t'}^k) \mathrm{d}t'$

$$\gamma_{t'}^{k*} = \underset{\gamma_{t'}^k, \gamma_0^k = \mathcal{O}_k, \gamma_1^k = \mathcal{O}_{k+1}}{\arg\min} \int_0^1 L_{\mathcal{M}}(\gamma_{t'}^k, \dot{\gamma}_{t'}^k) \mathrm{d}t',$$

$$\text{with} \quad \int_0^1 L_{\mathcal{M}}(\gamma_{t'}^k, \dot{\gamma}_{t'}^k) \mathrm{d}t' = \frac{1}{2} \int_0^1 \|\dot{\gamma}_{t'}^k\|_h^2, \tag{31}$$

where $L_{\mathcal{M}}(\gamma_{t'}^k, \dot{\gamma}_{t'}^k)$ denotes the Lagrangian. The minimising curve of this functional is the same as the minimiser of the curve length functional $\ell(\gamma_{t'})$ (Eq. 29), i.e., the geodesic (Do Carmo & Flaherty Francis, 1992).

By applying calculus of variations, the minimising curve of the functional $\mathcal{E}(\gamma_{t'}^k)$ can be obtained from the Euler-Lagrange equations, resulting in the following system of second order differential equations (Arvanitidis et al., 2017; Do Carmo & Flaherty Francis, 1992)

$$\ddot{\gamma_t}^k = -\frac{1}{2} H(\gamma_t^k)^{-1} \left(2 \left(I \otimes (\dot{\gamma_t}^k)^\top \right) \frac{\partial \mathrm{vec}[H(\gamma_t^k)]}{\partial \gamma_t^k} \dot{\gamma_t}^k - \frac{\partial \mathrm{vec}[H(\gamma_t^k)]^\top}{\partial \gamma_t^k} \left(\dot{\gamma_t}^k \otimes \dot{\gamma_t}^k \right) \right), \tag{32}$$

with boundary conditions $\gamma_0^k = \mathcal{O}_k$ and $\gamma_1^k = \mathcal{O}_{k+1}$, where $\otimes$ stands for the Kroenecker product, and $\mathrm{vec}[A]$ denotes the vectorisation operation of matrix A through stacking the columns of A into a vector. Arvanitidis et al. (2019) obtain the geodesics by approximating the solution of the boundary value problem of Eq. 32 with a probabilistic differential equation solver.

Extended free energy functional. We denote the collection of individual geodesics by $\Gamma_t \doteq \{\gamma_{t'}^k\}_{t=(k-1)\tau + t'\tau}$, where $\gamma_{t'}^k$ is the geodesic connecting $\mathcal{O}_k$ and $\mathcal{O}_{k+1}$, and $t' \in [0, 1]$ denotes a rescaled time variable. Additional to the constraints imposed in the previously explained setting (Sec A.1.1), here we add an extra term in the free energy $U_{\mathcal{G}}(x, t) \doteq \|\Gamma_t - x\|^2$ that accounts for the

local geometry of the invariant density, and guides the latent path towards the geodesic curves $\gamma^k_{t'}$ that connect consecutive observations

$$\mathcal{F}[Q] = \frac{1}{2}\int_0^T \int \left[\|g(x,t) - \hat{f}(x)\|^2_{\sigma^2} + U_{\mathcal{O}}(x,t) + \beta U_{\mathcal{G}}(x,t)\right] q_t(x)\, dx\, \mathrm{d}t. \tag{33}$$

Here we denote the observation term by $U_{\mathcal{O}}(x,t) \doteq -\sum_{t_k} \ln \mathrm{P}(\mathcal{O}_k|x)\delta(t-t_k)$, while β stands for a weighting constant that determines the relative weight of the geometric term in the control objective.

Following Opper (2019), for each inter-observation interval $[\mathcal{O}_k, \mathcal{O}_{k+1}]$ we identify the posterior path measure (minimiser of Eq. 33) by the solution of a stochastic optimal control problem (Maoutsa & Opper, 2022) with the objective to obtain a time-dependent drift adjustment $u(x,t) := g(x,t) - \hat{f}(x)$ for the system with drift $\hat{f}(x)$ with initial and terminal constraints defined by $U_{\mathcal{O}}(x,t)$, and additional path constraints $U_{\mathcal{G}}(x,t)$.

A.2 Approximate posterior over drift functions.

For a fixed path measure Q, the optimal measure for the drift Q_f is a Gaussian process given by

$$Q_f \propto \mathrm{P}_f \exp\left(-\frac{1}{2}\int \|f(x)\|^2_{\sigma^2} A(x) - 2\langle f(x), B(x)\rangle_{\sigma^2} \mathrm{d}x\right), \tag{34}$$

with

$$A(x) \doteq \int_0^T p_t(x)\mathrm{d}t,$$

and

$$B(x) \doteq \int_0^T p_t(x) g(x,t)\mathrm{d}t,$$

where $p_t(x)$ denotes the marginal constrained density of the state X_t. The function $g(x,t)$ denotes the effective drift.

We assume a Gaussian process prior for the unknown function f, i.e., $f \sim \mathrm{P}_0(f) = \mathrm{GP}(m^f, k^f)$ where m^f and k^f denote the mean and covariance function of the Gaussian process. Following Ruttor *et al.* Ruttor et al. (2013), we employ a sparse kernel approximation for the drift f by optimising the function values over a sparse set of S inducing points $\{Z_i\}_{i=1}^S$. We obtain the resulting drift from

$$\hat{f}_S(x) = k^f(x, \mathcal{Z})\,(I + \Lambda\,\mathcal{K}_S)^{-1}\,\mathbf{d}, \tag{35}$$

where we have defined introduced the notation $\mathcal{K}_S \doteq k^f(\mathcal{Z}, \mathcal{Z})$

$$\Lambda = \frac{1}{\sigma^2}\mathcal{K}_S^{-1}\left(\int k^f(\mathcal{Z}, x) A(x) k^f(x, \mathcal{Z})\mathrm{d}x\right)\mathcal{K}_S^{-1}. \tag{36}$$

$$\mathbf{d} = \frac{1}{\sigma^2}\mathcal{K}_S^{-1}\left(\int k^f(\mathcal{Z}, x) B(x)\mathrm{d}x\right)\mathcal{K}_S^{-1}, \tag{37}$$

We employ a sample based approximation of the densities in Eq. equation 34 resulting from the particle sampling of the path measure Q. Thus by representing the densities by samples, we can rewrite the density $p_t(x)$ in terms of a sum of Dirac delta functions centered around the particles positions

$$p_t(x) \approx \frac{1}{N}\sum_{j=1}^N \delta(x - X_j(t)),$$

and replace the Riemannian integrals with summation over particles. Here $X_j(t)$ represents the position of the j-th particle at time point t.

B DETAILS ON NUMERICAL EXPERIMENTS

We simulated a two dimensional Van der Pol oscillator with drift function

$$f_1(x, y) = \mu(x - \frac{1}{3}x^3 - y) \tag{38}$$

$$f_2(x, y) = \frac{1}{\mu}x, \tag{39}$$

starting from initial condition $x0 = [1.81, -1.41]$ and under noise amplitudes $\sigma = \{0.25, 0.50, 0.75, 1.00\}$ for total duration of $T = \{500, 1000\}$ time units. The employed inter-observation intervals $\tau \in [80, 320] \times dt$. The last inter-observation interval exceeds the half period of the oscillator and thus samples only a single state per period. This resulted in erroneous estimates. In this setting this indicates the upper limit of τ for which we can provide estimates. However for any inference method, if the observation process samples only one observation per period, identifying the underlying force field without additional assumptions is not possible with temporal methods. The discretisation time-step used for simulation of the ground truth dynamics, and path augmentation $\delta t = 0.01$. For sampling the controlled bridges we employed $N = 100$ particles evolving the associated ordinary differential equation as described in Maoutsa & Opper (2022); Maoutsa et al. (2020). The logarithmic gradient estimator used $M = 40$ inducing points. The sparse Gaussian process for estimating the drift was based on a sparse kernel approximation of $S = 300$ points. In the presented simulation we have employed a weighting parameter $\beta = 0.5$ (Eq. 33). This provides a moderate pull towards the invariant density. The example in Figure 1 was constructed with $\beta = 1$ and provides a better approximation of the transition density, than $\beta = 0.5$.

For identifying geodesic curves in settings where only one observation is collected per oscillation period, we have to bias the method that computes the geodesics to identify the geodesic towards the correct direction of movement, that in this case is not always the shortest curve on the manifold between two consecutive observations. To that end we assigned to each observation a phase-like variable and considered for the computation of the geodesics only the observations that had properly valued phases given the direction of movement and the phases of the two end-point observations. For example, for a counter-clockwise rotation and for phases at the end points of the bridge $\phi_k = \phi(\mathcal{O}_k)$ and $\phi_{k+1} = \phi(\mathcal{O}_{k+1})$, we construct the geodesics only on the observations that have phases within thin $[\phi_k, \phi_{k+1}]$ if $\phi_k < \phi_{k+1}$, or within $[\phi_k, 1] \cup [0, \phi_{k+1}]$ if $\phi_k > \phi_{k+1}$. With this approach we consider essentially only the relevant part of the manifold that aligns with the direction of movement. With the function $\phi(\cdot) : \mathcal{R}^D \rightarrow [0, 1]$ we denote the function that assigns a phase variable to each observation. Here for the Van der Pol oscillator we considered $\phi(\mathbf{x}) = \frac{\text{arctan2}(\frac{\mathbf{x}^{(2)}}{\mathbf{x}^{(1)}}) + \pi}{2\pi}$. An alternative option for assigning phase to each observation is measuring the angle between the Hilbert transform of $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(i)}$ itself, and rescaling the values to the $[0, 1]$ interval. Here $\mathbf{x}^{(i)}$ denotes one dimensional component of the observations.

C ADDITIONAL FIGURES

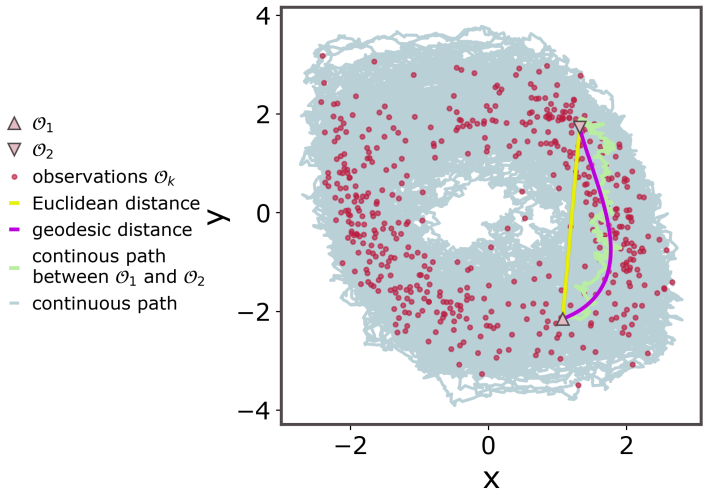

Figure 4: **Considered state increments for low frequency observations under Gaussian likelihood assumptions.** Euclidean distance (*yellow line*) - used to compute the state increments between successive observations - does not account for the curvature of the invariant density. The geodesic curve (*purple line*) provides a better approximation of the unobserved state of the system between successive observations (*light green line*).