Geometric path augmentation for inference of sparsely observed stochastic nonlinear systems

Dimitra Maoutsa*
Technical University of Berlin
Berlin, Germany
dimitra.maoutsa@tu-berlin.de

Abstract

Stochastic evolution equations describing the dynamics of systems under the influence of both deterministic and stochastic forces are prevalent in all fields of science. Yet, identifying these systems from sparse-in-time observations remains still a challenging endeavour. Existing approaches focus either on the temporal structure of the observations by relying on conditional expectations, discarding thereby information ingrained in the geometry of the system's invariant density; or employ geometric approximations of the invariant density, which are nevertheless restricted to systems with conservative forces. Here we propose a method that reconciles these two paradigms. We introduce a new data-driven path augmentation scheme that takes the local observation geometry into account. By employing non-parametric inference on the augmented paths, we can efficiently identify the deterministic driving forces of the underlying system for systems observed at low sampling rates.

1 Introduction

Stochastic differential equations are particularly expressive dynamical models naturally fit for representing systems evolving on multiple time-scales [1–4]. Extracting stochastic evolution equations from such systems has been of major interest in most sciences [5–15]. While identification of continuous time deterministic models has been largely resolved in the past [16–19], the same is not true for their stochastic counterparts. Inference of stochastic systems is particularly challenging in settings where observations are collected at low sampling rates (at large inter-observation intervals).

Most of the existing methods for identifying the deterministic driving forces of stochastic systems rely either on approximations of the **invariant density** (e.g., density estimation [20] or spectral methods like diffusion maps [21–25]) (***geometric methods***), or consider the **temporal structure** of the observations by computing conditional expectations of state increments [26–37] (***temporal methods***). However, geometric methods are limited only to systems with conservative forces by assuming either that the drift is the gradient of a potential [20, 24, 38], or that state variables are completely decoupled [21]. On the other hand, temporal methods perform poorly in settings with large inter-observation intervals (***sparse observations***) [39], since the state increments computed from the observations in those settings (see Appendix Eq. (11)) do not reflect the actual underlying dynamics (Fig. 1).

To mitigate the effect of sparse observations, a subset of the temporal methods employs **path augmentation**[2] to approximate the transition densities between successive observations by sampling **diffusion bridges**, i.e., diffusion processes constrained by their initial and terminal state [31–35]. Yet, the majority of the *non-parametric* approaches employ simplified bridge dynamics (e.g., Brownian [31,

*https://dimitra-maoutsa.gitlab.io/

[2]Here we employ the term **path augmentation** to refer to what is widely known as **data augmentation**. We resorted to this term because we consider it as more elegant and better descriptive of the proposed augmentation process, while the term 'data augmentation' is considerably vague.

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

35] or Ornstein-Uhlenbeck bridges [39]) that do not accurately reflect the underlying transition densities when the observed system is nonlinear (Fig. 2 c. and d.).

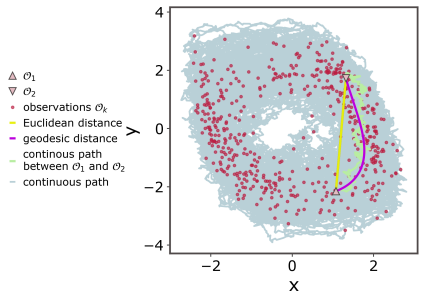

Figure 1: **Considered state increments for low frequency observations under Gaussian likelihood assumptions.** Euclidean distance (*yellow line*) - used to compute the state increments between successive observations - does not account for the curvature of the invariant density. The geodesic curve (*purple line*) provides a better approximation of the unobserved state of the system between successive observations (*light green line*).

An alternative path augmentation strategy would consider a coarse drift estimate (e.g., by assuming a Gaussian likelihood between observations, see Eq. (2)), and would subsequently employ a stochastic bridge sampler [40–42] to construct stochastic bridges with the estimated nonlinear drift. However, for large inter-observation intervals, the observations have zero probability under the law of the estimated diffusion (Fig. 2 b.). Consequently, any attempt to construct diffusion bridges between consecutive observations following the estimated dynamics will – depending on the employed framework – either show slow convergence rates, or fail altogether.

Here we propose an alternative approach. We postulate that the augmented paths should lie **in the vicinity** of the **geodesic curves** (Fig. 1b. magenta) that connect consecutive observations on the manifold induced by the invariant density of the system. To that end we introduce a path augmentation framework that constructs **geometrically constrained bridges** [3] by forcing the augmented paths towards the respective geodesics that connect consecutive observations (Fig. 2 e.). To that end we employ the stochastic control framework introduced in [40, 41] with path constraints that guide the augmented paths towards the geodesic curves that connect successive observations.

2 Setting

We consider stochastic systems described by stochastic differential equations (SDE) of the form

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

where $f(\cdot) : R^d \to R^d$ denotes the deterministic driving forces (*drift function*), and $\sigma\mathrm{d}\boldsymbol{\beta}_t$ represents the random forces acting on the system (*diffusion*). Here $\sigma \in \mathcal{R}^{d\times d}$ denotes the noise amplitude, and $\boldsymbol{\beta}$ the d-dimensional vector of independent Wiener processes. Onwards we consider Ito interpretation of stochastic integrals. We observe the system through an observation process $\mathcal{O}_k = \psi(X_{k\tau})$, where $X_{k\tau} \doteq X_t|_{t=\tau k}$, with $k = 1, 2, \ldots, K$ observations measuring the system state at **inter-observation intervals** τ. For simplicity, we consider identity functions for the observation process, i.e., $\psi(x) = x$, but the formalism easily generalises for more general functions.

High-frequency observations. For sufficiently fine observation timegrids, we assume that observations represent the system state in continuous time, i.e., that we observe the continuous path $X_{0:T}$. Thereby we can estimate the drift by approximating the first order Kramers-Moyal coefficient [43] through empirically estimating conditional expectations of state increments [27–29, 44]. Analogous Bayesian non-parametric methods [45] consider that the transition probabilities between observations are Gaussian for $\mathrm{d}t \to 0$, resulting in a (Gaussian) likelihood for the observations (see Sec. A Eq. 7)

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

and impose a Gaussian process prior on the function values f (Eq. (12)). In Eq. (2) we introduced the notation $\langle u, v\rangle \doteq u^\top \cdot \sigma^{-2} v$ and $\|u\|_{\sigma^2} \doteq u^\top \cdot \sigma^{-2} u$.

Low-frequency observations. As the inter-observation interval τ increases, the Gaussian likelihood (Eq. (2)) assumed between two successive observations is no longer valid if Eq. (1) is non-linear. Similarly, the state increments $X_{t+\tau} - X_t$ computed in this setting do not accurately represent the underlying dynamics (Fig. 1 a.). The likelihood for the drift $\mathrm{P}(\{\mathcal{O}_k\}_{k=1}^K|f)$ for such settings takes the form of a *path integral*

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

[3] Formally these constructs are no longer diffusion bridges but constrained diffusion paths. Here we overextend the notion of diffusion bridges to contrast it against the commonly employed diffusion bridges for path augmentation.

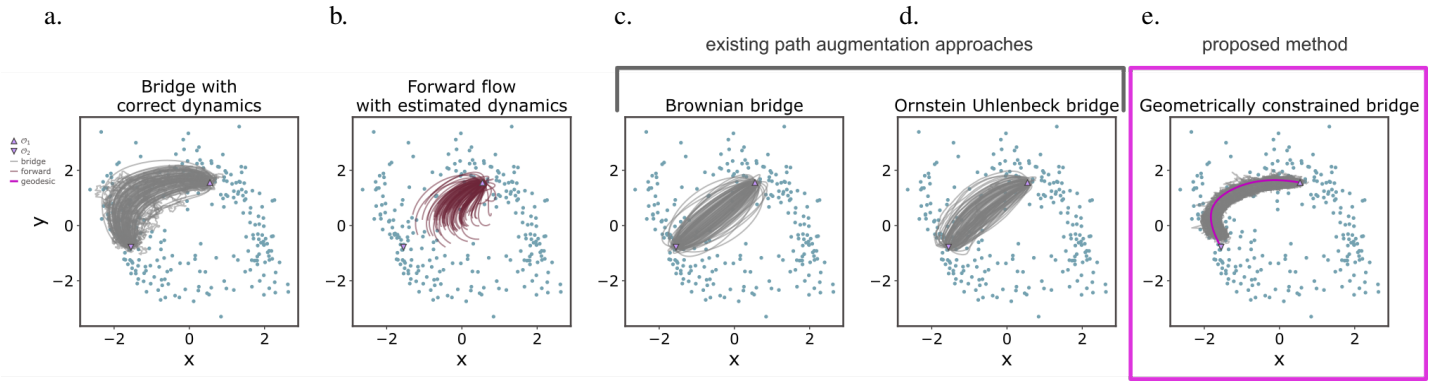

Figure 2: **Existing path augmentation strategies match poorly the underlying transition density between consecutive observations underestimating its curvature. a.)** Stochastic bridge marginal density (*grey*) between two successive observations $\mathcal{O}_1$ and $\mathcal{O}_2$ (*pink triangles*) 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.

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 resulting from the system 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 can consider that the invariant density of the system can be approximated with a (low dimensional) manifold induced by the nonlinear system dynamics. The observations are essentially 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 1).

3 Method

Since the likelihood of Eq. (3) is intractable, we consider the unobserved continuous path as latent random variables $X_{0:T}$, and employ Expectation Maximisation (EM) [46] to identify a maximum a posteriori estimate for the drift function. Similar parametric [33, 47] and non-parametric [39, 45] methods have addressed the drift inference in the past, targeting mainly high-frequency observation settings. Our approach here is inspired by the non-parametric method followed in [39, 45] with two key innovations:

- **(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 [39]).
- **(ii)** Importantly, we further **constrain the augmented paths to match the geometry of the invariant density** between consecutive observations (Fig. 1 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

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

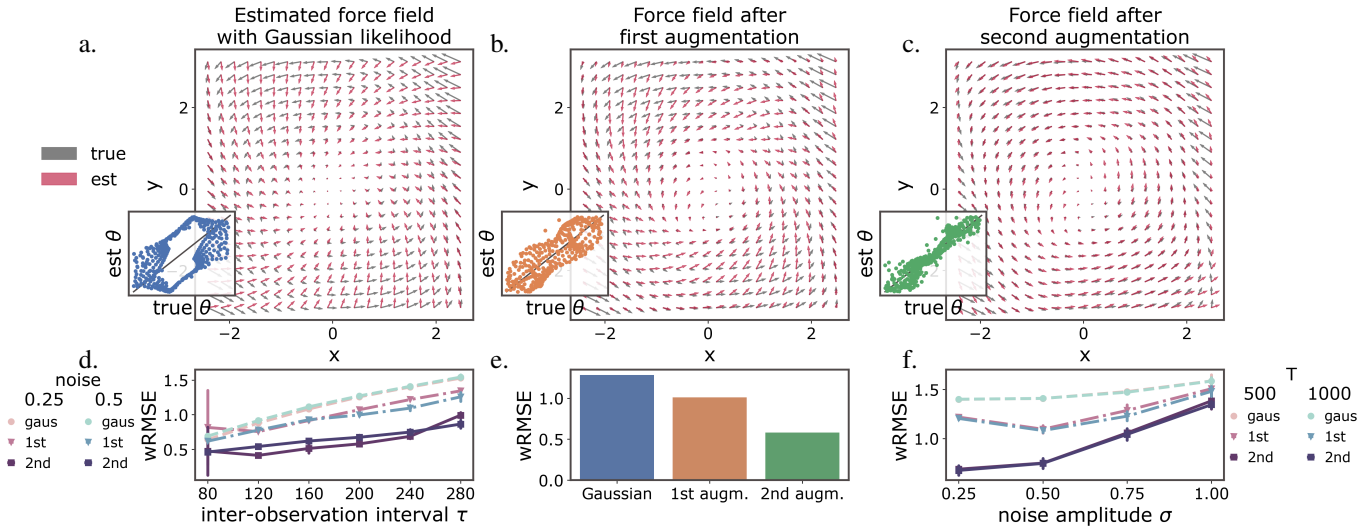

Figure 3: **Proposed path augmentation after two iterations already provides a good approximation of underlying drift.** Estimated (*red*) and true (*grey*) force field with **a.)** Gaussian likelihood **b.)** after one and **c.)** after second iteration of augmentations. (**insets**) Ground truth against 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.

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 $\gamma_{t'}^k$ that connect consecutive observations on the manifold $\mathcal{M}$ induced by the system's invariant density (Sec. A.1.2). Here we denote $\Gamma_t \doteq \{\gamma_{t'}^k\}_{t=(k-1)\tau+t'\tau}$, where $\gamma_{t'}^k$ is the geodesic connecting $\mathcal{O}_k$ and $\mathcal{O}_{k+1}$, and $t' \in [0,1]$. We identify the geodesic $\gamma_{t'}^k$ for each interval by learning the local metric of the manifold $\mathcal{M}$ (see Sec. A.1.2 and [48]).

Following [49], for each inter-observation interval $[\mathcal{O}_k, \mathcal{O}_{k+1}]$ we identify the posterior path measure (minimiser of Eq. (4)) by the solution of a stochastic optimal control problem [14, 40, 41] 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** $U_{\mathcal{G}}(x,t)$.

• **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 [39] (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. D). For the examined noise amplitudes (Fig. 3 f.) and for inter-observation intervals that result in more than one observation per oscillation period (Fig. 3d.), the proposed path augmentation algorithm improves the naive estimation with Gaussian assumptions within two iterations for most noise amplitudes(Fig. 3). For increasing noise the improvement contributed by our approach decreases (Fig. 3f.), but is nevertheless not negligible.

5 Conclusion and Discussion

We introduced a new method for identifying stochastic systems from sparse-in-time observations of the system's state. We proposed a path augmentation strategy that employs 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.

Geometric constraints for inference. Our method reconciles approaches that rely purely on the temporal structure of the observations with those that approximate the invariant density and ignore the temporal order of measurements. With the recent development of the field of geometric statistics [50, 51], and the surge of interest on the concept of manifold hypothesis [52, 53], i.e., the consideration that often the state of multi-dimensional dynamical systems is confined on low dimensional regions of the state space, several inference methods have tried to merge geometric and temporal perspectives for identification of stochastic systems. In the *Langevin regression* framework [54], Callaham *et al.* compute the Kramers-Moyal coefficients, and account for misestimation due to low sampling rate by solving the adjoint Fokker-Planck equation for the coefficients as proposed by Lade [55]. They incorporate geometric constraints by additionally regularising by moment matching between the observation density and the stationary Fokker-Planck probability density of the estimated SDE model. In [56] Tong *et al.* consider the manifold of the observations for inference of cellular dynamics. Their method employs dynamic optimal transport to interpolate between measured distributions constrained to lie in the vicinity of the observations. This approach has the same intuitions with our method, however Tong et al. do not employ stochastic differential equations to model the inherently stochastic cellular dynamics. Moreover they do not attempt any modeling of the underlying geometry of the data, but consider only constraints that penalise distances to individual observations. Shnitzer et al. [53, 57] employ diffusion maps to approximate the eigenfunctions of the backward Kolmogorov operator (the generator of the stochastic Koopman operator [58, 59]), and - since the eigenfunctions follow linear evolution equations - they evolve the dominant operator eigenspectrum with a Kalman filter to account for the temporal order of the observations. However their approach is limited to conservative systems, and assumes the existence of a spectral gap on the spectrum of the approximated operator, excluding thereby systems with continuous spectra, e.g. chaotic systems [60, 61].

Geodesic curves and the most probable path in the Onsager-Machlup sense for stochastic processes. The theoretical underpinnings of our work can be traced back to the work of Onsager and Machlup [62] and the computation of the **most probable path** (**MPP**) of a diffusion process between two predetermined states. Earlier work has employed the **Onsager-Machlup (OM) function** as Lagrangian to derive an expression for the **MPP** in terms of state variables and the drift of the diffusion process [63–69]. The resulting Lagrangian involves the energy of the path (see Sec. B), which is the same objective used to identify geodesics (see Sec. A.1.2). In our framework, to identify the geodesics between successive observations we assumed as smooth manifold the $\mathcal{R}^d$ with associated Riemannian metric h learned from the data. However the underlying SDE is defined on $\mathcal{R}^d$ under the Euclidean metric [70]. Different metrics in the definition of diffusion processes result in different generators, and thus in different path probabilities for each process. It would be an interesting theoretical result to calculate the transformation induced by the changing the Riemannian metric in the definition of the process. To the best of our knowledge such a result is not available in the literature, but is also not trivial. The change of metric induces a change in the diffusivity of the process, so a direct Girsanov transformation is not feasible. Yet, it may be possible to employ an inverse Lamperti transformation [71] to express the drift of the process in terms of a diffusion with multiplicative noise that would have induced the change in metric learned from the observations (see Sec. B). Finally, the connection to the **OM** functional already hints to an alternative method to obtain an estimate for the unknown drift with geometric considerations that obviates the computationally costly simulation of continuous paths.

Limitations. Our approach is limited to systems where the invariant density can be approximated by a manifold on which one can identify geodesics. Additionally by construction the method implicitly assumes that the invariant system's density is approximately uniformly sampled. However we foresee that by employing more advanced tools from geometric statistics [50, 51] the framework may be applicable for non-uniformly sampled invariant densities. Our experiments have shown that the proposed approach is better suited for systems with *cyclic balance* [72], i.e., with stationary fluctuating probability currents in the steady state. Systems with non-fluctuating currents in the stationary state are nevertheless effectively recovered with existing methods that rely on assumptions of conservative forces.

Acknowledgements

We are indebted to Nina Miolane for providing detailed information on how to compute geodesics from manifolds approximated with Variational Autoencoders. We further thank Georgios Arvanitidis for maintaining a publicly available repository with the algorithm employed here to construct geodesics, Stefan Sommer for answering questions on transformations of diffusion processes on manifolds, and Prof. Manfred Opper for prompting us to work on this problem and for providing initial guidance. We further acknowledge that previous work from the Python [73], numpy [74], scipy [75], matplotlib [76], seaborn [77], GPflow [78], pyEMD [79], and pytorch [80] communities facilitated the implementation of the computational part of this work.

An implementation of this work will be released in the following repository: https://github.com/dimitra-maoutsa/Geometric-path-augmentation-for-SDEs once the article gets submitted for archival publication.

Broader Impact Statement

We introduced a new path augmentation method that allows for efficient inference of stochastic systems observed at large inter-observation intervals. Our contribution aims to highlight the need for incorporating notions from the rapidly developing field of geometric statistics into the area of model discovery of stochastic systems. While geometric and topological properties of invariant densities for deterministic systems have been exceedingly studied in the past, the same is not true for their stochastic counterparts, and in particular of systems described by stochastic differential equations.

Our work aims further to highlight that data augmentation frameworks in settings where the amount of augmented data dominates the number of observations may lead to more accurate inferences by incorporating domain knowledge or other type of information in the augmentation (like here information regarding the geometry of the system's invariant density). Many of the algorithms employed with data augmentation frameworks exhibit only **local convergence**, e.g., the Expectation Maximisation algorithm employed here [81]. In settings where the initial estimate strongly deviates from its true value, naive data augmentation strategies might therefore converge to sub-optimal solutions, that do not reflect the ground truth.

We do not foresee any direct social impact of our work. However we acknowledge that stochastic systems may be used for military purposes and financial engineering, however the proposed method does not directly propose interventions to the observed system that may lead to unfavourable consequences.

Diffusive systems are prevalent in several scientific fields, such as parts of physics, biology, neuroscience, and ecology. We foresee that this work may benefit these disciplines by providing a tool for identifying systems of interest.

References

[1] Stevan J Arnold. *Phenotypic Evolution: The Ongoing Synthesis: (American Society of Naturalists Address)*. 2014.

[2] Russell Lande. "Natural selection and random genetic drift in phenotypic evolution". In: *Evolution* (1976), pp. 314–334.

[3] Subrahmanyan Chandrasekhar. "Stochastic problems in physics and astronomy". In: *Reviews of modern physics* 15.1 (1943), p. 1.

[4] Philip Nelson. *Biological physics*. WH Freeman New York, 2004.

[5] Tom Kuusela. "Stochastic heart-rate model can reveal pathologic cardiac dynamics". In: *Physical Review E* 69.3 (2004), p. 031916.

[6] Philip Sura. "Stochastic analysis of Southern and Pacific Ocean sea surface winds". In: *Journal of the atmospheric sciences* 60.4 (2003), pp. 654–666.

[7] Jose Casadiego, Dimitra Maoutsa, and Marc Timme. "Inferring network connectivity from event timing patterns". In: *Physical review letters* 121.5 (2018), p. 054101.

[8] Philip Sura and Sarah T Gille. "Interpreting wind-driven Southern Ocean variability in a stochastic framework". In: *Journal of marine research* 61.3 (2003), pp. 313–334.

[9] TD Frank, R Friedrich, and PJ Beek. "Stochastic order parameter equation of isometric force production revealed by drift-diffusion estimates". In: *Physical Review E* 74.5 (2006), p. 051905. DOI: https://doi.org/10.1103/physreve.74.051905.

[10] Anke M van Mourik, Andreas Daffertshofer, and Peter J Beek. "Deterministic and stochastic features of rhythmic human movement". In: *Biological cybernetics* 94.3 (2006), pp. 233–244.

[11] Juraj Bergman, Dominik Schrempf, Carolin Kosiol, and Claus Vogl. "Inference in population genetics using forward and backward, discrete and continuous time processes". In: *Journal of Theoretical Biology* 439 (2018), pp. 166–180.

[12] Raphael Sarfati, Jerzy Bławzdziewicz, and Eric R Dufresne. "Maximum likelihood estimations of force and mobility from single short Brownian trajectories". In: *Soft Matter* 13.11 (2017), pp. 2174–2180.

[13] Georg A. Gottwald, Daan T. Crommelin, and Christian L. E. Franzke. "Stochastic climate theory". In: *Nonlinear and Stochastic Climate Dynamics* (2017), pp. 209–240. DOI: https://doi.org/10.1017/9781316339251.009.

[14] Dimitra Maoutsa. "Revealing latent stochastic dynamics from single-trial spike train observations". Bernstein Conference for Computational Neuroscience, Berlin 2022. Sept. 2022. DOI: 10.12751/nncn.bc2022.250.

[15] Nazareno Campioni, Dirk Husmeier, Juan Morales, Jennifer Gaskell, and Colin J Torney. "Inferring microscale properties of interacting systems from macroscale observations". In: *Physical Review Research* 3.4 (2021), p. 043074.

[16] J Cremers and A Hübler. "Construction of differential equations from experimental data". In: *Zeitschrift für Naturforschung A* 42.8 (1987), pp. 797–802.

[17] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". In: *Proceedings of the national academy of sciences* 113.15 (2016), pp. 3932–3937. DOI: 10.1073/pnas.1517384113.

[18] Bryan C Daniels and Ilya Nemenman. "Automated adaptive inference of phenomenological dynamical models". In: *Nature communications* 6.1 (2015), pp. 1–8. DOI: 10.1038/ncomms9133.

[19] Kevin McGoff, Sayan Mukherjee, and Natesh Pillai. "Statistical inference for dynamical systems: A review". In: *Statistics Surveys* 9 (2015), pp. 209–252.

[20] Philipp Batz, Andreas Ruttor, and Manfred Opper. "Variational estimation of the drift for stochastic differential equations from the empirical density". In: *Journal of Statistical Mechanics: Theory and Experiment* 2016.8 (2016), p. 083404.

[21] Amit Singer and Ronald R Coifman. "Non-linear independent component analysis with diffusion maps". In: *Applied and Computational Harmonic Analysis* 25.2 (2008), pp. 226–239.

[22] Feliks Nüske, Péter Koltai, Lorenzo Boninsegna, and Cecilia Clementi. "Spectral properties of effective dynamics from conditional expectations". In: *Entropy* 23.2 (2021), p. 134. DOI: https://doi.org/10.3390/e23020134.

[23] Edward L Ionides, Carles Bretó, and Aaron A King. "Inference for nonlinear dynamical systems". In: *Proceedings of the National Academy of Sciences* 103.49 (2006), pp. 18438–18443. DOI: https://doi.org/10.1073/pnas.0603181103.

[24] Ronen Talmon and Ronald R Coifman. "Intrinsic modeling of stochastic dynamical systems using empirical geometry". In: *Applied and Computational Harmonic Analysis* 39.1 (2015), pp. 138–160.

[25] Carmeline J Dsilva, Ronen Talmon, C William Gear, Ronald R Coifman, and Ioannis G Kevrekidis. "Data-driven reduction for a class of multiscale fast-slow stochastic dynamical systems". In: *SIAM Journal on Applied Dynamical Systems* 15.3 (2016), pp. 1327–1351.

[26] Rudolf Friedrich and Joachim Peinke. "Description of a turbulent cascade by a Fokker-Planck equation". In: *Physical Review Letters* 78.5 (1997), p. 863. DOI: https://doi.org/10.1103/PhysRevLett.78.863.

[27] Silke Siegert, R Friedrich, and J Peinke. "Analysis of data sets of stochastic systems". In: *Physics Letters A* 243.5-6 (1998), pp. 275–280. DOI: 10.1016/S0375-9601(98)00283-7.

[28] Mario Ragwitz and Holger Kantz. "Indispensable finite time corrections for Fokker-Planck equations from time series data". In: *Physical Review Letters* 87.25 (2001), p. 254501. DOI: 10.1103/physrevlett.87.254501.

[29] Lorenzo Boninsegna, Feliks Nüske, and Cecilia Clementi. "Sparse learning of stochastic dynamical equations". In: *The Journal of chemical physics* 148.24 (2018), p. 241723. DOI: 10.1063/1.5018409.

[30] David Lamouroux and Klaus Lehnertz. "Kernel-based regression of drift and diffusion coefficients of stochastic processes". In: *Physics Letters A* 373.39 (2009), pp. 3507–3512.

[31] Andrew Golightly and Darren J Wilkinson. "Bayesian inference for nonlinear multivariate diffusion models observed with error". In: *Computational Statistics & Data Analysis* 52.3 (2008), pp. 1674–1693.

[32] Omiros Papaspiliopoulos, Yvo Pokern, Gareth O Roberts, and Andrew M Stuart. "Nonparametric estimation of diffusions: a differential equations approach". In: *Biometrika* 99.3 (2012), pp. 511–531.

[33] Giorgos Sermaidis, Omiros Papaspiliopoulos, Gareth O Roberts, Alexandros Beskos, and Paul Fearnhead. "Markov chain Monte Carlo for exact inference for diffusions". In: *Scandinavian Journal of Statistics* 40.2 (2013), pp. 294–321.

[34] Alexandros Beskos, Omiros Papaspiliopoulos, and Gareth O Roberts. "Retrospective exact simulation of diffusion sample paths with applications". In: *Bernoulli* 12.6 (2006), pp. 1077–1098.

[35] Siddhartha Chib, Michael K Pitt, and Neil Shephard. "Likelihood based inference for diffusion driven state space models". In: *Por Clasificar* (2006), pp. 1–33.

[36] Cagatay Yildiz, Markus Heinonen, Jukka Intosalmi, Henrik Mannerstrom, and Harri Lahdesmaki. "Learning stochastic differential equations with gaussian processes without gradient matching". In: *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2018, pp. 1–6.

[37] Rudolf Friedrich, Joachim Peinke, Muhammad Sahimi, and M Reza Rahimi Tabar. "Approaching complexity by stochastic methods: From biological systems to turbulence". In: *Physics Reports* 506.5 (2011), pp. 87–162.

[38] Tyrus Berry and John Harlim. "Iterated diffusion maps for feature identification". In: *Applied and Computational Harmonic Analysis* 45.1 (2018), pp. 84–119.

[39] Philipp Batz, Andreas Ruttor, and Manfred Opper. "Approximate Bayes learning of stochastic differential equations". In: *Physical Review E* 98.2 (2018), p. 022109.

[40] Dimitra Maoutsa and Manfred Opper. "Deterministic particle flows for constraining stochastic nonlinear systems". In: *Phys. Rev. Research* 4 (4 Oct. 2022), p. 043035. DOI: 10.1103/PhysRevResearch.4.043035. URL: https://link.aps.org/doi/10.1103/PhysRevResearch.4.043035.

[41] Dimitra Maoutsa and Manfred Opper. "Deterministic particle flows for constraining SDEs". In: *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.

[42] Valentin De Bortoli, Arnaud Doucet, Jeremy Heng, and James Thornton. "Simulating diffusion bridges with score matching". In: *arXiv preprint arXiv:2111.07243* (2021).

[43] M Tabar. "Kramers–Moyal Expansion and Fokker–Planck Equation". In: *Analysis and Data-Based Reconstruction of Complex Nonlinear Dynamical Systems*. Springer, 2019, pp. 19–29.

[44] Rudolf Friedrich, Silke Siegert, Joachim Peinke, Marcus Siefert, Michael Lindemann, Jan Raethjen, Güntner Deuschl, Gerhard Pfister, et al. "Extracting model equations from experimental data". In: *Physics Letters A* 271.3 (2000), pp. 217–222.

[45] 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* 26 (2013).

[46] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[47] Ola Elerian, Siddhartha Chib, and Neil Shephard. "Likelihood inference for discretely observed nonlinear diffusions". In: *Econometrica* 69.4 (2001), pp. 959–993.

[48] 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*. PMLR. 2019, pp. 1506–1515.

[49] Manfred Opper. "Variational inference for stochastic differential equations". In: *Annalen der Physik* 531.3 (2019), p. 1800233.

[50] Nina Miolane, Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, et al. "Geomstats: a Python package for Riemannian geometry in machine learning". In: *Journal of Machine Learning Research* 21.223 (2020), pp. 1–9.

[51] Stefan Sommer. "Probabilistic approaches to geometric statistics: Stochastic processes, transition distributions, and fiber bundle geometry". In: *Riemannian Geometric Statistics in Medical Image Analysis*. Elsevier, 2020, pp. 377–416. DOI: https://doi.org/10.1016/B978-0-12-814725-2.00018-2.

[52] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. "Testing the manifold hypothesis". In: *Journal of the American Mathematical Society* 29.4 (2016), pp. 983–1049.

[53] Tal Shnitzer, Ronen Talmon, and Jean-Jacques Slotine. "Manifold Learning for Data-Driven Dynamical System Analysis". In: *The Koopman Operator in Systems and Control*. Springer, 2020, pp. 359–382.

[54] Jared L Callaham, J-C Loiseau, Georgios Rigas, and Steven L Brunton. "Nonlinear stochastic modelling with Langevin regression". In: *Proceedings of the Royal Society A* 477.2250 (2021), p. 20210092. DOI: https://doi.org/10.1098/rspa.2021.0092.

[55] Steven J Lade. "Finite sampling interval effects in Kramers–Moyal analysis". In: *Physics Letters A* 373.41 (2009), pp. 3705–3709.

[56] 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*. PMLR. 2020, pp. 9526–9536.

[57] Tal Shnitzer, Ronen Talmon, and Jean-Jacques Slotine. "Manifold learning with contracting observers for data-driven time-series analysis". In: *IEEE Transactions on Signal Processing* 65.4 (2016), pp. 904–918. DOI: https://doi.org/10.1109/TSP.2016.2616334.

[58] Dimitrios Giannakis. "Data-driven spectral decomposition and forecasting of ergodic dynamical systems". In: *Applied and Computational Harmonic Analysis* 47.2 (2019), pp. 338–396.

[59] Nelida Črnjarić-Žic, Senka Maćešić, and Igor Mezić. "Koopman operator spectrum for random dynamical systems". In: *Journal of Nonlinear Science* 30.5 (2020), pp. 2007–2056.

[60] Bernard O Koopman and J v Neumann. "Dynamical systems of continuous spectra". In: *Proceedings of the National Academy of Sciences* 18.3 (1932), pp. 255–263.

[61] Igor Mezić. "Spectral properties of dynamical systems, model reduction and decompositions". In: *Nonlinear Dynamics* 41.1 (2005), pp. 309–325.

[62] Lars Onsager and Stefan Machlup. "Fluctuations and irreversible processes". In: *Physical Review* 91.6 (1953), p. 1505.

[63] Takahiko Fujita and Shin-ichi Kotani. "The Onsager-Machlup function for diffusion processes". In: *Journal of mathematics of Kyoto University* 22.1 (1982), pp. 115–130.

[64] Y Takahashi and S Watanabe. "The probability functionals (Onsager-Machlup functions) of diffusion processes". In: *Stochastic Integrals*. Springer, 1981, pp. 433–463.

[65] Detlef Dürr and Alexander Bach. "The Onsager-Machlup function as Lagrangian for the most probable path of a diffusion process". In: *Communications in Mathematical Physics* 60.2 (1978), pp. 153–170.

[66] Robert Graham. "Path integral formulation of general diffusion processes". In: *Zeitschrift für Physik B Condensed Matter* 26.3 (1977), pp. 281–290.

[67] Hidemi Ito. "Probabilistic construction of Lagrangean of diffusion process and its application". In: *Progress of Theoretical Physics* 59.3 (1978), pp. 725–741.

[68] Ruslan Leontievich Stratonovich. "On the probability functional of diffusion processes". In: *Selected Trans. in Math. Stat. Prob* 10 (1971), pp. 273–286.

[69] Stefan Sommer. "Anisotropic distributions on manifolds: Template estimation and most probable paths". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2015, pp. 193–204.

[70] Mireille Capitaine. "On the Onsager-Machlup functional for elliptic diffusion processes". In: *Séminaire de Probabilités XXXIV* (2000), pp. 313–328.

[71] Bernt Øksendal. "Stochastic differential equations". In: *Stochastic differential equations*. Springer, 2003, pp. 65–84.

[72] Kazuhisa Tomita and Hiroyuki Tomita. "Irreversible circulation of fluctuation". In: *Progress of theoretical physics* 51.6 (1974), pp. 1731–1749.

[73] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*. Vol. 620. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

[74] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[75] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[76] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing In Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[77] Michael L Waskom. "Seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (2021), p. 3021.

[78] Alexander G. de G. Matthews, Mark van der Wilk, Tom Nickson, Keisuke. Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. "GPflow: A Gaussian process library using TensorFlow". In: *Journal of Machine Learning Research* 18.40 (Apr. 2017), pp. 1–6. URL: http://jmlr.org/papers/v18/16-537.html.

[79] Ofir Pele and Michael Werman. "Fast and robust earth mover's distances". In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE. Sept. 2009, pp. 460–467.

[80] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. "Automatic differentiation in pytorch". In: (2017).

[81] Orlando Romero, Sarthak Chatterjee, and Sérgio Pequito. "Convergence of the expectation-maximization algorithm through discrete-time Lyapunov stability theory". In: *2019 American Control Conference (ACC)*. IEEE. 2019, pp. 163–168.

[82] Robert S Liptser and Albert N Shiryaev. *Statistics of random processes II: Applications*. Vol. 6. Springer Science & Business Media, 2013. DOI: 10.1007/978-3-662-10028-8.

[83] Carl Edward Rasmussen. "Gaussian processes in machine learning". In: *Summer School on Machine Learning*. Springer-Verlag. 2003, pp. 63–71.

[84] Christoph Honisch and Rudolf Friedrich. "Estimation of Kramers-Moyal coefficients at low sampling rates". In: *Physical Review E* 83.6 (2011), p. 066701.

[85] Raphaël Chetrite and Hugo Touchette. "Variational and optimal control representations of conditioned and driven processes". In: *Journal of Statistical Mechanics: Theory and Experiment* 2015.12 (2015), P12001.

[86] Satya N Majumdar and Henri Orland. "Effective Langevin equations for constrained stochastic processes". In: *Journal of Statistical Mechanics: Theory and Experiment* 2015.6 (2015), P06039.

[87] Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*. Vol. 6. Springer, 1992.

[88] John M Lee. *Introduction to Riemannian manifolds*. Vol. 176. Springer, 2018.

[89] Stephen Wiggins. *Normally hyperbolic invariant manifolds in dynamical systems*. Vol. 105. Springer Science & Business Media, 1994.

[90] Salah-Eldin A Mohammed and Michael KR Scheutzow. "The stable manifold theorem for stochastic differential equations". In: *Annals of probability* (1999), pp. 615–652. DOI: https://doi.org/10.1214/aop/1022677380.

[91] TV Girya and Igor Dmitrievich Chueshov. "Inertial manifolds and stationary measures for stochastically perturbed dissipative dynamical systems". In: *Sbornik: Mathematics* 186.1 (1995), pp. 29–45.

[92] Neil Fenichel and JK Moser. "Persistence and smoothness of invariant manifolds for flows". In: *Indiana University Mathematics Journal* 21.3 (1971), pp. 193–226.

[93] Ludwig Arnold. "Stochastic differential equations as dynamical systems". In: *Realization and Modelling in System Theory*. Springer, 1990, pp. 489–495.

[94] Andrew Carverhill. "Flows of stochastic dynamical systems: ergodic theory". In: *Stochastics: An International Journal of Probability and Stochastic Processes* 14.4 (1985), pp. 273–317.

[95] Christian Fröhlich, Alexandra Gessner, Philipp Hennig, Bernhard Schölkopf, and Georgios Arvanitidis. "Bayesian Quadrature on Riemannian Data Manifolds". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 3459–3468. DOI: https://doi.org/10.48550/arXiv.2102.06645.

[96] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. "Latent space oddity: on the curvature of deep generative models". In: *arXiv preprint arXiv:1710.11379* (2017). DOI: https://doi.org/10.48550/arXiv.1710.11379.

[97] Tooru Taniguchi and EGD Cohen. "Onsager-Machlup theory for nonequilibrium steady states and fluctuation theorems". In: *Journal of Statistical Physics* 126.1 (2007), pp. 1–41.

[98] Artur B Adib. "Stochastic actions for diffusive dynamics: Reweighting, sampling, and minimization". In: *The Journal of Physical Chemistry B* 112.19 (2008), pp. 5910–5916.

[99] Robert Graham. "Onsager-Machlup Function of Nonlinear Non-Equilibrium Thermodynamics". In: *Functional Integration*. Springer, 1980, pp. 263–280.

[100] Erlend Grong and Stefan Sommer. "Most probable flows for Kunita SDEs". In: *arXiv preprint arXiv:2209.03868* (2022).

[101] Dimitra Maoutsa, Sebastian Reich, and Manfred Opper. "Interacting particle solutions of Fokker–Planck equations through gradient–log–density estimation". In: *Entropy* 22.8 (2020), p. 802.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We provide numerical results to justify our claims.
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] In the impact statement.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Not yet. The work will be submitted for archival publication, therefore for now we do not open source our code. However, our work combines already published frameworks [39, 40,

48]. The articles [40, 48] provide github repositories with available implementations, which are the ones we used. For [39] we re-implemented the drift inference as described in the main paper and the supplement. For archival reasons, and to make our code discoverable for people who will read the camera-ready version of the article in the future, we will provide a link to a non-populated repository, that will host the implementation once our full article is ready for submission.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] One run of our framework for a two dimensional system requires around 4-6 hours (depending on the amount of data) on a laptop with Intel Core i7@ 1.80GHz CPU.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [No] Not yet. We will release the framework once the paper will be submitted for archival publication.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendices

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 drift-less 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 $[0,\, T]$ [82]

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

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 of 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 [45]. 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 [45] 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 $\mathrm{d}X_t$ [83]. 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 [26, 45] becomes inaccurate. To mitigate this issue Lade [55] 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 [84]. 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 [31–35]. 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 2 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 [39] 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 without 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 $\mathcal{O}_{1:K}$. We consider that the prior process (the process without considering the observations $\mathcal{O}_{1:K}$) is described by the equation

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

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 [85, 86]. We identify the approximate posterior measure Q with the posterior measure induced by an approximating process that is conditioned by the observations $\mathcal{O}_{1:K}$ [49], with governing equation

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

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

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

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

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

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

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 $V_T = \int_0^T \frac{\hat{f}(X_t)-g(X_t,t)}{\sigma^2}\mathrm{d}\boldsymbol{\beta}_t$. 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 [49]

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

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 [49] 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 [40, 41])

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

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. [40, 41] 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 with 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** [87, 88] $(\mathcal{M}, h)$ 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 [89–94]. 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 [95]. 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. in [48]. 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 [48]

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

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 [87].

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 [87, 96]

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

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. [48] 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 $\gamma_{t'}^k$ that connect consecutive observations

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

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 [49], 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 [40] 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.* [45], 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

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

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

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

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

The associated variance results similarly from the equation

$$\Sigma_S^2(x) = k^f(x,x) - k^f(x, \mathcal{Z})\left(I + \Lambda\, \mathcal{K}_S\right)^{-1} \Lambda\, k^f(\mathcal{Z}, x). \tag{38}$$

We employ a sample based approximation of the densities in Eq. (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 Theoretical evidence that supports the use of geodesics as geometric constraints

The Onsager-Machlup functional for diffusion processes has been known in theoretical physics as a characteriser of the most probable path (MPP) between two pre-defined states of the process. In [62], Onsager and Machlup used the thermal fluctuations of a diffusion process to show that the probability

density of a path $\gamma \in C^1\left([0,T],\mathcal{R}^d\right)$ in $\mathcal{R}^d$ over finite interval can be expressed as a Boltzmann factor

$$\mathrm{P}(\gamma) \sim \exp\left[-\int_0^T L(\gamma(t),\dot{\gamma}(t))dt\right], \tag{39}$$

where

$$L(\gamma(t),\dot{\gamma}(t)) = \frac{1}{2}\|\frac{\dot{\gamma}(t)-f(\gamma(t))}{\sigma}\|^2 + \frac{1}{2}\nabla\cdot f(\gamma(t)).^{4} \tag{40}$$

The function $L(\gamma(t),\dot{\gamma}(t))$ is known as the **Onsager-Machlup** function (action), while its integral over time is known as Onsager-Machlup action functional. It has been used as Lagrangian in Euler-Lagrange minimisation schemes to identify the **most probable path (MPP)** of a diffusion process between two given points in the state space [66, 68].

Stratonovich [68] considered the probability that a sample of a multidimensional diffusion process will lie in the vicinity of (within a tube of infinitesimal thickness around) an idealised smooth path in the state space. To compute this probability he constructed a probability functional which is identical to the Onsager-Machlup functional considered as Lagrangian for the diffusion process. Duerr et al. [65] considered scalar diffusion processes and constructed the Onsager-Machlup function from the asymptotic limit of the transition probability between the starting and end state of the path using a Girsanov transformation.

Considering diffusion processes defined on a Riemannian manifold $(\mathcal{M},g)$ with associated Riemannian metric g, the Onsager-Machlup functional can be expressed as the integral over the Lagrangian [64, 99, 100]

$$L(\gamma,\dot{\gamma}) = \frac{1}{2}\|\dot{\gamma}(t)\|_g^2 - \frac{1}{12}S(\gamma(t)), \tag{41}$$

where $\|\cdot\|_g$ denotes the Riemannian norm on the tangent space $\mathcal{T}_X\mathcal{M}$ of the manifold with respect to the metric g, and $S(\cdot)$ stands for the scalar curvature of the manifold at each point. The first term is the Lagrangian used to identify geodesic curves on manifolds (c.f. Eq. (A.1.2))

In our proposed formalism, for computational purposes we have assumed the entire $\mathcal{R}^d$ as smooth manifold. Thus the curvature of the manifold is everywhere zero for our setting, and we can identify the remaining term of Eq. (41) with the Lagrangian we optimised for computing the geodesics on the manifold $(\mathcal{R}^d,g)$, where g is the metric learned from the observations.

However the system we observed was a diffusion process defined in $\mathcal{R}^d$ with an Euclidean metric. Constructing a path augmentation scheme that guides the augmented paths towards the geodesics of a diffusion defined with respect to a different metric raises questions about the validity of our approach. Here we should note that diffusions with a general state dependent diffusion coefficient $\sigma \in \mathcal{R}^{m\times d}$, with $m \leq d$ can be considered as evolving on the manifold $(\mathcal{R}^m,g)$, with the associated metric $g = \left(\sigma\sigma^\top\right)^{-1}$ [70]. Thus it may be possible to associate the metric learned from the data with the metric arising from a state dependent diffusion by applying a transformation akin to an inverse Lamperti transform [71] to transform our learned SDE to one that would have induced the learned metric due to the state dependent diffusion. The existence of such a transformation would justify the proposed method. Our empirical results demonstrate that such a transformation may be possible.

C Does the proposed approach invalidate the Markovian property of the diffusion process?

The proposed path augmentation seemingly invalidates the Markovian property of the diffusion process. According to the Markov property of the diffusion of Eq. (1), the system state $X_{k\tau+\delta t}$ should depend only the state $X_{k\tau}$, i.e., the observation $\mathcal{O}_k$. The proposed augmentation makes the state $X_{k\tau+\delta t}$ depending not only on the next observation $\mathcal{O}_{k+1} = X_{(k+1)\tau}$, but also on past and future states that lie in the vicinity of these observations.

We effectively construct the augmented paths to compute the likelihood of a drift estimate. To compute this likelihood we require to evaluate the transition probabilities between consecutive observations.

[4] Onsager and Machlup's initial work concentrated around linear processes and therefore the functional initially introduced by the did not include the second term with the divergence of f as this is a constant for linear f. It was later added to the OM function to account for trajectory entropy corrections [97, 98]

Since for general nonlinear systems the transition probabilities are in general intractable, we have to resort to numerical approximations. Ideally we would approximate the transition density with a bridge sampler that would consider the nonlinear estimated SDE conditioned to pass though consecutive observations. However for coarse drift estimates, the observations have zero probability under the law of the estimated SDE, and construction of those bridges would result either in very taxing computations or would fail altogether. Instead, here, we compute the likelihood of a "corrected" estimate (the correction resulting from the invariant density) under which the observations have non-zero probability, and subsequently re-estimate the drift on the augmented path with this "corrected" estimate. By taking into account the local geometry of the observations, we provide systematic corrections for the misestimated drift function to generate the augmented paths. This effectively nudges the augmentation process towards the second observation of each inter-observation interval through the path constraint that forces the augmented paths towards the geodesics.

D Details on numerical experiments

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

$$f_1(x, y) = \mu(x - \frac{1}{3}x^3 - y) \tag{42}$$

$$f_2(x, y) = \frac{1}{\mu}x, \tag{43}$$

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 $\tau = \{80, 120, 160, 200, 240, 280, 320\}$. 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 [40, 101]. 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 2 was constructed with $\beta = 1$ and provides a better approximation of the transition density, than $\beta = 0.5$.