LOCALIZED SCHRÖDINGER BRIDGE SAMPLER

GEORG A. GOTTWALD AND SEBASTIAN REICH

Abstract. We consider the generative problem of sampling from an unknown distribution for which only a sufficiently large number of training samples are available. In this paper, we build on previous work combining Schrödinger bridges and Langevin dynamics. A key bottleneck of this approach is the exponential dependence of the required training samples on the dimension, d, of the ambient state space. We propose a localization strategy which exploits conditional independence of conditional expectation values. Localization thus replaces a single high-dimensional Schrödinger bridge problem by d low-dimensional Schrödinger bridge problems over the available training samples. As for the original approach, the localized sampler is stable and geometric ergodic. The sampler also naturally extends to conditional sampling and to Bayesian inference. We demonstrate the performance of our proposed scheme through experiments on a Gaussian problem with increasing dimensions and on a stochastic subgrid-scale parametrization conditional sampling problem.

Keywords: generative modeling, Langevin dynamics, Schrödinger bridges, conditional independence, localization, Bayesian inference, conditional sampling, multi-scale closure

AMS: 60H10,62F15,62F30,65C05,65C40

1. Introduction

In this paper, we consider the problem of sampling from an unknown probability measure $\nu(\mathrm{dx})$ on $\mathbb{R}^d$ for which we only have access to a finite set of training samples $x^{(j)} \sim \nu$, $j = 1, \ldots, M$. This problem has recently attracted widespread interest in the context of score-generative or diffusion modeling [18, 7, 17, 19, 22]. If the probability measure $\nu(\mathrm{d}x)$ possesses a probability density function $\pi(x)$, then a popular non-parametric approach to generative modeling is to estimate the score function $s_\theta(x) \approx \nabla \log \pi(x)$ by minimizing an appropriate loss function such as

$$\mathcal{L}(\theta) = \int_{\mathbb{R}^d} \pi(x) \| s_\theta(x) - \nabla \log \pi(x) \|^2 \mathrm{d}x \tag{1}$$

in the parameters $\theta \in \mathbb{R}^{d_\theta}$ [8]. This estimate can then be used in combination with overdamped Langevin dynamics to yield

$$\dot{X}(\tau) = s_\theta(X(\tau)) + \sqrt{2}\, \dot{W}(\tau), \tag{2}$$

where $W(\tau)$ denotes standard d-dimensional Brownian motion [13]. The stochastic differential equation is typically discretized by the Euler–Maruyama (EM) method to yield an iterative update of the form

$$X(n+1) = X(n) + \epsilon s_\theta(X(n)) + \sqrt{2}\, \Xi(n), \qquad \Xi(n) \sim \mathrm{N}(0, \epsilon I), \tag{3}$$

for $n \geq 0$, where $\epsilon > 0$ denotes the step size and $X(n)$ provides the numerical approximation to the solution of (2) at time $\tau_n = n\,\epsilon$. The EM algorithm is initialized at one of the training data

Date: September 12, 2024.

points; i.e., $X(0) = x^{(j^*)}$ with $j^* \in \{1, \ldots, M\}$ appropriately chosen, and the resulting discrete trajectory $X(n)$, $n \geq 1$, delivers approximate samples from the target distribution $\nu(\mathrm{d}x)$.

Instead of first estimating the score function from samples and then discretizing (2) in time, it has been proposed in [5] to employ Schrödinger bridges and to directly estimate the conditional expectation value

$$\mu(x;\epsilon) := \mathbb{E}[X(\epsilon)|X(0) = x] \tag{4}$$

from the given samples $\{x^{(j)}\}_{j=1}^M$ for given time-step $\epsilon > 0$. We denote the Schrödinger bridge approximation obtained from the samples by $m(x;\epsilon) : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$ and obtain the iteration scheme

$$X(n+1) = m(X(n);\epsilon) + \sqrt{2}\,\Xi(n), \qquad \Xi(n) \sim \mathrm{N}(0, \epsilon I). \tag{5}$$

Remark 1. We emphasize that the discrete-time formulation (5) can be considered even in case the probability measure $\nu(\mathrm{d}x)$ does not possess a probability density function $\pi(x)$ with respect to the Lebesgue measure on $\mathbb{R}^d$; e.g., the measure ν is concentrated on a submanifold $\mathcal{M} \subset \mathbb{R}^d$, as long as the conditional expectation value (4) can be defined appropriately. The Schrödinger bridge approximation to $m(x;\epsilon)$ allows for such an extension [5].

While it has been demonstrated in [5] that (5) works well for low-dimensional problems, the required number of training samples, M, increases exponentially in the dimension, d, of the samples [21]. In order to remedy this manifestation of the curse of dimensionality, we propose to utilize conditional independence in order to replace the Schrödinger bridge estimator for the conditional expectation value $m(x;\epsilon) \in \mathbb{R}^d$ by appropriately localized Schrödinger bridge estimators in each of the d components of $m(x;\epsilon)$. While the proposed localization strategy resembles localization strategies used in the ensemble Kalman filter (EnKF) [3, 16, 2], it is fundamentally different in at least two ways: (i) For Gaussian measures with covariance matrix Σ, the EnKF would localize the empirical estimator of Σ while our approach relies on a localization of Σ^{-1} as dictated by conditional independence. (ii) Localized Schrödinger bridge estimators are not restricted to Gaussian measures as long as conditional independence can be established.

While (5) can be used in the general context of score-generative or diffusion modeling, our main motivation is in Bayesian inference and conditional sampling with applications to multi-scale processes. Applications to Bayesian inference, for which $\pi(x)$ takes the role of the prior for given likelihood function $\pi(y|x)$, immediately suggest the modified update

$$X(n+1) = m(X(n);\epsilon) - \epsilon\nabla \log \pi(y|X(n)) + \sqrt{2}\,\Xi(n), \qquad \Xi(n) \sim \mathrm{N}(0, \epsilon I). \tag{6}$$

Furthermore, a particular choice of $\pi(y|x)$ can be used for conditional sampling [5].

The paper is organized as follows. The Schrödinger bridge formulation for $m(x;\epsilon)$ in (5) is summarized in the subsequent Section 2. The localized variant is developed in Section 3 first for a Gaussian distribution for which Σ^{-1} has a tri-diagonal structure and then for general target measure $\nu(\mathrm{d}x)$ for which conditional independence holds. An algorithmic summary is provided in Algorithm 1 and a discussion of numerical properties is provided in Section 3.3. As an application, we consider conditional sampling for a closure problem arising from the multi-scale Lorenz-96 model [9] in Section 4. The paper closes with some conclusions and suggestions for further work.

2. Schrödinger bridge sampler

In this section, we briefly recall how to approximate the conditional estimate (4) using Schrödinger bridges. One first introduces the symmetric matrix $T \in \mathbb{R}^{M\times M}$ of (unnormalized) transition probabilities

$$(T)_{jk} = \exp\left(-\frac{1}{4\epsilon}\|x^{(k)} - x^{(j)}\|^2\right) \tag{7}$$

for $j, k = 1, \ldots, M$. See [5] for a more general definition involving a scaling matrix $K(x)$.

One next introduces the uniform probability vector $w^* = (1/M, \ldots, 1/M)^\top \in \mathbb{R}^M$ over the samples $\{x^{(j)}\}_{j=1}^M$. The associated Schrödinger bridge problem can be reformulated into finding the non-negative scaling vector $v \in \mathbb{R}^M$ such that the symmetric matrix

$$P = D(v)\, T\, D(v) \tag{8}$$

is a Markov chain with invariant distribution w^*, i.e.,

$$P\, w^* = w^*. \tag{9}$$

Here $D(v) \in \mathbb{R}^{M\times M}$ denotes the diagonal matrix with diagonal entries provided by $v \in \mathbb{R}^M$. We remark that the standard scaling used in Schrödinger bridges would lead to a bistochastic matrix $\tilde{P}$, which is related to (8) by $\tilde{P} = M^{-1}P$.

The next step is to extend the discrete Markov chain (8) to all $x \in \mathbb{R}^d$. For that purpose one introduces the vector-valued functions $t(x) \in \mathbb{R}^M$ with entries

$$t^{(j)}(x) = \exp\left(-\frac{1}{4\epsilon}\|x^{(j)} - x\|^2\right) \tag{10}$$

for $j = 1, \ldots, M$. One then defines the probability vector $w(x) \in \mathbb{R}^M$ using the Sinkhorn weights, v, obtained in (8), i.e.,

$$w(x) = \frac{D(v)\, t(x)}{v^\top t(x)} \in \mathbb{R}^M \tag{11}$$

for all $x \in \mathbb{R}^d$. This vector gives the transition probabilities from any x to the data samples, which we collect in the data matrix of samples

$$\mathcal{X} = (x^{(1)}, \ldots, x^{(M)}) \in \mathbb{R}^{d\times M}. \tag{12}$$

Hence, the desired sample-based approximation of the conditional mean is given by

$$m(x;\epsilon) := \mathcal{X}\, w(x), \tag{13}$$

which provides a finite-dimensional approximation of the conditional expectation value $\mu(x;\epsilon)$ of the true underlying diffusion process (2).

It has been found advantageous in [5] to replace the time-stepping method (5) by the split-step scheme

$$X(n+1/2) = X(n) + \sqrt{2}\,\Xi(n), \qquad \Xi(n) \sim \mathrm{N}(0, \epsilon I), \tag{14a}$$

$$X(n+1) = \mathcal{X}\, w(X(n+1/2)), \tag{14b}$$

which can be viewed as sequential noising and denoising steps. The key property of the Schrödinger bridge sampler is that the final step of the Langevin sampler (14)b amounts to a projection into the convex hull of the samples, independent of the outcome of the noising step (14)a. This renders the sampling scheme numerically stable for any finite sample size M. This is in contrast to traditional Langevin samplers such as score generative models which directly solve the typically stiff Langevin equation (2); e.g., in case the probability measure $\nu(\mathrm{d}x)$ is concentrated on a submanifold $\mathcal{M} \subset \mathbb{R}^d$, simulating the Langevin equation necessitates computationally costly sufficiently small time steps to resolve the fast attraction toward the submanifold.

While (14) works well for low-dimensional problems, applications to medium- and high-dimensional problems have remained an open challenge since accurate approximations of the Schrödinger bridge problem require an increasing number of samples as the dimension d of the sample space $\mathbb{R}^d$ increases (cf. [21]).

The key observation of this paper is that the approximation of conditional expectations (4) via Schrödinger bridges does not necessarily require an accurate approximation of the Markov chain

(8) and that localization can be applied provided conditional independence can be established. This idea will be developed in the following section.

3. Localized Schrödinger bridge sampler

To introduce the main idea of localizing the Schrödinger bridge sampler developed in [5] we first consider an illustrative example of sampling from a multivariate Gaussian distribution. We will see that localization allows to significantly reduce the number of samples required to achieve a certain accuracy. In particular, the number of samples required to achieve a certain accuracy does not depend on the intrinsic dimension of the samples but rather on the correlation length which typically is much smaller.

3.1. **Motivational example: Gaussian setting.** Let $\Delta_h \in \mathbb{R}^{d\times d}$ denote the standard discrete Laplacian over a periodic domain $[0,1]$ with mesh-size $h = 1/d$. We assume that the sampling distribution $\pi(x)$ is Gaussian with zero mean and covariance matrix

$$\Sigma = (I - \Delta_h)^{-1}. \tag{15}$$

Instead of the distribution $\pi(x)$, we are given M samples $x^{(j)} \sim \mathrm{N}(0,\Sigma)$, $j = 1,\ldots,M$, and denote their α-th entry by $x_\alpha^{(j)}$ for $\alpha = 1,\ldots,d$. The goal is to produce more samples from $\mathrm{N}(0,\Sigma)$ using the time-stepping scheme (14) without making explicit reference to the unknown covariance matrix Σ. This particular setting of a generative model can become arbitrarily challenging for decreasing mesh-sizes $h = 1/d$.

In order to gain some insight into the problem, we first consider the standard EM sampler in case the distribution is known; i.e.,

$$X(n+1) = X(n) - \epsilon(I - \Delta_h)X(n) + \sqrt{2}\,\Xi(n), \qquad \Xi(n) \sim \mathrm{N}(0,\epsilon I). \tag{16}$$

Because of the structure of Δ_h, we can rewrite the EM update in the components of $X(n)$ in the form

$$X_\alpha(n+1) = w_{-1}\,X_{\alpha-1}(n) + w_0\,X_\alpha(n) + w_1\,X_{\alpha+1}(n) + \sqrt{2\epsilon}\,\Xi_\alpha(n), \qquad \alpha = 1,\ldots,d, \tag{17}$$

with weights

$$w_{\pm 1} = \frac{\epsilon}{2h^2}, \qquad w_0 = 1 - \epsilon\left(1 + \frac{1}{h^2}\right) \tag{18}$$

and periodic extension of X_α for $\alpha = 0$ and $\alpha = d+1$. We assume that the step size ϵ is chosen such that $w_0 \geq 0$. The EM update (17) shows the conditional expectation value of $X_\alpha(n+1)$ only depends on the value of the neighboring grid points of $X(n)$ with weights w_0 and $w_{\pm 1}$; i.e.,

$$\mathbb{E}[X_\alpha(n+1)\,|\,X(n)] = \mathbb{E}[X_\alpha(n+1)\,|\,(X_{\alpha-1}(n), X_\alpha(n), X_{\alpha+1}(n))] \tag{19a}$$

$$= w_{-1}\,X_{\alpha-1}(n) + w_0\,X_\alpha(n) + w_1\,X_{\alpha+1}(n). \tag{19b}$$

It is convenient to introduce the short-hand

$$X_{[\alpha]} := (X_{\alpha-1}, X_\alpha, X_{\alpha+1})^{\mathrm{T}} \in \mathbb{R}^{d_\alpha}, \tag{20}$$

$d_\alpha = 3$, to denote the set of neighboring grid points of X_α.

To help the reader navigating the various indices and sub- and superscripts we summarize here our notation. Superscripts (j) are reserved to denote samples $j = 1,\ldots,M$ as well as components of vectors in $\mathbb{R}^M$. For example, the components of the probability vector $w \in \mathbb{R}^M$ are denoted by $w^{(j)}$. The Greek subscript α with $\alpha = 1,\ldots,d$ is reserved to denote components of a vector x in state space $\mathbb{R}^d$, i.e. y_α for $\alpha = 1,\ldots,d$. Subscripts $[\alpha]$ are reserved to denote localisation around a component α; i.e., $x_{[\alpha]} \in \mathbb{R}^{d_\alpha}$.

The dependency of the conditional expectation value (19) on the neighboring points is to be exploited in the update step (14b), which we recall here in its component-wise formulation as

$$X_\alpha(n+1) = \sum_{j=1}^{M} x_\alpha^{(j)} w^{(j)}(X(n+1/2)), \tag{21}$$

for $\alpha = 1, \ldots, d$. We know from our previous considerations that the conditional expectation value of $X_\alpha(n+1)$ should depend on $X_{[\alpha]}(n+1/2)$ only. Hence the question arises if we can find appropriately localized probability vectors $w(x)$ for the Schrödinger bridge sampler (14)? The following formal argument can be made. We restrict $\mathrm{N}(0,\Sigma)$ to $\mathbb{R}^{d_\alpha}$ and truncate the samples $x^{(j)}$, $j = 1, \ldots, M$, accordingly to yield $x_{[\alpha]}^{(j)}$. The covariance matrix $\Sigma_{\mathrm{r}} \in \mathbb{R}^{d_\alpha \times d_\alpha}$ of the reduced random variables $X_{[\alpha]} \in \mathbb{R}^{d_\alpha}$ is simply the restriction of Σ to the corresponding sub-space, which in this particular example is independent of α. Furthermore, using the Schur complement, one finds

$$\Sigma_{\mathrm{r}}^{-1} = \begin{pmatrix} * & -\frac{1}{2h^2} & * \\ -\frac{1}{2h^2} & 1+\frac{1}{h^2} & -\frac{1}{2h^2} \\ * & -\frac{1}{2h^2} & * \end{pmatrix}, \tag{22}$$

where $*$ denotes entries which differ from the matrix which would be obtained by restricting Σ^{-1} to the corresponding sub-space. The important point is that the central elements remain identical (cf. (18)) and that only those entries enter the approximation of the conditional expectation value (19).

We now describe a localized Schrödinger bridge approach for this specific problem. One replaces the matrix $T \in \mathbb{R}^{M\times M}$ with entries (7) by localized matrices $T_\alpha \in \mathbb{R}^{M\times M}$ with entries

$$(T_\alpha)_{jk} = \exp\left(-\frac{1}{4\epsilon}\|x_{[\alpha]}^{(j)} - x_{[\alpha]}^{(k)}\|^2,\right), \qquad i,j = 1,\ldots,M, \tag{23}$$

for fixed $\alpha \in \{1,\ldots,d\}$. For each of these localized matrices T_α we employ the Sinkhorn algorithm to obtain the Sinkhorn weights $v_\alpha \in \mathbb{R}^M$ for $\alpha = 1,\ldots,d$. The key point is that the Euclidean inner product in $\mathbb{R}^d$, $d \gg 1$, is replaced by an inner product in $\mathbb{R}^{d_\alpha}$ with $d_\alpha = 3$. Furthermore, in this particular example, the corresponding Schrödinger bridge couples the restricted Gaussian distribution $\mathrm{N}(0,\Sigma_r)$ with itself. Next the single M-dimensional probability vector (11) is replaced by d M-dimensional probability vectors

$$w_\alpha(x_{[\alpha]}) := \frac{D(v_\alpha)\, t_\alpha(x_{[\alpha]})}{v_\alpha^\top t_\alpha(x_{[\alpha]})}, \qquad \alpha = 1,\ldots,d, \tag{24}$$

which depend on $x_{[\alpha]} \in \mathbb{R}^{d_\alpha}$ and where the vector-valued function $t_\alpha(x_{[\alpha]}) \in \mathbb{R}^M$ has entries

$$t_\alpha^{(j)}(x_{[\alpha]}) = \exp\left(-\frac{1}{4\epsilon}\|x_{[\alpha]}^{(j)} - x_{[\alpha]}\|^2\right), \qquad j = 1,\ldots,M. \tag{25}$$

Note that (24) depends on the restricted vectors $x_{[\alpha]}^{(j)} \in \mathbb{R}^{d_\alpha}$, $j = 1,\ldots,M$, only. It can be verified by explicit calculation that the interpolation property

$$w_\alpha^{(j)}(x_{[\alpha]}^{(i)}) = (P_\alpha)_{ij}, \qquad i,j = 1,\ldots,M, \tag{26}$$

holds.

We finally obtain the localized approximation

$$m_\alpha(x_{[\alpha]}) = \mathcal{X}_\alpha\, w_\alpha(x_{[\alpha]}), \qquad \alpha = 1,\ldots,d, \tag{27}$$

of the conditional expectation values, where

$$\mathcal{X}_\alpha = (x_\alpha^{(1)},\ldots,x_\alpha^{(M)}) \in \mathbb{R}^{1\times M}, \tag{28}$$

and the update (14b) is replaced by the localized update

$$X_\alpha(n+1) = m_\alpha(X_{[\alpha]}(n+1/2)), \qquad \alpha = 1, \ldots, d. \tag{29}$$

In other words, we have replaced a single Schrödinger bridge update in $\mathbb{R}^d$ by d Schrödinger bridge updates in $\mathbb{R}^{d_\alpha}$. We also introduce the localized data matrix

$$\mathcal{X}_{[\alpha]} = (x^{(1)}_{[\alpha]}, \ldots, x^{(M)}_{[\alpha]}) \in \mathbb{R}^{d_\alpha \times M}, \tag{30}$$

which enters into the computation of T_α.

3.1.1. *Numerical illustration.* To illustrate how well the localized sampling strategy is able to generate samples from a multivariate Gaussian, we generate M training samples of a d-dimensional multivariate Gaussian with

$$x^{(j)} \sim \mathrm{N}(0, \Sigma) \tag{31}$$

for $j = 1, \ldots, M$ with a $d \times d$ covariance matrix of the form (15) with a tridiagonal precision matrix with $\Sigma^{-1}_{i,i} = 2$, $\Sigma^{-1}_{i,i\pm1} = -0.5$, and periodic conditions $\Sigma^{-1}_{1,d} = \Sigma^{-1}_{d,1} = -0.5$. The corresponding entries of the covariance matrix are $\Sigma_{i,i} \approx 5.8$, $\Sigma_{i,i\pm1} \approx 1.6$, $\Sigma_{i,i\pm2} \approx 0.04$, and $\Sigma_{i,i\pm3} \approx 0.01$.

We employ the localized Schrödinger bridge sampler with a localisation set comprised of four neighbouring grid points, i.e. $d_\alpha = 5$. We further employ a standard Schrödinger bridge sampler without localisation. In Figure 1 we compare the generated new samples with the given samples for both sampling strategies. We show the resulting empirical histograms as well as the rows of the empirical covariance matrix $\hat{\Sigma}$. The rows are centered at the middle point using periodicity. To achieve reasonable accuracy for the standard sampler we require $M = 5,000$ samples and a bandwidth of $\epsilon = 0.25$ for $d = 9$. The localised sampler instead achieves superior samples with only a tenth of the data with $M = 500$ and $\epsilon = 0.025$. Moreover, the accuracy does not depend on the dimension for the localized Schrödinger bridge sampler. We show an example with dimension $d = 101$ which allows for the same accuracy as the smaller $d = 9$-dimensional case. The agreement with the associated entries $\Sigma_{i,i+j}$, $i = 51$, $j \in \{0, \pm1, \pm2, \pm3\}$ of the covariance matrix Σ is remarkable. We remark that $d = \mathcal{O}(10^2)$ is out of range for the standard unlocalised sampler.

3.2. **Localized Schrödinger bridge sampler for general measures.** The strategy of constructing a dimension-reduced localised Schrödinger bridge sampler as presented in the previous example of a multivariate Gaussian readily extends to general priors.

We need to introduce some further notation. For each α-th entry in the state vector $x \in \mathbb{R}^d$ we introduce a subset $\Lambda(\alpha) \subset \{1, \ldots, d\}$ and the associated restriction $x_{[\alpha]} \in \mathbb{R}^{d_\alpha}$ of $x \in \mathbb{R}^d$ of dimension $d_\alpha = \mathrm{card}\,(\Lambda(\alpha))$. The complementary part of the state vector is denoted by $x_{\backslash[\alpha]} \in \mathbb{R}^{d-d_\alpha}$. In the example of the multivariate Gaussian introduced in Section 3.1, we have $\Lambda(\alpha) = \{\alpha - 1, \alpha, \alpha + 1\}$ with the obvious periodicity extensions for $\alpha = 1$ and $\alpha = d$. With this notation in place, the implementation of the localized Schrödinger bridge sampler proceeds as described in Section 3.1.

The key assumption we make is that of conditional independence of x_α on $x_{\backslash[\alpha]}$, which allows for the dimension reduction. The conditional expectation value (27) turns out to be a Monte-Carlo approximation of the conditional expectation under this assumption. More precisely, given the transition density of overdamped Langevin dynamics, denoted here by $p(x'|x;\epsilon)$, we obtain

$$\mathbb{E}[X_\alpha(\epsilon)|X(0) = x] = \int x'_\alpha \, p(x'|x;\epsilon)\,\mathrm{d}x' = \int x'_\alpha \, p_\alpha(x'_\alpha|x;\epsilon)\,\mathrm{d}x'_\alpha \tag{32a}$$

$$= \int x'_\alpha \, p_\alpha(x'_\alpha|x_{[\alpha]};\epsilon)\,\mathrm{d}x'_\alpha = \mathbb{E}[X_\alpha(\epsilon)|X_{[\alpha]}(0) = x_{[\alpha]}], \tag{32b}$$

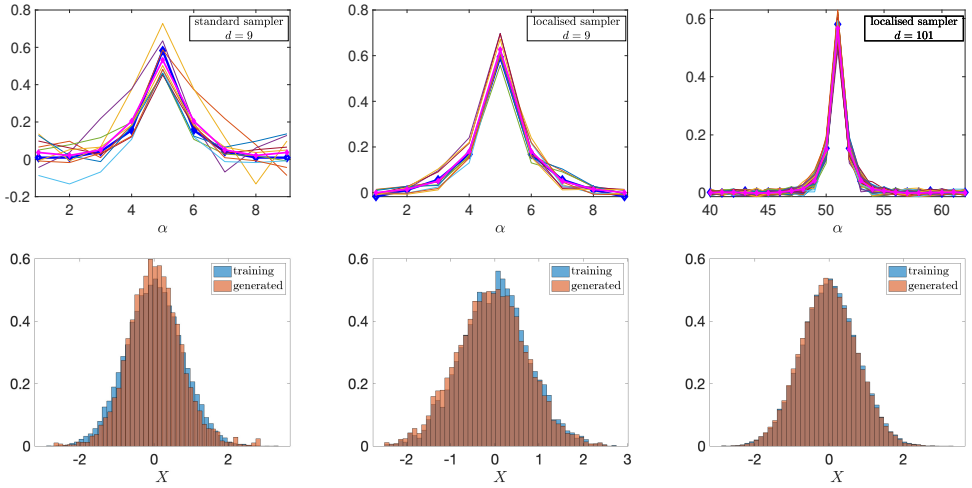

FIGURE 1. Comparison of the samples obtained from the standard and from the localized Schrödinger bridge sampler and given samples drawn from a d-dimensional multivariate Gaussian with a localized tridiagonal precision matrix. We show the centred rows of the empirical covariance matrix $\hat{\Sigma}$ (top row) and empirical histograms obtained from using all d components (bottom row). The blue markers denote the empirical covariance for the given samples; the magenta markers show the average over all d rows. For the $d = 9$ dimensional multivariate Gaussian we show results for the standard sampler trained on $M = 5,000$ samples (left column) and the localised sampler trained on $M = 500$ samples (middle column). The localised sampler was also run for a $d = 101$-dimensional multivariate Gaussian with only $M = 500$ samples (right column).

where the second line follows from the conditional independence assumption. Here $p_\alpha(x_\alpha|x;\epsilon)$ denotes the conditional distribution in x_α. It is reasonable to assume that we can construct a reversible overdamped Langevin process with invariant distribution $\pi_\alpha(x_{[\alpha]})$ and transition kernel $p_\alpha(x'_{[\alpha]}|x_{[\alpha]};\epsilon)$ on $\mathbb{R}^{d_\alpha}$. Here $\pi_\alpha(x_{[\alpha]})$ denotes the marginal distribution of $\pi(x)$ in $x_{[\alpha]}$. Then detailed balance of this dimension-reduced Langevin process is given by

$$p_\alpha(x'_{[\alpha]}|x_{[\alpha]};\epsilon)\,\pi_\alpha(x_{[\alpha]}) = p_\alpha(x_{[\alpha]}|x'_{[\alpha]};\epsilon)\,\pi_\alpha(x'_{[\alpha]}). \tag{33}$$

Note that in our localized Schrödinger bridge sampler detailed balance is ensured by the Sinkhorn algorithm which renders the Markov chain reversible. Detailed balance then implies for the conditional expectation value

$$\mathbb{E}[X_\alpha(\epsilon)|X_{[\alpha]}(0) = x_{[\alpha]}] = \int x'_\alpha\, p_\alpha(x'_{[\alpha]}|x_{[\alpha]};\epsilon)\,\mathrm{d}x'_{[\alpha]} \tag{34a}$$

$$= \int x'_\alpha\, \frac{p_\alpha(x_{[\alpha]}|x'_{[\alpha]};\epsilon)}{\pi_\alpha(x_{[\alpha]})}\,\pi_\alpha(x'_{[\alpha]})\,\mathrm{d}x'_{[\alpha]} \tag{34b}$$

$$= \int x'_\alpha\, \rho_\alpha(x'_{[\alpha]}|x_{[\alpha]};\epsilon)\,\pi(x'_{[\alpha]})\,\mathrm{d}x'_{[\alpha]}, \tag{34c}$$

with

$$\rho_\alpha(x'_{[\alpha]}|x_{[\alpha]};\epsilon) := \frac{p_\alpha(x_{[\alpha]}|x'_{[\alpha]};\epsilon)}{\pi_\alpha(x_{[\alpha]})}. \tag{35}$$

We note that $\rho_\alpha(x'_{[\alpha]}|x_{[\alpha]};\epsilon)$ is a density with respect to the reference measure induced by $\pi_\alpha(x_{[\alpha]})$. We finally approximate the integral in (34c) via Monte Carlo approximation using the restricted data samples $x^{(j)}_{[\alpha]} \sim \pi_\alpha$, $j = 1, \ldots, M$; i.e.,

$$\mathbb{E}[X_\alpha(\epsilon)|X_{[\alpha]}(0) = x_{[\alpha]}] \approx \sum_{j=1}^{M} x^{(j)}_\alpha \, w^{(j)}_\alpha(x_{[\alpha]}), \qquad x^{(j)}_\alpha \sim \pi_\alpha, \tag{36}$$

with $w^{(j)}_\alpha(x_{[\alpha]}) \propto \rho_\alpha(x^{(j)}_{[\alpha]}|x_{[\alpha]};\epsilon)$ such that $\sum_j w^{(j)}_\alpha(x) = 1$. This expression is of the form used in the localized Schrödinger bridge sampler (27). However, since the transition kernels $\rho_\alpha(x'_{[\alpha]}|x_{[\alpha]};\epsilon)$ are not available in general, the weight vector $w_\alpha(x_{[\alpha]}) \in \mathbb{R}^{d_\alpha}$ is approximated by the Schrödinger bridge approach as in (24). Algorithm 1 summarizes the localized Schrödinger bridge sampler.

Algorithm 1: Localised Schrödinger bridge sampler

Input: Input: Samples $\mathcal{X} \in \mathbf{R}^{d\times M}$.

Parameters : Bandwidth ϵ. Localisation dimension d_α. Desired number of new samples N. Number of decorrelation steps n_c.

Output: Output: New samples $x^{(j)}_{\mathrm{s}}$ for $j = 1, \ldots, N$.

Step 1: Construct transition Sinkhorn weights $v_{[\alpha]}$
for $\alpha \leftarrow 1$ **to** d **do**
  construct localised data $\mathcal{X}_{[\alpha]}$;
  construct kernel matrix $T_\alpha \in \mathbf{R}^{M\times M}$ from localised data;
  construct Sinkhorn weights v_α from T_α;
end
Step 2: Generate N new samples $x^{(j)}_{\mathrm{s}}$ using the Sinkhorn weights v_α
for $j \leftarrow 1$ **to** N **do**
  each new sample is started from a random initial sample
  $X(0) \leftarrow x^{(j^\star)}$ for $1 \le j^\star \le M$ and random $j^\star$;
  for $n \leftarrow 0$ **to** n_c **do**
    for $\alpha \leftarrow 1$ **to** d **do**
      $X_{[\alpha]}(n) \leftarrow X(n)$;
      $\Xi_{[\alpha]}(n) \sim \mathrm{N}(0, \epsilon I_{d_\alpha})$;
      $X_{[\alpha]}(n+1/2) = X_{[\alpha]}(n) + \sqrt{2}\,\Xi_{[\alpha]}(n)$; `/* noising step */`
      construct vector $t_\alpha(X_{[\alpha]}(n+1/2)) \in \mathbf{R}^M$;
      construct conditional probability $w_\alpha(X_{[\alpha]}(n+1/2)) \in \mathbf{R}^M$ using v_α ;
      construct localized data $\mathcal{X}_\alpha$;
      $X_\alpha(n+1) = \mathcal{X}_\alpha \, w_\alpha(X_{[\alpha]}(n+1/2))$; `/* projection step */`
    end
  end
  $x^{(j)}_{\mathrm{s}} \leftarrow X(n_c)$;
end

Remark 2. We have assumed here a strong form of conditional independence by requesting that (32) holds for all $\epsilon > 0$. In general, such a condition will be satisfied approximately for sufficiently small ϵ only. Compare the EM sampler (16), which provides an accurate approximation to the true transition densities $p(x'|x;\epsilon)$ of the underlying diffusion process for $\epsilon > 0$ sufficiently small by ignoring higher-order dependencies. In practice, this requires a careful choice of the dependency sets $\Lambda(\alpha)$ which define $x_{[\alpha]} \in \mathbb{R}^{d_\alpha}$.

3.3. **Algorithmic properties.** We briefly discuss a few important results on the stability and the ergodicity of the proposed localized Langevin samplers, which they essentially inherit from the unlocalized Schrödinger bridge sampler [5].

The following lemma establishes that, since each $w_\alpha(x_{[\alpha]})$, $\alpha = 1, \ldots, d$, is a probability vector for any $\epsilon > 0$, the localized update step (29) is stable. In order to simplify notations, we denote by $m_{\rm loc}(x;\epsilon) \in \mathbb{R}^d$ the vector of localized expectation values with components $m_\alpha(x_{[\alpha]})$, $\alpha = 1, \ldots, d$, defined by (27).

Lemma 1. Let us introduce the set $\mathcal{C}_M \subset \mathbb{R}^d$ defined by

$$\mathcal{C}_M = \{x \in \mathbb{R}^d : |x_\alpha| \leq |\mathcal{X}_\alpha|_\infty\}. \tag{37}$$

It holds that the vector $m_{\rm loc}(x;\epsilon) \in \mathbb{R}^d$ of localized expectation value satisfies

$$m_{\rm loc}(x;\epsilon) \in \mathcal{C}_M \tag{38}$$

for all choices of $\epsilon > 0$ and all $x \in \mathbb{R}^d$.

Proof. The lemma follows from the fact that the α-component of $m_{\rm loc}(x;\epsilon)$ is given by (27) and the fact that $w_\alpha(x_{[\alpha]})$ is a probability vector for all $\epsilon > 0$ and all $x \in \mathbb{R}^d$. □

Lemma 1 also establishes stability of the general Langevin sampler defined by (5) with localized $m_{\rm loc}(x;\epsilon)$, i.e.

$$X(n+1) = m_{\rm loc}(X(n);\epsilon) + \sqrt{2}\,\Xi(n), \qquad \Xi(n) \sim \mathrm{N}(0, \epsilon I), \tag{39}$$

for all step sizes $\epsilon > 0$. Note that $X(n+1)$ is no longer in the convex hull of the data as the original unlocalized Schrödinger bridge (cf. (14)b), but instead is confined to $\mathcal{C}_M$ in expectation. The next lemma shows that this Langevin sampler is also geometrically ergodic.

Lemma 2. Let us assume that the data generating density π has compact support. Then the localized time-stepping method (39) possesses a unique invariant measure and is geometrically ergodic.

Proof. Consider the Lyapunov function $V(x) = 1 + \|x\|^2$ and introduce the set

$$\mathcal{C} = \{x : \|x\| \leq R\} \tag{40}$$

for suitable $R > 0$. Since $m_{\rm loc}(x;\epsilon) \in \mathcal{C}_M$ and π has compact support, one can find a radius $R > 0$, which is independent of the training data $\mathcal{X}$, such that $\mathcal{C}_M \subset \mathcal{C}$ and

$$\mathbb{E}[V(X(n+1)|X(n)] \leq \lambda V(X(n)) \tag{41}$$

for all $X(n) \notin \mathcal{C}$ with $0 \leq \lambda < 1$. Furthermore, because of the additive Gaussian noise in (39), there is a probability density function $\rho(x) > 0$ on $\mathcal{C}$ and a constant $\delta > 0$ such that

$$\mathrm{n}(x'; m_{\rm loc}(x;\epsilon), 2\epsilon I) \geq \delta\rho(x') \tag{42}$$

for all $x, x' \in \mathcal{C}$. Here $\mathrm{n}(x; m, \Sigma)$ denotes the Gaussian probability density function with mean m and covariance matrix Σ. In other words, $\mathcal{C}$ is a small set in the sense of [12]. Geometric ergodicity follows from Theorem 15.0.1 in [12]. See also the self-contained presentation in [11]. □

Remark 3. We emphasize that, contrary to the unlocalized Schrödinger bridge sampler, the localized $m_{\rm loc}(x;\epsilon)$ is not restricted to the linear subspace spanned by the training data $\mathcal{X}$ in case $M < d$. The localized sampler shares this desirable property for very high-dimensional problems with the localized EnKF [16, 2, 3].

4. Conditional localised Schrödinger bridge sampler

As for the standard Schrödinger bridge sampler [5] the localized sampler lends itself to conditional sampling. Consider samples $x^{(j)} = (z^{(j)}, \psi^{(j)})^T$ for $j = 1, \ldots, M$ with $\psi = \psi(z)$. The localized Schrödinger bridge sampler described in Section 3 and Algorithm 1 allows us to learn the joint probability $p(z, \psi)$. To draw samples from the conditional probability $p(\psi|z)$ we may use the localized conditional probability vector $t_\alpha(x_{[\alpha]})$ and the Sinkhorn weights v_α obtained from the samples $x^{(j)}$, i.e. executing lines 1-6 in Algorithm 1. Conditional sampling is achieved by ensuring that at each sampling step the z-component of the generated samples X are set to the value z^* on which we wish to condition. To achieve this we add the conditioning assignment, $Z(n+1) \leftarrow z^*$, between lines 20 and 21 of Algorithm 1. We demonstrate the performance of the conditional sampler for the multiscale Lorenz-96 model in the next subsection.

4.1. Conditional sampling for a closure problem.

We apply the conditional localized Schrödinger bridge sampler to the multiscale Lorenz-96 model for K slow variables z_k which are each coupled to J fast variables $y_{j,k}$ with

$$\frac{d}{dt} z_k = -z_{k-1}(z_{k-2} - z_{k+1}) - z_k + F - \frac{hc}{b} \sum_{j=1}^{J} y_{j,k}, \tag{43a}$$

$$\frac{d}{dt} y_{j,k} = -cby_{j+1,k}(y_{j+2,k} - y_{j-1,k}) - cy_{j,k} + \frac{hc}{b} z_k. \tag{43b}$$

with periodic boundary conditions with $z_{k+K} = z_k$, $y_{j,k+K} = y_{j,k}$ and $y_{j+J,k} = y_{j,k+1}$. This $d = K(J+1)$-dimensional model was introduced as a caricature for the mid-latitude atmospheric dynamics [9]. The degree of time-scale separation is controlled by the parameter c. The ratio of the amplitudes of the large-scale variables z_k and the small-scale variables $y_{j,k}$ is controlled by the parameter b. The slow and fast dynamics are coupled with coupling strength h. The parameter F denotes external forcing. As equation parameters we choose $K = 12$ and $J = 24$, i.e. $d = 300$, and $F = 20$, $c = b = 10$ and $h = 1$ as in [20, 1, 4]. These parameters lead to chaotic dynamics with a maximal Lyapunov exponent of $\lambda_{\max} \approx 18.29$ in which the fast variables experience temporal fluctuations which are 10 times faster and 10 times smaller than those of the slow variables. This corresponds to the regime of strong coupling in which the dynamics is driven by the fast sub-system [6].

In the climate science and other disciplines one is typically only interested in the slow large-scale dynamics. A direct simulation of the multiscale system (43), however, requires a small time step adapted to the fastest occurring time scale, making long term integration to resolve the slow dynamics computationally infeasible. Scientists hence aim to design a computationally tractable model for the slow variables only in which the effect of the fast dynamics is parameterized. This is the so called closure or subgrid-scale parameterization problem. In particular, we seek a model of the form

$$\frac{d}{dt} z_k = G_k(z) + \psi_k(z), \tag{44}$$

with $z = (z_1, z_2, \ldots, z_K)$, $G_k(z) = -z_{k-1}(z_{k-2} - z_{k+1}) - z_k + F$. We assume here that scientists have prior physics-based knowledge about the resolved vector field $G_k(z)$ but lack knowledge of the closure term $\psi_k(z)$ which parametrizes the effect of the fast unresolved dynamics. The

closure term may be deterministic or stochastic, depending on the choice of equation parameters in (43). We will employ the localized Schrödinger bridge sampler to generate samples of the closure term $\psi(z)$ conditioned on the current model state $z(t)$. The sampler will be trained on M samples $x^{(j)} = (z^{(j)}, \psi^{(j)})$, $j = 1, \ldots, M$.

We obtain $M = 40,000$ samples $z^{(j)} \in \mathbb{R}^K$ by integrating (43) using a fourth-order Runge-Kutta method with a fixed time step $\Delta t = 5 \times 10^{-4}$. Samples of the closure term $\psi^{(j)} \in \mathbb{R}^K$ are then determined from the samples $z^{(j)}$ via

$$\psi^{(j)} := \frac{z^{(j+1)} - z^{(j)}}{\Delta t} - G(z^{(j)}), \tag{45}$$

for $j = 1, \cdots, M$. This defines our M samples $x^{(j)} = (z^{(j)}, \psi^{(j)}) \in \mathbb{R}^{2K}$ for $j = 1, \ldots, M$ to be used to train the Schrödinger bridge sampler.

To numerically integrate the closure model (44) using an Euler discretization

$$z(m+1) = z(m) + (G(z(m)) + \psi(m))\, \delta t \tag{46}$$

in terms of the state vector $z \in \mathbb{R}^K$ with a time step $\delta t = 10\Delta t$, we generate at each time step $m \geq 0$ a sample $\psi(m)$ conditioned on the current state $z(m)$. These samples should be uncorrelated to the samples drawn at the previous time step. This is achieved by running the localized Schrödinger bridge sampler conditioned on $z^* = z(m)$ at each time step m for $n_c = 100$ decorrelation steps (cf. Algorithm 1).

For the localized Schrödinger bridge sampler we employ a time step of $\epsilon = 0.1$ and consider a nearest neighbour localisation with $\Lambda(\alpha) = \{\alpha - 1, \alpha, \alpha + 1, \alpha, K + \alpha - 1, K + \alpha, K + \alpha + 1\}$ with the obvious periodic extensions for $\alpha = 1$ and $\alpha = d$. To account for the varying ranges of z and ψ when estimating the matrices (23) and (25) for fixed time step ϵ, we replace the standard Euclidean product with a scaled one where we divide in the inner product the z-variables by σ_z and the ψ-variables by σ_ψ, where σ_ψ^2 and σ_x^2 denote the climatic variances of the slow variables and the closure term, respectively, estimated from the samples $x^{(j)}$.

Figure 2 shows a comparison of the outputs of the localized Schrödinger bridge sampler with data obtained from simulating the full multi-scale Lorenz-96 system (43). We show results for the covariance of the slow variables z, obtained from the samples $z^{(j)}$ of the full multi-scale Lorenz 96 system (43), and of the discretization of the closure scheme (46) over $40,000$ time steps with time step δt. Samples $\psi(m)$ are generated at each time step using the conditional sampling algorithm. In particular, we show the entries of the rows of the empirical covariance matrix, centred about $k = 6$ employing periodicity of the system. It is seen that our localized sampler reproduces the covariance structure of the full system very well. We further show a comparison of the empirical histograms obtained by integrating the closure model (46) with the original samples $z^{(j)}$ which were obtained from a simulation of the full multi-scale Lorenz 96 system (43). The learned non-stiff stochastic closure model (46) is able to reproduce the actual histogram well.

5. Conclusions

The construction of the Schrödinger bridge is fraught with an unfavorable dependency on the dimension d. The required number of samples scales for a desired accuracy exponentially on the underlying intrinsic dimensionality of the data [21]. This is exactly the advantage of our proposed localization: Localization requires less samples for a desired accuracy. We have demonstrated numerically the advantage of localization by means of a Gaussian measure for which the inverse covariance matrix has tri-diagonal structure and for a nonlinear close problem arising from the multi-scale Lorenz-96 model. We have also established theoretically that the proposed sampler is stable and geometric ergodic under relatively mild conditions. The stability of our sampler

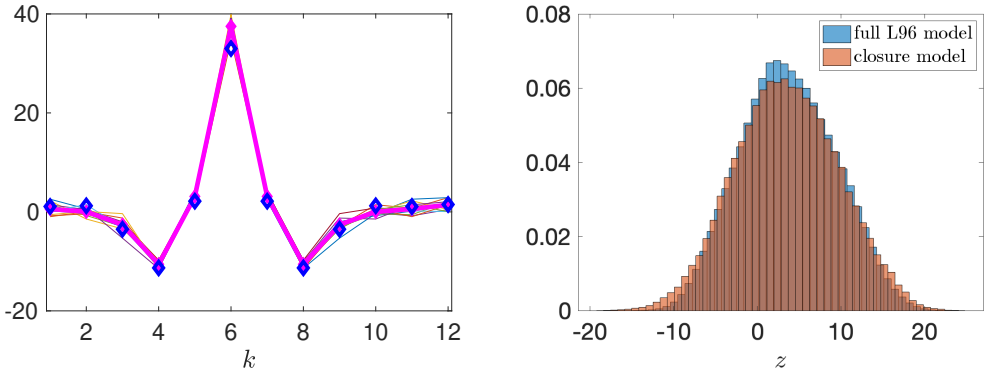

FIGURE 2. Comparison of the samples obtained from the localized Schrödinger bridge sampler and given samples drawn from the multi-scale Lorenz-96 system (43) using nearest neighbour localization with $\Lambda(\alpha) = \{\alpha - 1, \alpha, \alpha + 1\, K + \alpha - 1, K + \alpha, K + \alpha + 1\}$ with the obvious periodic extensions for $\alpha = 1$ and $\alpha = d$. We consider $40,000$ new and given samples. Left: Centred rows of the empirical covariance matrix for z. The magenta line denotes the mean over all rows. The blue markers denote the empirical covariance for the given samples. Right: Empirical histograms. The closure model used at ten times larger time step.

allows for applications to data drawn from a singular measure. This sets it apart from score-generative models which rely on the differentiability of the measure. While this work has focused on overdamped Langevin dynamics as a mean of sampling from a distribution, the methodology generalizes to more general formulations of score-generative and diffusion modeling [7, 17, 19, 22]. Further potential applications include sequential data assimilation [16, 2, 3], feedback particle filter and homotopy methods [23, 15, 14], which are implemented utilizing Schrödinger bridges, and interacting particle sampling methods, which rely on grad-log density estimators [10].

Acknowledgements. This work has been funded by Deutsche Forschungsgemeinschaft (DFG) - Project-ID 318763901 - SFB1294. GAG acknowledges funding from the Australian Research Council, grant DP220100931.

REFERENCES

[1] H. M. Arnold, I. M. Moroz, and T. N. Palmer. Stochastic parametrizations and model uncertainty in the Lorenz 96 system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1991):20110479, 2013. doi: 10.1098/rsta.2011.0479.

[2] M. Asch, M. Bocquet, and M. Nodet. *Data assimilation: Methods, algorithms, and applications*. SIAM, 2016.

[3] G. Evensen, F. Vossepoel, and P. van Leeuwen. *Data Assimilation Fundamentals: A unified Formulation of the State and Parameter Estimation Problem*. Springer Nature Switzerland AG, Cham, Switzerland, 2022.

[4] G. A. Gottwald and S. Reich. Supervised learning from noisy observations: Combining machine-learning techniques with data assimilation. *Physica D: Nonlinear Phenomena*, 423: 132911, 2021.

[5] G. A. Gottwald, F. Li, S. Reich, and Y. Marzouk. Stable generative modeling using Schrödinger bridges. Technical report, arXiv:2401.04372, 2024.
[6] S. Herrera, J. Fernández, M. Rodríguez, and J. Gutiérrez. Spatio-temporal error growth in the multi-scale Lorenz'96 model. *Nonlinear Processes in Geophysics*, 17(4):329, 2010.
[7] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.
[8] A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6, 2005.
[9] E. N. Lorenz. Predictability: A problem partly solved. In T. Palmer, editor, *Proc. Seminar on predictability Vol. 1*, pages 1–18, Reading, UK, 1996. ECMWF.
[10] D. Maoutsa, S. Reich, and M. Opper. Interacting particle solutions of Fokker–Planck equations through gradient-log-density estimation. *Entropy*, 22(8), 2020.
[11] J. Mattingly, A. Stuart, and D. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101:185–232, 2002.
[12] S. Meyn and R. T. Tweedy. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, 2nd edition, 2009.
[13] G. A. Pavliotis. *Stochastic Processes and Applications*. Springer Verlag, New York, 2016.
[14] J. Pidstrigach and S. Reich. Affine-invariant ensemble transform methods for logistic regression. *Found. Comput. Math*, 2022. published online 21 January 2022.
[15] S. Reich. A dynamical systems framework for intermittent data assimilation. *BIT Numerical Mathematics*, 51(1):235–249, 2011.
[16] S. Reich and C. Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, 2015.
[17] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
[18] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.
[19] Y. Song, J. N. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. Technical report, arXiv:2011.13456, 2020.
[20] D. S. Wilks. Effects of stochastic parametrizations in the Lorenz '96 system. *Quarterly Journal of the Royal Meteorological Society*, 131(606):389–407, 2005. doi: 10.1256/qj.04.03.
[21] C. Wormell and S. Reich. Spectral convergence of diffusion maps: Improved error bounds and an alternative normalisation. *SIAM J. Numer. Anal.*, 59:1687–1734, 2021. doi: 10.1137/20M1344093.
[22] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. Technical report, arXiv:2209.00796, 2022.
[23] T. Yang, P. G. Mehta, and S. P. Meyn. Feedback particle filter. *IEEE Trans. Automat. Control*, 58(10):2465–2480, 2013. ISSN 0018-9286. doi: 10.1109/TAC.2013.2258825.

School of Mathematics and Statistics, University of Sydney
Email address: georg.gottwald@sydney.edu.au

Department of Mathematics, University of Potsdam
Email address: sebastian.reich@uni-potsdam.de